PARTIALLY OCCLUDED OBJECT DETECTION BY FINDING THE VISIBLE FEATURES AND PARTS

Kai Chi Chan[*], *Alper Ayvaci*[†] *and Bernd Heisele*[†]

[*] Purdue University [†] Honda Research Insitute, USA

ABSTRACT

We address the partially occluded object detection problem by implementing a model which includes latent visibility flags that are attached to cells and parts of a Deformable Part Model (DPM) [1]. A visibility flag indicates whether an image portion is part of a pedestrian or part of an occluder. To compute the visibility flags and the score of the detector simultaneously, we maximize a concave objective function that is composed of the following four parts: (1) the detection scores of visible cells and parts, (2) a cell-to-cell consistency term which encourages neighboring cells to have the same visibility flags, (3) a cell-to-part consistency term which encourages compatible labeling among overlapping cells and parts, and (4) a penalty term for cells and parts that are labeled as occluded. The maximization of the concave objective function is done using the Alternating Direction Method of Multipliers (ADMM). By removing scores of occluded cells and parts from the final detection score we significantly improve detection performance on partially occluded pedestrians. In experiments we show that our system outperforms the standard DPM and other state-of-art methods on a benchmark database of partially occluded pedestrians.

Index Terms— Partially Occluded Object Detection, Deformable Part Models, ADMM

1. INTRODUCTION

The deployment of vision-based driver assistance systems in mass produced cars has been a major driving force behind computer vision research on object detection and recognition. Even though commercial systems for traffic participant detection exist, e.g. *Mobileye*, their inability to handle partial occlusions significantly limits their applicability. Developing detection algorithms that are robust against partial occlusions remains an active area of research with high impact on future technology for advanced driver assistance systems.

Despite the significant effort that has been put on developing robust object detection systems, most of the generic object detectors developed to date [2, 1, 3] expect that the objects are fully visible in the scene. Even for small partial occlusions the performance of these systems drops dramatically [4].

Another important challenge for object detection is shape variations. Non-rigid models [5, 6, 7, 1] address this issue by allowing the parts of an object deviate from their expected positions unlike a rigid holistic approach [2, 8]. Discriminatively trained deformable part models (DPM)[1] increased the popularity of non-rigid approaches. These models contain a root detector which operates on a coarse scale and a set of part detectors operating at a finer scale. Since all detectors are linear SVM classifiers, they are referred to as "filters". Based on the deformable part model (DPM), we propose a novel model which contains binary visibility flags to indicate if a root-cell or a parts is occluded or visible. The basic idea is to detect occluded cells and parts and remove their contributions to the detection score. Our model tries to find a visibility configuration that maximizes the detection score by neglecting the contribution of occluded regions. We also use prior knowledge about the properties of occlusions typically found in images of pedestrians in urban scenes: (1) partial occlusions occur more often on the lower body of a pedestrian than on the upper body [4] and (2) occluded pixels tend to be spatially clustered and form connected regions. To account for the first property, we introduce a penalty term that encourages a solution with occlusions at the bottom of the detection window. For the second property, we introduce a consistency term which is applied to the visibility flags of neighboring cells. Visibility configurations in which neighboring cells share the same visibility flags are preferred over solutions in which they disagree. The proposed model also

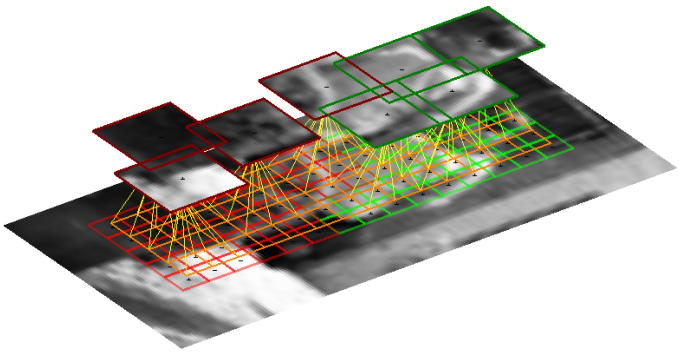


Fig. 1: *Consistency graph: Root cells are represented by the squares on the image, and parts are drawn above. Edges that are represented by orange lines indicate the cell-to-cell consistency while yellow lines indicate the cell-to-part consistency.*

favors a solution where parts and the cell of the root filter have consistent visibility flags wherever they overlap. To find the visibility maps, we solve a concave optimization problem enforcing the requirements listed above.

