WNet: Joint multiple head detection and head pose estimation from a spectator crowd image

Yasir Jan[0000−0002−9232−1703], Ferdous Sohel[0000−0003−1557−4907], Mohd Fairuz Shiratuddin[0000−0002−9529−6320], and Kok Wai Wong[0000−0001−8767−1031]

Murdoch University, Perth, Australia
{Y.Jan, F.Sohel, F.Shiratuddin, K.Wong}@murdoch.edu.au

Abstract. Crowd image analysis has various application areas such as surveillance, crowd management and augmented reality. Existing techniques can detect multiple faces in a single crowd image, but small head/face size and additional non facial regions in the head bounding box makes the head detection (HD) challenging. Additionally, in existing head pose estimations (HPE) of multiple heads in an image, individual cropped head image is passed through a network one by one, instead of estimating poses of multiple heads at the same time. The proposed WNet, performs both HD and HPE jointly on multiple heads in a single crowd image, in a single pass. Experiments are demonstrated on the spectator crowd S-HOCK dataset and results are compared with the HPE benchmarks. WNet proposes to use lesser number of training images compared to number of cropped images used by benchmarks, and does not utilize transferred weights from other networks. WNet not just performs HPE, but joint HD and HPE efficiently i.e. accuracy for more number of heads while depending on lesser number of testing images, compared to the benchmarks.

Keywords: Head detection · Head pose estimation · Crowd Analysis.

1 Introduction

Research in crowd analysis is focused on various aspects of crowd such as crowd counting [9], [26], face detection [15], [16], body detection [19], [5] and HPE [15], [16], [19], [2], [1]. These tasks become more challenging with the increasing number of people in the crowd, body occlusions and the low resolution features [26]. This paper focuses on joint multiple HD and multiple HPE of small sized heads present in a crowd image.

Previously, research has aimed towards single task of face detection [8], HD or HPE, while others propose joint HD and HPE [15], [16], [2], [1], [22], [24] but with few limitations. Some techniques can do face alignment and HPE in a joint manner [24] but are applicable for images with single heads only. Other techniques do multiple HD but do not perform multiple HPE [15], [16], [19]. Multiple HPE is done by cropping individual faces/heads and then individually estimating the pose of each face/head one by one. Head dimensions is also an issue for some techniques [15]. They can detect faces of varying dimensions but perform weakly in small sized faces, because they are dependent on other pretrained networks [6]. To detect small sized faces upscaling of low resolution images to high resolution images is proposed [6], which also increases computation. Yet these architectures are only applicable for single HPE at a time. Therefore, as far as we know, there is no existing technique which performs joint HD and HPE of multiple small sized heads.

The proposed WNet architecture, consists of two cascaded UNets [17], [7], followed by a grouping module. Previously, UNets have been used for image transformation techniques such as image de/colorization, sketch to image conversion, aerial image to map generation, facades generation, image de-noising / de-snowing, image segmentation and vice versa [17], [7], [23]. In tasks such as aerial to map generation and facades generation, detailed aerial and facade images are transformed into simplified unicolor regions. Based on these previous ideas, in the proposed architecture, UNets are first used to transform cluttered crowd images into simplified color blocks based images. These simplified images give a better visual understanding of the scene rather than the original cluttered scene. Therefore, using the transformed simplified images, simple color search and euclidean distance based techniques can be utilized to do HD and HPE.

The main contributions in this paper are as follows:

- To the best of our knowledge, this is the first CNN based technique to perform joint HD and HPE of multiple heads.
- It does not need each head to be cropped and passed to the network individually for HPE.
- WNet generates intermediate secondary output images, which reduces the complete black box effect of the end to end network pipeline.

The rest of the paper is organized as follows. Section 2 discusses the previous work related to head detection and head pose estimations. Section 3 discusses our methodology, as well as architecture. Section 4 discusses the S-HOCK dataset. Section 5 explains our experimental protocol while in Section 6 the results are shown. Section 7 concludes the paper, while Section 8 discusses the future possible improvements in this work.

2 Related Work

There are various CNN based approaches for face detection [15], [16]. Some techniques are focused on single person face datasets [13], and do not target crowd images. For multiple face datasets, such as WIDER Face [25], CNN based networks trained on the Imagenet dataset, are used for face detection. Since Imagenet have dimensions larger than 40 pixels [6], therefore, these networks cannot successfully locate and identify objects of smaller dimensions [6], [10]. For detecting images with tiny faces, image upscaling [6] is proposed, so that smaller faces become bigger in size and are then detected. Another solution is to have a network which can detect small dimensional objects, rather than images scaled to higher dimension and then detecting. Instead of upscaling, “super-resolved” CNN features [10] can be used to detect small objects. Super-resolving convolves small object CNN features and makes them similar to large object CNN features. Other than HD, some techniques do body detection as well, in which network can locate center of body parts, including heads [3]. Such body detection techniques may locate heads but are not aimed for HPE.

