Pedestrian Occlusion Level Classification using Keypoint Detection and 2D Body Surface Area Estimation

Shane Gilroy [1,2], Martin Glavin[2], Edward Jones[2] and Darragh Mullins[2]

[1]Institute of Technology Sligo, Ireland
[2]National University of Ireland Galway, Ireland
gilroy.shane@itsligo.ie, {martin.glavin,edward.jones,darragh.mullins}@nuigalway.ie

Abstract

Effective and reliable pedestrian detection is among the most safety-critical features of semi-autonomous and autonomous vehicles. One of the most complex detection challenges is that of partial occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground object. A number of current pedestrian detection benchmarks provide annotation for partial occlusion to assess algorithm performance in these scenarios, however each benchmark varies greatly in their definition of the occurrence and severity of occlusion. In addition, current occlusion level annotation methods contain a high degree of subjectivity by the human annotator. This can lead to inaccurate or inconsistent reporting of an algorithm's detection performance for partially occluded pedestrians, depending on which benchmark is used. This research presents a novel, objective method for pedestrian occlusion level classification for ground truth annotation. Occlusion level classification is achieved through the identification of visible pedestrian keypoints and through the use of a novel, effective method of 2D body surface area estimation. Experimental results demonstrate that the proposed method reflects the pixel-wise occlusion level of pedestrians in images and is effective for all forms of occlusion, including challenging edge cases such as self-occlusion, truncation and inter-occluding pedestrians.

1. Introduction

Pedestrian detection is one of the most safety-critical features of driver assistance systems and autonomous vehicles. Pedestrian detection is particularly challenging due to the deformable nature and irregular profile of the human body in motion and the inconsistency of color information due to clothing, that can enhance or camouflage any part of a pedestrian. Pedestrian detection systems have improved significantly in recent years with the proliferation of deep learning based solutions and the availability of larger and more diverse datasets. Despite this, many challenges still exist before we reach the detection capabilities required for safe autonomous driving. One of the most complex scenarios is that of partial occlusion, where a target object is only partially available to the sensor due to obstruction by another foreground object. The frequency and variety of occlusion in the automotive environment is substantial and is impacted by both natural and man-made infrastructure as well as the presence of other road users. Pedestrians can be occluded by static or dynamic objects, may inter-occlude (occlude one another) such as in crowds, and self-occlude - where parts of a pedestrian overlap. State of the art pedestrian detection solutions claim a detection performance of approximately 65%-75% of partially and heavily occluded pedestrians respectively using current benchmarks [8]. However, the definition of the occurrence and severity of occlusion varies greatly, and a high degree of subjectivity is used to categorize pedestrian occlusion level in each benchmark as shown in Table 1. In addition to this, oc-

Occlusion Level	Low	Partial	Heavy
EuroCity Persons [2]	<40%	40-80%	>80%
CityPersons [22]	-	<35%	35-75%
Kitti [7]	"Fully Visible"	"Partially Occluded"	"Difficult to See"
Caltech Pedestrian [5]	-	1-35%	35-80%
Multispectral Pedestrian [11], OVIS [16]	-	$\leq 50\%$	>50%
Daimler Tsinghua [13]	<10%	10-40%	41-80%

Table 1. Categories of occlusion levels by dataset.

currences of self occlusion, where one part of the body occludes another, has typically been overlooked entirely when categorizing occlusion level. This can lead to inaccurate or inconsistent reporting of a pedestrian detection algorithm's performance, depending on which dataset is used to verify detection performance [8][9]. In order to address this issue, an objective, repeatable method of occlusion level classification is required for ground truth annotation so that algorithms can be evaluated and compared on an equal scale.

This research proposes a novel, objective and consistent method for pedestrian occlusion level classification for ground truth annotation of partially occluded pedestrians. The proposed method more accurately represents the pixel-wise occlusion level than the current state of the art and works for all forms of occlusion including challenging edge cases such as self-occlusion, inter-occluding pedestrians and truncation.

The contributions of this research are threefold: 1. A novel, objective method for pedestrian occlusion level classification for ground truth annotation is presented. 2. A novel method for estimating the visible 2D body surface area of pedestrians in images. 3. The proposed method is the first occlusion level classifier to infer the level of pedestrian self-occlusion.

2. Related Work

This section provides an overview of current occlusion level classification methods for pedestrian detection, pedestrian occlusion level analysis for flood level assessment and commonly used methods for estimating total body surface area.

