Supplementary Material:
Boosting Detection in Crowd Analysis via Underutilized Output Features

A. Detailed Evaluation Metrics

Crowd Counting Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are widely used as counting metrics, and they are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|e_i - gt_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(e_i - gt_i)^2} \tag{2}$$

Here, e_i and gt_i represent the estimated count and ground truth count of crowds, respectively, and N is the total number of images.

Crowd Localization F1 Measure, Precision, and Recall are commonly used as metrics for crowd localization, as proposed in [4]. We denote the two point sets of prediction results as P_p and ground truth as P_g, and construct a Bipartite Graph $G_{p,s}$ for the two sets. Then, we compute the distance matrix of P_p and P_g. If the distance between $p_p \in P_p$ and $p_g \in P_g$ is less than a predefined distance threshold σ, we consider p_p and p_g to be successfully matched, and obtain a boolean match matrix (True and False denote matched and non-matched) corresponding to each element of the distance matrix. Finally, by implementing the Hungarian algorithm, we obtain a Maximum Bipartite Matching for $G_{p,s}$. Based on the counts of True Positive (TP), False Positive (FP), and False Negative (FN), we can compute Precision (P), Recall (R), and F1 Measure (F_1) as follows:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = \frac{2PR}{P+R} \tag{3}$$

Crowd Detection Following standard practices, we adopt the average precision (AP) as the detection metric, with an Intersection over Union (IOU) threshold of 0.5.

B. Datasets

WIDER-Face [5] is a dense face detection dataset consisting of 32,203 images and 393,703 face labels, which exhibit high variability in terms of scale, pose, and occlusion.
ShanghaiTech [6] consists of two independent subsets, Part A and Part B. Part A contains highly congested images collected from the Internet, while Part B is comprised of images taken from the busy streets of metropolitan areas in Shanghai.
UCF-QNRF [1] contains 1535 images, which exhibit a much wider range of crowd counts compared to the previous datasets, making it a more challenging dataset for crowd analysis
JHU-Crowd++ [3] is a large-scale unconstrained dataset that comprises a total of 4,372 images, containing 1,515,005 head annotations and captured under a variety of conditions. The dataset includes challenging images captured under various weather-based degradation, as well as some negative samples that may be detected as false positives.
NWPU-Crowd [4] is the largest crowd analysis dataset, consisting of 5,109 images and 2,133,375 annotated heads with varying crowd densities. For an authentic evaluation of crowd counting and localization, we report our results from the official website of NWPU-crowd.

C. Visualization of Compression

Figure 1 provides additional visualization of the 2D-1D feature compression achieved by the **PSDNN + Crowd Hat**. Heat map is adopted where 2D compression is on the left and 1D compression is on the right.

D. Visualization of Detection

Figures 2 and 3 depict the detection results of SDNet with and without our proposed Crowd Hat module. Following the practice in [2,4], we use green boxes to indicate true positives based on ground truth annotations, red boxes for false negatives, and yellow boxes for false positives, for better clarity.

