Adaptive Pattern-Parameter Matching for Robust Pedestrian Detection Technical Appendix

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

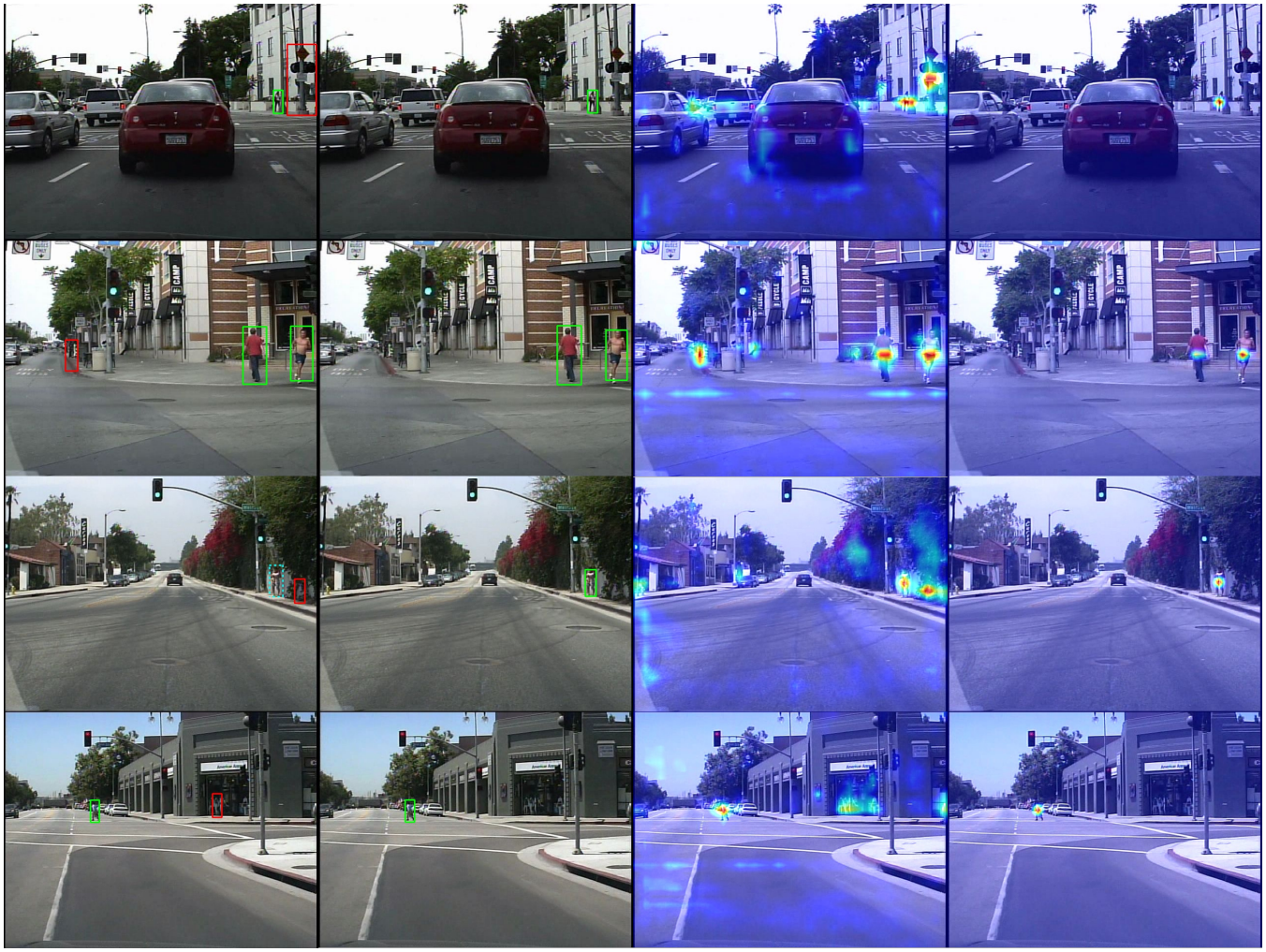

Figure 5: Qualitative results of proposed AP^{2}M and baseline CSP on hard patterns of different scales. The 1^{st} and 2^{nd} columns are detection results of baseline and our AP^{2}M presented by bounding boxes. **Green** are true positives, **Red** are false positives and **Dashed Cyan** are missing detections. 3^{rd} and 4^{th} columns are CAMs corresponding with detection results, generated from the 1^{st} layer of "Detection Head" to process "Output Feature" in Figure 1. The warmer the higher activation for detecting pedestrians.

Qualitative Analysis

In this section, extra qualitative results between our proposed **AP^{2}M** and baseline CSP (Liu et al. 2019) are provided, in the manner of not only detection results but also Class Activation Map (CAM) by Grad-CAM (Selvaraju et al. 2017) to better understand the external behavior and internal mechanism of AP^{2}M.

