Density Map guided Object Detection in Aerial Images

Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, Shanyue Guan

Motivation

- Aerial images **have large object scale variance** due to different viewpoints, making detection challenging.
- Objects are unevenly distributed, leading to ineffective uniform **cropping strategies that miss important contextual information**.

Uniform cropping

Density map-based cropping helps generate more accurate regions for object detection, preserving context and reducing truncation.

Density cropping

Dataset

The benchmark **VisDrone** dataset consists of 288 video clips formed by **261,908 frames and 10,209 static images**, captured by various drone-mounted cameras, covering a wide range of aspects including location with 9 different classes

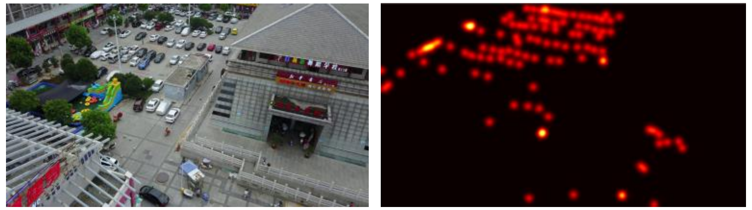
Goal

Train a **CNN model to produce density maps** that closely match ground truth, effectively identifying object-dense regions.

Use **density thresholds to selectively crop high-density areas**, minimizing background and capturing key object details.

Density-based cropping to improve detection accuracy and efficiency, especially for small, densely-packed objects.

Actual Image and Density Map

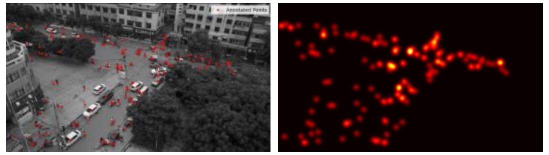


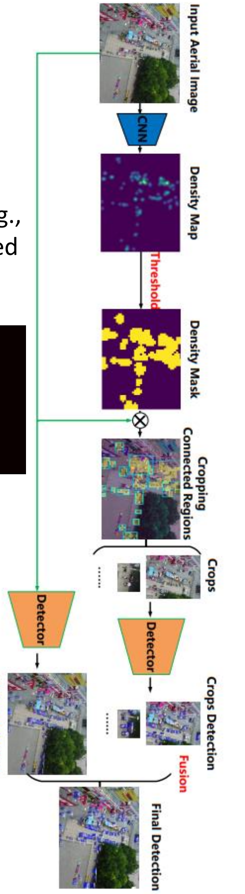
Methodology

DMNet Architecture: Combines density map generation, cropping, and fusion for accurate object detection.

Ground Truth Density Maps

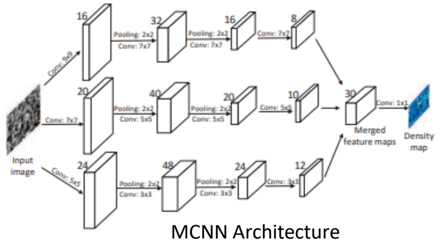
Class-wise Kernel: Adapts to large objects (e.g., buses) to avoid crop truncation instead of fixed kernel size.





Density Map Generation

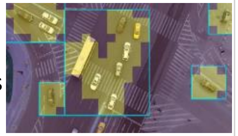
Uses Multi-column CNN (MCNN) for multi-scale feature extraction. Trained with pixel-wise error for generating density maps.



MCNN Architecture

Density Mask & Cropping

Sliding window and thresholding create a binary mask to identify dense regions. Adjusts density threshold to refine boundaries and reduce noise.



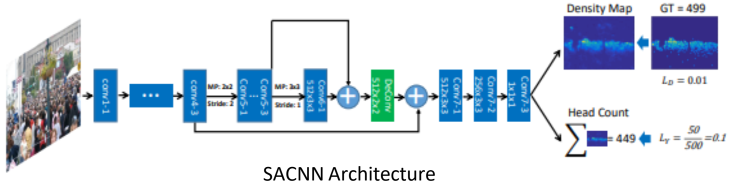
Fusion Detection

Merges detection results from both cropped regions and the full image. Applies non-maximum suppression for final, refined detection.

