Video Individual Counting for Moving Drones

Supplementary Material

The supplementary provides more details for the paper "Video Individual Counting for Moving Drones", including the following aspects.

- Details about Training and Testing.
- More Visualization Results.
- More Examples of MovingDroneCrowd.
- Limitations.

A. Details about Training and Testing

Training details: Since MovingDroneCrowd videos have been sufficiently downsampled to eliminate redundancy, we randomly select frame interval δ in the range of $3 \sim 8$ to guarantee the training pairs contain diverse inflow and outflow pedestrian variations. For data augmentation, training images are downsampled so that the longer side does not exceed 2560 pixels and the shorter side does not exceed 1440 pixels, ensuring that the cropped images contain enough pedestrians. The cropping, flipping, and scaling strategies follow those in [1]. The initial learning rate is set as 1e-5 with a weight decay of 1e-6 and follows a polynomial decay with a power of 0.9. We use VGG16, initialized with ImageNet pre-trained weights, as the backbone for feature extraction. The model is implemented with PyTorch and trained on A800 GPUs.

Test details: Our model can receive images with irregular resolutions during testing. To reduce computational cost, the longer side and shorter sides of the input image are limited to no more than 1920 and 1080 pixels, respectively.

Fig. 4 shows that our method maintains reasonable performance across a wide range of frame intervals, demonstrating its robustness to interval variations. It achieves the best performance when $\delta = 4$, so we set the frame interval δ to 4 during testing on MovingDroneCrowd.

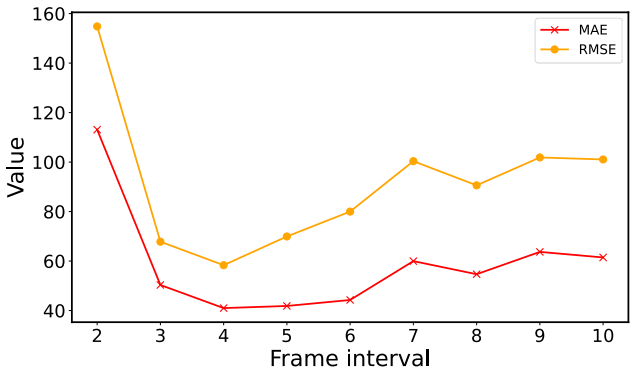

Figure 4. Ablation study of test frame interval δ on MovingDroneCrowd.

B. More Examples of MovingDroneCrowd

Fig. 1 presents additional video samples from our dataset MovingDroneCrowd, with each frame annotated with head bounding boxes and identity IDs. These examples highlight the key characteristics of our dataset: dense crowds, complex motion patterns, varying lighting conditions, and diverse camera heights and angles.

C. More Visualization Results

Fig. 2 presents additional visualization results of our method on MovingDroneCrowd. The first scene is a densely crowded scene with significant drone movement, while the second scene captures a sparsely populated area during low-altitude drone flight. Both scenes were recorded under low-light conditions. These results demonstrate that our method accurately predicts the inflow density map for each frame relative to its previous frame. This demonstrates that our method is sufficiently robust, achieving strong performance in complex environments, including dense, sparse, and low-light conditions.

Fig. 3 presents the visualization results of our method on the previous dataset UAVVIC. This scene was captured by a hovering drone with minimal camera movement, demonstrating that our method still performs well in static scenes.

D. Limitations

The visualization results on the test set show that the shared density map is not perfectly learned and still contains many erroneous responses, leading to some errors in the inflow and outflow density maps as well. Due to the similarity in pedestrian appearance, directly learning shared pedestrian features across two frames remains a challenging task. Computing cross attention between two frames is computationally expensive and time-consuming. These issues will be addressed in our future work.

References

[1] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. Dr.vic: Decomposition and reasoning for video individual counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3083–3092, 2022. 1

