VisDrone-CC2021: The Vision Meets Drone Crowd Counting Challenge Results

Zhihao Liu[1], Zhijian He[2], Lujia Wang[2], Wenguan Wang[3], Yixuan Yuan[4],
Dingwen Zhang[5], Jinglin Zhang[6], Pengfei Zhu[1], Luc Van Gool[3], Junwei Han[5],
Steven Hoi[7], Qinghua Hu[1], Ming Liu[2], Junwen Pan[1],
Baoqun Yin[11], Binyu Zhang[10], Chengxin Liu[12], Ding Ding[8], Dingkang Liang[12],
Guanchen Ding[8], Hao Lu[12], Hui Lin[9], Jingyuan Chen[8], Jiong Li[9], Liang Liu[12]
Lin Zhou[8], Min Shi[12], Qianqian Yang[11], Qing He[11], Sifan Peng[11],
Wei Xu[10], Wenwei Han[8], Xiang Bai[12], Xiwu Chen[12], Yabin Wang[9],
Yinfeng Xia[11], Yiran Tao[8], Zhenzhong Chen[8], Zhiguo Cao[12]

[1]Tianjin University, Tianjin, China.
[2]The Hong Kong University of Science and Technology, Hong Kong, China.
[3]ETH Zurich, Zurich, Switzerland.
[4]City University of Hong Kong, Hong Kong, China.
[5]Northwestern Polytechnical University, Xian, China.
[6]Nanjing University of Information Science and Technology, Nanjing, China.
[7]Singapore Management University, Singapore.
[8]Wuhan University, Wuhan, China.
[9]Xi'an Jiaotong University, Xi'an, China.
[10]Beijing University of Posts and Telecommunications, Beijing, China.
[11]University of Science and Technology of China, Hefei, China
[12]Huazhong University of Science and Technology, Wuhan, China

Abstract

Crowding counting research evolves quickly by the leverage of development in deep learning. Many researchers put their efforts into crowd counting tasks and have achieved many significant improvements. However, current datasets still barely satisfy this evolution and high quality evaluation data is urgent. Motivated by high quality and quantity study in crowding counting, we collect a drone-captured dataset formed by 5,468 images(images in RGB and thermal appear in pairs and 2,734 respectively). There are 1,807 pairs of images for training, and 927 pairs for testing. We manually annotate persons with points in each frame. Based on this dataset, we organized the Vision Meets Drone Crowd Counting Challenge(Visdrone-CC2021) in conjunction with the International Conference on Computer Vision (ICCV 2021). Our challenge attracts many researchers to join, which pave the road of speed up the milestone in crowding counting. To summarize the competition, we select the most remarkable algorithms from participants' submissions and provide a detailed analysis of the evaluation results. More information can be found at the website: `http://www.aiskyeye.com/`.

1. Introduction

Crowd counting represents a wide range of counting tasks including but not limited to people, vehicles, animals, and commodities. It's a branch of crowd analysis and plays an important role in the field of video surveillance. In the past few years, an increasing number of researchers devote their efforts to crowding counting applications, *e.g.*, video surveillance, urban planning, as well as traffic control.

With increasing demands of accuracy and real-time analysis in crowding counting, many high-efficiency algorithms have sprung up like mushrooms recently. For example, Basic CNN [8] adopts basic CNN layers to construct the architecture. That includes convolutional layers, pooling layers,

and fully connected layers. It's an initial work using CNN to address problems of crowd counting or density estimation. After that methods like MCNN [30] start to adopt multi-column architecture models to solve the problem and have brought excellent performances in crowding counting tasks.