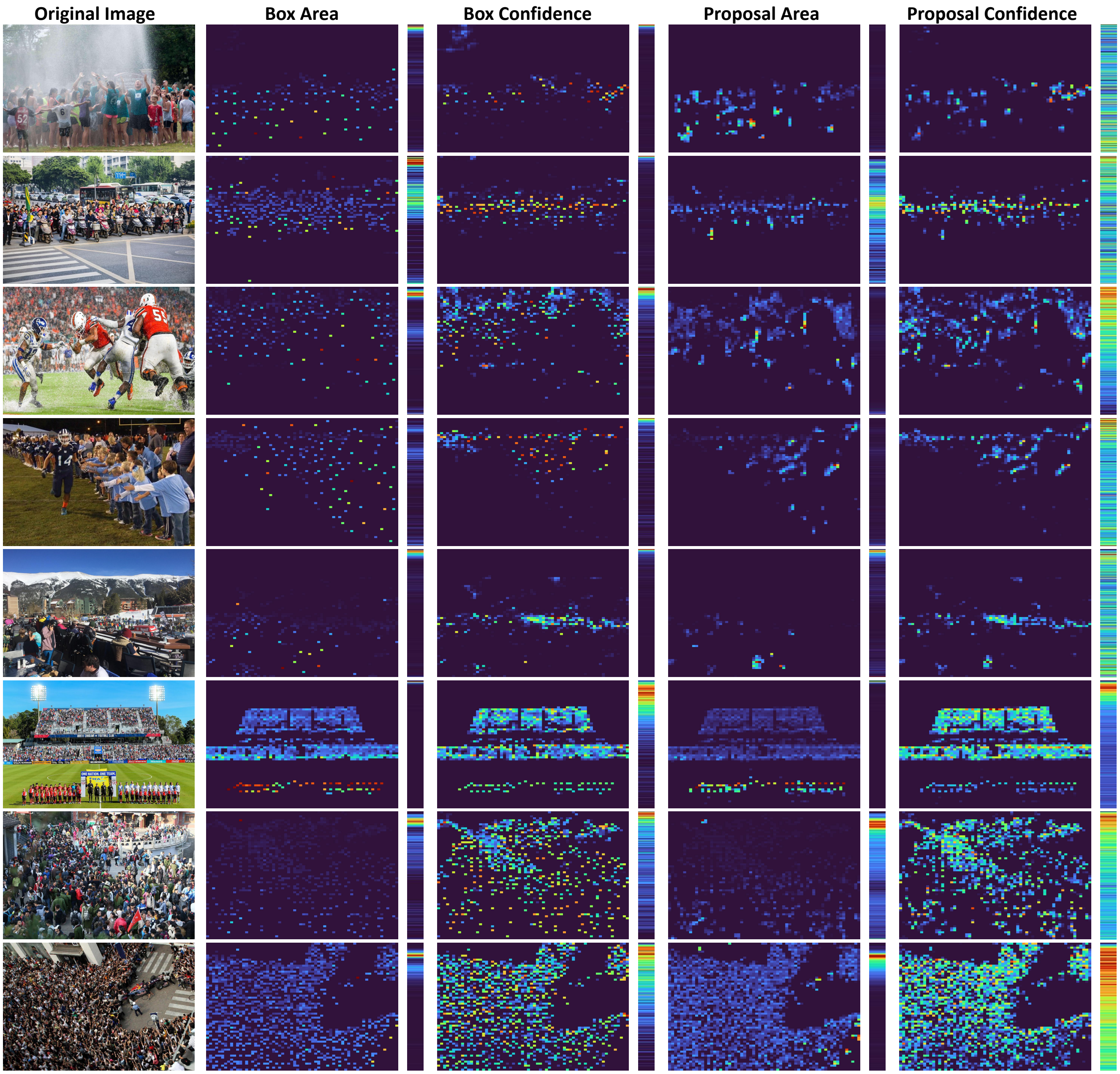

Figure 1. To visualize the 2D-1D feature compression, we present both the 2D compression matrices (on the left) and 1D distribution vectors (to their right) for each output feature. In the 1D distribution vectors, we denote 0 at the top and 1 at the bottom. Additionally, we include the original image in the leftmost column for reference. *Zoom in for better visualization.*

