



Efficient feature fusion network based on center and scale prediction for pedestrian detection

Tao Zhang¹ · Yahui Cao¹ · Le Zhang¹ · Xuan Li¹

Accepted: 30 April 2022 / Published online: 21 July 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Center and scale prediction (CSP) is an anchor-free pedestrian detector with good performance. However, there are lots of parameters in the detector, which seriously limits the speed. In this paper, a new network is designed for the improvement of the detector speed, which contains less parameters, named Feature Fusion: Center and Scale Prediction (F-CSP). F-CSP fuses multi-scale feature maps with two efficient feature fusion networks: Feature Pyramid Networks (FPN) and Balanced Feature Pyramid (BFP). Specifically, FPN is used to reduce the channel of feature maps, and BFP is used to fuse multiple feature maps into a single one. This way, the proposed detector achieves competitive accuracy and higher speed on the challenging pedestrian detection benchmark. The performance of F-CSP is demonstrated on the Caltech dataset. Compared with CSP, under the premise of ensuring accuracy, the speed is increased from 45.1 to 32.9 ms/img.

Keywords Pedestrian detection · Convolutional neural network · Feature fusion · Center and scale prediction

1 Introduction

In recent years, with the rapid development of assisted driving [1–3], autonomous driving [4–6], and intelligent monitoring [7–9], pedestrian detection [10–13] has received more and more attention. In the field of autonomous driving, it is important to avoid pedestrians on the lane effectively. In the intelligent monitoring system, the pedestrian behavior accurate analysis is also essential. Both competitive accuracy and satisfying speed are significant for the pedestrian detector.

Pedestrian detection can be divided into anchor-based and anchor-free [14]. Anchor-based includes Faster R-CNN [15], single-shot multiBox detector (SSD) [16], YOLO9000 [17], and anchor-free includes CornerNet [18], ExtremeNet [19], center and scale prediction (CSP) [14], etc. Both of

them have excellent performance in detection accuracy, especially anchor-free. However, they have lots of parameters while improving the accuracy, which leads to the expensive computation and limits the computation speed. So anchor-based and anchor-free are difficult to be applied to hardware devices.

To solve this problem, this paper proposes a new network with high-speed: Feature Fusion Network based on Center and Scale Prediction (F-CSP) detector. CSP is one of the anchor-free classical pedestrian detectors for pedestrian detection. CSP is proposed for a higher-level abstraction, which learns central points of the objects. The object for detection is motivated as a high-level semantic feature detection task. In many pedestrian detection tasks, CSP shows its outstanding performance with high accuracy. To reduce the complexity of the network structure and improve the processing speed, F-CSP improves the feature fusion module in CSP according to the two efficient feature fusion networks: Feature Pyramid Network (FPN) [20] and Balanced Feature Pyramid Network (BFP) [21]. FPN aims to reduce the channel of feature maps. BFP aims to fuse multi-scale feature maps into a single one. And there are no additional parameters to be introduced. Besides, the channel of feature maps keep the same when they propagate in the two modules. The experimental results show that the proposed network performs well in the speed.

