A Novel Efficient Dual-Gate Mixed Dilated Convolution Network for Multi-Scale Pedestrian Detection

Etikala Raja Vikram Reddy
Agnel Charities' Fr. Conceicao Rodrigues Institute of Technology, India
vicky.raj43@gmail.com (corresponding author)

Sushil Thale
Agnel Charities' Fr. Conceicao Rodrigues Institute of Technology, India
sushil.thale@fcrit.ac.in

Received: 29 August 2023 | Revised: 13 September 2023 | Accepted: 16 September 2023

ABSTRACT

With the increasing use of onboard high-speed computing systems, vehicle manufacturers are offering significant advanced features of driver assistance systems. Pedestrian detection is one of the major requirements of such systems, which commonly use cameras, radar, and ultrasonic sensors. Image recognition based on captured image streams is one of the powerful tools used for the detection of pedestrians, which exhibits similarities and distinguishing features compared to general object detection. Although pedestrian detection has advanced significantly along with deep learning, some issues still need to be addressed. Pedestrian detection is essential for several real-world applications and is an initial step in outdoor scene analysis. Typically, in a crowded situation, conventional detectors are unable to distinguish persons from each other successfully. This study presents a novel technique, based on the Dual Gate Mixed Dilated Convolution Network, to address this problem by adaptively filtering spatial areas where the patterns are still complicated and require further processing. The proposed technique manages obscured patterns while offering improved multiscale pedestrian recognition accuracy.

Keywords-deep learning; image recognition; mixed dilated convolution; multiscale pedestrian recognition; spatial regions

I. INTRODUCTION

Several computer-vision applications, including autonomous vehicles, person identification, and surveillance, depend on the ability to identify pedestrians. Pedestrian detection has been substantially improved by recent deep Convolutional Neural Network (CNN)-based approaches [1-5]. Since pedestrians regularly block each other, this problem remains challenging. Detectors must be able to recognize various items in a crowded area but also detect pedestrians even when they are largely hidden. According to [6-10], the identification of human body parts can help modern detection models cope with occlusion. Compared to other methods, deep learning methods show better performance and a high level of accuracy [11]. In many areas, such as the healthcare sector, CNNs are the most widely used deep learning methods [12]. In [13], a deep-learning CNN was trained to predict crystal orientation. Although the effectiveness of the current approaches has improved, the question of how to fit such challenging pedestrian structures correctly and specifically achieve the necessary detection parameters is still a challenge. Most detectors perform poorly because they evaluate pedestrian occurrences uniformly using identical parameters, which is in a way insufficient for those with sophisticated designs. Methods based on the Feature Pyramid Network (FPN) integrate high-level semantics with low-level facts, making them suitable for detecting smaller elements such as pedestrians. However, this requires lowering the spatial resolution to identify larger objects, which affects performance. Some pedestrian detection techniques use parallel branches for multiple scales, but only coarse-grained pattern-parameter matching can be employed due to fixed parameter sizes. Current multiscale detection techniques also result in higher computing costs and impaired real-time effectiveness.

High-performance pedestrian recognition relies on extracting discriminative features from local patterns. Local image descriptors or pre-trained CNN backbones are used to identify the local patterns, while powerful classifier layers extract high-level semantic patterns. The processing of challenging patterns requires additional traits such as multiple scales or occlusion. In [14], a zoom-in-zoom-out component and a scale-sensitive pedestrian attention mask were suggested to improve the ability of feature maps to recognize small-sized pedestrians. In [15], a novel viewpoint was provided, in which pedestrian detection was driven as a rising semantic feature identification job using simple convolutions for the center but also scale forecasts. In [16], a new two-stream CNN model was proposed for semantic segmentation that explicitly connects shape information as a distinct processing branch that examines input simultaneously with the standard stream.