*Corresponding author

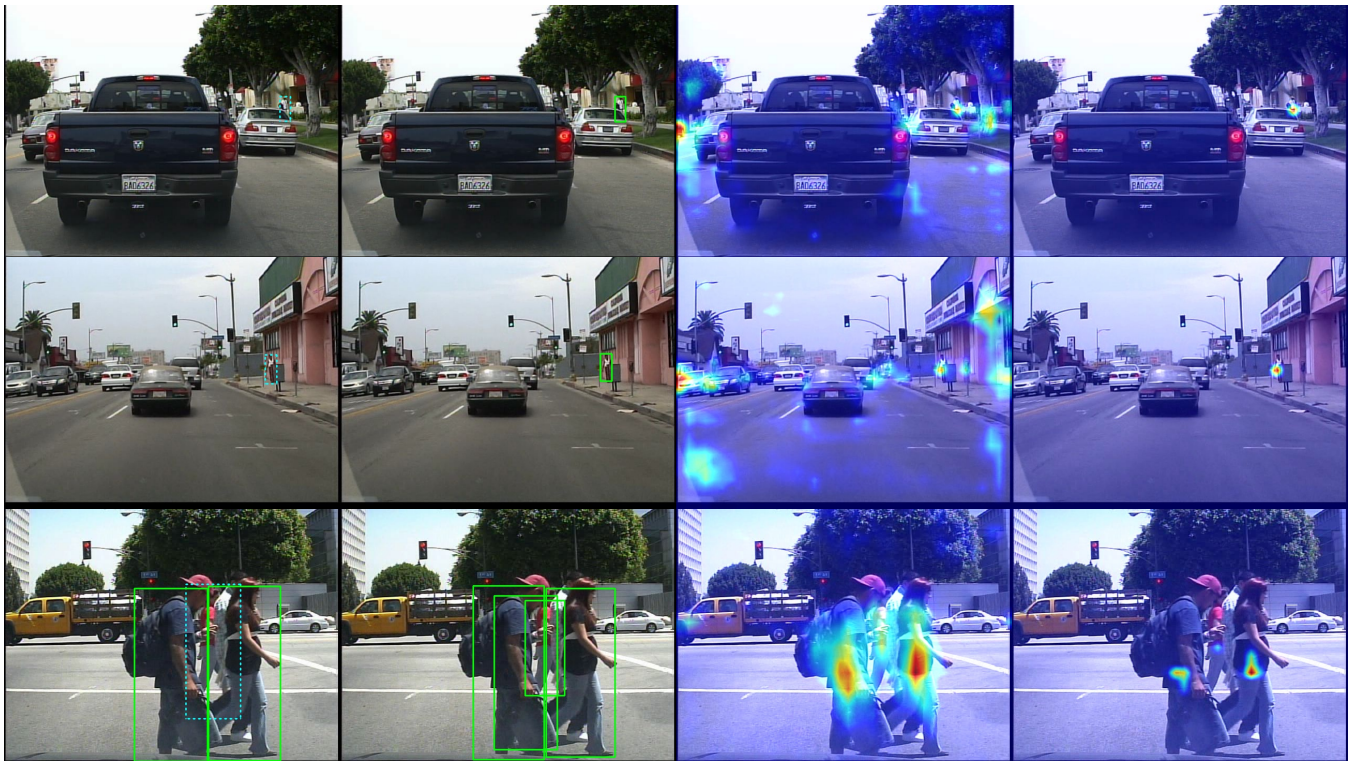

Figure 6: Qualitative results of proposed AP^{2}M and baseline CSP on hard patterns of different occlusion types. The 1^{st} and 2^{nd} columns are detection results of baseline and our AP^{2}M presented by bounding boxes. **Green** are true positives and **Dashed Cyan** are missing detections. 3^{rd} and 4^{th} columns are CAMs corresponding with detection results, following the settings of Figure 5. The warmer the higher activation for detecting pedestrians.

Comparison on Hard Patterns of Scales

As illustrated in Figure 5, hard patterns of different scales are mainly some human-like objects, e.g. traffic lights, trunks of trees, mannequins in shop window. Their appearances, scales and aspect ratios are closer to human beings than other objects and thus share various similar patterns with pedestrians, which requires detectors to process them with more parameters and distinguish them from real pedestrians.

The 1^{st} row of Figure 5 is an example that detectors handle the hard patterns of large scales. The traffic light is mistakenly detected by the baseline CSP as a pedestrian, the scale of which is far larger than that of the people nearby. Although it comprises the even sharper pattern, the baseline still produces a wrong detection as well as a high activation towards it, because such kind of objects share similar patterns of pedestrian. With the help of parameter-pattern matching, our AP^{2}M are more capable of dealing with such complicated patterns via specialized parameter size. Consequently, AP^{2}M only outputs the correct bounding box and high activation map for the pedestrian of smaller scale.

