Adaptive Pattern-Parameter Matching for Robust Pedestrian Detection

Mengyin Liu, Chao Zhu*, Jun Wang, Xu-Cheng Yin
School of Computer and Communication Engineering
University of Science and Technology Beijing, Beijing, China
blean@live.cn, chaozhu@ustb.edu.cn, wj_fm0604@foxmail.com, xuchengyin@ustb.edu.cn

Abstract

Pedestrians with challenging patterns, e.g. small scale or heavy occlusion, appear frequently in practical applications like autonomous driving, which remains tremendous obstacle to higher robustness of detectors. Although plenty of previous works have been dedicated to these problems, properly matching patterns of pedestrian and parameters of detector, i.e., constructing a detector with proper parameter sizes for certain pedestrian patterns of different complexity, has been seldom investigated intensively. Pedestrian instances are usually handled equally with the same amount of parameters, which in our opinion is inadequate for those with more difficult patterns and leads to unsatisfactory performance. Thus, we propose in this paper a novel detection approach via adaptive pattern-parameter matching. The input pedestrian patterns, especially the complex ones, are first disentangled to simpler patterns by parallel branches in Pattern Disentangling Module (PDM) with various receptive fields. Then, Gating Feature Filtering Module (GFFM) dynamically decides the spatial positions where the patterns are still not simple enough and need further disentanglement by the next-level PDM. Cooperating with these two key components, our approach can adaptively select the best matched parameter size for the input patterns according to their complexity. Moreover, to further explore the relationship between parameter sizes and their performance on the corresponding patterns, two parameter selection policies are designed: 1) extending parameter size to maximum, aiming at more difficult patterns for different occlusion types; 2) specializing parameter size by group division, aiming at complex patterns for scale variations. Extensive experiments on two popular benchmarks, Caltech and CityPersons, show that our proposed method achieves superior performance compared with other state-of-the-art methods on subsets of different scales and occlusion types.

Introduction

In recent years, pedestrian detection has made great progress with the success of convolutional neural networks (CNNs) (Simonyan and Zisserman 2014; He et al. 2016), and has been widely applied in person re-identification, video surveillance and autonomous driving, etc. Nevertheless, especially in realistic circumstances, pedestrian detectors suffer an inevitable accuracy loss from challenging patterns, e.g. small scale or heavy occlusion. Despite of improved performance achieved by the existing approaches in the literature (Li et al. 2017; Lin et al. 2018; Liu et al. 2019; Huang et al. 2020), the problem of how to properly match these difficult pedestrian patterns and the required amount of detector parameters, i.e., constructing a detector with proper parameter sizes for certain pedestrian patterns of different complexity, has been seldom investigated intensively. Instead, most detectors treat pedestrian instances equally with the same parameter size, which in our opinion is inadequate for those with harder patterns, leading to unsatisfactory performance.

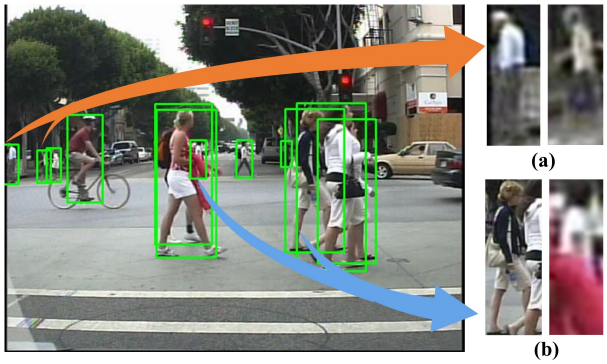

Figure 1: Sample image from Caltech pedestrian dataset. (a) Pedestrians of small scale. (b) Pedestrians with different occlusion types. Both of them possess more difficult patterns requiring more detector parameters to be detected.

For instance, as shown in Figure 1(a), in autonomous driving scenario, some small-scale pedestrians comprise blurry and noisy patterns, which require a detector with more parameters to discriminate them from other human-like small objects or background. Instead, large-scale pedestrians usually contain numerous sharper patterns, which is painless for detectors to distinguish them with far less parameters.

Among those methods tried matching patterns and parameters for multi-scale detection, FPN (Lin et al. 2017) introduces the lateral connection to fuse high-level semantics and low-level details, and provides more parameters for detecting smaller objects. Hence, it has been widely adopted in detecting objects with scale variations like pedestrians (Cao et al. 2019). Less spatial resolution by repeatedly down-