To improve the development of crowd counting algorithms, various datasets were collected by researchers to train and test, such as UCF_CC_50 [14], UCF-QNRF [15]. However, drone-vision datasets are still vacant due to challenging issue such as small objects and cluttered background. To step forward the progress in crowd counting methods, we organized the Vision Meets Drone Crowd Counting Challenge(Visdrone-CC2021) in conjunction with the International Conference on Computer Vision (ICCV 2021). We provide a drone-captured RGB-Thermal crowd counting dataset(DroneRGBT), which is a large-scale dataset and collected in many different scenes, with 168,583 annotated instances. Participants are invited to submit their results of crowd counting algorithms as well as their method description to avoid cheating in our competition. There are 8 algorithms from 5 institutions that are taken into consideration in this paper and we will provide a detailed analysis of the whole result. The detailed evaluation results and the leader-board can be found on our challenge website:`http://www.aiskyeye.com/`.

2. Related Work

In this section, we review the related crowd counting datasets and algorithms briefly.

2.1. Crowd Counting Datasets

To promote the development of crowd counting methods and create more challenges for counting tasks, more and more counting datasets have been created and used to train and test. There are several common used datasets such as UCSD [4], Mall [5], UCF-QNRF[15], NWPU-Crowd [27], and UCF_CC_50 [14]. UCSD [4] is the first dataset collected for crowd counting, it contains 2000 frames but there's one frame annotated in every five frames. Compared with UCSD [4], Mall [5] also contains 2000 frames but more pedestrians are annotated and the density of the crowd is considered in it. However, there exists more perspective distortion, which causes more scale change of the target objects. UCF_CC_50 [14] covers a wide range of densities and different scenes while it has only 50 images. More recently datasets such as UCF-QNRF [15] have considered the advantages mentioned above. And apart from that, it also contains diverse sets of lighting variations. NWPU-Crowd [27] contains $5,109$ images with $2,133,238$ annotated instances. Compared with previous datasets, it has the largest density range $0 \sim 20,033$ and contains various illumination scenes. It also has advantages such as negative samples and fair evaluation. Apart from the commonly used datasets mentioned above, there are also specialized datasets, which can be only used in certain scenarios. TS [7], STF [7] is about train station and EBP [31] for bridge. Furthermore, there are also many counting datasets for other objects. CARPK [12] is a car counting dataset, and TRANCOS [10] is the first dataset for vehicle counting. A detailed comparison of the datasets is shown in Table 1(Some statistics are cited from [9]).

2.2. Crowd Counting Methods

Recently a large number of researchers put their efforts into optimizing their crowd counting methods to tackle problems caused by variations in object shapes, heavy occlusion, and large variation on camera viewpoints. Zhang *et al*. [30], proposes a Multi-column Convolutional Neural Network (MCNN), whose Multi-column CNN architecture makes it adaptive to the various scales of people heads. Viresh Ranjan *et al*. [23] puts up with a two-branch CNN architecture that can further promote the resolution of the density map. In contrast with using multi-scale CNN architectures, Deepak Babu Sam *et al*. [1] presents a switching convolutional neural network that leverages the difference of crowd density in a frame to improve the accuracy of crowd counting. Although CNN-based methods have achieved unprecedented performances on crowd counting, the overfitting problem still exists in many architectures. Thus, Shi *et al*. [24] proposes a new method called D-ConvNet which uses a deep negative correlation learning strategy to produce generalizable features and independent of the backbone architectures. They compare their method with VGGNet and other state-of-the-art methods which proves its superiority. Furthermore, many researchers take the training loss function into consideration to achieve higher accuracy. Cao *et al*. [3] creates a novel encoder-decoder network and also proposes a novel training loss which is a combination of Euclidean loss and local pattern consistency loss. Ma *et al*. [20] uses the Bayesian loss function, which can construct a density contribution probability model to improve the performances of their model.

3. The VisDrone-CC2021 Challenge

The Vision Meets Drone Crowd Counting Challenge requires participants to count pedestrians from drone-captured video frames. Participants need to train their algorithms with the provided training dataset and submit their results with a method description file. They are allowed to train their model with an external training dataset while each team has 5 evaluation opportunities in total.