2.1. Occlusion Level Classification in Autonomous Driving Datasets

A number of publicly available datasets provide annotation of the level of pedestrian occlusion in the automotive environment. Table 1 provides an overview of the categories used to define the severity of occlusion in current popular datasets. Analysis of current benchmarks demonstrate the range of inconsistency and subjectivity in the definition of low, partial and heavy occlusion.

The Eurocity Persons Dataset [2] categorizes occlusion into three distinct levels: low occlusion (10%-40%), moderate occlusion (40%-80%), and strong occlusion (larger than 80%). Classification is carried out by human annotators. The full extent of the occluded pedestrian is estimated, and the approximate level of occlusion is then estimated to be within one of the three defined categories. This process is also used to classify the level of truncation of pedestrians near the image border. A similar approach is undertaken in the Caltech Pedestrian Dataset [5] in which pedestrians are annotated with two bounding boxes that denote the visible and full pedestrian extent. In the case of occluded pedestrians, the location of hidden parts of the full pedestrian were estimated by the human annotator in order to calculate the occlusion ratio. Cases of occluded pedestrians are then categorized into partial occlusion (1-35% occluded) and heavy occlusion (35-80% occluded). Further analysis of this dataset determined that the probability of occlusion in the automotive environment is not uniform, but rather has a strong bias for the lower portion of the pedestrian to be occluded and for the top portion to be visible.

Classification of occluded pedestrians in the CityPersons dataset [21] [22] is achieved by drawing a line from the top of the head to the middle of the two feet of the occluded pedestrian. Human annotators are required to estimate the location of the head and feet if these are not visible. A bounding box ("$BB-full$") is then generated for the full pedestrian area using a fixed aspect ratio of 0.41(width/height). A visible pedestrian area bounding box ("$BB-vis$") is also annotated and the occlusion ratio is calculated as $Area(BB-vis)/Area(BB-full)$. These estimates of occlusion level are then categorized into two levels in the Citypersons benchmark, Reasonable (<=35% occluded) and Heavy Occlusion (35%-75%).

Occluded Video Instance Segmentation (OVIS) [16] estimates the degree of occlusion by calculating the ratio of intersecting areas of overlapping bounding boxes to the total area of the respective bounding boxes. The authors acknowledge that although this proposed "Bounding Box Occlusion Rate" can be a rough indicator for the degree of occlusion, it can only reflect the occlusion between objects in a partial way and it does not accurately represent the pixel-wise occlusion level of the target objects.

A more semantic approach to determining the occlusion level was taken in the Kitti Vision Benchmark [7], where human annotators were simply asked to mark each bounding box as "visible", "semi-occluded", "fully-occluded" or "truncated". A similar approach was used in the Multispectral Pedestrian Dataset [11] where pedestrians "occluded to some extent up to one half" are tagged as partial occlusion; and those whose contour is "mostly occluded" were tagged as heavy occlusion during ground truth annotation.

2.2. Occlusion Analysis for Flood Level Estimation

Chaudhary et al [4], propose a method of flood level classification from social media images based on the visibility of pedestrians in the image. In this research the average height of a human adult is estimated to be 170cm. The flood level classifier detects pedestrians in an image and estimates how much of the pedestrian is occluded by flood water by vertically subdividing the pedestrian into 11 distinct levels. The highest level of the pedestrian occluded by the water indicates the flood height in the image location.

Feng et al [6] estimates flood level based on the relative

proportions of occluded semantic body parts which are perceived to be below the water line. Pedestrian detection is carried out on images using MaskRCNN [10] and keypoint detection is applied to images using Openpose [3] in order to identify the location of the ankle, knee, hip and chest of the detected pedestrian. The water line in the image is then hypothesized to be at the bottom line of the bounding box of a person. The relative proportions of semantic body parts which have been identified as below the water line are then used to estimate the height of the water level from the ground. A similar approach is taken by Quan et al [17] in which keypoint detection is correlated with a binary mask output of a pedestrian detector. Analysis is then carried out to determine if keypoints which represent the hip or knees are outside of the detected binary mask area due to occlusion by flood water in the image, thereby indicating a relative flood level.