*Corresponding author

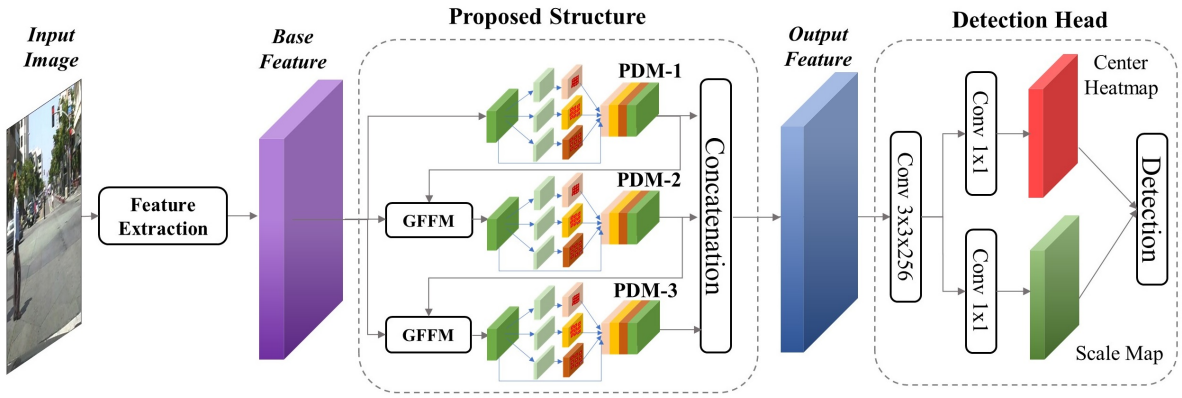

Figure 2: The overall architecture of our proposed AP^{2}M network. "Feature Extraction" and "Detection Head" blocks follow the way in CSP (Liu et al. 2019). The proposed PDMs disentangle the input pattern into simpler patterns by parallel branches with various receptive fields, corresponding with patterns of various object scales or occluded areas. The proposed GFFMs adaptively filter the spatial position where the patterns are still not simple enough and need further disentanglement by the next-level PDM, according to the patterns from the previous-level PDM as well as the patterns from the base feature.

sampling, whereas, is inevitable in detecting larger objects in FPN, leading to higher performance loss (Guo et al. 2020). In a divide-and-conquer manner, some pedestrian detectors exploit parallel branches to handle different scales separately, e.g. GADF (Lin et al. 2018) and SAF-RCNN (Li et al. 2017). Yet only coarse-grained pattern-parameter matching are applied due to the fixed parameter size on each branch.

Meanwhile, as a result of the annotation protocol of mainstream pedestrian detection datasets, pedestrian bounding boxes often have some other objects inside when occlusion occurs, as illustrated in Figure 1(b). Consequently, the detectors are forced to classify other instances or objects as a part of current pedestrian. To handle such complex patterns, different parameter sizes have to be considered. Quantitate analysis has also been conducted by Cao et al. (2019) that occluded areas are diversified due to the scales and visible ratios. Different sizes of occluded areas also require different parameter sizes to match the patterns. Currently, most of the occlusion handling strategies mainly focus on part-aware feature extraction and special training or testing scheme, e.g. OR-CNN (Zhang et al. 2018), RepLoss (Wang et al. 2018) and PBM+R^{2}NMS (Huang et al. 2020), which treat occluded patterns equally with the same parameter size.

Moreover, there are also detectors handling both types of hard patterns. CSP (Liu et al. 2019) embraces anchor-free detector with multi-scale feature fusion, which is a "Entanglement" that prosesses the fused feature by a single branch with fixed receptive field and shared parameters, thus providing a mixture of easy/still-hard-handled patterns to detection head. A "Disentanglement" design is able to improve the performance on still-hard-handled patterns.

In conclusion, we have observed that: adaptive pattern-parameter matching problem is remained to be comprehensively investigated for both scale variations and occlusions. Therefore, we propose a novel detection approach to adaptively match pedestrian patterns and detector parameter sizes in manifold cases. Our main contributions are listed below:

- We introduce a novel network named **AP^{2}M** for **A**daptive **P**attern-**P**arameter **M**atching on patterns of different occluded areas and object scales. Specifically, Pattern Disentangling Module (PDM) adopts parallel branches with various receptive fields, processing corresponding areas of hard patterns into easier ones for detection head as a "Disentanglement" in a divide-and-conquer manner.
- Multi-level gating mechanism is designed to dynamically filter the spatial positions where the patterns are still not simple enough and need further disentanglement by the next-level PDM. Therefore, it makes the proposed method capable of adaptively selecting the best matched parameter size for input patterns according to their complexity.
- To further explore the relationship between parameter sizes and their performance on certain patterns, two parameter selection policies are designed based on Vanilla PDM: 1) Extended Parameter Policy maximizes parameter size to tackle more difficult patterns for different occlusion types; 2) Grouped Parameter Policy obtains specialized parameter size by group division, aiming at complex patterns for scale variations. Complementing their advances, our model achieves new state-of-the-art performance in the Caltech and CityPersons benchmarks.

Related Works

Extracting discriminative patterns serve as one of the essential elements for high performance pedestrian detection. Generally speaking, from a given image or video frame, local patterns are distinguished by local image descriptors (Ojala, Pietikäinen, and Mäenpää 2000; Dalal and Triggs 2005) or pre-trained CNN backbones, (Simonyan and Zisserman 2014; He et al. 2016). Cooperatinging with powerful classifiers e.g. SVM (Smola and Schölkopf 1998) or deeper layers of CNNs, patterns with high-level semantics are extracted from local ones to facilitate detection. Among these patterns the hard ones require more parameters to process, e.g. pedestrians with small scales, heavy occlusion or even a

mixture. Such a relationship between patterns of pedestrian and parameter sizes of detector can be regarded as "pattern-parameter matching" problem.