W3Net [17] was proposed to respond to these challenges, dividing the pedestrian detection task into where, what, and whether subtasks that target pedestrian localization, scale prediction, and classification, respectively. AugFPN [18] is an innovative feature pyramid network created to maximize the potential of multiscale features by adding three straightforward but valuable elements: Consistent supervision, residual feature augmentation, and soft Region-of-Interest (RoI) selection. In [19], the part-aware multi-scale fully convolutional network was proposed to handle occlusion and wide-scale variation concerns. As a result, a pedestrian occurrence that is only partially visible may receive a good recognition confidence score, reducing the chances of not being noticed. In [20], a gated multi-layer convolutional feature extraction network was proposed to improve the accuracy of pedestrian detection, using a squeeze unit to reduce the feature dimension before the gated unit. In [21], a prior-based receptive field block was presented, which directs the network to focus on the pedestrian by taking into account the pedestrian's shape prior, in addition to a bidirectional feature enhancement module that improves semantic features and localization information. In [22], an innovative detection technique was introduced, based on adaptive pattern-parameter matching. The Pattern Disentangling Module (PDM) first divides input pedestrian sequences, particularly complicated ones, into smaller structures to detect heads. The Gated Path Selection Network (GPSNet) [23] aims to learn adaptive receptive fields. The SuperNet two-dimensional multiscale network, which tightly integrates traits from growing receptive fields, serves as the basis for GPSNet.

In contrast to the methods mentioned above, this study proposes a method using a gating mechanism for pixel-level feature quality in addition to several branches of dilated convolution to retain spatial resolution.

II. THE PROPOSED METHOD

Pedestrian detection is the initial step in outdoor scene analysis. Even after utilizing the advantages of deep learning algorithms from generic object sensors, strong occlusion and dense crowding make it extremely difficult to recognize pedestrians. Occlusion patterns can also be addressed through part-aware feature extraction. Combining multiscale feature maps to collect additional information is an efficient method for multiscale pedestrian identification. As an alternative, this study created a unique gating-based module, termed the Dual Gate-Mixed Dilated Convolution Network, which adaptively filters spatial regions where the patterns are always complex but need to be processed further. A novel ResNet-50-based dual-gating technique was used to reduce model complexity at run-time. Dual gating determines the key characteristics of each convolutional block along the spatial and channel dimensions, allowing dynamic skips of superfluous channels and unimportant sections. FPNs address multiscale variation difficulties in object detection with minimal processing load. Mixed Dilated Convolution (MDC) is used to expand feature maps for better pedestrian identification in long distances, addressing obscured patterns and improving multiscale pedestrian recognition accuracy. Figure 1 shows the architecture of the proposed scheme.

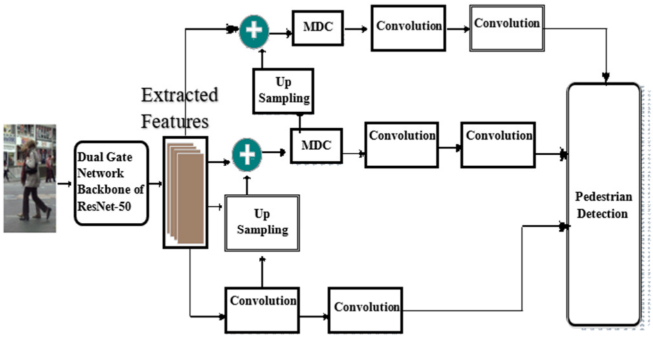

Fig. 1. Architecture of the proposed scheme

A. Dual-Gate Network

Figure 2 shows the basic dual-gating. Every convolutional block's spatial and channel-gating modules anticipate the informative features in two distinct dimensions, using the intermediate feature maps. Thus, unnecessary computations may be avoided during execution.

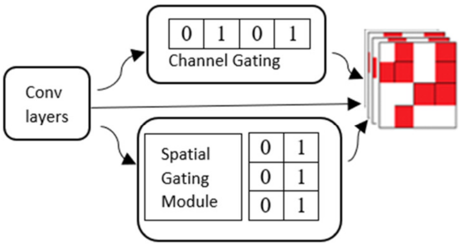

Fig. 2. Basic model for dual gate network.

1) Spatial Gating Module

Spatial sparsity is typically present in feature maps. According to [24], the backdrop information in shots has a smaller impact on the results, showing that not all geographical locations have the same significance. The spatial gating component aims to calculate the informative zones throughout the spatial dimension. Figure 3 shows the organizational layout of the training spatial gating component. In the first step, an Adaptive Average Pooling operation (AdaAvgPooling) is used to aggregate the local data of the input feature maps, but then a conventional 3×3 ResNet is applied to yield a 2D spatial attention map.