3.1. Dataset

The dataset for our challenge VisDrone-CC2021 is a drone-captured RGB-Thermal crowd counting dataset (DroneRGBT) which consists of 2734 pairs of images. It

Table 1. We summarize the minimal, maximal, average and total annotations of the crowd counting datasets.

Dataset	Type	Year	Frames	Min	Max	Ave	Total
UCSD [4]	surveillance	2008	2,000	11	46	24.9	49,885
UCF_CC_50 [14]	surveillance	2013	4,543	50	94	1,279.5	63,974
Mall [5]	surveillance	2013	2,000	13	53	31.2	62,316
CARPK [12]	aerial	2017	1,448	1	188	62.0	89,777
UCF-QNRF [15]	surveillance	2018	1,535	49	12,865	815.4	1,251,642
DroneCrowd [29]	aerial	2019	33,600	25	455	144.8	4,864,280
NWPU-Crowd [27]	surveillance	2020	5109	0	20,033	418	2,133,375
DLR-ACD [2]	aerial	2019	33	285	24368	6857	226,291
VisDrone-CC2020	aerial	2020	3,360	25	421	144.7	486,155
VisDrone-CC2021	aerial	2021	5,468	1	403	57.7	168,583

contains images taken from different locations, viewpoints, and various background clutters. As shown in Fig. 1, our dataset covers a wide range of scenarios such as streets, parks, playgrounds, and plazas. We divide the dataset into three parts according to its illumination, dark, dusk, and light. And two-thirds of each part is picked to form the training set, and one third for the testing set. Roughly, our dataset has these 3 properties:

- **Scale** Different scales of the object are considered in our dataset. We define three categories according to the height of the camera which significantly affect the diameter of objects: *High altitude* ($30-50$ meters high) and *Medium* ($50-80$ meters) and *Low* ($80-100$ meters).

- **Illumination** Our dataset is captured in various scenarios with 3 kinds of illumination conditions, *Dark*, *Dusk*, and *Light*.

- **Density** Density represents the number of objects in each frame. The density of our dataset lies in $1-403$ which covers most scenarios in our daily life. And different densities appear basically according to the probability in life.

3.1.1 Comparison to previous datasets.

Compared with the previous crowd counting datasets mentioned above, our proposed dataset DroneRGBT brings more challenges to modern counting methods.

- Our dataset has more high-perspective pictures which means that most objects are smaller(the diameter of objects ≤ 15 pixels) and more difficult to recognize.

- More low illumination scenes are considered. More than half of the pictures are in the dark or dusk environments both in the train and test dataset.

- The density of people covers a wide range and it's closer to reality.

- we provide RGB and Thermal pictures in pairs, which can extract more features than single RGB images.

3.2. Evaluation Protocol

In our evaluation work, we adopt mean absolute error(MAE) and mean square error(MSE) between the ground truth and the predicted number of people as our evaluation metrics, which are defined as follows.

$$\text{MAE} = \frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M} \sum_{j=1}^{N_i} |f_{i,j} - \hat{f}_{i,j}|, \tag{1}$$

$$\text{MSE} = \sqrt{\frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M} \sum_{j=1}^{N_i} |f_{i,j} - \hat{f}_{i,j}|^2}, \tag{2}$$

In the equation above, M is the number of video clips in our evaluation dataset, i represents the i-th video clip. N_i is the number of frames in i-th video clip. j represents the j-th frame in each video clip. $\hat{f}_{i,j}$ and $f_{i,j}$ are the predicted number of people heads and the ground truth in i-th video clip, j-th frame. MAE and MSE are calculated respectively but we adopt MAE as the basis of ranking.

4. Results and Analysis

In the following section, we evaluate the submitted algorithms in the Vision Meets Drone Crowd Counting Challenge(VisDrone-CC2021) and present a detailed analysis of these methods.