2. RELATED WORK

Occlusion handling has been a prime research area for computer vision community. To improve the robustness of a HOG-based detector against occlusions, Wang et al. [9] decide the visibility of a cell by thresholding the linear classifier response to HoG features. Similarly, Vedaldi and Zisserman [10] propose to use binary variables to determine whether a cell is occluded, but their approach is limited to occlusions due to truncation of objects at image borders. Following the steps of [10], Gao et al. [11] employ a structural SVM to learn cell-level visibility models. However, their model needs to be trained on features of both the occluded and occlude object classes. This might increase the complexity of the training stage since an occluder object can be anything.

Instead of using a holistic detector, the approaches proposed in [12, 13, 14, 15] construct a part-based detectors and associate a hidden visibility variable to each part detector. Ouyang et al. [12] use a discriminative deep network to learn the visibility relationships on a hierarchical part model and infer the visibility configuration using the correspondent back-propagation network. Niknejad et. al [13] construct a conditional random field (CRF) to achieve the same task. Unlike these methods, our approach identifies occlusion patterns at a finer cell level. Similarly, [16, 17] identify the occlusions on close up shots of faces by inferring the visibility of the parts on a hierarchical deformable part model.

The depth from stereo vision is another cue to handle occlusions. Baumgartner et al. [18] separate occluded and occluder objects by projecting their 3D point clouds to the ground plane. Enzweiler et al. [19] train a set of part-based expert classifiers on features extracted from intensity values, optical flow vectors and depth estimates. They also infer the degree of visibility by examining discontinuities in motion and depth. Unlike these methods, our approach uses only a single image from a single camera.

Another approach to occlusion handling is to train a set of occlusion-specific classifiers. The main drawback of this approach is that the test time grows linearly with the size of the classifier set. Mathias et al. [20] show that a sub-linear cost-growth can be achieved using Franken-classifiers. Yet another approach is to learn models including the occluder classes. Tang et al. [21] train a double-person detector to handle pedestrian-to-pedestrian occlusions.

3. MODEL

A linear deformable part model, trained following the steps of [1], consists of a root filter $F_0 \in \mathbb{R}^{w_0 \times h_0 \times d}$, its bias term b_0, a set of part filters $F_p \in \mathbb{R}^{w_p \times h_p \times d}$, $p = 1 \dots P$, the associated bias terms $\{b_p\}_{p=1}^{P}$ and deformation coefficients δ_p. Here, d is the number of HOG features computed in a cell, w_0 and h_0 are the width and height of the detection window in terms of cells, and similarly w_p and h_p are the width and height of the detector for the part p.

Given a detection window at position x with scale s, the contribution of the root filter is computed as

$$R(x,s) = F_0^T \cdot H(x,s) + b_0. \tag{1}$$

where $H(x,s) \in \mathbb{R}^{w_0 \times h_0 \times d}$ is the concatenation of the HOG features that are computed on the cells. Since this is a linear operation, we can decompose the root detector response to cell level responses and rewrite the root detector scores as

$$R_0(x,s) = \sum_{c=1}^{C} \left[(F_0^c)^T \cdot H^c(x,s) + b_0^c) \right] = \sum_{c=1}^{C} R_0^c(x,s). \tag{2}$$

where $C = w_0 \times h_0$ is the number of cells and $b_0 = \sum_c b_0^c$. Also, $F_0^c \in \mathbb{R}^d$ and $H^c(x,s) \in \mathbb{R}^d$. To decompose b_0, we follow the lines of [9].