2.3. Body Surface Area Estimation

Wallace [19] proposed a method of classification of body surface area for the purposes of diagnosing the severity of burn damage of the average adult burn victim [12]. This method, known as the "Wallace Rule of Nines", is commonly used by emergency medical providers and first responders to assess the total affected body surface area of burn patients [1][18]. The Rule of Nines estimates total body surface area by assigning percentages, in multiples of 9% to semantic body areas, based on the relative physical dimensions of the average adult. The head is estimated to be 9% of the total body surface area (4.5% for the front and 4.5% for the rear). The chest, abdomen, upper back and lower back are each assigned 9%. Each leg is assigned 18%, each arm is assigned a total of 9% and the groin is assigned the remaining 1%. Further research such as [1][15] validate the Rule of Nines for use in the assessment of total body surface area for the average adult, however, provide amendments to more accurately reflect body proportions in specific edge cases such as obese adults and infant children.

3. Methodology

An objective method for occlusion level classification is proposed, which removes the subjectivity of the human annotator and more accurately reflects the pixel wise occlusion level than the current state of the art [2][5][7][11][13][16][22]. Occlusion level classification consists of 3 steps: 1. Keypoint detection is applied to the input image in order to identify the presence and visibility of specific semantic parts of each pedestrian instance. 2. A visibility threshold is applied to determine which keypoints are occluded within the image. 3. Visible keypoints are then grouped into larger semantic parts and the total visible surface area is calculated using the 2D body surface area estimation method outlined in Section 3.2 and Figure 1. The proposed method classifies occlusion level for all forms of pedestrian occlusion, including challenging edge cases such as self occlusion, inter-occluding pedestrians and truncation. An overview of the classification pipeline is shown in Figure 2 and qualitative examples of the classifier output for multiple scenarios can be seen in Figure 3.

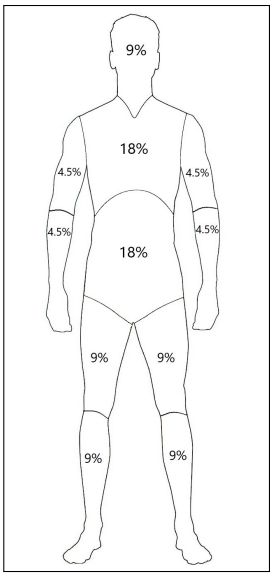

Figure 1. 2D Body Surface Area.

Body Part (% BSA)	Related Keypoints
Head (9%)	Nose or Eyes or Ears
Upper Torso (18%)	L Shoulder + R Shoulder
Upper Left Arm (4.5%)	L Shoulder + L Elbow
Lower Left Arm (4.5%)	L Elbow + L Wrist
Upper Right Arm (4.5%)	R Shoulder + R Elbow
Lower Right Arm (4.5%)	R Elbow + R Wrist
Lower Torso (18%)	L Hip + R Hip
Upper Left Leg (9%)	L Hip + L Knee
Lower Left Leg (9%)	L Knee + L Ankle
Upper Right Leg (9%)	R Hip + R Knee
Lower Right Leg (9%)	R Knee + R Ankle

Table 2. Percentage of total Body Surface Area (BSA) and related keypoints for each semantic body part.

3.1. Keypoint Detection

Keypoint detection is carried out by a Faster RCNN based keypoint detector using pretrained weights from Detectron2 [20]. The model uses a ResNet-50-FPN backbone and is trained using the COCO keypoints dataset [14]. The keypoint detector outputs 17 keypoints on the human body in addition to a visibility score for each predicted keypoint. Predicted keypoints include shoulders, elbows, wrists, hips, knees and ankles as well as facial characteristics such as

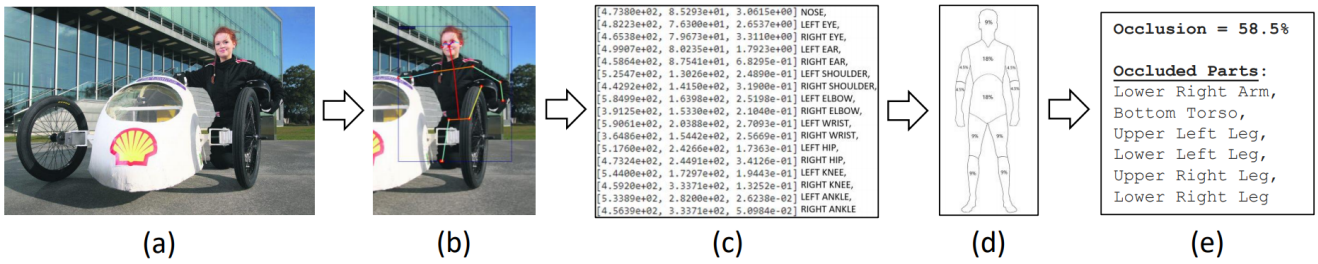

Figure 2. Occlusion level classification overview. (a) Read input image (b) Apply keypoint detection to each pedestrian instance (c) Assess keypoint visibility to identify occluded keypoints (d) Calculate total visible surface area (e) Output occlusion level classification.