Pedestrian Detection with Hard Pattern Handling

Multi-Scale Pattern Handling Some pedestrian detectors are devoted to higher accuracy on the patterns of scale variations. In a divide-and-conquer manner, Li et al. (2017) proposes an adapted version of Faster R-CNN (Ren et al. 2015) for handling the smaller and larger scale by separate branches. GADF (Lin et al. 2018) performs feature extraction from different levels of backbone network.Although parallel branches are effective on scale variations, they can only maintain a coarse-grained pattern-parameter matching due to the fixed parameter size on each branch. Differently, some approaches (Cao et al. 2019) are based on the FPN (Lin et al. 2017)that provide more parameters for detecting smaller objects with longest bottom-up and top-down streams. However, the FPN-like networks detect larger objects with repeatedly down-sampling operation, leading to certain performance loss (Guo et al. 2020).In comparison, our proposed method adopts not only multiple branches of dilated convolution (Yu and Koltun 2015) to keep spatial resolution but also gating mechanism for adaptive pattern-parameter matching.

Occlusion Pattern Handling Occlusion pattern handling has become another research hotpot in pedestrian detection. Some researchers have created new training or testing schemes. Wang et al. (2018) propose RepLoss that attracts predicted boxes closer to their ground truth and repulses others. Huang et al. (2020) adopts a new post-processing algorithm to preserve highly overlapped pedestrians. Another solution to occlusion pattern is part-aware feature extraction. Occlusion-aware RCNN (Zhang et al. 2018) extracts feature by body parts of every proposal and re-scores the original proposal to highlight the visible parts. PAMS-FCN (Yang et al. 2020) introduces parallel branches for larger and smaller objects with part-aware RoI Poolings. However, most of these approaches treat occluded patterns equally with the same parameter size.

Generic Hard Pattern Handling The central key to handling hard patterns of scale variation and occlusion is accurate localization. ALFNet (Liu et al. 2018), AR-Ped (Brazil and Liu 2019) and PRNet (Song et al. 2020) introduce multi-phase regression refinement. Employing multi-scale feature fusion, TLL (Song et al. 2018) is trained by a novel type of ground truth based on somatic topological line of pedestrian. Following the former one, CSP (Liu et al. 2019) further embraces the popular anchor-free detector with less hyper-parameters for anchor setting and more flexibility in estimating the scale and aspect ratio of bounding boxes. LBST (Cao et al. 2019) further introduces finer regression for both smaller scale and occlusion problem.

Unfortunately, progressive refining strategies are lacking in filtering the location for further processing. And multi-scale feature fusion approaches provides a mixture of easy and hard patterns to detection head without pattern-parameter matching and causes potential performance loss.

Gating Mechanism

As an effective and explicit strategy to improve the pixel-wise quality of feature, gating mechanism is firstly used in segmentation and then promoted to other tasks. GPSNet (Geng et al. 2020) adopts the gate mechanism to adaptively select pattern passing path, assembling the best receptive field to segment current image. Gated SCNN (Takikawa et al. 2019) preserves shape information by gating mechanism for sharper segmentation. Liu et al. (2020) apply gating mechanism in pedestrian detector for scale variation and occlusion. Nevertheless, its gate modules are only applied in separate layers, which ignores the interactions between deeper high-level patterns and shallower low-level patterns that are claimed and evaluated by FPN.

Differently, we have managed to design a novel gating module that can adaptively filter spatial positions where the patterns are still complex and need further processing.

Proposed Method

The overall architecture of our proposed **A**daptive **P**attern-**P**arameter **M**atching (**AP^{2}M**) network is illustrated in Figure 2, which is an anchor-free detection framework following the way as in the baseline CSP (Liu et al. 2019). A pre-trained CNN processes the input image into features at different depth. They are concatenated into "Base Feature", a mixture of patterns with various complexities which is processed with fixed parameter size in CSP.

In order to achieve adaptive pattern-parameter matching, our architecture consists of two key components: Pattern Disentangling Module (PDM) and Gating Feature Filtering Module (GFFM). PDM is designed for disentangling the input patterns (especially the complex ones) into simpler patterns, corresponding with various scales or occluded areas. GFFM is designed for sieving the input features and selecting the locations where require more parameters to process. These two modules cooperate to adaptively match different patterns and parameter sizes. More details will be introduced in the following sections.

Pattern Disentangling Module

An excellent strategy for dealing with decomposable problems of high complexity as it is, divide-and-conquer separates the whole problem into sub-problems that easier to be solved. Similarly, the feature extraction network of a detector should process a mixture of harder patterns with different areas into easier ones for detection head to predict.

Given the input feature $\mathbf{T}$ from shallower level of the backbone network, e.g. in urban driving scenario, $\mathbf{T}$ is likely to comprise a mixture of complex patterns of pedestrian with different scales or sizes of occlusion area, denoted as $\{\mathbf{t}_i|\mathbf{t}_i \in \mathbf{T},\ i \in \mathbf{P}\}$ where $\mathbf{P}$ represents all the locations of the input feature map. The aim of our proposed Pattern Disentangling Module (PDM) is to disentangle such a mixture $\mathbf{T}$ into a series of simpler patterns $\mathbf{t}_i$.