$$A_{s1} \in R\left[\frac{h_l+1}{t}\right] \times \left[\frac{w_l+1}{t}\right],$$

$$A_{s1} = f_{3\times3}(AdaAvgPooling(X_l)) \qquad (1)$$

where $f_{3\times3}$ represents the 3×3 ResNet, and A_{s1} (i,j) reflects the significance of the tile at position (i,j) on the final feature map. In this scenario, the use of a tile-based mask was decided for two reasons. Based on the spatial attention A_{s1}, the execution of those unimportant tiles can be immediately concealed during the inference phase. It is possible to write the binary spatial gating mask M_{s1} as follows:

$$M_{s1}(i,j) = \begin{cases} 1 & A_{s1}(i,j) \geq 0 \\ 0 & \text{Otherwise} \end{cases} \qquad (2)$$

Now, the discrete binary masks are relaxed to continuous variables using the Gumbel-Softmax re-parameterization algorithm [25]. Specifically, given the spatial attention A_{s1}, it is possible to define the probability that the spatial tiles will be executed $P_{s1}^{1} = \sigma(A_{s1})$, where σ is the sigmoid function. The probability that the tiles will not be implemented is $P_{s1}^{0} = 1 - \sigma(A_{s1})$. The spatial gating M_{s1} may therefore be represented by probabilistic binary random variables $(M_{s1}(i,j) = 1) = P_{s1}^{1}(i,j)$, and then, the sampling process of M_{s1} can be reparametrized as:

$$M_{s1} = \arg\max\left(\log(P_{s1}^{k}) + gk\right), \ \ \forall k = 0{,}1 \qquad (3)$$

where $\{\ gk\}k = \{0{,}1\}$ are independent and identically distributed random variables that follow the Gumbel distribution. The Gumbel-Softmax approach substitutes a softmax for the arg max since the latter is not continuous. The variational sample M'_{s1} from the Gumbel-Softmax relaxation can be stated for such a binary special case as follows:

$$M'_{s1} = \frac{\exp\left(\frac{\log\left(P_{s1}^{1}\right)+g1}{T}\right)}{\sum_{k\in\{0,1\}} \exp\left(\frac{\log\left(P_{s1}^{k}\right)+gk}{T}\right)} = \sigma\left(\frac{A_{s1}+g0-g1}{T}\right) \qquad (4)$$

The sampling process and the variation among the two Gumbels could be described as a Logistic distribution $g0 - g1 = \log U - \log(1-U)$, where $U \sim \text{Uniform}\ (0{,}1)$. Therefore:

$$M'_{s1} = \sigma\left(\frac{A_{s1}+\log U-\log(1-U)}{T}\right) \qquad (5)$$

where T is the softmax's temperature, which regulates how the softmax and argmax functions differ from one another. In this study, the formula T =2/3 was used, as given by [25]. A step function was added to M'_{s1} to get a binary mask all through the forward pass and use the continuous M'_{s1} to generate the gradient during the backward pass. Figures 3 and 5 show the features of the training's spatial gating and the channel gating components, respectively.

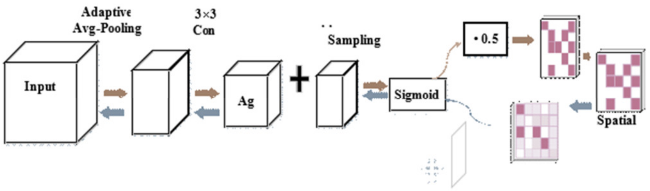

Fig. 3. Spatial gating module.

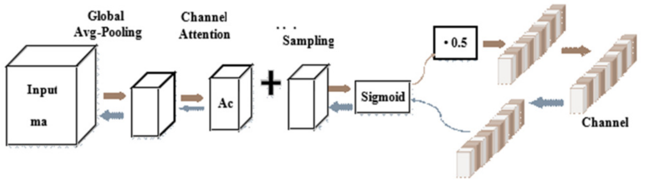

Fig. 4. Channel gating module.

2) *Channel Gating Module*

Since a channel's saliency varies depending on the input, dynamically choosing essential channels to execute is a promising way to reduce computation while retaining as much of the model's representational power as possible. Consequently, this study proposed and developed a channel gating module to detect superfluous channels that can be ignored during inference using the input pictures. Figure 4 shows the structural layout of the channel gating module. Initially, a global average pooling method was used to aggregate the spatial data from the feature maps to produce a context descriptor (*GlbAvgPooling*). To gain channel attention $A_{c1} \in R^{c_l+1}$, the descriptor is sent to a lightweight network next. The channel attention network, which is similar to the SE block [26], is made up of two consecutive fully connected layers that have *c*/*r* neurons in the hidden layer to minimize computations, where *r* is the reduction ratio. The channel attention is illustrated as follows:

$$A_{c1} = W_1 * \delta[\text{norm}(W_0 * GlbAvgPooling(X_l))] \qquad (6)$$

where $W_0 \in R^{c_l+1\times\frac{c_l}{r}}$, δ is the ReLU function, and norm represents the batch normalization. In this study's experiments, *r* was set to 4. Then the channel attention is employed to construct the binary channel mask M_{c1} during inference, much like with the spatial gating module. The channel gating module additionally employs the Gumbel-Softmax relaxation during the training phase to provide the continuous mask M'_{c1} for back-propagation. The channel gating module may also be linked with ResNets to fully understand the crucial channels. Figure 4 shows how the dual gating was integrated into the leftover blocks. The inference paths are indicated by dashed arrows.

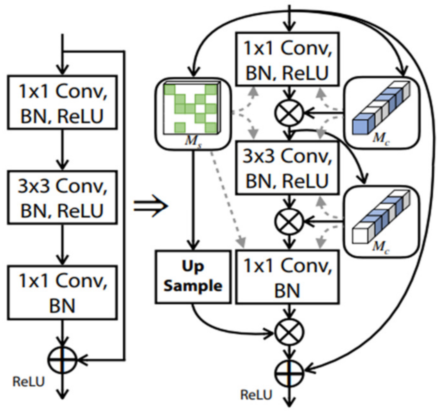

Fig. 5. Integration of dual gating into the residual blocks.

B. Mixed Dilated Convolution

Although the human head with shoulder region has low form variation and is relatively stable, it is a small target in comparison to other objects for a single image and the pixel per inch ratio is lower. The model's shortcomings in handling multi-scale variation issues in object detection tasks are mostly addressed by FPNs [27]. The dilated convolution [28] may also extract various feature maps with various semantic information at various dilated rates. In the FPN phase of Dual-Gate ResNet-50, the context information is again fused using mixed dilated convolution to increase the receptive field of the feature maps and improve the recognition of small objects. In [29], it was shown that mixed dilated convolution creates a gridding impact as a result of the dilated rate stacking, resulting in some missing pixels and a break in information continuity. To implement the dilated rate stacking, the following equation must be fulfilled:

$$M_i = \max[M_{i+1} - 2r_i, M_{i+1} - r_i, r_i] \qquad (7)$$

where r_i is the dilated rate of layer i and M_i is the maximum dilated rate of layer i. Assuming that there are n layers in total, to satisfy $M_n = r_n$, a simple example is r = 1, 2, 4. As a result, the proposed MDC architecture is made up of dilated rates r = 1, 2, 4. Figure 5 shows the MDC structure of the module after the initial FPN fusion. In this MDC module, three distinct dilated convolution rates are used in parallel, with the magnitude of the dilated rate reflecting the diameter of the receptive field. Initially, the amount of feature map channels $C \in \mathbb{R}^{W\times H\times L}$ are decreased after feature fusion by convolution, a feature map $C_1 \in \mathbb{R}^{H\times w\times L}$ is generated. The dilated convolution layers are again sampled on the feature map C_1 at various dilated rates (r=1, 2, 4) to obtain $C_2, C_3, C_4 \in \mathbb{R}^{H\times W\times L}$. Finally, C_1, C_2, C_3, C_4 are connected to attain the feature map $M \in \mathbb{R}^{H\times W\times L}$, $M = [C_1, C_2, C_3, C_4]$. The model may take in information from a range of receptive fields, which can then be merged to extract specialized semantic information, particularly feature information for small targets.

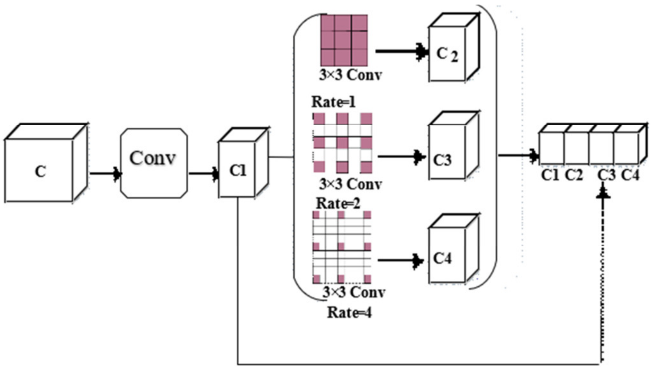

Fig. 6. Structure of MDC

In front of the Dual-Gate ResNet-50, two MDC modules are interconnected. When the connectivity of the spatial feature pyramid is completed, the four contextual data-aware module features are fused using MDC and sent to Dual-Gate ResNet-50 for identification. The picture features are removed using DR-Net, resulting in feature maps with more in-depth semantic data. The MDC module samples and merges the feature maps with dilated convolutions of varied dilated rates to broaden the receptive field as well as utilize finer-grained feature data.