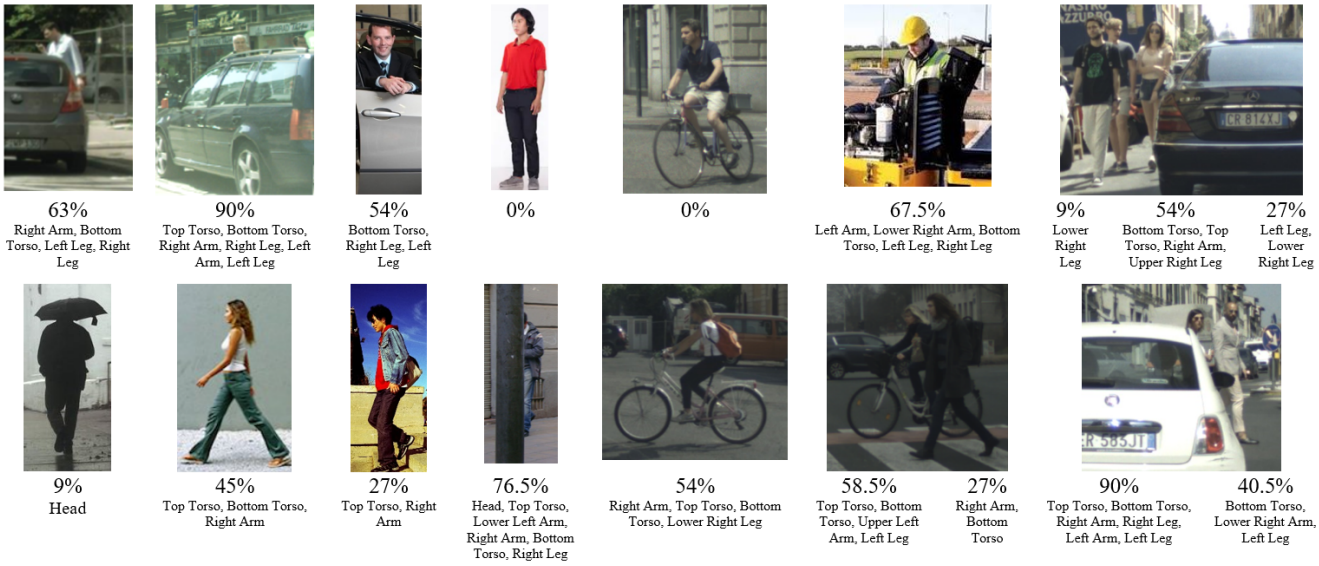

Figure 3. Qualitative validation results. Occlusion level is displayed below each image in addition to a list of occluded semantic parts. Examples are shown for cases of inter-class occlusion, self occlusion and inter-occluding pedestrians. Images containing multiple pedestrian instances read from left to right. All images are compiled from publicly available sources.

nose, eyes and ears. A threshold is then applied to the visibility score to determine which keypoints are occluded within the image. The presence of specific grouped keypoints indicates the presence of sematic body parts as outlined in Table 2.

3.2. 2D Body Surface Area Estimation

The "Wallace Rule of Nines" [19] is a time-tested method for determining total body surface area of the average adult. Although effective in the assessment of the body surface area of physical pedestrians, the Rule of Nines is not suitable for assessing the visible surface area of pedestrians in 2D images due to the 3D nature of the human body. An adapted version of the Rule of Nines is proposed for use in determining the visible body surface area of 2D pedestrian images for occlusion level classification. The original proportions of the Rule of Nines have been adjusted respectively to compensate for only one side of the body being visible at any one time, as in the case of 2D images. The proposed method for 2D body surface area estimation is shown in Figure 1. Detected keypoints are related to the semantic body areas in the method shown in Table 2. Examples of the classification output is shown in Figure 3.