As is shown in Figure 3(a), after the first convolution layer for down-sizing input channel size, a series of branches with multiple receptive fields are introduced into the Vanilla PDM (V-PDM), on the basis of dilated convolution and identity

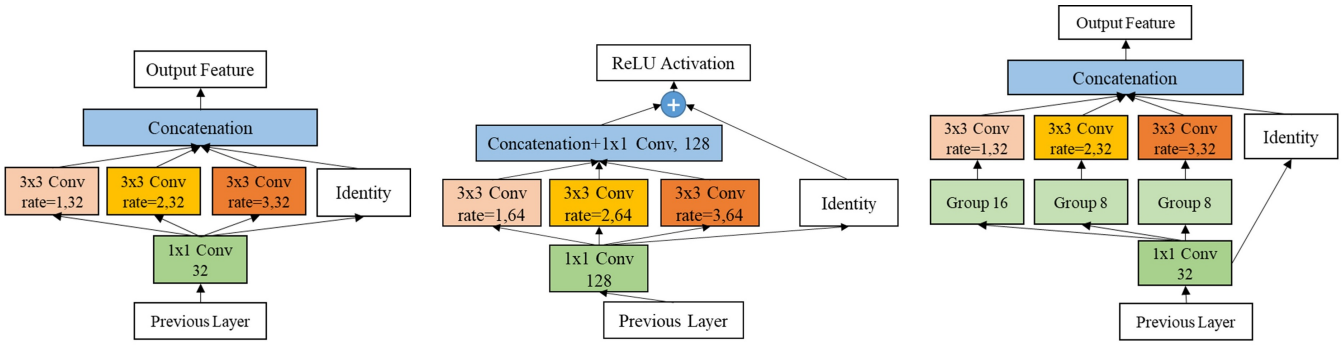

Figure 3: Pattern Disentangle Modules. "*k*x*k* Conv rate=*r*, *c*" means the convolution with dilation rate *r*, kernel size *k* and channel *c*. "group *g*" means dividing feature into a *g*-channel group. Extended and Grouped PDM are under two different parameter selection policies based on Vanilla PDM.

connection (dilation rate=0), the former of which are capable of enlarging receptive field, free from information loss by down-sampling in FPN.

With dilation rates $\{0,\ 1,\ 2,\ 3\}$, receptive field sizes $\{0\times 0,\ 3\times 3,\ 5\times 5,\ 7\times 7\}$, each branch j only processes pattern at a specific spatial range, corresponding with a range of object scales or occluded areas. Finally, the disentangled pattern $\mathbf{t}_i^j$ on branch j and position i denoted as:

$$\mathbf{t}_i^j = \delta\left(\mathbf{W}_r^j\left(\delta\left(\mathbf{W}_d\left(\mathbf{t}_i\right)+\mathbf{b}_d\right)\right)+\mathbf{b}_j\right) \tag{1}$$

where $\mathbf{W}$, $\mathbf{b}$ are marked as the weights and biases of convolution layers, $\mathbf{W}_d \in \mathbb{R}^{1\times 1\times C\times \phi}$, $\mathbf{W}_r^j \in \mathbb{R}^{3\times 3\times \phi\times \eta}$, $\mathbf{b}_d \in \mathbb{R}^{\phi}$, $\mathbf{b}_j \in \mathbb{R}^{\eta}$. C is the input channel, ϕ is the output channel size of the convolution layer for downsizing the C, and η is the output channel size of branch j with dilation rate r. The d marks the 1^{st} convolution layer that down-scales the channel size of input feature. δ refers to ReLU function (Hahnloser et al. 2000). The final output pattern $\mathbf{T}_{PDM}^*$ consists of $\mathbf{t}^j$ on each branch by concatenation:

$$\mathbf{T}_{PDM}^* = Concate\left(\left[\mathbf{t}^1, \mathbf{t}^2, ..., \mathbf{t}^j\right]\right) \tag{2}$$

However, for some very hard patterns, e.g. very small pedestrians or heavily occluded pedestrians, it is still inadequate even they are already disentangled once. Therefore, it is necessary to introduce gating mechanism, by which each pixel of these disentangled patterns will be decided whether to be passed to the next-level PDM for further processing, with the aid of our proposed GFFM, which will be described in detail in the following section.

Different Parameter Selection Policies for PDM

To further investigating the relationship between the amount of parameters and the complexity of patterns, we specially design two parameter selection policies based on V-PDM:

Extended Parameter Policy We consider that the patterns with different occlusion types are harder to be handled than non-occluded ones because detectors are forced to classify the other objects as part of current pedestrian instance, illustrated in Figure 1(b). Based on V-PDM, Extended PDM (E-PDM) is designed under this policy, as shown in Figure 3(b). The output channel size of the first convolution is increased to 128, those of every branch are raised up to 64, and an extra convolution is performed to decrease the output channel size from branches.

Grouped Parameter Policy We consider that the patterns with smaller scales are relatively more difficult than larger scale ones, due to their blurry and noisy information. Nor is this all, Table 3 shows the insufficient capability of V-PDM on small scale pedestrians. So it is reasonable to raise up the parameter size for smaller receptive fields. We keep equal channel size of the output from each branch, so that none of them will dominate the final output. Group division strategy is employed to the input features for different parameters, so that the input channel size 32 can be divided into $\{16,\ 8,\ 8\}$ for the branches with receptive fields $\{3\times 3,\ 5\times 5,\ 7\times 7\}$ respectively, and the new PDM under this policy is named "Grouped PDM" (G-PDM), as shown in Figure 3(c).