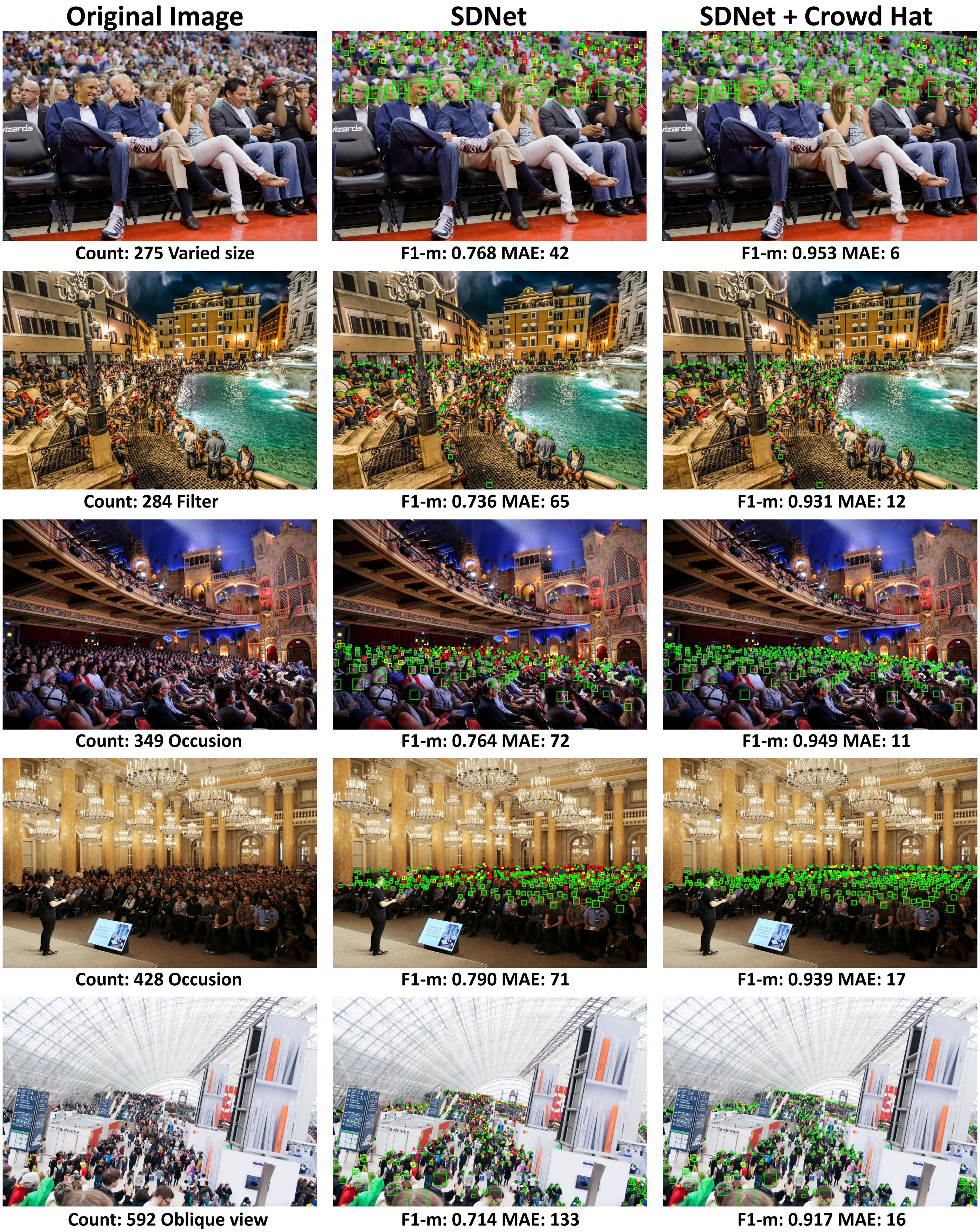

Figure 2. Visualization of Detection in Low Density Crowd.