✉ Xuan Li
lixuantju@tju.edu.cn

Tao Zhang
zhangtao@tju.edu.cn

Yahui Cao
caoyahui@tju.edu.cn

Le Zhang
Polaris963@163.com

¹ School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

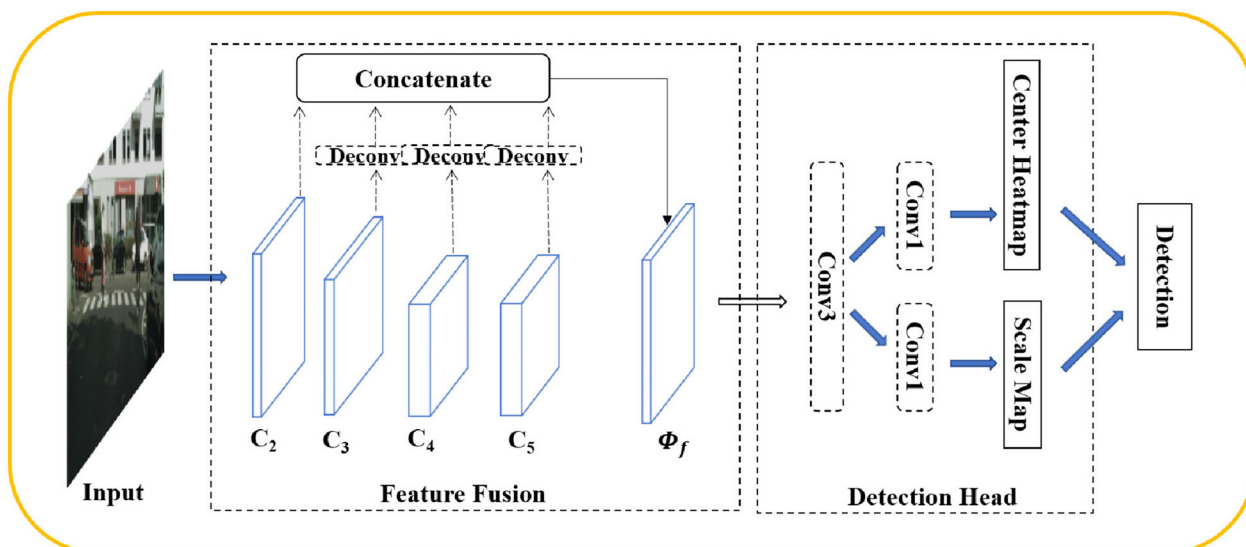


Fig. 1 Overall architecture of CSP [14]. CSP is composed of two components: the feature fusion module and the detection head. The feature fusion module contains lots of learnable parameters which limit the speed

In summary, the main contributions of this work are as follows: We design a new network which contains fewer parameters, and it replaces the feature fusion module in CSP with FPN and BFP. F-SCP meets the computation speed requirement of the pedestrian detector. The proposed F-CSP achieves the new state-of-the-art speed on the challenging pedestrian detection benchmarks, Caltech.

2 Related work

2.1 Object detection algorithm

Object detection algorithm [22–24] is the basis of computer vision task, which is used to identify objects and their positions in images. Traditional object detection algorithms without convolutional neural network (CNN) mainly include histogram of oriented gradient (HOG) [25], deformable part model (DPM) [26] and non-maximum suppression (NMS) [27].

After the rise of CNN, CNN is used to complete the object detection tasks instead of traditional manual algorithms [28,29]. Now, the state-of-the-art object detection algorithm can be divided into the two-stage algorithm and the one-stage algorithm according to whether to generate region proposals [30]. The two-stage algorithm first generates region of interests that may contain objects, and then classifies samples through CNN. It mainly includes R-CNN, Fast R-CNN, Faster R-CNN. R-CNN [31] uses the region proposal method to create the region for detection, and then adopts the CNN for classification. However, this algorithm has a lot of candidate solutions that results in high com-

putational complexity. Fast R-CNN [32] extracts the image features by a feature extractor instead of extracting multiple times for each image from the beginning. This algorithm avoids extracting redundant features and improves the training speed. Faster R-CNN [15] applies a new region proposal network based on the Fast R-CNN, and it is more effective and efficient. The two-stage algorithm has high accuracy in position and recognition because of correcting the candidate boxes continuously. However, it needs to run the detection and classification process multiple times, so the detection speed is low [33].

The one-stage algorithm does not need to generate region proposals. It transforms the object detection problem into a regression and classification problem. The test speed is fast because the result can be directly output after the images input into the network. In 2015, YOLO [17] algorithm was first proposed, which obtained the coordinates of the bounding box, the confidence and the class probabilities of the object through images directly. This algorithm detects objects very fast and can learn generalized features, but it is easy to make mistakes in detecting small objects. In 2016, SSD [16] was proposed, which combined the regression idea in YOLO and the anchor mechanism in Faster R-CNN. SSD performs regression by using the multi-scale regional features of each position in images, which not only maintains the high-speed characteristics of YOLO, but also ensures the detection accuracy of Faster R-CNN. One of the reasons why the performance of the one-stage algorithm is inferior to that of the two-stage algorithm is the imbalance between positive and negative samples in the region proposal. In 2017, the Retinanet [34] proposed the focal loss function to automatically reduce the weight of categories that are easy to be