4.1. Submitted Methods

More than 70 participants submitted their results in the VisDrone-CC2021 Challenge. Here 8 submissions with complete counting results as well as method description files are selected to be discussed. In the following part, we

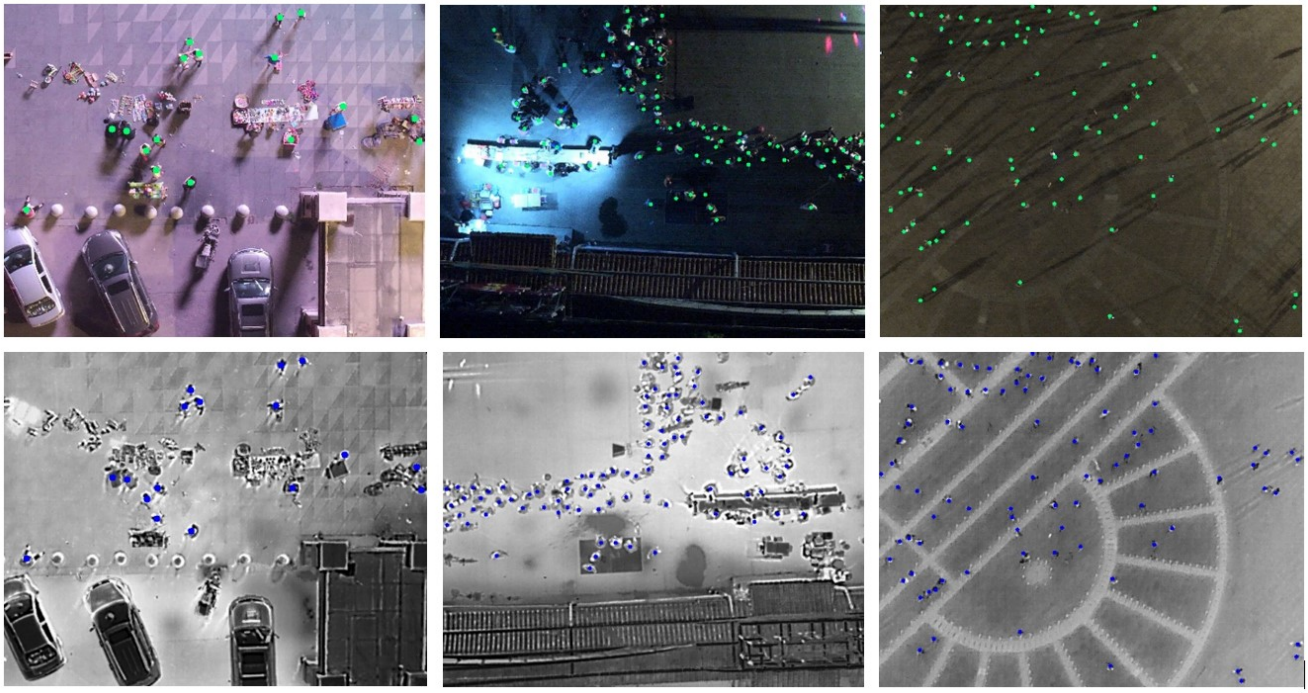


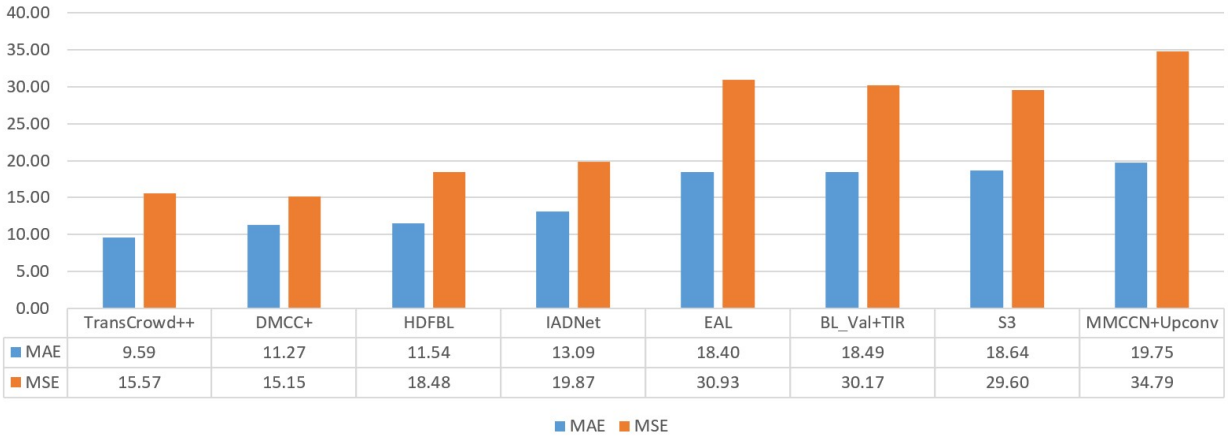
Figure 1. Annotation exemplars in the VisDrone-CC2021 dataset. Blue and Green dots represent person heads.

generally analyze these methods and try to tell the differences among them. More details about these methods description can be found in the Appendix A.

The majority of the submit methods have adopted various fusion mechanisms to fuse the extracted features from the dataset. TransCrowd++(A.1) takes Swin Transformer[19] as its backbone, and adopts a top-bottom fusion mechanism to make full use of the information extracted from each stage of the model. HDFBL(A.3) uses HDFNet [21] as their base model and Bayesian Loss [20] as their loss function. HDFNet [21] is a hierarchical dynamic filtering network and the TIR and RGB features are properly combined in this network. IADNet(A.4) uses a fusion mechanism called Information Aggregation-Distribution Module(IADM [17]). And the complete of IADNet(A.4) is a combination of IADM [17], BCNet [18], and OT Loss [26]. EAL(A.5) especially proposes a crowd head edge detector to generate auxiliary supervision signals to assist in generating a crowd density map. In detail, they use the pyramid fusion method to fuse extracted various features from RGB and TIR images, and then, concatenate the crowd head edge features and the density features to generate high-quality density maps. Apart from that, DMCC+(A.2) was developed based on MMCCN [22] and Density Map Learning(DML) [13] which