Unlike the root detector score, computation of the scores for each part requires a search where deviation from the expected location is penalized proportional to the amount of the deviation. The response of each part detector is computed as

$$R_p(x,s) = \max_{dx \in X} \left(F_p^T \cdot H^p(x+dx,s) + b_p - (\delta_p)^T \cdot \phi(dx) \right), \tag{3}$$

where $X \subset \mathbb{Z}^2$ is the search neighborhood around the expected position of the part p, and $\phi(dx) = [dx_1, dx_1^2, dx_2, dx_2^2]^T$.

The final score for the detection window combines the responses from the root and the part detectors:

$$R(x,s) = \sum_{c=1}^{C} R_0^c(x,s) + \sum_{p=1}^{P} R_p(x,s). \tag{4}$$

When an object is partially occluded, root and part filters in the occluded regions may produce negative responses that mislead the detector and prevent it from recognizing the target object. Ideally, the occluded regions should not contribute to the overall detection score. Therefore, we propose to attach a visibility flag to each root-cell and part which takes the value 0 if the associated component is occluded or 1 otherwise. This helps us to aggregate scores only if they are computed on the visible regions of the object, not on the occluders.

To find out which cells and parts are visible, we solve an optimization problem that maximizes the detection score:

$$\widehat{v_0}, \{\widehat{v_p}\}_{p=1}^{P} = arg \max_{v_0, \{v_p\}} \sum_{c=1}^{C} R_0^c v_0^c + \sum_{p=1}^{P} R_p v_p. \tag{5}$$

where $v_0 = [v_0^0 \dots v_0^C]^T$ and $\{v^p\}_{p=1}^{P}$ are root cell and part visibility flags. Note that we omit the parameters x, s for simplicity as in this section, we focus on a single detection window. Note that this problem can be solved simply by assigning 0 to the visibility flag if the corresponding detector

response is less than 0, and 1 otherwise. However, the likelihood of a region to be occluded is not evenly distributed over the detection window. For example, pedestrians are typically occluded from below in traffic scenes, [4]. To encourage that the visibility map estimate follows this statistic, we introduce two penalty terms

$$R_{occ} = \max_{v_0,\{v_p\}_{p=1}^{P}} \sum_{c=1}^{C} [R_0^c v_0^c - \alpha\lambda_0^c(1-v_0^c)] + \sum_{p=1}^{P} [R_p v_p - \beta\lambda_p(1-v_p)] \quad (6)$$

where λ_0^c and λ_p are the penalties to be paid for neglecting the detector responses of the cell c and part p, respectively and α and β are the weights of the penalty terms. R_{occ} is the new detection score that we propose to handle partial occlusions. Note that DPM is a special case of our model when all root cells and parts are marked as visible.

Since occlusions usually happen on continuous parts of an object, it is likely that a root-cell and its neighbors shares the same visibility flag. Furthermore, it is also likely that a part and the root cells which overlap with it would share the same visibility flag. To be able to estimate visibility configurations following these expectations, we add two new terms to Eq. (6) which are called: cell-to-cell consistency and part-to-cell consistency. With these final additions, the optimization problem takes the following form,

$$\max_{v_0,\{v_p\}} \sum_{c=1}^{C} [R_0^c v_0^c - \alpha\lambda_0^c(1-v_0^c)] + \sum_{p=1}^{P} [R_p v_p - \beta\lambda_p(1-v_p)] - \gamma \sum_{c_i \sim c_j} |v_0^{c_i} - v_0^{c_j}| - \gamma \sum_{c_i \approx p_j} |v_0^{c_i} - v_{p_j}| \quad (7)$$

where $c_i \sim c_j$ denotes that c_i and c_j are adjacent cells, $c_i \approx p_j$ denotes that the cell c_i and the part p_j overlaps and γ is the regularization parameter. Fig. 1 depicts the cell-cell adjacency and cell-part overlap relations.