Since these polices are specialized to different types of hard patterns, to achieve the best performance on as more patterns as possible, it is reasonable for us to ensemble the advantages of two policies by fusing their detection results with NMS, marked as a final version of our proposed AP^{2}M.

Gating Feature Filtering Module

Being popular used in plenty of advanced vision tasks, gating mechanism generates a heatmap and re-weights some part of the input by an activation function. With the purpose of filtering the spatial locations where the patterns need further disentanglement by the next-level PDM, we design a simple but powerful gating module, named Gating Feature Filtering Module (GFFM).

As is shown in Figure 4, GFFM receives two kinds of input: disentangled pattern $\mathbf{T}_{PDM}$ from the previous level of PDM, and the original pattern named "Base Feature" $\mathbf{T}_{Base}$ as in Figure 2. The disentangled pattern from PDM contains richer semantic information, e.g. higher activation value towards certain scale or occlusion type, but considerable details in the original pattern are lost through the disentangle-

ment. Inspired by FPN that provides interaction of deeper high-level patterns and shallower low-level patterns, both the input patterns take equal part in generating the spatial heatmap $\mathbf{H} \in (0,\ 1)$, which helps GFFM make comprehensive decision on where to activate:

$$\mathbf{H} = \sigma\left(\mathbf{W}_g\left(Concate\left(\left[\mathbf{T}_{Base}, \mathbf{T}_{PDM}\right]\right)\right) + \mathbf{b}_g\right) \quad (3)$$

where $\mathbf{W}_g \in \mathbb{R}^{1\times1\times(C_B+C_P)\times1}$, $\mathbf{b}_g \in \mathbb{R}$. C_B and C_P is the channel size of $\mathbf{T}_{Base}$ and $\mathbf{T}_{PDM}$ respectively. The gate activation function σ is widely used Sigmoid function. Its input range is $(-\infty,\ +\infty)$ and output is $(0,\ 1)$, which is suitable to re-weight the patterns. Hence, the weighted patterns will be further disentangled in the next level PDM. Moreover, the original pattern is also combined with the activated disentangled patterns to generate the output $\mathbf{T}^*_{GFFM}$, providing the next level PDM with more detailed information:

$$\mathbf{T}^*_{GFFM} = Concate\left(\left[\mathbf{T}_{Base}, \mathbf{H}\cdot\mathbf{T}_{PDM}\right]\right) \quad (4)$$

For example, if the pattern $\mathbf{t}_\theta$ at position θ is very hard to disentangle, GFFM will multiply it with $\mathbf{h}_\theta \to 1$ at position θ inside heatmap $\mathbf{H}$ at each level. So $\mathbf{t}_\theta$ can be disentangled into simpler paterns sequentially with their total parameters. If the pattern $\mathbf{t}_\omega$ at position ω is extremely simple enough, GFFM will not pass it to the next PDM by multiply it with $\mathbf{h}_\omega \to 0$ at position ω inside heatmap $\mathbf{H}$. Thus only one-level PDM with its parameters will be employed to disentangle it. If the pattern needs a mixture of different processing scheme, e.g. a few deeper processing for extracting important semantics like center of a pedestrian, and some simpler processing for preserving details like the texture of clothing, the value of $\mathbf{h} \in (0, 1)$ will be decided by GFFM at each level adaptively. Hence, our proposed method can choose the best matched parameter sizes by selecting a pattern passing stream according to the complexity of the input patterns and adaptively matches parameter sizes and patterns with ease.

Detection Head

After our proposed adaptive pattern-parameter matching, "Output Feature" are further processed by "Detection Head". Following the anchor-free style of CSP (Liu et al. 2019), "Detection Head" first down-scales input feature to less channels. Then multiple branches generate feature maps: "Center Heatmap" to classify the centers of pedestrians, "Scale Map" to predict height scales on fixed aspect ratio 0.41 and "Offset Map" to adjust the localization horizontally and vertically (omitted for simplicity). Finally, these maps are assembled into bounding boxes of pedestrians. More details can be found in the CSP paper.

Experiments

In this section, extensive experiments are conducted on two popular pedestrian detection benchmarks, i.e. Caltech (Dollár et al. 2009) and CityPersons (Zhang, Benenson, and Schiele 2017), to evaluate our proposed method. Ablation study is performed to validate the effectiveness of key components in the proposed framework. Furthermore, we also report the sate-of-the-art comparison on both benchmarks.

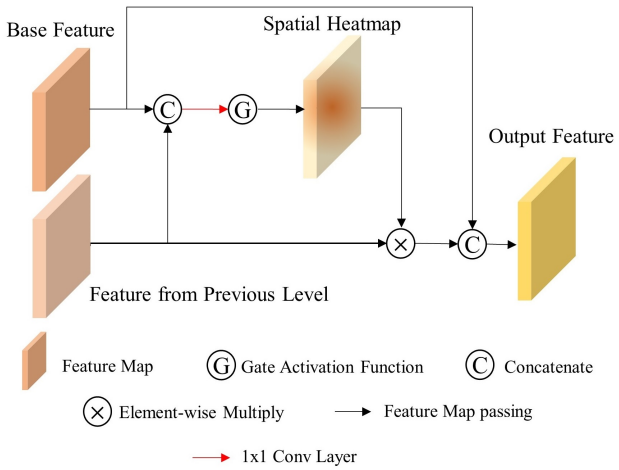

Figure 4: Gating Feature Filtering Module. Spatial Heatmap is generated from both the input features for re-weighting, activated by Gate Activation Function Sigmoid.