uses a modified multi-branch Res-Net [11] to extract multiscale features. Additionally, different loss functions are adopted in the submitted algorithms. BL_Val+TIR(A.6) uses Bayesian loss as the loss function [20] and a standard backbone network VGG19 [25] without external detectors as well as multi-scale architectures. The novel loss function calculates a density contribution probability and have a good performance in our counting dataset. MMCCN+Upconv(A.8) uses MSELoss to supervise the result on pixel level and proposes a Block MAELoss to supervise the result on local. S3(Semi-balanced Sinkhorn with Scale Consistency,A.7) proposes a new measure-based counting approach to directly regress the predicted density maps to the scattered point-annotated ground truth. They formulated the counting task as a measure matching problem and designed a Sinkhorn counting loss to measure matching. In short, all of the algorithms mentioned above have done an excellent job and improved the counting performance from multiple perspectives.

4.2. Overall Results

The overall evaluation results are shown in Fig. 2. As is illustrated, the MAE score rises from 9.59 to 19.75 while the MSE score is between 15.15 and 34.79. TransCrowd++(A.1) gets the best MAE score profits from its Swin Transformer backbone and the top-bottom fusion mechanism. DMCC+(A.2) ranks second but its MSE gets the best score. It uses MMCCN [22] and Density Map Learning(DML) [13] to do the counting task. MMCCN is used to extract and fuse the features while DML is for generating better density map. Then follows HDFBL(A.3)



	TransCrowd++	DMCC+	HDFBL	IADNet	EAL	BL_Val+TIR	S3	MMCCN+Upconv
MAE	9.59	11.27	11.54	13.09	18.40	18.49	18.64	19.75
MSE	15.57	15.15	18.48	19.87	30.93	30.17	29.60	34.79