The other rows are examples of hard patterns of smaller scale. In the 2^{nd} row, the baseline detector does a good job in detecting two pedestrians of larger scale but recoginizes a trunk of tree as a pedestrian in the mean time. However, our AP^{2}M is not puzzled by such hard patterns of small scales and achieves more accurate detection and more concentrated activation map than baseline's. Meanwhile, both false-negative and false-positive results are predicted by the baseline in the 3^{rd} row. We observe that the false-positive sample (a little block of wall) stimulates the baseline detector to output higher activation in CAM than the false-negative one, which results in the wrong detection. Dissimilarly, our AP^{2}M only focuses on the pedestrian rather than other part of the background and thus produces the correct detection results.

Furthermore, the 4^{th} row demonstrates a scenario with higher difficulty that there are lots of mannequins in shop window by the street. Being more similar to real pedestrian especially in the distance, mannequins consists of harder patterns for detector than other objects aforementioned. As a result, the baseline detector are activated by nearly all the mannequins, shown by the CAM, and even outputs a false-positive bounding box for one of them. Our AP^{2}M, instead, is robust towards these harder patterns and performs an excellent detection.

Comparison on Hard Patterns of Occlusion

In Figure 6, typical examples of processing hard paterns of occlusion are given for qualitative analysis. These pedestrians are more likely to be ignored by detectors, because some part of them are occluded by other objects and thus force the