Datasets

The Caltech pedestrian dataset (Dollár et al. 2009) contains 2.5 hours of video data captured on the streets of Los Angeles. Over 70% of annotated pedestrian instances are less than 100 pixels high, including particularly small pedestrian instances that are less than 50 pixels. The standard test set includes 4024 images. By fixing the inconsistency and box misalignment, Zhang et. al. (2016b) have released new annotations to correct the official ones. To compare fairly with the baseline and other state-of-the-art methods, following evaluations will be performed based on new annotations.

CityPersons (Zhang, Benenson, and Schiele 2017) is a recently published large-scale pedestrian detection dataset. We train the model on an official training set with 2975 images and test it on validation set with 500 images. We follow the standard evaluation metric: log miss rate which is averaged over the false positive per image (FPPI) in $[10^{-2};\ 100]$, denoted as MR^{-2}. All tests are applied on original data without resizing, visible body or head boxes for fair comparison.

Implementation Details

Our proposed method is implemented on the basis of a powerful pedestrian detector CSP (Liu et al. 2019) and Keras framework. Adam (Kingma and Ba 2014) is used for optimization. ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) serves as the backbone network. For Caltech dataset, one Nvidia P100 GPU is utilized for training, with 1×10^{-4} learning rate. For CityPersons, two P40 GPUs are applied to training, with 2×10^{-4} learning rate. All the tests are conducted on a single 1080Ti GPU. The size of training images is 336×448 for Caltech and 640×1280 for CityPersons. For the best ensemble, IoU threshold of NMS after fusing detection results of two policies are 0.54 for Caltech and 0.59 for CityPersons. Note that if not mentioned, every convolution layer is followed with Batch Normalization (Ioffe and Szegedy 2015) and ReLU activation function (Hahnloser et al. 2000).

Method	Reasonable	All	Large	Near	Medium	Heavy
CSP (Liu et al. 2019)	4.54	56.94	7.84†	10.42†	51.75†	45.81
V-PDM	4.45	56.86	7.73	9.79	51.79	45.04
E-PDM	4.12	56.02	6.70	9.91	50.61	44.05
G-PDM	3.99	56.17	6.73	9.58	50.78	45.66
AP^{2}M	**3.30**	**55.89**	**5.30**	**8.71**	**50.12**	**42.20**

Table 3: Comparison among the variants of PDMs on Caltech. "†" means the result evaluated by us with trained model weights published by CSP. **Bolden** are the best results and underlined the 2^{nd} best. GFFMs of each model are omitted for simplicity.

Method	Reasonable	Heavy	Partial	Bare	Small	Medium	Large
CSP (Liu et al. 2019)	11.0	49.3	10.4	7.3	16.0	3.7	6.5
E-PDM	11.3	**48.3**	10.4	7.2	17.2	3.7	6.8
G-PDM	10.9	50.8	10.7	6.8	15.5	5.1	6.2
AP^{2}M	**10.4**	48.6	**9.7**	**6.2**	**15.3**	**3.5**	**5.3**

Table 4: Comparison among the variants of PDMs on CityPersons. **Bolden** are the best results and underlined the 2^{nd} best. GFFMs of each model are omitted for simplicity.

Levels	PDMs	GFFMs	E-PDM	G-PDM
0			4.54	4.54
1	✓		4.45	4.45
2	✓✓	✓	4.21	4.15
3	✓✓✓	✓✓	**4.12**	**3.99**

Table 1: Ablation study for the levels. ✓ marks the number of PDMs or GFFMs at certain amount of levels.

Levels	PDMs	GFFM	Conc	MR^{-2}
1	✓			4.45
2	✓✓		✓	4.24
		✓		**4.15**

Table 2: Ablation study for GFFM. ✓ marks the number of PDMs, GFFMs or Concatenations at certain amount of levels. "Conc" refers to Concatenation.

Ablation Study

The ablation study is first performed on Caltech dataset. The most widely-used and comprehensive Reasonable subset is used for comparisons.

Table 1 illustrates the results for different levels. Both E-PDM and G-PDM gain a lower miss rate after adding the 1^{st} level. On the 2^{nd} level, G-PDM enjoys more performance raising by 0.3% while E-PDM only 0.24%. Finally, 3-level model of G-PDM achieves miss rate of 3.99%, 0.13% lower than E-PDM. Due to the larger size of parameter, in our opinion, E-PDM tends to over-fit on certain hard patterns more than G-PDM. Therefore, G-PDM are more robust to the comprehensive subset Reasonable.