Figure 2. Comparison of submissions in the VisDrone-CC2021 Challenge.

and IADNet(A.4). HDFBL(A.3) uses HDFNet [21] as its Neural Network architecture and they further processed the dataset while training the model. In IADNet(A.4), a novel fusion mechanism named IADM [17] is used to fusion the features extracted from RGB images and thermal images. Then, the following three methods get almost the same grades around 18.5. However, there are big differences among them in detail. EAL(A.5) especially proposes a crowd head edge detector to generate auxiliary supervision signals to assist in generating a crowd density map. BL_Val+TIR(A.6) cites BL [20] algorithm, which calculates a density distribution probability model from annotations to do the counting task. And S3(A.7) also considered the inaccuracy of Gaussian assumption while generating density ground truth but proposes a new measure-based counting approach which directly regresses the predicted density map to the point-annotated ground truth. At the end of this table is MMCCN+Upconv(A.8) but it also gets a fine score compared with other submissions not mentioned in this paper. MMCCN+Upconv(A.8) still adopts Gaussian kernel to represent people but applied a small sigma to constrain crowded people from each other. As a conclusion, both of HDFBL(A.3) and BL_Val+TIR(A.6) get good performance using Bayesian Loss. Different fusion mechanisms are adopted in TransCrowd++(A.1), DMCC+(A.2), HDFBL(A.3), IADNet(A.4), EAL(A.5), and all of them have improved the counting performance to some extent.

4.3. Discussion

From the submit papers we can conclude that there are lots of aspects that can be further improved. Even the best result still has a 9.59 mean absolute error. It's not satisfying when we aim to apply the counting algorithms to real-world applications. Here, we propose several suggesting research directions of crowd counting:

- There are large-scale variations of crowd heads in reality. We can design a scale-aware model to further improve the accuracy of crowd estimation.

- The density distributions are very different for two modalities, and therefore how to automatically fuse the complementary multi-modal information is much challenging.

- Background clutter may cause the performance of the basic counting method to decrease to a large extent. There are massive unlabeled data from the web, which we can use for self-supervised or weakly-supervised learning.

5. Conclusions

In this paper, we summarize the VisDrone-CC2021 Challenge and give a detailed analysis of the results in this challenge. We collect a drone-based dataset consists of 5468 frames for participants to train and test their models. Compared with previous datasets, the VisDrone-CC2021 dataset provides RGB and Thermal pictures in pairs, which means that more information can be extracted and it raises new challenges to original counting methods. We select the top 8 results submitted on our website. Among them, TransCrowd++(A.1) gets the best MAE score of 9.59, from which we can see that there is still a lot of room for improvement. To further perfect our challenge and stimulate the development of crowd counting methods, we will expand the dataset with more original attributes and propose a fairer evaluation system. We sincerely hope our work can attract more researchers to join crowd counting algorithms redesign on drone-captured scenes, and further boost the development of computer vision.

A. Detailed Crowd Counting Algorithms considered in our paper

A detailed description of all the counting methods considered in the VisDrone-CC2021 Challenge is provided in this appendix.

A.1. RGB-T images crowd counting using Swin Transformer (TransCrowd++)

Dingkang Liang, Xiwu Chen, Wei Xu, Xiang Bai
{*dkliang, xiwuchen, xuwei2020, xbai*}*@hust.edu.cn*

TransCrowd++ adopts the Swin Transformer [19] as the backbone, and a top-bottom fusion mechanism is used to make full use of the variant spatial information extracted from different stages of the model(see Fig. 3). During the training phase, we train an independent TransCrowd++ model for RGB images and TIR images, respectively. We choose the Euclidean distance to measure the estimation difference at pixel level between the estimated density map and the ground truth. External dataset: According to the test guideline, we can utilize external dataset. The VisDrone2020 dataset [29] is also captured from the aerial view, which can be used as an external dataset for pretraining. This dataset only contains RGB images without TIR images. Hence, we adopt the DM-CycleGAN [22] to transform the RGB images into the TIR-style images. Data processing: For both RGB and TIR images, we set the Gaussian kernel as 4. We augment the training data by horizontal flipping, random brightness, saturation and contrast changing, adding noise. Additionally, we scale all images by 1.5 times offline to add more samples.