III. RESULTS AND DISCUSSION

A. Dataset Description

The proposed method was implemented on a large-scale dataset named CityPersons [30].

B. Confusion Matrix

Figure 6 shows the confusion matrix for the proposed method. Predicted values are shown on the X-axis, while actual values are shown on the Y-axis. The confusion matrix contains four features. TN, FP, TP, and FN. The selected values for TN, TP, FP, and FN were 74, 47, 11, and 28, respectively.

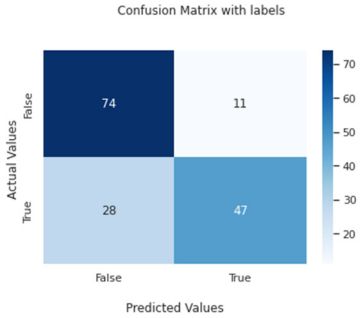

Fig. 7. Confusion matrix.

C. Precision-Recall

Figure 7 shows the precision-recall curve, determined based on the network's output.

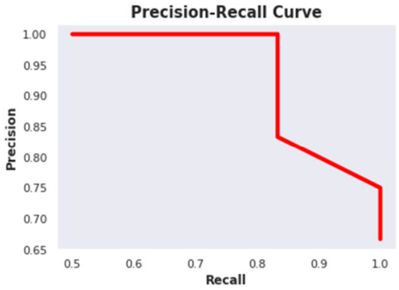

Fig. 8. Precision-Recall curve.

Every object's class AP is determined using the Precision and Recall curve. The percentage of all ground truths that are derived from true positives is known as recall. The ratio of all forecasts which are precise to positives is known as precision.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (9)$$

An average precision value of 95.83% was achieved using the proposed approach, with a mean of 50%.

D. Log-average Miss Rate

The log-average Miss Rate (MR), which is somewhat parallel to recall, is used to determine how many undetected objects there are. MR is the ratio of False Negatives (FN) to Ground Truth (GT) in the dataset:

$$\mathrm{MR} = \frac{\mathrm{FN}}{\mathrm{GT}} \tag{10}$$

MR was found to be 20%.

IV. COMPARISON RESULTS

A. Training and Validation Loss

Figure 8 shows the training and validation loss of the proposed method. The blue line signifies the training loss and the red line denotes the validation loss. The validation loss shows how effectively the model fits new information while the training loss shows how well it matches training data. For the proposed strategy, the validation loss was substantial compared to the training loss.

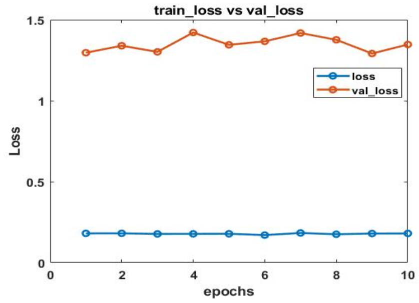

Fig. 9. Training and validation loss.

B. Training and Validation Accuracy

Figure 9 shows the training and validation accuracy of the proposed method, displayed by the blue and red lines, respectively. The accuracy of the training set was greater than that of the validation set.

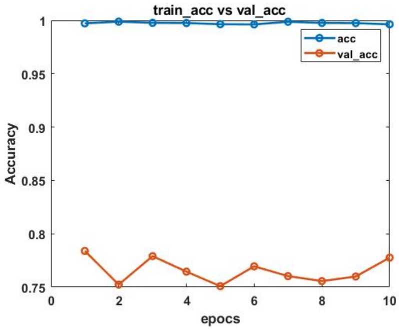

Fig. 10. Training and validation accuracy

C. Image Classification

Figure 10 shows how the proposed approach was visually represented. In the first image, the people in the city are hidden while the second shows the results of the proposed approach in the identification of the people.

