Clustered Object Detection in Aerial Image –Supplementary Material–

Anonymous ICCV submission

Paper ID 4194

1. Cluster Ground Truth Generation

(a) object boxes (b) clusters (c) cluster boxes

Figure 1: Illustration of generating cluster bounding box annotations. The red boxes and dots are the ground truth bounding boxes and center points of objects. Green and yellow boxes are the clusters after applying the meanshift[1] algorithm. Green and yellow dashed boxes are the generated cluster bounding box annotations.

Given the objects bounding box annotations, $G = \{g_i\}_{i=1}^{N}$, where i is the index of the object, $g_i = \{x1_i, y1_i, x2_i, y2_i\}$, $x1, y1$ and $x2, y2$ are the location of top-left and bottom-right points of the box, respectively. Based on the G (red boxes in Figure 1a), we can obtain the center points of these objects, G_c (red dots in Figure 1a). Then, the meanshift[1] algorithm is employed on G_c to seek the clusters center points C_c (green and yellow points in Figure 1c). The object bounding box annotations belong to a cluster j are denoted as G_{cj}. The cluster bounding box annotations are produced by finding $min(\{x1_i | i \in G_{cj}\})$, $min(\{y1_i | i \in G_{cj}\})$ and $max(\{x2_i | i \in G_{cj}\})$, $max(\{y2_i | i \in G_{cj}\})$ in cluster j (green and yellow dashed boxes in Figure1c).

2. Non-Max Merging (NMM)

The concrete process of the non-max merging is presented in Alg.1. In specific, given a set of cluster bounding boxes, $\mathcal{B} = \{B_i\}_{i=1}^{N_\mathcal{B}}$, the corresponding confidence scores, $\mathcal{S} = \{S_i\}$, and overlap threshold, τ_{op}, we first determine if the $\mathcal{B}$ is empty or not, if it is, no operation will be conducted. When the $\mathcal{B}$ is not empty, we find the index, m, of the box with the highest score and remove it from the $\mathcal{B}$. Among the rest boxes, we select the boxes, whose overlap with B_m are equal and larger than τ_{op}, then merge the these boxes with B_m. Here, the merge operation is to find the minimum of the top-left point and maximum of the bottom-right point among the boxes and B_m. After merging operation, we eliminate the merged boxes from $\mathcal{B}$ and their score from $\mathcal{S}$, then combine the generated box, M, with $\mathcal{B}'$. We repeat the aforementioned steps until the $\mathcal{B}$ is empty. Thus, the $\mathcal{B}'$ and $\mathcal{S}'$ are the boxes and corresponding scores of the merged cluster detections.

Algorithm 1: Non-Max Merging (NMM)

Input: $\mathcal{B} = \{B_i\}_{i=1}^{N_\mathcal{B}}$,$\mathcal{S} = \{S_i\}_{i=1}^{N_\mathcal{B}}$, τ_{op};
Output: $\mathcal{B}' = \{B'_i\}_{i=1}^{N_{\mathcal{B}'}}$,$\mathcal{S}' = \{S'_i\}_{i=1}^{N_{\mathcal{B}'}}$;
begin
  $\mathcal{B}' \leftarrow \{\}$;
  while $\mathcal{B} \neq \emptyset$ **do**
    $m \leftarrow \arg\max S$;
    $M \leftarrow B_m$;
    $\mathcal{B} \leftarrow \mathcal{B} - B_m$;
    for $B_i \in \mathcal{B}$ **do**
      if $overlap(B_m, B_i) \geq \tau_{\text{op}}$ **then**
        $M \leftarrow merge(M, B_i)$; $\mathcal{B} \leftarrow \mathcal{B} - B_i$;
        $\mathcal{S} \leftarrow \mathcal{S} - S_i$
      end
    end
    $\mathcal{B}' \leftarrow \mathcal{B}' \cup M$;
  end
  $\mathcal{S}' \leftarrow \mathcal{S}$;
  return $\mathcal{B}', \mathcal{S}'$;
end

3. Qualitative Results

We compare the detection results between Faster RCNN[3] and the proposed ClusDet in Fig 2 and show some qualitative results on VisDrone[6], UAVDT[2], and DOTA[4] in Fig 3, Fig 4, and Fig 5, respectively. The detection bounding boxes with score higher than 0.8 are displayed. Different colors of the bounding boxes denote different object categories.