HPE is achieved using hand crafted features as well as deep learning based approach [15], [11], [14]. Deep learning based approaches in this aspect are very scarce as research is focused on facial landmark detection which indirectly is used for HPE [16], [18]. In low resolution images, the techniques depending on facial landmarks fail [18]. Therefore, HPE techniques should be independent of facial landmark detection. Existing HPE techniques have various other limitations. They are applicable to single person images, or they focus on multiple persons by individually cropping each face for pose estimations [15], [16]. Multiple faces pose estimations are not performed in a single forward pass of the network.

The network [15] proposes to fuse intermediate layer features and train separate sub networks for face detection, landmark localization and pose estimation. Similarly, [18] proposes to train a separate sub network for each head angle separately. Both these networks are aimed to calculate single head angles at a time, therefore cannot perform multiple HPE at the same time. UNets have also been used previously in a cascaded/stacked manner[4], [20] , but for different applications.

3 Method

In the proposed WNet method, we follow a 3 step approach, as shown in Figure 1. First step is to transform a crowd image into a less cluttered image. Aim is to remove all the non head pixels from the image, because they may cause clutter and errors in further steps. Therefore at first step a UNet [7] converts a crowd image to a head-region-masked (HRM) image. The generated HRM image will have all the non head pixel values suppressed to 0, while other head region pixels remain unaffected. Second step is to have an image with color markers. To reduce the effect of color variation in HRM heads, the HRM image is first converted to grayscale and then passed onto second UNet block. The second UNet generates a color coded head (CCH) image from grayscaled HRM image. In CCH image individual head regions and their centers are marked with different colors. The centers are marked with 3×3 white pixels, while the head regions are colored based on their pose. In the third step, CCH image is used and based on colors and euclidean distance, the head centers and the head regions are identified. Based on the color of the head region head poses are also identified. The details of the architecture is explained below.

3.1 WNet Architecture

The proposed WNet architecture, is composed of two UNet blocks, followed by a grouping module, as shown in Figure 1. Each UNet block layers are similar

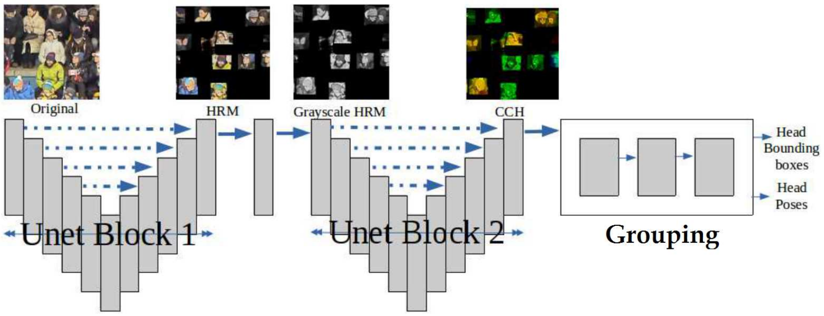

Fig. 1. The proposed WNet architecture, consists of two UNet blocks [7], followed by a grouping block. The UNet blocks transform the original image to a color coded head image. The grouping block locates the marked head centers, then bounding boxes and then head poses. (Best viewed in color)

to [7] . Each block consists of 8 encoder and 8 decoder layers. The number of output filters at each encoder layer are in the sequence: 64, 128, 256, 512, 512, 512, 512. There are skip connections between encoder and corresponding decoder layers. Therefore at each decoder layer input, the output from previous layer is concatenated with the corresponding encoder layer output. The number of filters in the decoder are in the sequence: 512, 512, 512, 512, 256, 128, 64, 3. Encoder layers are activated with LeakyReLU while decoder has ReLU activation functions. At all layers filter size is 4×4, with stride 2 and padding 1 in each direction. Both UNets are trained separately, using conditional Generative Adversarial Networks (cGANs), similar to [7].

UNet block 1: The first UNet block is trained to convert a cluttered RGB crowd image into a simplified HRM RGB image. The output HRM image is generated by suppressing all R,G,B channel values of the background and body part pixels, and making them 0. The head region pixels remain unchanged. The HRM image is then converted to grayscale by averaging all R,G and B channel values. As a result the inter channel color variation is finished and all the heads become almost similar in color, irrespective of their original color shades. This reduces the negative effect caused due to skin or hair color.