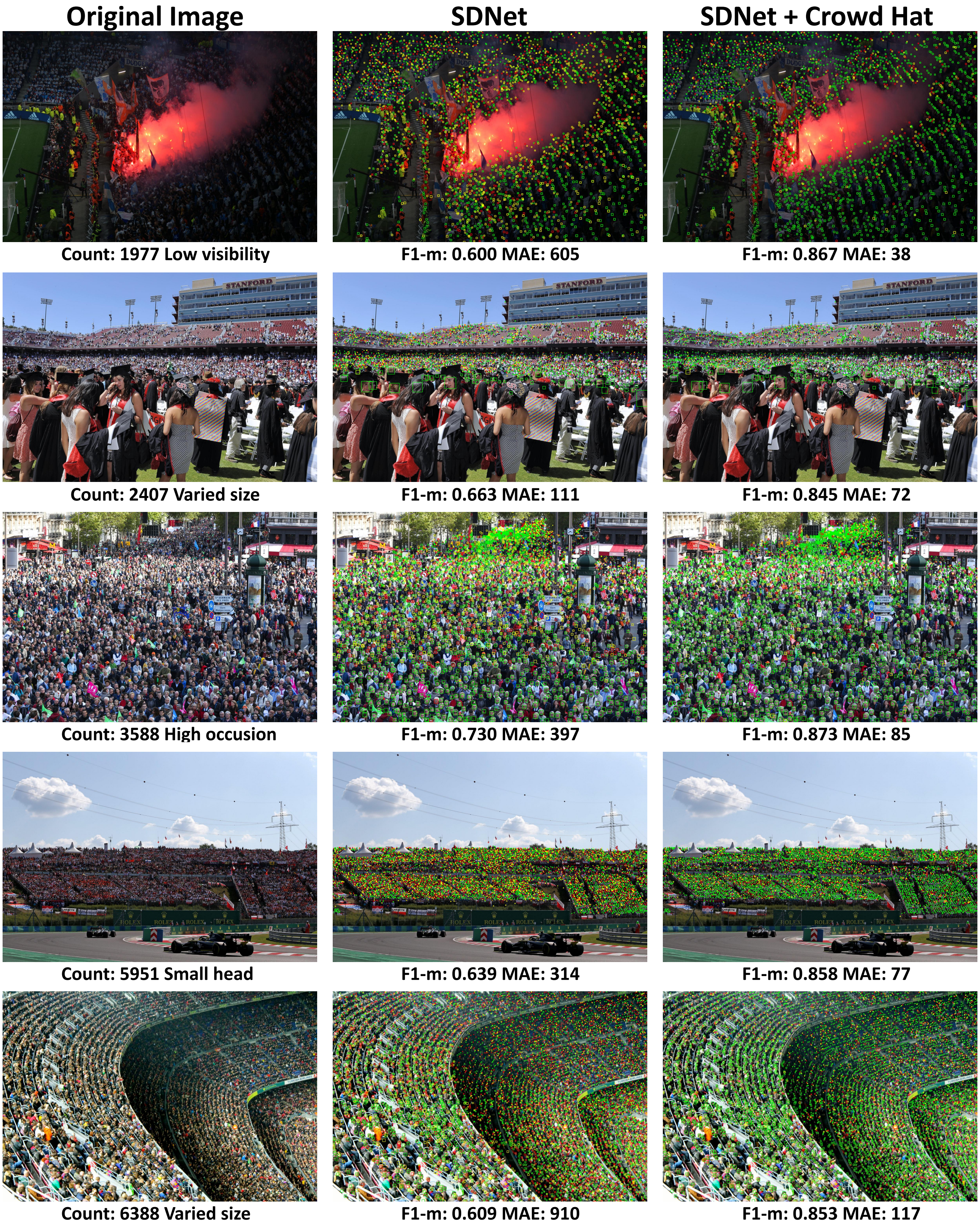

Figure 3. Visualization of Detection in High Density Crowd.

References

[1] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Máadeed, Nasir M. Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 544–559. Springer, 2018. 1

[2] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6469–6478. Computer Vision Foundation / IEEE, 2019. 1

[3] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. JHU-CROWD++: large-scale crowd counting dataset and A benchmark method. *CoRR*, abs/2004.03597, 2020. 1

[4] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6):2141–2149, 2021. 1

[5] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5525–5533. IEEE Computer Society, 2016. 1

[6] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 589–597. IEEE Computer Society, 2016. 1