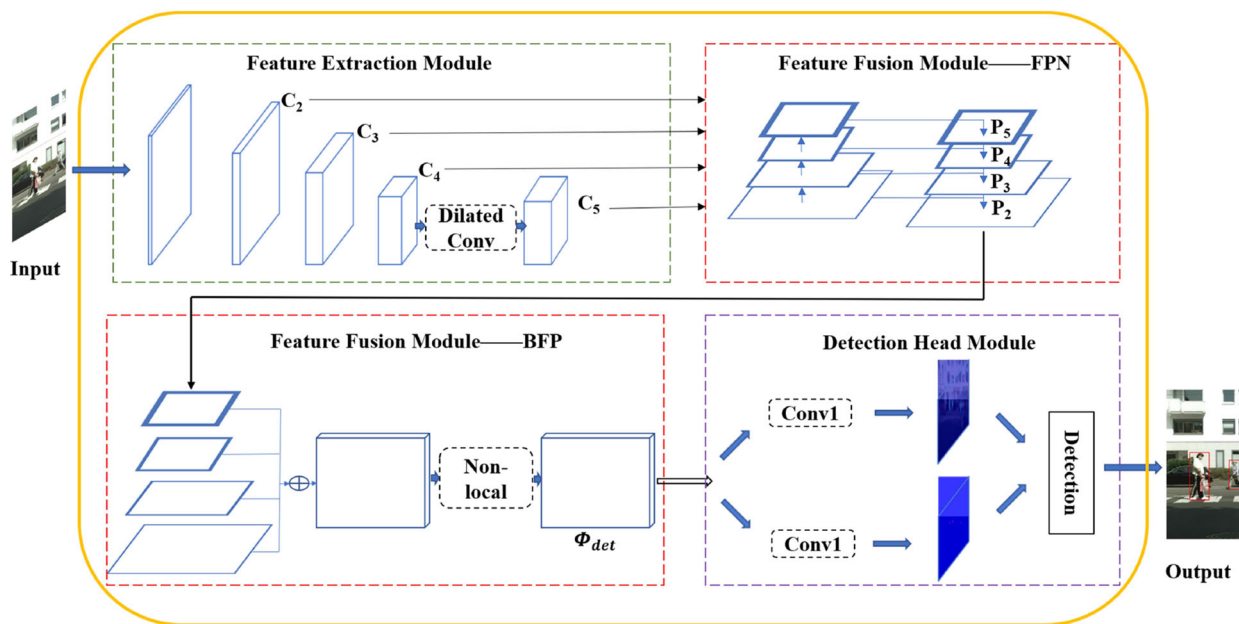


Fig. 2 Overall architecture of F-CSP, which mainly comprises three modules, i.e., the feature extraction module, the feature fusion module(FPN and BFP) and the detection head module. The feature extraction module which is ResNet-50 outputs four feature maps. The feature

fusion module which consists of FPN and BFP networks fuses above four feature maps into a single one. The detection head module contains two 1×1 convolution layers, one for the center location and the other for the corresponding scale

classified, and increase the focus of categories that are difficult to be classified. As a one-stage algorithm, Retinanet makes a trade-off between the detection accuracy and the test speed. It achieves comparable accuracy compared with the two-stage algorithms while ensuring the fast speed.

2.2 Pedestrian detection algorithm based on the key points

The framework of CSP is shown in Fig. 1. In the feature extraction module (the outputs of the module is defined as *stage 2, 3, 4, 5* respectively), CSP uses dilated convolution [35] to expand the receptive field in the last residual block, so that the accuracy of small object detection can be improved. Feature map with the higher resolution might contain richer details, so it is suitable for small object detection, such as pedestrians. In the feature fusion module, CSP firstly applies deconvolution layers [36] to adjust the size of feature maps. Then, CSP fuses the feature maps into a single one, denoted as Φ_f , by concatenation. In the detection head, CSP reduces the channels of Φ_f through a 3×3 convolution layer and predicts the center and scale by two 1×1 convolution layers, respectively.

Although the network is succinct to reduce the complexity effectively, it has lots of parameters. There are two reasons for this: (i). The convolution kernels of the deconvolution layers are 512, 1024, and 2048, respectively, which are the

same as the amounts of the *stage3, 4, 5* channels. (ii). When CSP attaches a single 3×3 convolution layer on Φ_f , the fused channel is 3584, so the size of convolution kernels must be 3584 in depth. Therefore, the network has expensive computation and limited speed.

3 Methodology

3.1 Module analysis

The feature fusion module must follow two premises to improve the speed: (i). When the feature fusion module adjusts the feature map size, it should avoid introducing many parameters to the network structure. (ii). When feature maps are propagated in the network, few channels should be maintained.