UNet block 2: The second UNet block takes the grayscaled HRM image as input, and an output CCH image is generated. In CCH image different head pose will be represented by activating only a combination of RGB color channels, and suppressing the rest to value 0. For example, a right pose head region will have only R channel activated, while G and B channels are suppressed to 0. Similarly, a left head pose will have only G channel activated and rest of the channels will be suppressed to 0. Therefore, for the five head poses i.e. right, left, front, away and down, the respective activated color channel combinations used are R, G, B, RG and GB respectively. Additionally, a 3×3 white pixel block is created in the middle of each head region, which is later used to identify the head center.

Therefore, each head's pixels in the input image are replaced with the channel selected head's pixels with an additional white pixel block in the center of head.

Grouping: The CCH image generated from the previous UNet block is used to extract the head centers and head pose. This step can be divided into 3 sub steps. (*i*) The mid positions of 3×3 white pixels in the image are extracted. It identifies the head centers. Noise correction is done, by averaging head centers within the half head size range. (*ii*) Group all the head pixels, which are around their centers within the average half head size range. The minimum and maximum pixel positions of each group gives the bounding box values of that head. (*iii*) Identify head poses of each head. In a CCH image, all the pixels within the head bounding box are color coded based on the pose. Therefore, the color codes of all the pixels within the head region are aggregated, and the maximum color is chosen to identify the head pose.

WNet generates simplified uncluttered crowd images at intermediate steps in the pipeline before calculating head bounding boxes, while other techniques first generates numerical values for head bounding boxes and then mark the head boundaries. These simplified images can be useful in applications where exact head locations are not required but a clear visual understanding of the scene e.g. crowd surveillance, augmented reality, is required. This intermediate step also reduces the full blackbox effect of the networks. Each intermediate step of the network can be fine tuned for further improvement.

4 Dataset

The Spectator Hockey (S-HOCK) dataset [19] is used for the experiments. It is the only publicly available dataset for spectator crowd with small heads annotated with head bounding boxes and head poses. It has a total of 75 videos (5 camera views × 15 different matches). Video from only a single camera has been annotated with head bounding box and head pose values. Each videos is of 31 sec duration is split into 930 annotated frames. Out of the 15 matches, 2 are used for training, 2 for validation and 11 for testing. Each frame has a resolution of 1280 × 1024 pixels with the dense crowd.

5 Experiments

S-HOCK has a total of 1860 annotated training frames (930 frames of 2 training videos) and 10230 annotated testing frames (930 frames of 11 testing videos). To reduce redundancy and also reduce the number of frames, 9 frames out of 10 are skipped. A total of 93 frames are extracted from each 930 frame video ($5 : 10 : 930$). Therefore the interval between two consecutive frames is approx 1/3 of a second. It should be observed that only 186 frames are used for training which is far lesser number than the thousands of ImageNet data, and 1023 frames are used for testing. In 1023 testing frames there are a total of 155085 people, with average head dimensions of width 36.46 pixels and height 40.69 pixels. The range of head dimensions in testing images is shown in Figure 2.

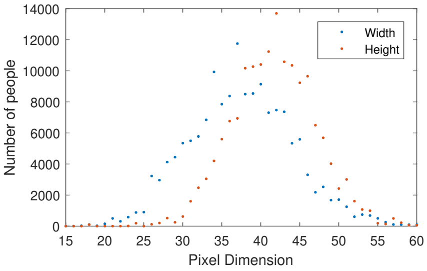

Fig. 2. Head dimensions of people in testing images.

Each frame is further sliced into 20 subframes to maintain a resolution of 256 × 256. The proposed WNet architecture consists of 2 UNets, each one of them trained separately. Both UNets are trained with their own input output image pair as discussed below.

UNet block 1: All the 186 frames are sliced into 20 each subframe (totaling 3720 subframes). Input 3720 subframes are paired with output 3720 HRM subframes, for generating input output training pairs. Training of UNEt block 1 with these input output pairs is done for 30 epochs.

UNet block 2: Block 2 is trained using the grayscaled HRM and respective RGB CCH frames of the 186 training frames of the dataset. Input HRM subframes are grayscaled and paired with respective RGB CCH subframes, for generating input output training pairs. But there is a class imbalance problem regarding the head pose i.e. left pose heads in the dataset are more than the right pose heads. Therefore, to solve the class imbalance issue, horizontally flipped 3720 sub frames are added into the training data. It makes a total of 7440 training subframes. Flipped and grayscaled HRM subframes are paired with respective flipped CCH subframes, for generating additional input output training pairs. For flipped images, the color of left and right pose will also be swapped. The 7440 input output pairs are used for training the UNet2 block with 10 epochs.