The effectiveness of GFFM is also evaluated. We choose G-PDM that performs better in former experiments as the target model. For comparison, GFFM is replaced with channel-wise concatenation. In Table 2, the miss rate of GFFM is lower than concatenation. We see that concatenation passes all patterns including simple ones unnecessary to further process and thus causes overfitting. In a word, the results prove that GFFM has better capability of adaptively selecting the best pattern processing stream than no filtering.

Comparison among the Variants of PDM

We then evaluate the effectiveness of the Vanilla PDM as well as its different parameter selection policies.

Table 3 shows the performance comparison on the Caltech subsets of different scale and occlusion. Cooperating with GFFM, V-PDM performs better than the baseline CSP on most of subsets, but is still inadequate for certain difficult patterns, e.g. on Medium and Heavy subset. Thus the results reveal necessity of our design of two parameter selection policies. G-PDM surpasses the V-PDM on all subsets including Medium with higher robustness to smaller scales. E-PDM also obtains lower miss rate on occlusion subset Heavy than V-PDM as is designed. More specifically, both of the two PDMs are specialized for other types of patterns and relatively complemented, i.e. E-PDM performs better on {All, Large, Medium, Heavy} and G-PDM does excellent jobs on {Reasonable, Near}. Effectiveness of two policies is fully proven by experiments above. Under ensemble strategy, AP^{2}M boosts the advances of them.

For CityPersons, the performance comparison on the subset {Reasonable}, {Small, Medium, Large} for scale and {Heavy, Partial, Bare} for occlusion are presented in Table 4. Similar to the results on Caltech, E-PDM is expert

† Due to problems in evaluation code on Caltech provided by CSP, we have to replace "vRng" from "[inf, inf]" to "[-inf, inf]" to get correct results on different Scale subsets. For visualizations, please refer to lmy98129.github.io/academic/src/AP2M-Appendix.pdf.

Method	Backbone	Reasonable	Heavy	Partial	Bare	Small	Medium	Large
Faster RCNN	VGG-16	15.4	-	-	-	25.6	7.2	7.9
Faster RCNN+Seg	VGG-16	14.8	-	-	-	22.6	6.7	8.0
TLL	ResNet-50	14.4	52.0	15.9	9.2	-	-	-
RepLoss	ResNet-50	13.2	56.9	16.8	7.6	42.6	-	-
OR-CNN	VGG-16	12.8	55.7	15.3	6.7	42.3	-	-
ALFNet	ResNet-50	12.0	51.9	11.4	8.4	19.0	5.7	6.6
Cascade R-CNN	VGG-16	12.0	53.6	-	-	38.4	-	-
LBST	ResNet-50	12.6	48.7	-	-	18.6	-	-
CSP	ResNet-50	11.0	49.3	10.4	7.3	16.0	3.7	6.5
Spatial-wise Gate	VGG-16	13.6	52.4	-	-	41.2	-	-
Channel-wise Gate	VGG-16	13.5	53.5	-	-	37.6	-	-
PBM+R^{2}NMS	VGG-16	11.1	53.3	-	-	-	-	-
PRNet	ResNet-50	10.8	53.3	10.0	6.8	-	-	-
AP^{2}M (Ours)	ResNet-50	**10.4**	**48.6**	**9.7**	**6.2**	**15.3**	**3.5**	**5.3**

Table 6: Comparison with the state-of-the-arts on CityPersons. **Bolden** are the best results.

Method	Backbone	Reason.	All	Heavy
RPN+BF	VGG-16	7.3	59.9	54.6
Faster RCNN	VGG-16	8.7	62.6	53.1
HyperLearner	VGG-16	5.5	61.5	48.7
ALFNet	ResNet-50	6.1	59.1	51.0
RepLoss	ResNet-50	5.0	59.0	47.9
OR-CNN	VGG-16	4.1	-	-
CSP	ResNet-50	4.5	56.9	45.8
AR-Ped	VGG-16	4.4	-	-
PAMS-FCN	ResNet-50	4.5	-	-
AP^{2}M (Ours)	ResNet-50	**3.3**	**55.9**	**42.2**

Table 5: Comparison with the state-of-the-arts on Caltech. **Bolden** are the best results. "Reason." refers to Reasonable.

for heavy occluded pattern and G-PDM is robust to patterns of smaller scales. These two different policies display complementary properties on different subsets. Complementing their advantages, AP^{2}M achieves best performance on most of the subsets.

In general, we observe that based on V-PDM, E-PDM is initially designed and surely robust for handling all kinds of occlusion, G-PDM also eliminates the weakness of V-PDM about smaller scale handling on Medium subset as is expected. Complementary of these two policies further inspires us to ensemble the PDMs and thus develop a more effective AP^{2}M for handling various types of hard patterns.

Comparison with the State-of-the-arts

For Caltech, we compare our AP^{2}M with state-of-the-arts: RPN+BF (Zhang et al. 2016a), Faster RCNN (Zhang, Benenson, and Schiele 2017), HyperLearner (Mao et al. 2017), RepLoss (Wang et al. 2018), ALFNet (Liu et al. 2018), CSP (Liu et al. 2019), PAMS-FCN (Yang et al. 2020), AR-Ped (Brazil and Liu 2019), OR-CNN (Zhang et al. 2018). RepLoss and OR-CNN are optimized for occlusion patterns. PAMS-FCN are dedicated to scare variation and occlusion. Since only the results on {Reasonable, All, Heavy} subsets are available for these methods, we follow the same protocol to present comparisons. In Table 5, our method achieves miss rate 3.3%, 55.9%, 42.2% on Reasonable, All and Heavy respectively, fully outperforming the best competitor OR-CNN by 0.8% on Reasonable, and CSP by 1.0% and 3.6% on All and Heavy.