FPN has been widely used in various object detectors due to the high efficiency. It involves a bottom-up pathway, a top-down pathway, and lateral connections. The bottom-up pathway is the feedforward computation of ResNet-50, which computes a feature hierarchy consisting of feature maps at several scales with a scaling step of 2. The top-down pathway hallucinates higher-resolution features by upsampling feature maps from higher pyramid levels. The feature maps contain coarse spatial information and stronger semantic information. Each lateral connection merges feature maps

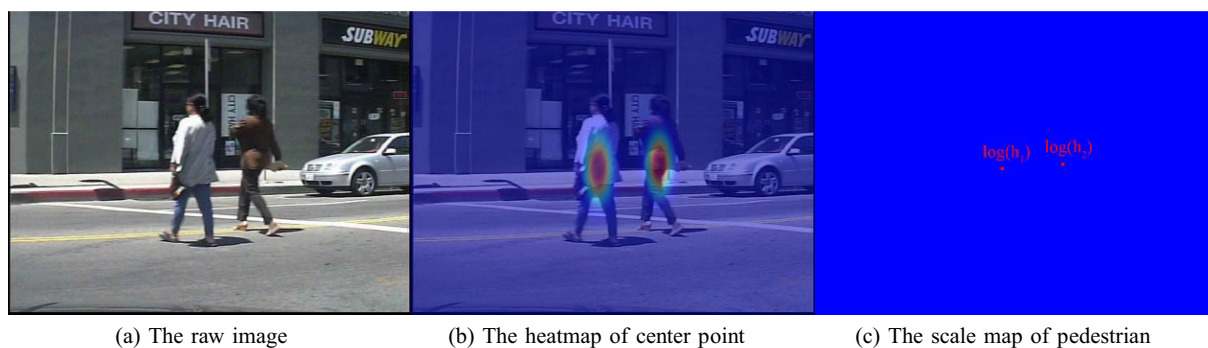


Fig. 3 The input image and its corresponding label image. **a** The original image to be detected. **b** The center point heatmap after calculating the two-dimensional Gaussian distribution at the object center point. **c** The scale map corresponding to each pedestrian expressed by $\log(h)$

of the same spatial size from the bottom-up pathway and the top-down pathway. Because of these components, FPN can be applied easily to other networks.

Many feature maps are produced although FPN follows the two premises introduced above. To further reduce the computation, BFP fuses these feature maps into a single one, closely following FPN. BFP is to strengthen the multi-scale feature maps using the same deeply integrated balanced semantic feature maps. It consists of three steps, rescaling, integrating, and refining. Specifically, BFP first resizes the multi-scale feature maps to an intermediate size (the same size as one of the feature maps) via interpolation and max-pooling. Then, it fuses the feature maps into a balanced semantic feature by weighted average. Note that this procedure does not contain any parameters. Finally, a non-local neural network is used to refine the balanced semantic feature map to be more discriminative. Through matching the two modules, F-CSP can achieve excellent speed.

3.2 Network structure

The framework of F-CSP is shown in Fig. 2. There are three sub-modules in the network: feature extraction module, feature fusion module, and detection head module.

In the feature extraction module, F-CSP uses ResNet-50 to produce four feature maps, which are denoted as $\{C_2, C_3, C_4, C_5\}$ (the last residual block is the dilated convolution), which is the same as CSP. The sizes of $\{C_2, C_3, C_4, C_5\}$ are downsampled by 4, 8, 16, 16 with respect to the input image. And the channels of $\{C_2, C_3, C_4, C_5\}$ are 256, 512, 1024 and 2048, respectively.¹

In feature fusion module, F-CSP concatenates FPN and BFP to fuse multi-scale feature maps. Separately, FPN out-

puts four semantic feature maps, denoted as $\{P_2, P_3, P_4, P_5\}$, corresponding to $\{C_2, C_3, C_4, C_5\}$ that are respectively of the same spatial sizes. BFP fuses $\{P_2, P_3, P_4, P_5\}$ into a balanced semantic feature. But notably, the size of the balanced semantic feature is a considerable issue. Mentioned in 3.1, when the balanced semantic feature has the same size as P_2, P_3, P_4 or P_5 , they all meet the requirement of the feature fusion module. In this paper, $\{P_2, P_3, P_4, P_5\}$ is resized into the same size as P_3 . Then the balanced semantic feature is refined by the embedded Gaussian non-local attention module. The output is denoted as Φ_{det} .

In detection head module, F-CSP applies a few convolution layers to parse Φ_{det} into detection results. Specifically, a single convolution layer is attached to reduce its channel dimensions to 256-dim. And then two sibling 1×1 convolution layers are appended to produce the center heatmap and scale map, respectively.² The visualization result is shown in Fig. 3.

4 Experiment

4.1 Experimental environment

The proposed method is implemented in PyTorch, and the model is trained on the Caltech with NVIDIA RTX 2080TI GPU and Intel(R) E5-2630 v4 @ 2.20GHz CPU.