6 Results

The proposed WNet technique is tested using 1023 testing frames sliced into resolution 256 × 256. The technique first detects heads and generates the bounding box. Then based on the bounding box it identifies the head pose of each detected head. After frame slicing, the heads at the edges of the subframes get sliced as well. HD and HPE results are generated by ignoring the sliced heads. Generally the Intersection over Union (IoU) of detected head bounding box region and ground truth head bounding box region greater than 0.5 is accepted as correct.

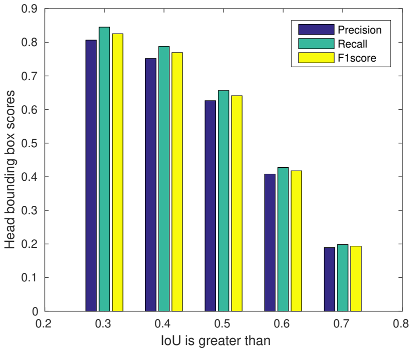

Fig. 3. Head bounding box accuracy.

Experiments are performed for multiple HD IoUs i.e. IoU greater than 0.3, 0.4, 0.5, 0.6, 0.7. The precision, recall and f1score of HD for varying IoU is given in the Figure (3). It achieves more than 0.6 precision, recall and F1 score for HD (IoU greater than 0.5).

The HPE benchmarks mentioned in the dataset paper [19] performs HPE on 34949 single head instances. An individual head is cropped and then head pose is estimated. Therefore, it takes 34949 cropped head images to generate the results of 34949 heads. Their accuracy ranges between 0.3 and 0.4, as shown in Table 1. In contrast, the proposed WNet architecture was tested with fewer number of frames (1023 frames / 20460 subframes). Although in these lesser testing frames, the total number of heads is 155085 which is ≈ 4.4 times more than the benchmark number of heads (34949). Therefore, WNet can do HPE for more heads with lesser frames. Additionally, this technique also performs HD, not just HPE. HPE accuracy is dependent on the HD accuracy. Therefore, the table 1 shows the accuracy for correctly estimated head poses with respect to correctly detected heads.

The accuracy of different IoU values is greater than 0.32. Our focus is on the number of accurately detected heads and head poses which is greater for WNet compared to other techniques. Results should be compared keeping in view that the benchmarks do not perform HD and can perform HPE on single cropped head. While proposed WNet jointly performs HD and HPE on multiple heads. Therefore, WNet uses lesser testing data frames and calculates higher number of accurate head poses.

7 Conclusion

Proposed WNet architecture, can perform joint multiple HD and HPE in a single forward pass. The technique is based on image to image transformation. First sub blocks of the architecture transforms the cluttered crowd images into color marked simplified images. Using the simplified images, head bounding box and

Table 1. S-HOCK dataset benchmark HPE average accuracy compared with WNet HPE accuracy. Benchmarks are tested on 34949 cropped heads images. While WNet is tested on only 1023 full scene frames (155085 heads). For IoU = (0.3, 0.4, 0.5, 0.6, 0.7), $HD \approx 0.8, 0.7, 0.6, 0.4, 0.2$.

Method	Avg Acc.	**Acc HPE headcount**	**Method type**
Orozco [12]	0.368	≈ 12861	Only single HPE
WArCo [21]	0.376	≈ 13140	Only single HPE
CNN [19]	0.346	≈ 12092	Only single HPE
SAE [19]	0.348	≈ 12162	Only single HPE
WNet (IoU = 0.3)	0.321	≈ **39825**	Joint multiple HD and HPE
WNet (IoU = 0.4)	0.323	≈ **35064**	Joint multiple HD and HPE
WNet (IoU = 0.5)	0.325	≈ **30241**	Joint multiple HD and HPE
WNet (IoU = 0.6)	0.337	≈ **20905**	Joint multiple HD and HPE
WNet (IoU = 0.7)	0.34	≈ 10545	Joint multiple HD and HPE

head pose information are extracted. The proposed technique has been tested on S-HOCK spectator crowd dataset. Results of HD and HPE show that the proposed WNet technique uses lesser number of frames and accurately estimates pose of more number of heads. WNet is an image to image transformation technique, which is also useful for applications where exact head dimensions are not required but simplified crowd images may be useful for a better visual understanding.e.g. crowd surveillance.

8 Future Work

This technique currently aims towards spectator crowd dataset, but it could b extended for heads/faces in the wild. It could benefit from heads segmented datasets, which could improve HD accuracy. This may result in sharper head extraction. Additionally a range of colors maybe assigned for a range of head poses which can be correlated. Instead of image to image transformation, head bounding box values can be extracted using other techniques, which can be further used for pose estimation.