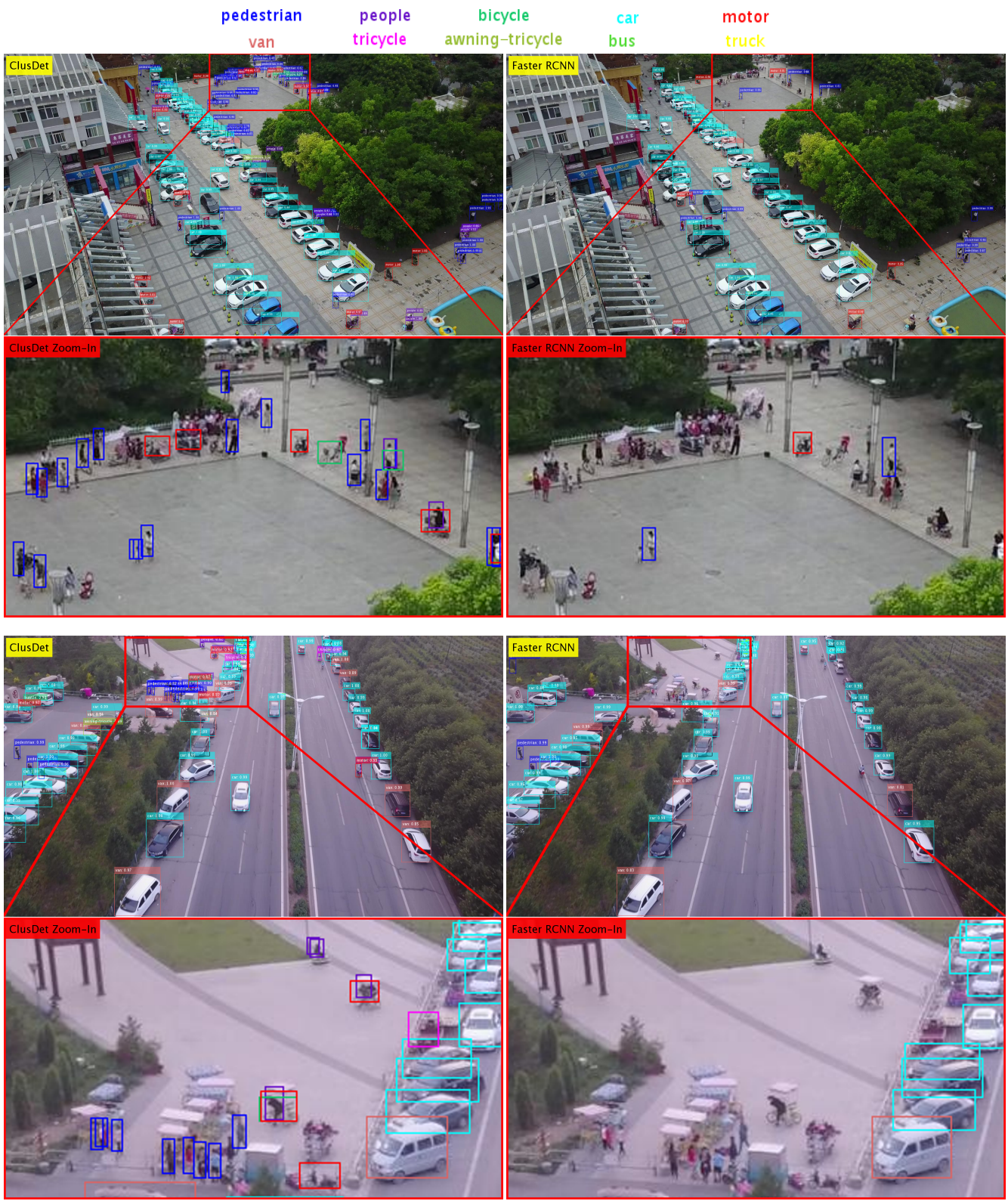

Figure 2: The comparison of the detection results between FRCNN[3] and the proposed ClusDet. The ResNeXt101[5] is used as backbone network. The red boxes indicate the cluster regions.

Figure 3: The visualization of the detection results of the proposed ClusDet on VisDrone[6] dataset. The ResNeXt101[5] is used as backbone network.

Figure 4: The visualization of the detection results of the proposed ClusDet on UAVDT[2] dataset. The ResNet50[5] is used as backbone network.