For CityPersons, more state-of-the-art methods are compared including: Faster RCNN+Seg (Zhang, Benenson, and Schiele 2017), TLL (Song et al. 2018), Spatial and Channel-wise Gate (Liu et al. 2020), Cascade R-CNN (Cai and Vasconcelos 2019), PBM+R^{2}NMS (Huang et al. 2020), PRNet (Song et al. 2020), LBST (Cao et al. 2019). TLL is specialized for scales. RepLoss, PBM+R^{2}NMS, OR-CNN, PRNet are for occlusion. LBST, Spatial and Channel-wise Gate cope with both scale variation and occlusion. Table 6 shows that our AP^{2}M surpasses all the other methods on all subsets and achieves promising performance as expected.

In conclusion, our AP^{2}M has performed as a new state-of-the-art on both benchmarks especially in handling hard patterns, which sufficiently validates its capability of adaptive pattern-parameter matching.

Conclusion

In this paper, we have proposed a novel pedestrian detection method AP^{2}M for adaptive pattern-parameter matching towrads challenging patterns e.g. small scale or heavy occlusion, which consists of two key components: PDM for disentangling hard pattern into simpler patterns and GFFM for adaptively filtering the spatial positions where patterns need further processing with more parameters. We further investigated the relationship between detector parameter size and complexity of certain patterns and designed two parameter selection policies: Extended Parameter Policy for patterns of occlusion and Grouped Parameter Policy for patterns of small objects. With the help of model ensemble, we obtain a powerful model for both two kinds of patterns and achieve new state-of-the-art results on two challenging pedestrian detection benchmarks Caltech and CityPersons.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61703039 and 62072032, and Beijing Natural Science Foundation under Grant 4174095.

References

Brazil, G.; and Liu, X. 2019. Pedestrian detection with autoregressive network phases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7231–7240.

Cai, Z.; and Vasconcelos, N. 2019. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Cao, J.; Pang, Y.; Han, J.; Gao, B.; and Li, X. 2019. Taking a look at small-scale pedestrians and occluded pedestrians. *IEEE transactions on image processing* 29: 3143–3152.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 886–893. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Dollár, P.; Wojek, C.; Schiele, B.; and Perona, P. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 304–311. IEEE.

Geng, Q.; Zhang, H.; Qi, X.; Yang, R.; Zhou, Z.; and Huang, G. 2020. Gated Path Selection Network for Semantic Segmentation. *arXiv preprint arXiv:2001.06819* .

Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; and Pan, C. 2020. Augfpn: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12595–12604.

Hahnloser, R. H.; Sarpeshkar, R.; Mahowald, M. A.; Douglas, R. J.; and Seung, H. S. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405(6789): 947–951.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, X.; Ge, Z.; Jie, Z.; and Yoshie, O. 2020. NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10750–10759.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; and Yan, S. 2017. Scale-aware fast R-CNN for pedestrian detection. *IEEE transactions on Multimedia* 20(4): 985–996.

Lin, C.; Lu, J.; Wang, G.; and Zhou, J. 2018. Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 732–747.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, T.; Huang, J.-J.; Dai, T.; Ren, G.; and Stathaki, T. 2020. 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)*, 3867–3871. IEEE.

Liu, W.; Liao, S.; Hu, W.; Liang, X.; and Chen, X. 2018. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 618–634.

Liu, W.; Liao, S.; Ren, W.; Hu, W.; and Yu, Y. 2019. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5187–5196.

Mao, J.; Xiao, T.; Jiang, Y.; and Cao, Z. 2017. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3127–3136.

Ojala, T.; Pietikäinen, M.; and Mäenpää, T. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, 404–420. Springer.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

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

Smola, A. J.; and Schölkopf, B. 1998. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. *Algorithmica* 22(1-2): 211–231.

Song, T.; Sun, L.; Xie, D.; Sun, H.; and Pu, S. 2018. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 536–551.

Song, X.; Zhao, K.; Zhang, W.-S. C. H.; and Guo, J. 2020. Progressive Refinement Network for Occluded Pedestrian Detection. In *the European Conference on Computer Vision (ECCV)*.

Takikawa, T.; Acuna, D.; Jampani, V.; and Fidler, S. 2019. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 5229–5238.

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

Yang, P.; Zhang, G.; Wang, L.; Xu, L.; Deng, Q.; and Yang, M.-H. 2020. A Part-Aware Multi-Scale Fully Convolutional Network for Pedestrian Detection. *IEEE Transactions on Intelligent Transportation Systems* .

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* .

Zhang, L.; Lin, L.; Liang, X.; and He, K. 2016a. Is faster R-CNN doing well for pedestrian detection? In *European conference on computer vision*, 443–457. Springer.

Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; and Schiele, B. 2016b. How far are we from solving pedestrian detection? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1259–1267.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3221.

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