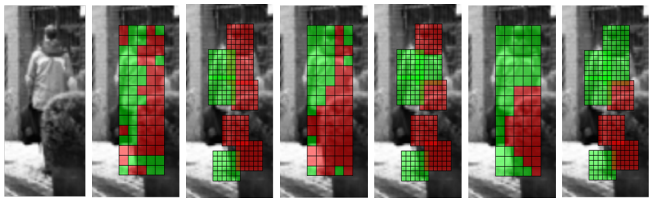
4. IMPLEMENTATION

4.1. Occlusion Penalty

Dollar et al. [4] observe that the lower part of a pedestrian would be occluded more often than the upper parts in a city-driving scenario. To have visibility flag configurations approximating these statistics, we set penalties at bottom parts of the detection window lower than the penalties at upper parts. The penalties are modeled with a sigmoid function as

$$\lambda_0^c = \frac{1}{2}\left[\frac{1}{1+e^{-\tau(h^c - h_0/2)}}\right] + \frac{1}{2} \quad (8)$$

where h^c is the height of the cell c from the bottom of the detection window and τ controls the steepness. λ_p is computed in the same way.



(a) input (b) initialization (c) first iteration (d) third iteration

Fig. 2: *The visibility map estimates: (a) Input image (b) The initialization passed to ADMM: To acquire this map, we threshold the cell-level and part-level detector responses at* 0. *Red and green indicate the variables with values* 0 *and* 1, *respectively. (c) The binarized visibility estimate at first iteration. (d) The solution at convergence (third iteration).*

4.2. Estimation of Visibility Flags

Since the number of detection windows in an image is large, estimating the visibility flags for all detection windows will be time consuming. Thus, we first run the DPM detector and pick the top 10% of the detection windows as candidates. Then, we use Alternating Direction Method of Multipliers (ADMM) [22] to speed up the solution of the optimization problem in Eq. (7). To be able to apply ADMM, we relax this problem by removing the integer constraints.

To simplify the notation, let $q = [v_0^0 \dots v_0^C\ v_1 \dots v_P]^T$ be a column vector that groups the visibility flags together and $\omega = [(R_0^0 + \alpha\lambda_0^0) \dots (R_0^C + \alpha\lambda_0^C)\ (R_1 + \beta\lambda_1) \dots (R_P + \beta\lambda_P)]^T$ be the stacked scores and occlusion penalties. Both q and ω are $C+P$ dimensional vectors. To convert the consistency terms (last two terms) in Eq. (7) into a matrix form, we construct the differentiation matrices D' and D'', and then stack them into $D = [(D')^T\ (D'')^T]^T$. In the ADMM form, the problem in Eq. (7) can be written as

$$\begin{aligned} \text{minimize} \quad & -\omega^T q + \lambda\|z_1\|_1 + J_{[0,1]}(z_2) \\ \text{subject to} \quad & Dq = z_1, q = z_2, q \in [0,1]^{C+P}, \end{aligned}$$

where $\|\cdot\|_1$ is the l_1-norm, $J_{[0,1]}(\cdot)$ is the indicator function which maps the input value that is between 0 and 1 to 0 and any other value to ∞. We form the augmented Lagrangian as

$$\begin{aligned} L(\omega, z_1, z_2, u_1, u_2) &= -\omega^T q + \lambda\|z_1\|_1 + J_{[0,1]}(z_2) \\ &+ \frac{\rho_1}{2}\|Dq - z_1 + u_1\|_2^2 \\ &+ \frac{\rho_2}{2}\|q - z_2 + u_2\|_2^2, \end{aligned} \quad (9)$$

where $\rho_1 > 0$, $\rho_2 > 0$ are penalty parameters. In our experiments, we allowed ADMM to converge at most in 20 iterations. If it does not converge by then, we use the q estimate of the last iteration. To initialize q, we threshold the detector responses $[R_0^0 \dots R_0^C\ R_1 \dots R_P]^T$, Fig. 2. The elements of the final q estimate are projected onto $\{0,1\}$ to estimate visibility flags. Fig. 2 depicts the visibility map estimates.