A.2. Density Map Learning based Crowd Counting Method for VisDrone RGB-T images (DMCC+)

Guanchen Ding, Lin Zhou, Ding Ding, Wenwei Han, Yiran Tao, Jingyuan Chen, Zhenzhong Chen
{*gcding, ramsey, 2018302130013, mikudiary, taoyiran*}*@whu.edu.cn, jingyuanchen1423@gmail.com, zzchen@whu.edu.cn*

DMCC+ was developed based on MMCCN [22] and Density Map Learning(DML) [13] for VisDrone 2021 Crowd Counting task. MMCCN used a modified multi-branch Res-Net [11] to extract multiscale features of visible and thermal modalities and model adaptive fusion for the extracted high-level features to predict the density maps. DML [13] presented the adaptive Gaussian density map generation network to obtain a better density map for the crowd counting applications. Besides, the modal-alignment method of MMCCN was considered to solve the misalignment problem of the visible and thermal modalities. During training, we choose the MSE loss to train our counting model. According to the permission of the challenge guideline, we used additional training data of MMCCN and used a variety of data augment methods, including random flipping, resizing, contrast enhancement, etc.

A.3. Hierarchical Dynamic Filtering Network with Bayesian Loss for RGB-TIR crowd counting (HDFBL)

Yabin Wang
iamwangyabin@stu.xjtu.edu.cn

We use modified multi-branch structure HDFNet[21] to do the counting, which we name HDF++. HDFNet is a simple yet effective hierarchical dynamic filtering network for RGB-D SOD tasks. Especially, The TIR and RGB features are combined to generate region-aware dynamic filters to guide the decoding in RGB stream. Bayesian loss is a novel loss function that constructs a density contribution probability model from the point annotations. Instead of constraining the value at every pixel in the density map, the proposed training loss adopts more reliable supervision on the count expectation at each annotated point.

We use additional training data to enhance our performance. The additional data is ECCV2020 Challenge DroneCrowd [29], but we only select night sequences (from 00069 to 00074). Since this challenge allows to use additional training data, and we indicate this issue here. Also, we pretrain the model using ImageNet and the RGBT Crowd Counting dataset provided, but I think almost all crowd counting methods need a pretraining process.

A.4. Information Aggregation-Distribution Network for RGBT crowd counting (IADNet)

Min Shi, Chengxin Liu, Liang Liu, Hao Lu, Zhiguo Cao
{*minshi, cx_liu, wing, hlu, zgcao*}*@hust.edu.cn*

Our method is a combination of IADM [17], BCNet [18], and OT loss [26]. We modify the backbone of IADM and use the paradigm of BCNet to count objects. Meanwhile, we also train density map-based models with OT loss. The final result is obtained by averaging the counting results from different models, including the BCNet-based model and density map-based method. Fig.4 gives an overview of the model structure. The network is composed of three branches: RGB branch, fusion branch, and thermal branch. RGB branch and thermal branch extract feature from RGB images and thermal images, respectively. A fusion branch is adopted to fusion the features from two different modalities. The fusion mechanism is called Information Aggregation-Distribution Module (IADM), which is applied between multiple intermediate feature maps of the RGB branch and thermal branch. This mechanism work on both fusing feature maps and distributing the in-