In order to verify the performance of the F-CSP algorithm, this paper selects several algorithms for comparison: HOG, Faster R-CNN, Retinanet and CSP, which are commonly used algorithm in manual object detection algorithms, two-stage algorithm, one-stage algorithm, and pedestrian detection algorithm, respectively.

At the same time, for fair comparison, this paper re-implements the Faster R-CNN, Retinanet and CSP algorithm. The backbone is ResNet-50, and the weight parameters are

¹ For ResNet-50, its Conv layers can be divided into five stages, in which the output feature maps of the five stages are downsampled by 2, 4, 8, 16, 32 with respect to the input image, respectively. As regular [37,38], the dilated convolutions are adopted in the last residual block to keep its output as 1/16 of the input image size.

² Optionally, to slightly adjust the center location, an extra offset prediction branch can be appended in parallel with the above two branches.

Table 1 Comparisons with the other algorithms on Caltech

| Method | Parameters | Test time | $MR^{-2}(\%)$ | |
|-----------------|----------------|--------------------|---------------|----------|
| | | | IoU=0.5 | IoU=0.75 |
| HOG [25] | – | – | 92.03 | 100.00 |
| FasterRCNN [15] | 39.82 MB | 76.4 ms/img | 7.52 | 64.08 |
| Retinanet [34] | 36.08 MB | 65.8 ms/img | 53.33 | 77.36 |
| CSP [14] | 38.0 MB | 45.1 ms/img | 5.13 | 26.5 |
| F-CSP | 25.9 MB | 32.9 ms/img | 6.04 | 37.03 |

pretrained on ImageNet. Adam is applied to optimize the network. This paper also uses the moving weight [39] to find the best feature map combination scheme and the best feature resolution. Besides, HOG algorithm is a traditional object detection algorithm, so it does not need to train CNN.

The evaluation follows the standard Caltech evaluation metric, that is, log-average Miss Rate over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^0]$ (denoted as MR^{-2}). Tests are applied to the original image size without enlarging for speed consideration.

4.2 Experimental results and analysis

Caltech comprises approximately 2.5 hours of auto-driving video with extensively labeled bounding boxes. We use the training data augmented by tenfolds (42782 frames) and test on the 4024 frames in the standard test set, and all exper-

iments are conducted on the new annotations provided by [40]. The size of each image is 480×640 , a mini-batch contains 16 images with one GPU, the learning rate is set as 10^{-4} , and training phase is stopped after 15K iterations.

Table 1 reports the comparison results with HOG, Faster R-CNN, Retinanet and CSP on the Caltech. It can be seen that F-CSP has the highest *Test Time*, and Faster R-CNN has the lowest *Test Time*. The detection accuracy of Faster R-CNN is higher than Retinanet, while its *Test Time* is lower than Retinanet. Besides, it is shown that F-CSP only requires 25.9MB parameters compared to 38.0MB of CSP. Due to fewer parameters, F-CSP achieves higher speed, the *Test Time* is 32.9 ms/img, increased by 28% to CSP. Meantime, the accuracy drops with lower resolution. (The MR^{-2} of F-CSP is 6.88%, compared to 5.84% of CSP.)

The visualization results of CSP and F-CSP are shown in Fig. 4. It demonstrates that CSP and F-CSP both can wholly