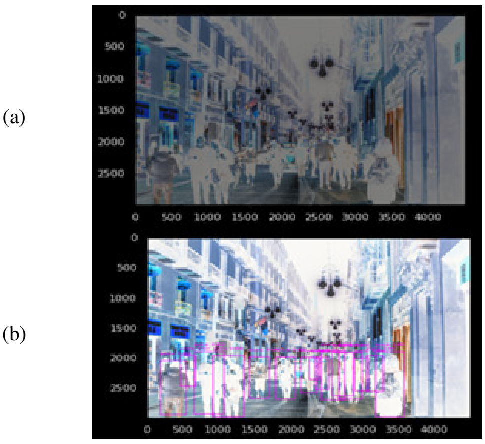

Fig. 11. Picture visualization: (a) City persons, (b) proposed method.

D. Comparison of Methods

Table I shows the results of the proposed and three other existing methods. The MR values were significantly lower than those of the other methods. On an Intel Pentium processor with 8GB of RAM, the processing speeds for Faster RCNN, SSD, and YOLOv3 were 1.7, 2.4, and 3.8 fps, respectively. The proposed method had an increase in processing speed of 2.9 fps. The results show that the proposed method performed significantly better than the others.

TABLE I. METHODS AND MISS RATE VALUES

Method	Reasonable	Reasonable_ small	Reasonable_ occ= heavy	All
Proposed	5.02 %	7.16%	26.89%	28.72%
APDpretrain [31]	7.31%	10.81%	28.07%	32.71%
Pedestron [32]	7.69%	9.16%	27.08%	28.33%
DVRNet [33]	11.17%	15.62%	42.52%	40.99%

V. CONCLUSION

This study strategically deployed the ResNet-50 Dual Gating Technique to streamline the model's runtime complexity. The proposed method harnesses the power of end-to-end trained spatial and channel gating components, allowing seamless integration with widely adopted CNN architectures. This integration serves the crucial purpose of eliminating superfluous computations during model execution, thereby enhancing efficiency and accelerating inference times. One of the significant advancements of the proposed design was the resolution of issues stemming from feature fusion in FPN through the innovative implementation of MDC. This strategy not only addresses existing challenges but also presents a more elegant and effective way to handle feature fusion, thereby contributing to the overall model's superior performance.

Compared to previous approaches, the results of the proposed method were truly remarkable. Training loss was minimized, showcasing the model's exceptional learning capabilities. Moreover, the accuracy reached impressive heights, confirming the effectiveness of the ResNet-50 Dual Gating technique. Another noteworthy achievement was observed in MR, which showed significant improvements over competing methods. This reduction in MissRate signifies that the proposed approach excels in correctly identifying pedestrians on various scales and challenging scenarios. The ResNet-50 Dual Gating technique, coupled with MDC, promises to revolutionize the field by optimizing model efficiency and robustness, setting a new standard for pedestrian recognition systems.

REFERENCES

[1] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion Loss: Detecting Pedestrians in a Crowd," presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 7774–7783, https://doi.org/10.1109/CVPR.2018.00811.

[2] G. Brazil, X. Yin, and X. Liu, "Illuminating Pedestrians via Simultaneous Detection and Segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Jul. 2017, pp. 4960–4969, https://doi.org/10.1109/ICCV.2017.530.

[3] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 966–974, https://doi.org/10.1109/CVPR.2018.00107.

[4] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV*, Munich, Germany, Jun. 2018, pp. 643–659, https://doi.org/10.1007/978-3-030-01264-9_38.

[5] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What Can Help Pedestrian Detection?," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6034–6043, https://doi.org/10.1109/CVPR.2017.639.

[6] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What Can Help Pedestrian Detection?," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Venice, Italy, Jul. 2017, pp. 6034–6043, https://doi.org/10.1109/CVPR.2017.639.

[7] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep Learning Strong Parts for Pedestrian Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Sep. 2015, pp. 1904–1912, https://doi.org/10.1109/ICCV.2015.221.

[8] S. Zhang, J. Yang, and B. Schiele, "Occluded Pedestrian Detection Through Guided Attention in CNNs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 6995–7003, https://doi.org/10.1109/CVPR.2018.00731.

[9] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd," in *Computer Vision – ECCV 2018*, Munich, Germany, 2018, pp. 657–674, https://doi.org/10.1007/978-3-030-01219-9_39.

[10] S. Wang, J. Cheng, H. Liu, and M. Tang, "PCN: Part and Context Information for Pedestrian Detection with CNNs." arXiv, Apr. 12, 2018, https://doi.org/10.48550/arXiv.1804.04483.