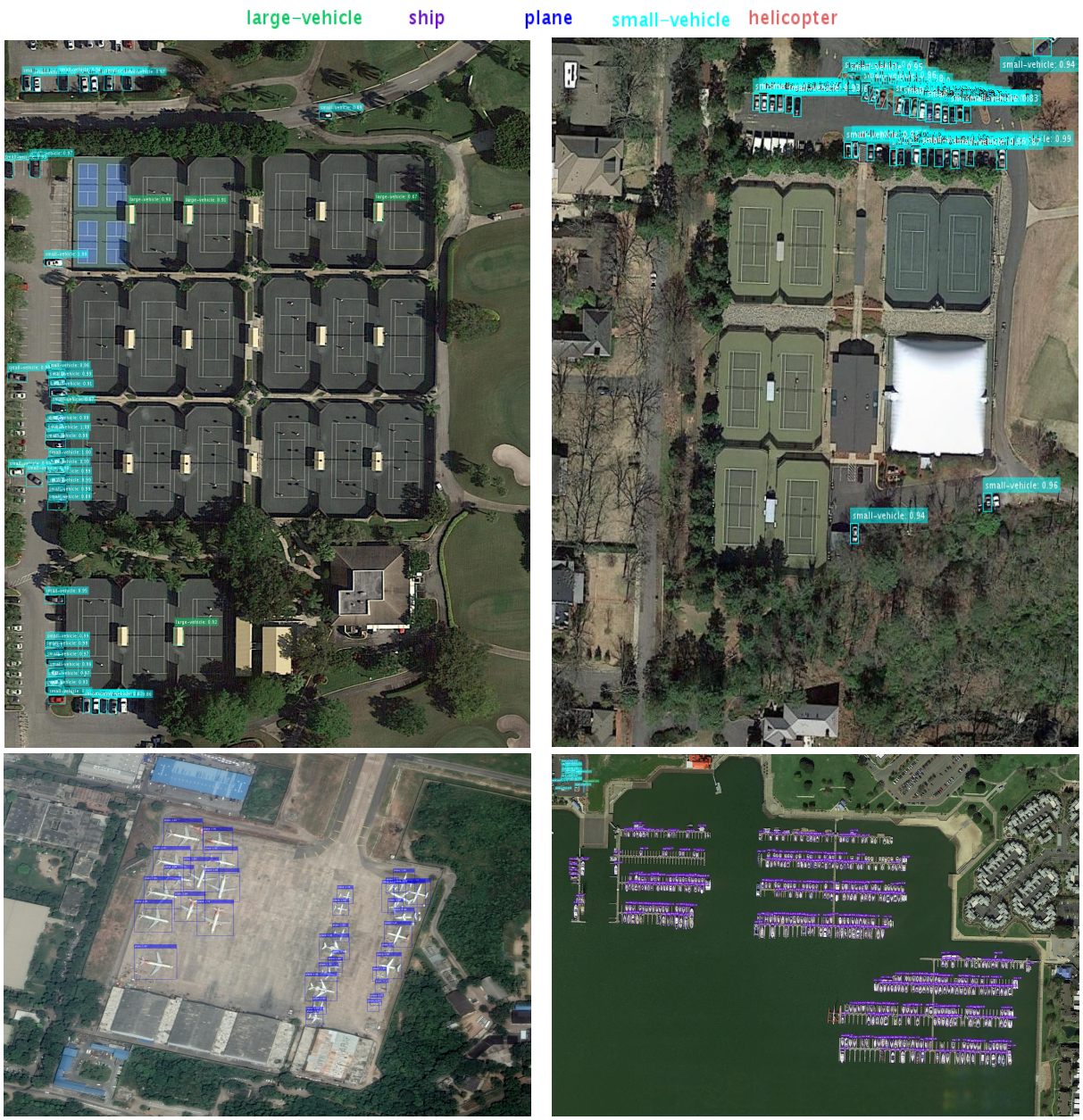

Figure 5: The visualization of the detection results of the proposed ClusDet on DOTA[4] dataset. The ResNet50[5] is used as backbone network.

References

[1] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, (5):603–619, 2002. 1

[2] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: object detection and tracking. In *ECCV*, 2018. 1, 4

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2

[4] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 1, 5

[5] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 2, 3, 4, 5

[6] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu. Vision meets drones: a challenge. *arXiv:1804.07437*, 2018. 1, 3