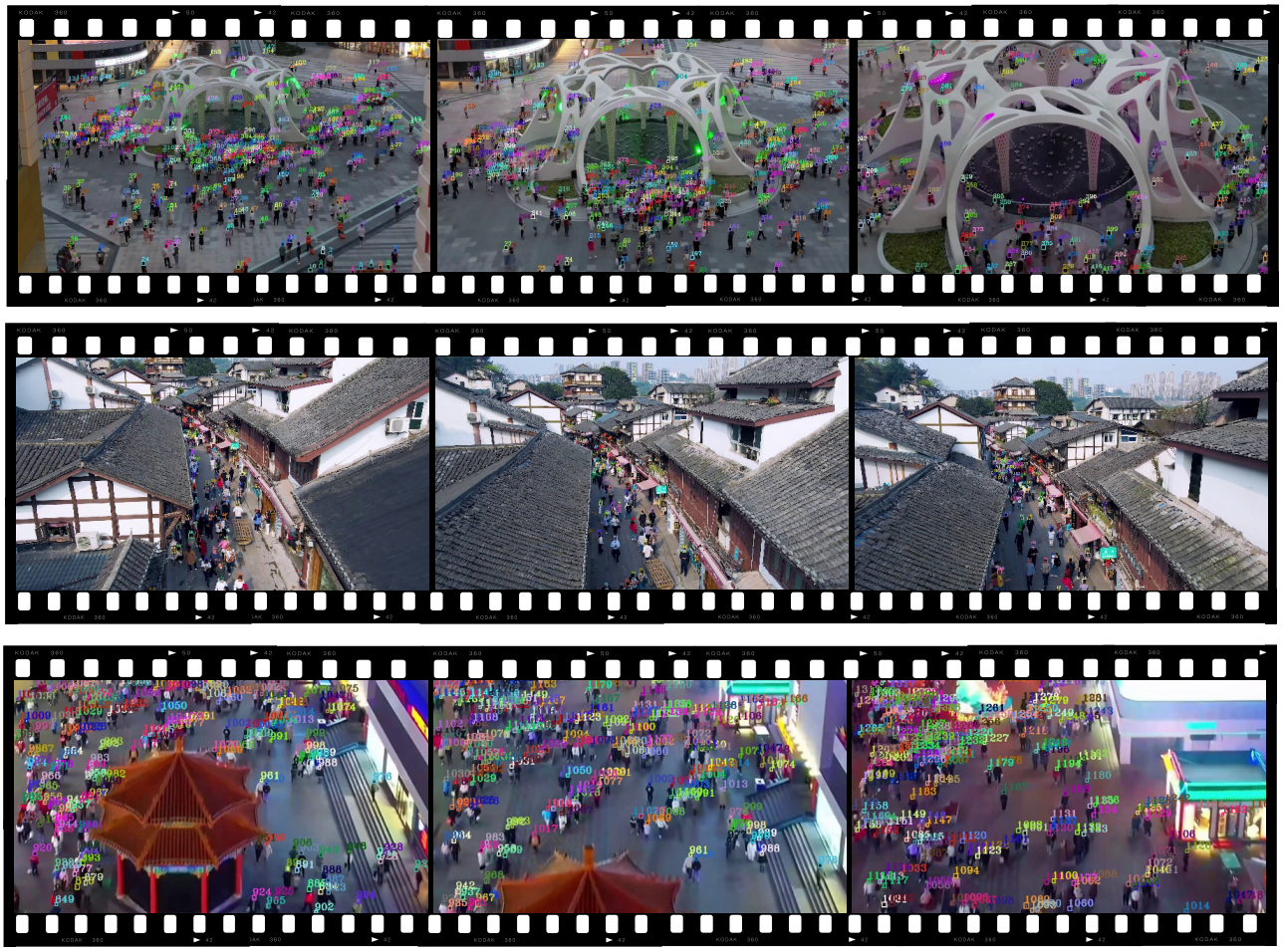

Figure 1. Additional samples from dataset MovingDroneCrowd. Due to space constraints, only three frames from each video are shown, with each frame annotated with head bounding boxes and ID labels.

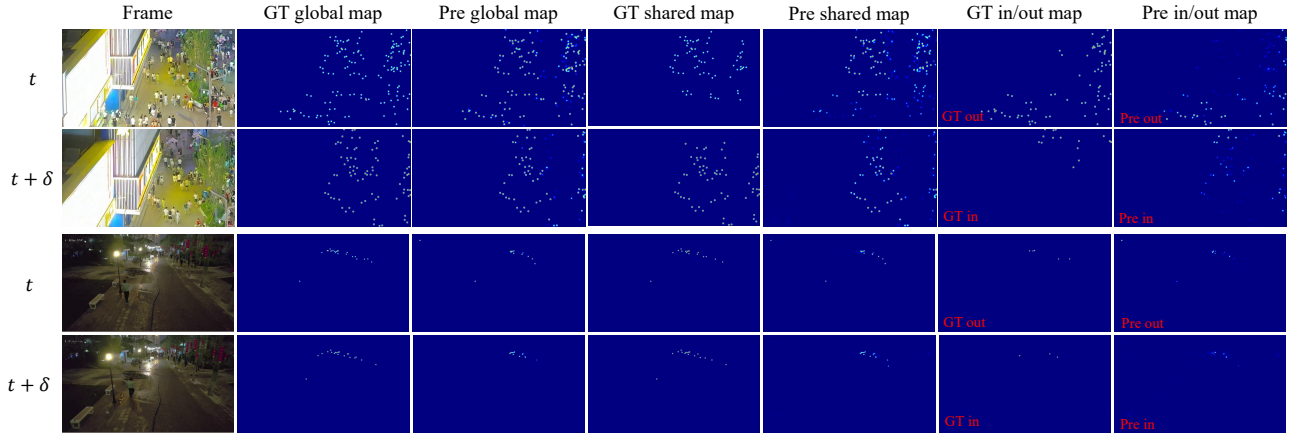

Figure 2. Additional visualization results of our method on dataset MovingDroneCrowd. These results demonstrate that our method performs well in low-light, dense, and sparse scenes.

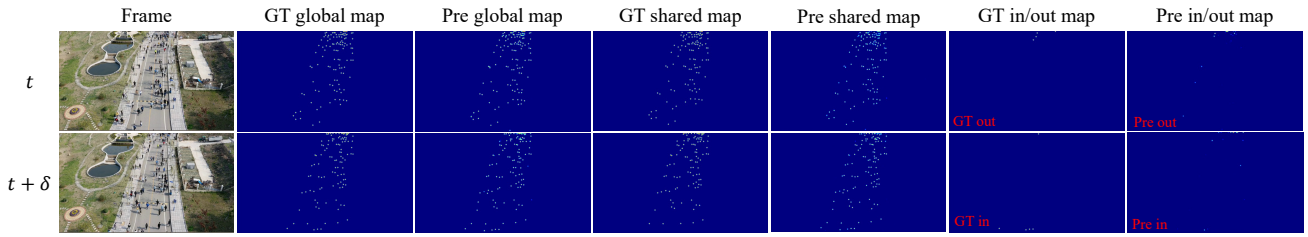

Figure 3. Additional visualization results of our method on dataset UAVVIC. It indicates that our method also achieves satisfactory performance in static scenes.