[11] D. Patil and S. Jadhav, "Road Segmentation in High-Resolution Images Using Deep Residual Networks," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9654–9660, Dec. 2022, https://doi.org/10.48084/etasr.5247.

[12] S. Rani, Y. Chabrra, and K. Malik, "An Improved Denoising Algorithm for Removing Noise in Color Images," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8738–8744, Jun. 2022, https://doi.org/10.48084/etasr.4952.

[13] D. K. Suker, "Deep Learning CNN for the Prediction of Grain Orientations on EBSD Patterns of AA5083 Alloy," *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8393–8401, Apr. 2022, https://doi.org/10.48084/etasr.4807.

[14] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-Aware Deep Feature Learning for Robust Pedestrian Detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3820–3834, 2020, https://doi.org/10.1109/TIP.2020.2966371.

[15] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5182–5191, https://doi.org/10.1109/CVPR.2019.00533.

[16] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated Shape CNNs for Semantic Segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Jul. 2019, pp. 5228–5237, https://doi.org/10.1109/ICCV.2019.00533.

[17] Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun, "Where, What, Whether: Multi-Modal Learning Meets Pedestrian Detection," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14065–14073.

[18] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving Multi-Scale Feature Learning for Object Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 12592–12601, https://doi.org/10.1109/CVPR42600.2020.01261.

[19] P. Yang, G. Zhang, L. Wang, L. Xu, Q. Deng, and M.-H. Yang, "A Part-Aware Multi-Scale Fully Convolutional Network for Pedestrian Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 1125–1137, Oct. 2021, https://doi.org/10.1109/TITS.2019.2963700.

[20] T. Liu, J. J. Huang, T. Dai, G. Ren, and T. Stathaki, "Gated Multi-Layer Convolutional Feature Extraction Network for Robust Pedestrian Detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Feb. 2020, pp. 3867–3871, https://doi.org/10.1109/ICASSP40776.2020.9054437.

[21] Y. Tan, H. Yao, H. Li, X. Lu, and H. Xie, "PRF-Ped: Multi-scale Pedestrian Detector with Prior-based Receptive Field," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, Jan. 2021, pp. 6059–6064, https://doi.org/10.1109/ICPR48806.2021.9412031.

[22] M. Liu, C. Zhu, J. Wang, and X. C. Yin, "Adaptive Pattern-Parameter Matching for Robust Pedestrian Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2154–2162, May 2021, https://doi.org/10.1609/aaai.v35i3.16313.

[23] Q. Geng, H. Zhang, X. Qi, G. Huang, R. Yang, and Z. Zhou, "Gated Path Selection Network for Semantic Segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 2436–2449, 2021, https://doi.org/10.1109/TIP.2020.3046921.

[24] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not All Pixels Are Equal: Difficulty-Aware Semantic Segmentation via Deep Layer Cascade," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6459–6468, https://doi.org/10.1109/CVPR.2017.684.

[25] C. J. Maddison, A. Mnih, and Y. W. Teh, "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables." arXiv, Mar. 05, 2017, https://doi.org/10.48550/arXiv.1611.00712.

[26] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, https://doi.org/10.1109/tpami.2019.2913372.

[27] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944, https://doi.org/10.1109/CVPR.2017.106.

[28] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions." arXiv, Apr. 30, 2016, https://doi.org/10.48550/arXiv.1511.07122.

[29] P. Wang *et al.*, "Understanding Convolution for Semantic Segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1451–1460, https://doi.org/10.1109/WACV.2018.00163.

[30] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A Diverse Dataset for Pedestrian Detection," in *2017 IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4457–4465, https://doi.org/10.1109/CVPR.2017.474.

[31] J. Zhang *et al.*, "Attribute-Aware Pedestrian Detection in a Crowd," *IEEE Transactions on Multimedia*, vol. 23, pp. 3085–3097, 2021, https://doi.org/10.1109/TMM.2020.3020691.

[32] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable Pedestrian Detection: The Elephant In The Room," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 11323–11332, https://doi.org/10.1109/CVPR46437.2021.01117.

[33] L. Shi, C. Livermore, and I. A. Kakadiaris, "DVRNet: Decoupled Visible Region Network for Pedestrian Detection," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Houston, TX, USA, Sep. 2020, pp. 1–9, https://doi.org/10.1109/IJCB48548.2020.9304883.