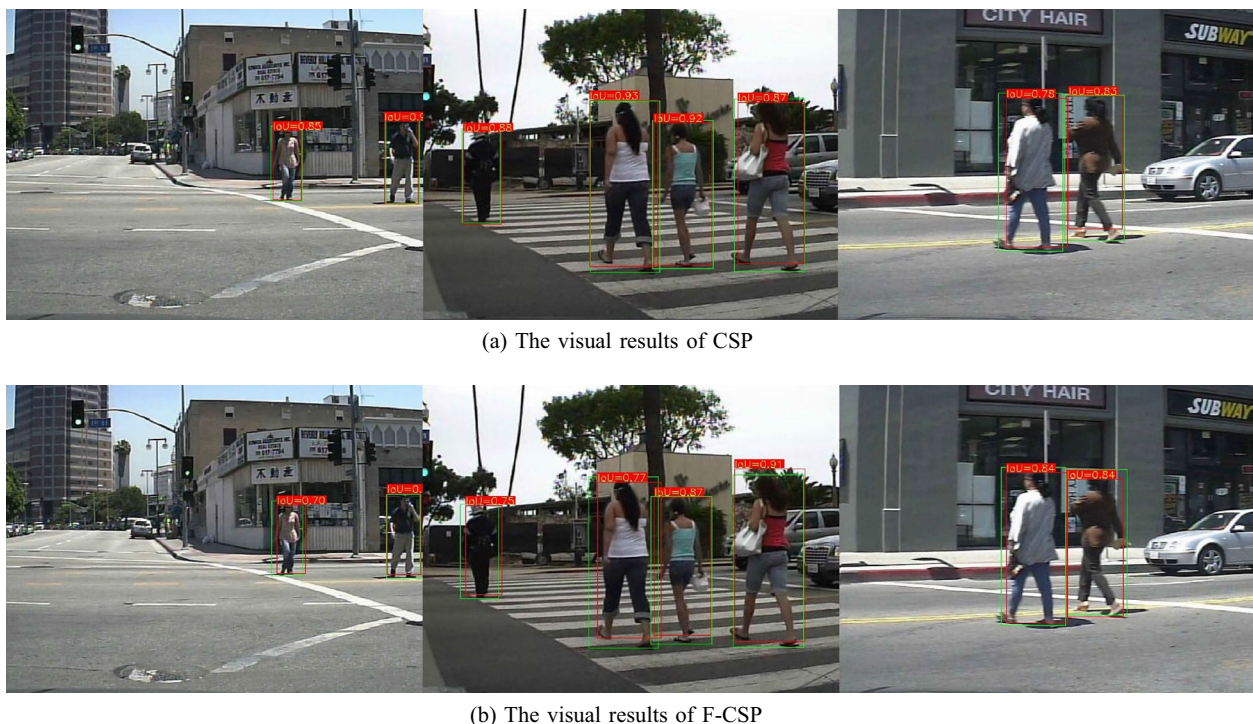


Fig. 4 Visualization results of CSP and F-CSP. The green boxes represent for ground-truth, and the red boxes represent for predictions. Otherwise, the overlap between ground-truth and predictions is expressed by IoU

Table 2 Performance comparison of F-CSP with different sizes of the balanced semantic feature. The leftmost column represents the balanced semantic feature which has the same size as P_2 , P_3 , P_4 or P_5 , respectively

| | Parameters | Test time | $MR^{-2}(\%)$ | |
|-------|------------|------------|---------------|--------------|
| | | | IoU=0.5 | IoU=0.75 |
| P_2 | 25.9MB | 80.0ms/img | 11.37 | 36.8 |
| P_3 | | 32.9ms/img | 6.88 | 37.02 |
| P_4 | | 30.0ms/img | 9.19 | 46.31 |
| P_5 | | 30.0ms/img | 10.61 | 38.44 |

frame pedestrians though there are some tiny differences. Therefore, the accuracy of F-CSP is dropped compared with CSP, but it is acceptable for the requirements of pedestrian detection.

Table 2 shows the results of the balanced semantic feature in different sizes. From Table 2, we can see that F-CSP achieves balanced performance when the balanced semantic feature has the same size as P_3 (MR^{-2} is 6.88% under IoU=0.5, *Test Time* is 32.9ms/img). It is worth noting that the *Test Time* is 80.0ms/img when the balanced semantic feature is resized into the same size as P_2 . The reason for the existing lowest speed is that the size of other three feature maps is required to be enlarged through interpolation. This procedure introduces expensive computation. Besides, the *Test Time* is 30.0ms/img when the balanced semantic feature has the same size as P_4 and P_5 , that is faster than 80.0ms/img. There is no extra computation, while feature maps are adjusted to the same size as P_4 and P_5 by max-pooling. Although this setting achieves the fastest speed, it has worse accuracy (the MR^{-2} is 9.19% and 10.61%, respectively).

Table 3 shows the results with different feature combination. It can be seen that the best accuracy ($MR^{-2} = 6.88\%$, *Test Time* = 32.9ms/img) comes from the combination of $\{C_2, C_3, C_4, C_5\}$. The combination of $\{C_2, C_3\}$ has fewer parameters and lower *Test Time*, but it brings more miss results. Notably, the MR^{-2} with the combi-

Table 3 Performance comparison of F-CSP under different feature combination schemes

| Feature maps | | | | Parameters | Test time | $MR^{-2}(\%)$ |
|--------------|-------|-------|-------|------------|------------|---------------|
| C_2 | C_3 | C_4 | C_5 | | | |
| ✓ | ✓ | | | 2.9MB | 5.2ms/img | 13.9 |
| | ✓ | ✓ | | 9.9MB | 27.8ms/img | 7.45 |
| | | ✓ | ✓ | 24.5MB | 30.7ms/img | 23.74 |
| ✓ | ✓ | ✓ | | 10.5MB | 28.0ms/img | 7.08 |
| | ✓ | ✓ | ✓ | 25.2MB | 32.0ms/img | 6.95 |
| ✓ | ✓ | ✓ | ✓ | 25.9MB | 32.9ms/img | 6.88 |

nation of $\{C_4, C_5\}$ is 23.74%. There are two reasons for the worst accuracy: (i). The shallower feature maps with rich details do not participate in the prediction. (ii). The semantic information on $\{C_4, C_5\}$ is damaged during upsampling.