4. Validation

Qualitative Validation was carried out by applying the proposed method to a wide range of images containing various pedestrian poses, backgrounds and multiple forms of occlusion, including cases of self-occlusion, inter-occluding pedestrians, and truncation. Occlusion level and the occluded semantic parts of each pedestrian instance was deduced using the proposed occlusion level classification method. Human visual inspection was then used to verify the performance of the occlusion level classifier in each case. A custom dataset of 320 images, compiled from multiple publicly available sources including [2][22][23], was

used in this validation step to ensure a wide diversity of pedestrian occlusion scenarios. Examples of the qualitative validation are provided in Figure 3.

4.1. Quantitative Validation

Quantitative validation was carried out by comparing the proposed method with the calculated pixel-wise occlusion level, derived using MaskRCNN [10], and the current state of the art as described in CityPersons [22] for both visible and progressively occluded pedestrians. In order to determine the pixel-wise occlusion, the total pixel area must be calculated for both the fully visible pedestrian and the same pedestrian under occlusion. To achieve this, a custom dataset of 200 images was created, including a wide range of occlusion scenarios and challenging pedestrian poses such as walking, running and cycling. MaskRCNN [10] was applied to a fully visible reference image and the masked pixel area ($MaskArea_{full}$) was calculated for each pedestrian instance. Occlusions were then superimposed on the reference image and the remaining visible pedestrian pixel area ($MaskArea_{occ}$) is calculated in order to determine the pixel-wise occlusion ratio, Eq.1.

$$Occ_{pixel} = \frac{MaskArea_{occ}}{MaskArea_{full}} \tag{1}$$

The proposed method was then compared with the pixel-wise occlusion level and the method described in CityPersons [22] to determine the pixel-wise accuracy of the proposed occlusion level classifier. More subjective occlusion level classification methods such as those used in [2][5][7][11] are omitted for the purposes of this testing. A visibility threshold of 0.15 was used for all keypoints to identify visible semantic parts using the method outlined in this document. A sample of the images used in these experiments can be seen in Figure 4. Quantitative validation results are provided in Figure 5.

5. Discussion

The method described in this document proposes an objective method for occlusion level classification. The qualitative validation results shown in Figure 3 demonstrate the capability of the proposed method for classifying occlusion level for all forms of occlusion, including challenging edge cases such as self-occlusion, truncation, and inter-occluding pedestrians. By removing the subjectivity of a human annotator, the proposed method is more robust and repeatable than the current state of the art and is suitable for the objective comparison of pedestrian detection algorithms, regardless of the benchmark used. Classification of pedestrian self-occlusion, heretofore ignored in the assessment of partially occluded pedestrians, may have a large impact on assessing the detectability of pedestrians using

Figure 4. Quantitative validation dataset sample images. The custom dataset consists of 200 images covering a wide range of pedestrian poses and superimposed occlusions. All images are compiled from publicly available sources.

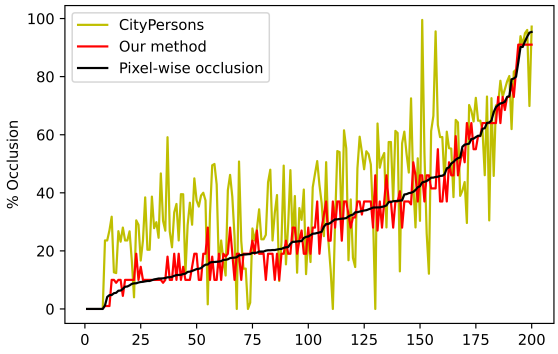

Figure 5. Quantitative Evaluation Results. Our proposed method is compared with the pixel-wise occlusion level as produced by MaskRCNN[10] and the current state of the art as described in CityPersons[22] for a dataset of 200 images. Results demonstrate that our method is a significant improvement over the state of the art when plotted against the pixel-wise occlusion level.

modern techniques. This is especially relevant in scenarios where detection confidence is linked to the presence of key salient features which may be self-occluded by the target pedestrian in the image. More detailed analysis of detection performance in cases of self-occlusion will increase our understanding of the behaviour of deep learning-based de-

tection routines. Characterization of detection performance for what were previously considered "visible" pedestrians, in cases where the algorithm specific informative value of a pedestrian is occluded will help identify potential failure modes of current state of the art pedestrian detection systems.