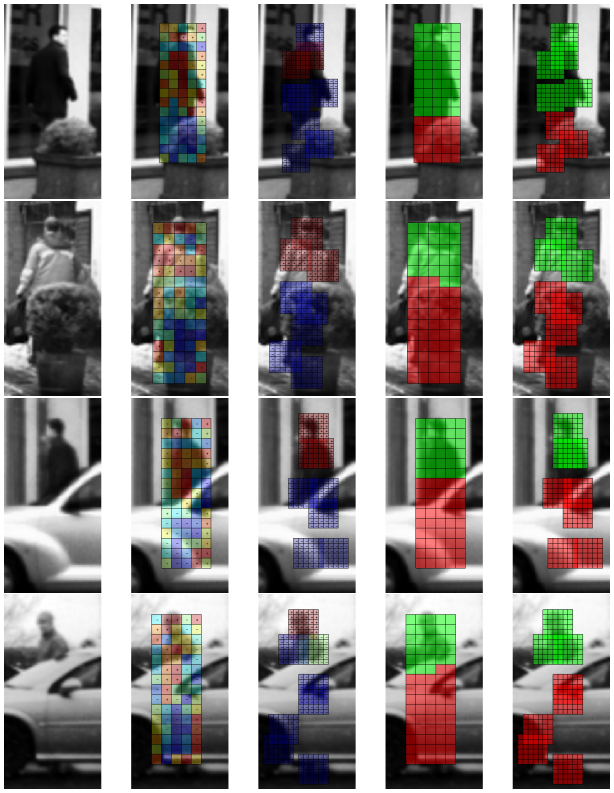


Fig. 3: *Visibility map estimates (low/medium/high level occlusions): (first column) The input image. (second and third column) The filter response of root cells and the parts: The plus and minus signs indicate the positive and negative detector responses and the color (varies from blue(low) to red(high)) represents the magnitude of the response. (forth and fifth column) the visibility flags (red:occluded, green:visible) of root cells and parts.*

5. EXPERIMENTS

The proposed approach is tested on the Daimler Multi-Cue Occluded Pedestrian Classification benchmark [19]. The training and test set contain non-occluded and partially occluded pedestrian images with resolution of 48×96 pixels. The negative set is picked from examples that pose a difficulty to a shape based pedestrian detector. We choose Daimler dataset because it has a large variety of examples of partially occluded pedestrians with (parked) cars or infrastructure nearby as occluders. Most of the relevant examples in ETH dataset [23] and Caltech Pedestrian Detection Benchmark [4] are pedestrian-pedestrian occlusions where a multi-pedestrian detector [24] would perform well. However, these examples are not a big challenge for a driver assistance system or a self-driving car since these systems already take action when they detect the occluder pedestrian which is fully visible in most of the cases and physically closer to the vehicle. A real challenge is, for example, detecting a pedestrian as early as possible before he/she rushes to the road from the behind of a parked car.

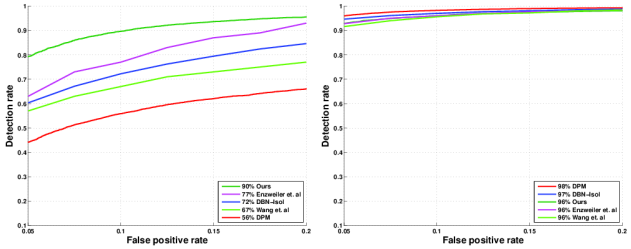


Fig. 4: *Quantitative results on Daimler Multi-Cue Occluded Pedestrian Classification benchmark: (left) Partially occluded pedestrian set and (right) Visible (non-occluded) pedestrian set. The performance of each method is listed at* 0.1 *false-positive rate.*

We used an existing object model previously trained on the INRIA person dataset [2] for experiments. However, we still need to learn the model parameters such as $\alpha, \beta, \gamma, \rho_1$ and ρ_2. Since the Daimler dataset only provides partially occluded pedestrian images on the test set, we built up a synthetic training set. To generate occluded pedestrian examples, we picked "cars" as the occluder object type since it is a common situation in urban traffic scenarios. The whole synthetic data set contains 2000 positive examples and 3000 negative examples. The optimal parameters are found with a grid search over the parameter space. To reduce the search space, we assumed that $\rho_1 = \rho_2$. Fig. 3 depicts the accuracy of the visibility map estimation on example images from Daimler dataset. Even though the individual cell and part responses would be misleading to determine whether a region is occluded, the combination of both supported by consistency and occlusion penalty terms leads to the robust estimation of the visibility maps even the spatial support for the pedestrian is small, Fig. 3.