The proposed method is extensively compared with the state of the arts on three settings: Reasonable, All and Heavy Occlusion. As shown in Fig. 5, F-CSP performs consistently for different occlusion levels compared to other detectors, which proves that the proposed method is effective.

5 Conclusion

This paper proposes a new pedestrian detection network named F-CSP. This network fuses multi-scale feature maps with two efficient feature fusion networks: FPN and BFP. FPN is used to fuse more semantic to the feature maps. BFP is used to balance multiple semantic feature maps. As a result, the proposed F-CSP detector achieves competitive accuracy and higher speed on the challenging pedestrian detection benchmark. However, the accuracy of F-CSP can be further improved, while its detection speed is very high.

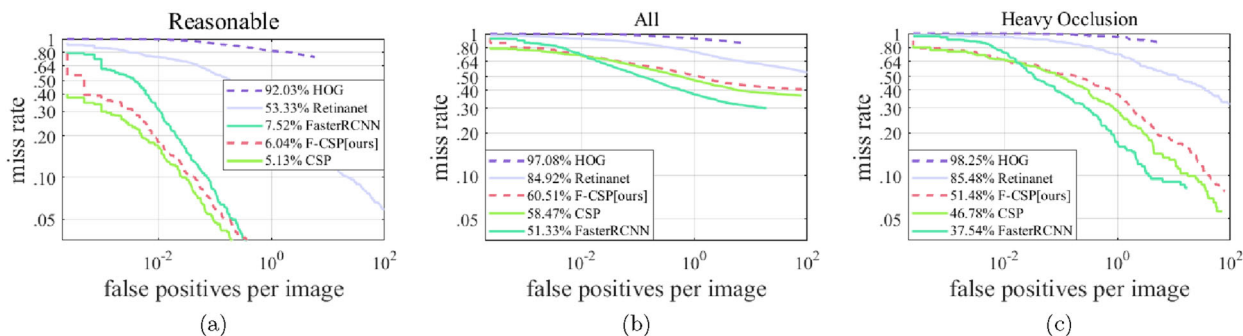


Fig. 5 Comparisons with other algorithms on Caltech using new annotations

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

- Moya, S., Grau, S., Tost, D.: The wise cursor: assisted selection in 3D serious games. *Vis. Comput.* **29**(6), 795 (2013). <https://doi.org/10.1007/s00371-013-0831-3>
- Sherstyuk, A., Jay, C., Treskunov, A.: Impact of hand-assisted viewing on user performance and learning patterns in virtual environments. *Vis. Comput.* **27**(3), 173 (2011). <https://doi.org/10.1007/s00371-010-0516-0>
- Ballit, A., Mougharbel, I., Ghaziri, H., Dao, T.T.: Computer-aided parametric prosthetic socket design based on real-time soft tissue deformation and an inverse approach. *Vis. Comput.* (2021). <https://doi.org/10.1007/s00371-021-02059-9>
- Chen, G., Qin, H.: Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *Vis. Comput.* (2021). <https://doi.org/10.1007/s00371-021-02067-9>
- Fan, X., Pan, G., Mao, Y., He, W.: A personalized traffic simulation integrating emotion using a driving simulator. *Vis. Comput.* **36**(6), 1203 (2020). <https://doi.org/10.1007/s00371-019-01732-4>
- Musse, S.R., Cassol, V.J., Thalmann, D.: A history of crowd simulation: the past, evolution, and new perspectives. *Vis. Comput.* (2021). <https://doi.org/10.1007/s00371-021-02252-w>
- He, Z., Li, Q., Feng, H., Xu, Z.: Fast and sub-pixel precision target tracking algorithm for intelligent dual-resolution camera. *Vis. Comput.* **36**(6), 1157 (2020). <https://doi.org/10.1007/s00371-019-01724-4>
- Bagheri Baba Ahmadi, S., Zhang, G., Wei, S., Boukela, L.: An intelligent and blind image watermarking scheme based on hybrid SVD transforms using human visual system characteristics. *Vis. Comput.* **37**(2), 385 (2021). <https://doi.org/10.1007/s00371-020-01808-6>
- Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* **29**(10), 983 (2013). <https://doi.org/10.1007/s00371-012-0752-6>
- Zhang, H., Hu, Z., Hao, R.: Joint information fusion and multi-scale network model for pedestrian detection. *Vis. Comput.* **37**(8), 2433 (2021). <https://doi.org/10.1007/s00371-020-01997-0>
- Khan, S.D., Basalamah, S.: Scale and density invariant head detection deep model for crowd counting in pedestrian crowds. *Vis. Comput.* **37**(8), 2127 (2021). <https://doi.org/10.1007/s00371-020-01974-7>
- Silveira, R., Dapper, F., Prestes, E., Nedel, L.: Natural steering behaviors for virtual pedestrians. *Vis. Comput.* **26**(9), 1183 (2010). <https://doi.org/10.1007/s00371-009-0399-0>
- Li, Z., He, S., Hashem, M.: Robust object tracking via multi-feature adaptive fusion based on stability: contrast analysis. *Vis. Comput.* **31**(10), 1319 (2015). <https://doi.org/10.1007/s00371-014-1014-6>
- Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5187–5196 (2019)
- Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural. Inf. Process. Syst.* **28**, 91 (2015)
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)
- Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
- Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
- Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 850–859 (2019)
- Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
- Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
- Singh, V.K., Kumar, N.: Saliency bagging: a novel framework for robust salient object detection. *Vis. Comput.* **36**(7), 1423 (2020). <https://doi.org/10.1007/s00371-019-01750-2>
- Xu, J., Cao, W., Liu, B., Jiang, K.: Object restoration based on extrinsic reflective symmetry plane detection. *Vis. Comput.* (2021). <https://doi.org/10.1007/s00371-021-02192-5>
- Wang, B., Chen, S., Wang, J., Hu, X.: Residual feature pyramid networks for salient object detection. *Vis. Comput.* **36**(9), 1897 (2020). <https://doi.org/10.1007/s00371-019-01779-3>
- Shu, C., Ding, X., Fang, C.: Histogram of the oriented gradient for face recognition. *Tsinghua Sci. Technol.* **16**(2), 216 (2011)
- Yan, J., Lei, Z., Wen, L., Li, S.Z.: The fastest deformable part model for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2497–2504 (2014)
- Rothe, R., Guillaumin, M., Van Gool, L.: Non-maximum suppression for object detection by passing messages between windows. In: Asian Conference on Computer Vision, pp. 290–306. Springer (2014)
- Papageorgiou, C.P., Oren, M., Poggio, T.: A general framework for object detection. In: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), pp. 555–562. IEEE (1998)
- Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627 (2009)
- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. *Int. J. Comput. Vision* **128**(2), 261 (2020)
- 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, pp. 580–587 (2014)
- Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
- Duan, K., Xie, L., Qi, H., Bai, S., Huang, Q., Tian, Q.: Corner proposal network for anchor-free, two-stage object detection. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 399–416. Springer (2020)
- Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
- Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 International Conference on Computer Vision, pp. 2018–2025. IEEE (2011)
- Liu, Y., Zhang, Y.M., Zhang, X.Y., Liu, C.L.: Adaptive spatial pooling for image classification. *Pattern Recogn.* **55**, 58 (2016)

37. Wang, S., Cheng, J., Liu, H., Tang, M.: Pcn: Part and context information for pedestrian detection with cnns. arXiv preprint [arXiv:1804.04483](https://arxiv.org/abs/1804.04483) (2018)
38. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6995–7003 (2018)
39. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint [arXiv:1703.01780](https://arxiv.org/abs/1703.01780) (2017)
40. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3127–3136 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Le Zhang is a graduate student in School of Electronic Information Engineering from Tianjin University, Tianjin, China. His current interests are computer vision and machine learning.



Tao Zhang received the M.S. degree in School of Electronic Information Engineering from Tianjin University, Tianjin, China, in 2001, the Ph.D. degree from the same University in 2004. He is currently an associate professor of the School of Electrical and Information Engineering, Tianjin University, China. His current interests include image, video, and acoustic signal processing, auditory model, speech enhancement and hardware/software partitioning.



Xuan Li received the Ph.D. degree in School of Electronic Information Engineering from Tianjin University, Tianjin, China. Her current interests are person re-identification, machine learning, computer vision and evolutionary computation.



Yahui Cao is a Ph.D. in School of Electronic Information Engineering from Tianjin University, Tianjin, China. Her current interests are evolutionary computation and machine learning.