The quantitative validation results shown in Figure 5, demonstrate the proposed method's capability in representing the "real world" or pixel-wise occlusion value for challenging pedestrian poses, regardless of the severity or form of occlusion. The proposed method of 2D body surface area estimation shown in Figure 1, derived from the "Wallace Rule of Nines", has proven effective in calculating the visible area of partially occluded pedestrians for a wide range of pedestrian poses and occlusion scenarios. Further analysis of the quantitative validation results clearly displays an improvement over the current state of the art [22] when compared to the pixel-wise occlusion value.

6. Conclusions

This research proposes an objective method of pedestrian occlusion level classification for ground truth annotation. The proposed method uses keypoint detection to identify the visible semantic parts of partially occluded pedestrians and calculates a percentage occluded body surface area using a novel method for 2D body surface area estimation. The proposed method removes the subjectivity of the human annotator used by the current state of the art, in turn increasing the robustness and repeatability of pedestrian occlusion level classification. Qualitative and quantitative validation demonstrates the effectiveness of the proposed method for all forms of occlusion including challenging edge cases such as self-occlusion and inter-occluding pedestrians. Experimental results show a significant improvement over the current state of the art when plotted against the pixel-wise pedestrian occlusion level. Widespread use of the proposed method will improve the accuracy and consistency of occlusion level annotation in pedestrian detection benchmarks. Detailed analysis of edge cases such as self-occlusion, previously overlooked in popular pedestrian detection datasets, will increase our understanding of deep learning-based detection routines and help to identify potential failure modes in current technology. This in turn will inform the development of more robust pedestrian detection systems for semi-autonomous and autonomous vehicles.

References

[1] Kaveh Borhani-Khomani, Søren Partoft, and Rikke Holmgaard. Assessment of burn size in obese adults; a literature review. *Journal of plastic surgery and hand surgery*, 51(6):375–380, 2017. 3

[2] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019. 1, 2, 3, 4, 5

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 3

[4] Priyanka Chaudhary, Stefano D'Aronco, Matthew Moy de Vitry, João P Leitão, and Jan D Wegner. Flood-water level estimation from social media images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(2/W5):5–12, 2019. 2

[5] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311. IEEE, 2009. 1, 2, 3, 5

[6] Yu Feng, Claus Brenner, and Monika Sester. Flood severity mapping from volunteered geographic information by interpreting water level from images containing people: A case study of hurricane harvey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:301–319, 2020. 2

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1, 2, 3, 5

[8] Shane Gilroy, Edward Jones, and Martin Glavin. Overcoming occlusion in the automotive environment-a review. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 1, 2

[9] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337, 2021. 2

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 5

[11] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 1, 2, 3, 5

[12] GEORGE A KNAYSI, GEORGE F CRIKELAIR, and BARD COSMAN. The rule of nines: its history and accuracy. *Plastic and reconstructive surgery*, 41(6):560–563, 1968. 3

[13] Xiaofei Li, Fabian Flohr, Yue Yang, Hui Xiong, Markus Braun, Shuyue Pan, Keqiang Li, and Dariu M Gavrila. A new benchmark for vision-based cyclist detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1028–1033. IEEE, 2016. 1, 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[15] Edward H Livingston and Scott Lee. Percentage of burned body surface area determination in obese and nonobese patients. *Journal of surgical research*, 91(2):106–110, 2000. 3

[16] Jiyang Qi, Yan Gao, Xiaoyu Liu, Yao Hu, Xinggang Wang, Xiang Bai, Philip HS Torr, Serge Belongie, Alan Yuille, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 1, 2, 3

[17] Khanh-An C Quan, Vinh-Tiep Nguyen, Tan-Cong Nguyen, Tam V Nguyen, and Minh-Triet Tran. Flood level prediction via human pose estimation from social media images. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 479–485, 2020. 3

[18] Ilaria Tocco-Tussardi, Benjamin Presman, and Fredrik Huss. Want correct percentage of tbsa burned? let a layman do the assessment. *Journal of Burn Care & Research*, 39(2):295–301, 2018. 3

[19] AB Wallace. The exposure treatment of burns. *The Lancet*, 257(6653):501–504, 1951. 3, 4

[20] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[21] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1259–1267, 2016. 2

[22] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 1, 2, 3, 4, 5, 6

[23] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 4