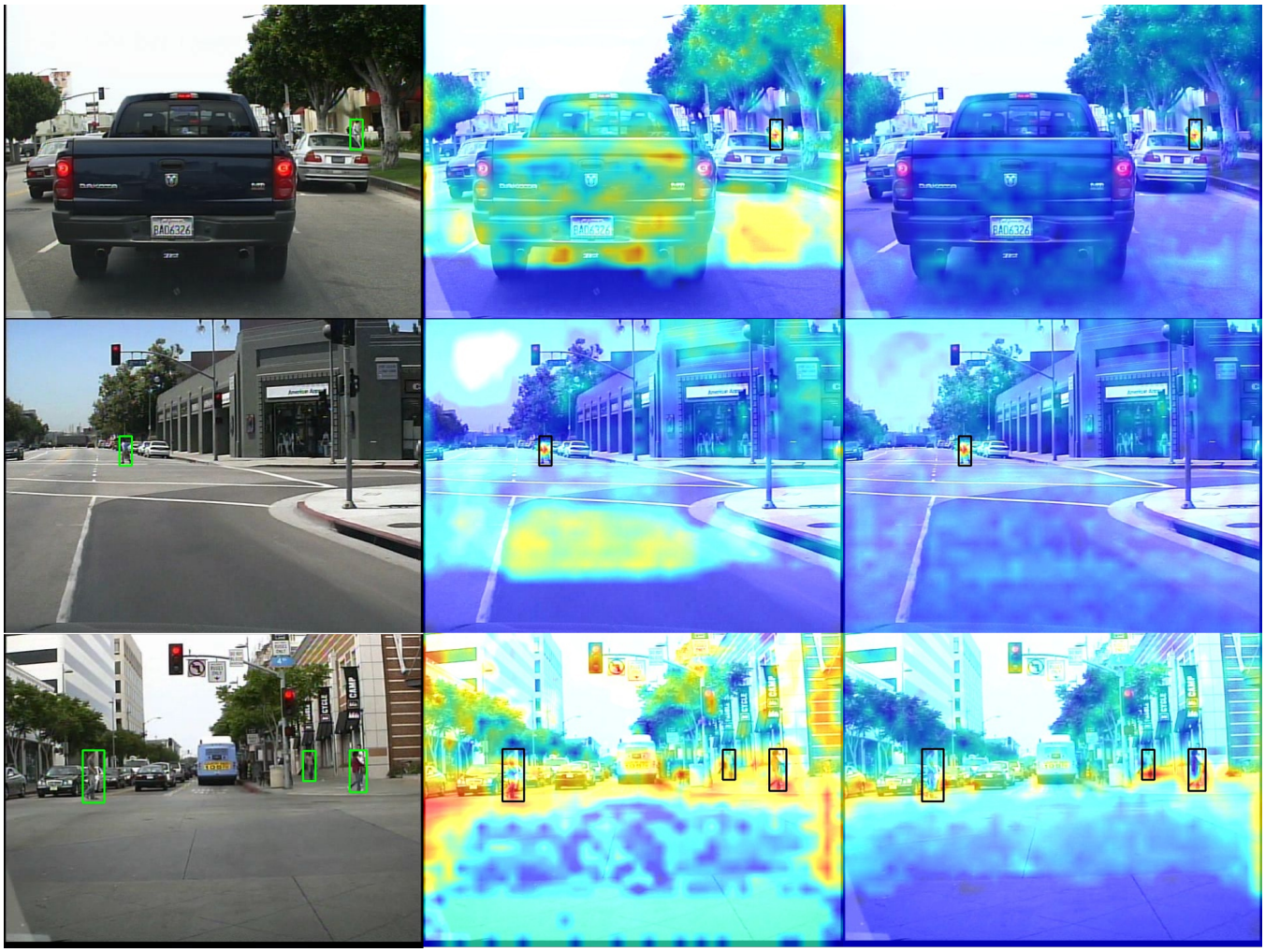

Figure 7: Qualitative analysis of the gating mechanism of our proposed AP^{2}M on hard patterns of scale and occlusion. The 1^{st} is detection results of our AP^{2}M presented by bounding boxes. **Green** or **Black** are true positives. The 2^{nd} and 3^{rd} column are "Spatial Heatmap" of GFFM at the 2^{nd} and 3^{rd} level, which determine the certain location where patterns are in need of further processing with more parameters by re-weighting them. The warmer the higher re-weight value in spatial heatmap $\mathbf{H} \in (0, 1)$.

detectors to classify foreign objects as part of the pedestrian.