Fig. 4 shows that our approach outperforms the baseline method, DPM, significantly on Daimler dataset in case the pedestrians are occluded. The difference is approximately 34% at 0.1 false-positive rate. We also compare our approach quantitatively against [19, 12, 9]. Even though the dataset provides optical flow and stereo depth estimates for test images, to have a fair comparison, the results in Fig. 4 rely only on intensity images. Our approach performs superior to existing methods on the occluded pedestrian subset and gets comparable results on the non-occluded pedestrian subset, Fig. 4.

6. CONCLUSION

We propose a method that detects the partially occluded pedestrians by determining the visible parts of the object. Once the visibility flags are determined by solving Eq. (7), the responses from occluded regions are suppressed at computation of the detection score. Experimental results shows that the proposed method significantly outperforms the baseline approach and the existing methods on partially occluded pedestrian images from the Daimler Multi-Cue Occluded Pedestrian Classification benchmark [19], and performs com-

parably on the images that contain fully visible pedestrians.

7. REFERENCES

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2005, vol. 1, pp. 886–893.

[3] Mohammad Norouzi, Mani Ranjbar, and Greg Mori, "Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[5] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, 2001.

[6] Bernd Heisele, Thomas Serre, Massimiliano Pontil, Thomas Vetter, and Tomaso Poggio, "Categorization by learning and combining object parts," in *NIPS*, 2001, pp. 1239–1245.

[7] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. of European Conference on Computer Vision*. 2004.

[8] Shanshan Zhang, Christian Bauckhage, and Armin B Cremers, "Informed haar-like features improve pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[9] Xiaoyu Wang, T.X. Han, and Shuicheng Yan, "An hog-lbp human detector with partial occlusion handling," in *IEEE International Conference on Computer Vision*, Sep. 2009, pp. 32–39.

[10] Andrea Vedaldi and Andrew Zisserman, "Structured output regression for detection with partial truncation," in *Advances in Neural Information Processing Systems*, 2009, pp. 1928–1936.

[11] Tianshi Gao, B. Packer, and D. Koller, "A segmentation-aware object detection model with occlusion handling," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1361–1368.

[12] Wanli Ouyang and Xiaogang Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3258–3265.

[13] H.T. Niknejad, T. Kawano, Y. Oishi, and S. Mita, "Occlusion handling using discriminative model of trained part templates and conditional random field," in *IEEE Intelligent Vehicles Symposium*, Jun. 2013, pp. 750–755.

[14] Hossein Azizpour and Ivan Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. of European Conference on Computer Vision*, 2012.

[15] Ross Girshick, Pedro Felzenszwalb, and David McAllester, "Object detection with grammar models," in *Proc. of Advances in Neural Information Processing Systems*, 2011.

[16] Xiang Yu, Zhe Lin, Jonathan Brandt, and Dimitris N Metaxas, "Consensus of regression for occlusion-robust facial feature localization," in *Proc. of European Conference on Computer Vision*, 2014.

[17] Golnaz Ghiasi and Charless C Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] T. Baumgartner, D. Mitzel, and B. Leibe, "Tracking people and their objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3658–3665.

[19] M. Enzweiler, A. Eigenstetter, B. Schiele, and D.M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 990–997.

[20] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool, "Handling occlusions with franken-classifiers," in *IEEE International Conference on Computer Vision*, Dec. 2013, pp. 1505–1512.

[21] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele, "Detection and tracking of occluded people," in *Proceedings of the British Machine Vision Conference*. 2012, pp. 9.1–9.11, BMVA Press.

[22] B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang, "An admm algorithm for a class of total variation regularized estimation problems," *ArXiv e-prints*, Mar. 2012.

[23] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *International Conference on Computer Vision (ICCV'07)*, 2007.

[24] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele, "Detection and tracking of occluded people," *International Journal of Computer Vision*, 2012.