Improvements/Suggestions

1. Scale – Adaptive CNN (SACNN) over MCNN

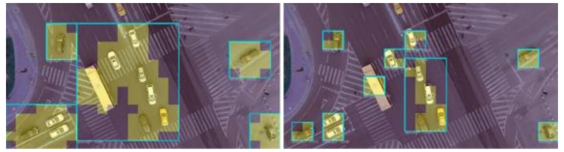
Single Column Backbone, SaCNN uses a single-column CNN with one filter size. **Scale Adaptation,** combines feature maps from multiple layers to handle scale variations. **Reduced Parameters,** by sharing low-level features across scales, SaCNN has fewer parameters.



SACNN Architecture

2. Image based adaptive thresholding

Instead of using a fixed threshold, this approach dynamically adjusts the threshold by first **performing object detection on the entire image to identify average bounding box sizes**. This information is then used to adaptively modify the threshold, which should improve cropping on smaller scale objects.



Threshold = 0.01 Threshold = 0.1

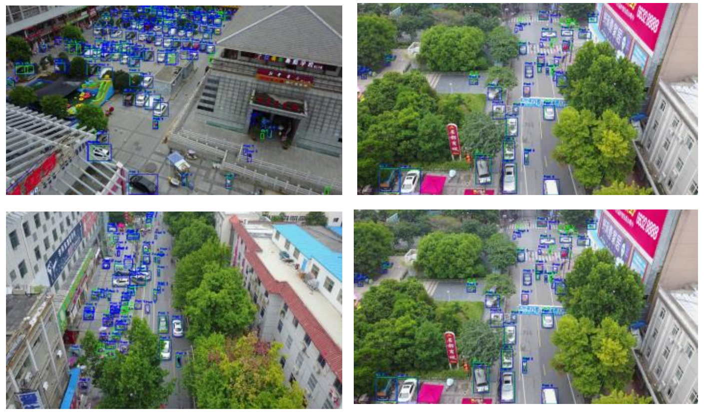
3. Speed improvements in density based cropping

The manual BFS approach has a worst-case complexity of $O(n^2)$ due to neighbor exploration and visit tracking, while **scipy.ndimage.label uses the Union-Find algorithm** with a near-constant time complexity of $O(n\alpha)$, which has better memory efficiency and faster performance.

4. YOLOv9 over YOLOv5 for object detection

YOLOv9 outperforms YOLOv5 **with improved accuracy, faster inference, and better model efficiency**, with advancements in architecture and optimization techniques and better detection performance, especially for smaller objects.

Results



MCNN obtained a pixel wise loss value of **19.843 on training data and 3.412 on validation data**

Speed-based improvements: Reduced time to process density cropping from 1830 sec to 30 sec on validation data, using the improvement suggested.

YOLOv5, v9 with adaptive threshold and RCNN based Resnet50 have been compared in the following table –

Metric	YOLOv5 Model	YOLOv9 Model	Best from Paper
AP	0.500	0.510	0.294
AP_{50}	0.504	0.520	0.532
AP_{75}	0.498	0.490	0.306
AP_{small}	0.608	0.620	0.216
AP_{medium}	0.493	0.500	0.412
AP_{large}	0.473	0.460	0.571

References

Li, M., Xu, Y., Wu, X., Jiang, Y., & Hu, Q.
"Density Map Guided Object Detection in Aerial Images."
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 10-11.

Zhang, L., Shi, M., & Chen, Q.
"Crowd Counting via Scale-Adaptive Convolutional Neural Network."
IEEE Winter Conference on Applications of Computer Vision (WACV), 2018.
Available at: https://doi.org/10.48550/arXiv.1711.04433

data, and code at github.com/Monochrome901/DMNet_ee798_2nd_part