In the 1^{st} row, a pedestrian with his/her bottom half of body occluded occurs on the right side of input image. The CAM of baseline detector is attracted by the left side of the road rather than the pedestrian, leading to the missing detection marked by dashed cyan box. Our AP^{2}M is concentrated to the occluded pedestrian with higher activation in warmer color, so that it accurately detects the pedestrian without any distraction. The 2^{nd} row shows another small-scale pedestrian like the former one, but this instance are occluded on the right leg by a mail box which is a different occlusion type compared with the former one. Despite of that, our method also outperforms the baseline with robust detection and sharper activation.

Differently, the 3^{rd} row presents a group of large-scale pedestrian among which two pedestrians are occluded (in the red and white clothes). Due to the evaluation protocol of pedestrian datasets, a predicted bounding box is regraded as true positive if the Intersection over Union (IoU) between it and ground truth is higher than threshold (0.5). So the occluded pedestrian in white is also detected by the box of the right-most pedestrian. However, the pedestrian dressed in the red is treated in different ways by two detectors. Although larger area of activation is generated, the bouding boxes from the baseline detector are more sparse spatially and thus more likely to be supressed by Non-Maximum Supression in post-processing. Our AP^{2}M, instead, produces a more concentrated activation that avoids the overlapping with actviation of other pedestrians. Hence, bouding box for detecting this pedestrian in red is successfully predicted by AP^{2}M.

In conclusion, the experimental results reverals that our proposed method AP^{2}M strongly improves the robustness of baseline towards manifold hard patterns of differnt scales

and occlusion types. Additionally, our methods succeeds in eliminating unneccesary activaion for some part of background brought by baseline method, as is distinctly shown in Figure 5 and 6. Via introducing pattern-parameter matching, our method is able to select the best matched parameter size according to complexity of input patterns and thus facilitates pedestrian detection with higher accuracy and robustness.

Qualitative Analysis of Gating Mechanism

Typical examples are provided in Figure 7 for qualitative analysis of gating mechaism adopted by GFFM in our proposed AP^{2}M. Ideally, the GFFM is expected to determine the position where patterns are complex and need more parameters to handle with by re-weighting them with a spatial heatmap $\mathbf{H} \in (0, 1)$. The spaial heatmap is generated on the basis of the disentangled pattern from previous level of PDM and the original one named "Base Feature" in Figure 2.

The 1^{st} row shows that the bottom part of a pedestrian is occluded and thus our AP^{2}M also passes the patterns of this occluded part through the level $\{1,\ 2,\ 3\}$. By means of gating mechanism, all the complex patterns are selected by GFFMs at level 2 and 3, including the occluded part of pedestrian and the surface of cars with analogous texture of human clothing. Following the selection, our AP^{2}M processes them with more parameters and thus detects the pedestrian and eliminates potential false positives with ease. Similarly, the small scale pedestrian with blurry patterns in the 2^{nd} row is also paid attention to by GFFMs and detected successfully by our method.

For some pedestrian instances with sharper patterns, as shown in the 3^{rd} row, GFFM makes a right decision that it is only processed at shallower level 2 rather than deeper level 3. Along with this pattern are some patterns of potential false positives, e.g. traffic lights, cars, buildings and road surface. But all of them are omitted at level 3, because AP^{2}M distinguish them from real pedestrians by disentangling such complex patterns into simpler ones with PDMs. Other pedestrians with smaller scale, instead, are further disentangled at level 3, of which more patterns at margin are concentrated for more accurate localization.

Generally speaking, our proposed AP^{2}M benefits from the gating mechanism a lot that specific pattern passing streams with different paramater sizes are constructed according to the complexity of input patterns. Therefore, on chanllenging patterns including small scale and heavy occlusion, our AP^{2}M achieves the superior performance and high robustness compared with other counterparts.

References

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.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.