9 Acknowledgment

This work was partially supported by an Australian Research Council grant DE120102960 and Murdoch University grant.

References

1. Asteriadis, S., Karpouzis, K., Kollias, S.: Face tracking and head pose estimation using convolutional neural networks. In: Proceedings of the SSPNET 2Nd

International Symposium on Facial Analysis and Animation. pp. 19–19. FAA '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1924035.1924046, http://doi.acm.org/10.1145/1924035.1924046

2. Bao, J., Ye, M.: Head pose estimation based on robust convolutional neural network. Cybern. Inf. Technol. **16**(6), 133–145 (Dec 2016). https://doi.org/10.1515/cait-2016-0083, https://doi.org/10.1515/cait-2016-0083
3. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1302–1310 (July 2017). https://doi.org/10.1109/CVPR.2017.143
4. Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., Sommer, W.H., Ahmadi, S.A., Menze, B.H.: Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. pp. 415–423. Springer International Publishing, Cham (2016)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (Sept 2010). https://doi.org/10.1109/TPAMI.2009.167
6. Hu, P., Ramanan, D.: Finding tiny faces. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1522–1530 (July 2017). https://doi.org/10.1109/CVPR.2017.166
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5967–5976 (2017)
8. Jiang, H., Learned-Miller, E.: Face detection with the faster r-cnn. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). pp. 650–657 (May 2017). https://doi.org/10.1109/FG.2017.82
9. Kang, D., Ma, Z., Chan, A.B.: Beyond counting: Comparisons of density maps for crowd analysis tasks - counting, detection, and tracking. CoRR **abs/1705.10118** (2017), http://arxiv.org/abs/1705.10118
10. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1951–1959 (July 2017). https://doi.org/10.1109/CVPR.2017.211
11. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(4), 607–626 (April 2009). https://doi.org/10.1109/TPAMI.2008.106
12. Orozco, J., Gong, S., Xiang, T.: Head pose classification in crowded scenes. In: Proceedings of the British Machine Vision Conference. pp. 120.1–120.11. BMVA Press (2009), doi:10.5244/C.23.120
13. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 41.1–41.12. BMVA Press (September 2015). https://doi.org/10.5244/C.29.41, https://dx.doi.org/10.5244/C.29.41
14. Patacchiola, M., Cangelosi, A.: Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recognition **71**(Supplement C), 132 – 143

(2017). https://doi.org/https://doi.org/10.1016/j.patcog.2017.06.009, http://www.sciencedirect.com/science/article/pii/S0031320317302327

15. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2018). https://doi.org/10.1109/TPAMI.2017.2781233
16. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). pp. 17–24 (May 2017). https://doi.org/10.1109/FG.2017.137
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, pp. 234–241. Springer International Publishing, Cham (2015)
18. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. CoRR **abs/1710.00925** (2017), http://arxiv.org/abs/1710.00925
19. Setti, F., Conigliaro, D., Rota, P., Bassetti, C., Conci, N., Sebe, N., Cristani, M.: The s-hock dataset: A new benchmark for spectator crowd analysis. Computer Vision and Image Understanding **159**(Supplement C), 47 – 58 (2017). https://doi.org/https://doi.org/10.1016/j.cviu.2017.01.003, http://www.sciencedirect.com/science/article/pii/S1077314217300024, computer Vision in Sports
20. Shah, S., Ghosh, P., Davis, L.S., Goldstein, T.: Stacked u-nets: A no-frills approach to natural image segmentation. CoRR **abs/1804.10343** (2018), http://arxiv.org/abs/1804.10343
21. Tosato, D., Spera, M., Cristani, M., Murino, V.: Characterizing humans on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1972–1984 (Aug 2013). https://doi.org/10.1109/TPAMI.2012.263
22. Vu, T., Osokin, A., Laptev, I.: Context-aware CNNs for person head detection. In: International Conference on Computer Vision (ICCV) (2015)
23. Wang, C., Xu, C., Wang, C., Tao, D.: Perceptual adversarial networks for image-to-image transformation. IEEE Transactions on Image Processing **27**(8), 4066–4079 (Aug 2018). https://doi.org/10.1109/TIP.2018.2836316
24. Xu, X., Kakadiaris, I.A.: Joint head pose estimation and face alignment framework using global and local cnn features. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). pp. 642–649 (May 2017). https://doi.org/10.1109/FG.2017.81
25. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 833–841 (June 2015). https://doi.org/10.1109/CVPR.2015.7298684