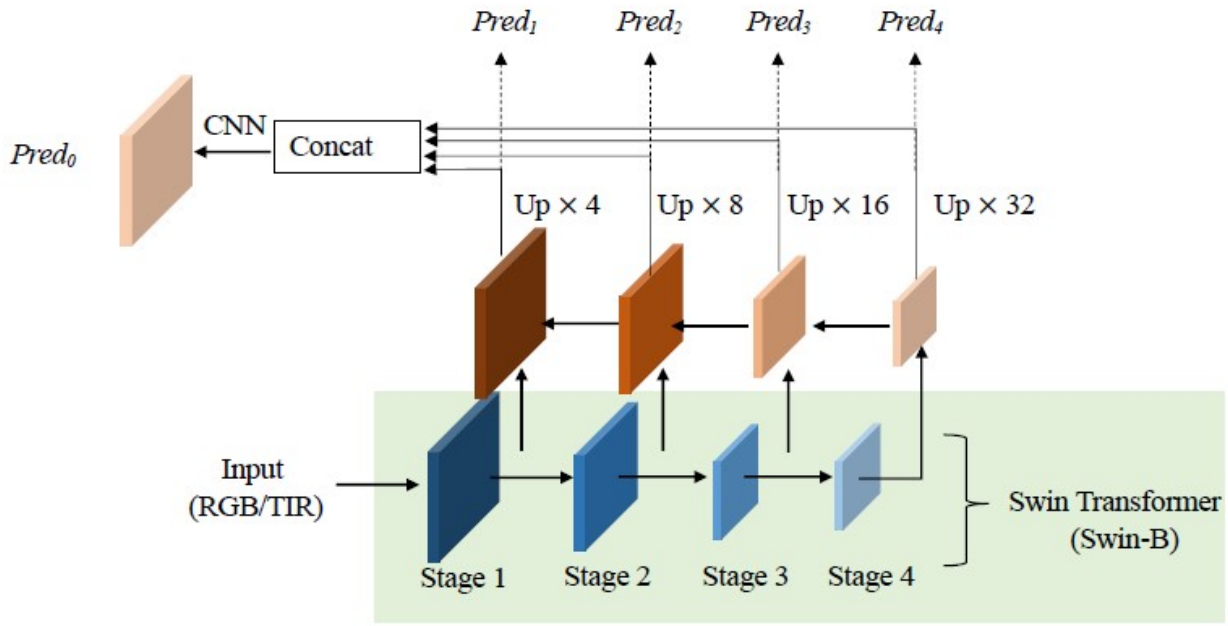


Figure 3. The pipeline of TransCrowd++.

formation in the fused feature maps to the RGB branch and thermal branch. Model architecture: The architectures of the RGB branch and thermal branch are kept the same with VGG19 [25]with batch normalization. These two branches are initialized by parameters pre-trained on ImageNet [6].

A.5. Edge Auxiliary Learning for Still Image Crowd Density Estimation (EAL)

Sifan Peng, Yinfeng Xia, Qianqian Yang, Qing He, Baoqun Yin
{sifan, julyxia, yanghcqq, heqing2020, bqyin}@mail.ustc.edu.cn

We propose a novel multi-task learning method named Edge Auxiliary Learning for Still Image Crowd Density Estimation (EAL). Aimed at the fact of background interference, we propose a novel crowd head edge map that contains all of the human head edges in an image. We present an algorithm that only detects crowd head edges. Then, the head edge map is treated as an auxiliary supervision signal to assist in distinguishing the crowd heads from the background.

We regard the sub-networks of the first nineteen layers of the VGG [25] divided by five pooling layers as the five Blocks in the network. These blocks extract a variety of features from the input RGB image and TIR image accordingly. We adopt the first three blocks to extract features from TIR and RGB images, and we fuse the two features by adding an operation. Then we use the pyramid fusion method to fuse the high-level features and low-level features output by each block. Next, the proposed crowd edge regression task supervises different blocks to learn crowd head edge features. Finally, we concatenate the crowd head edge features and the crowd density features to generate a high-quality crowd density map.

A.6. BL_Val +TIR

Jiong Li
421982539@qq.com

In crowd counting, due to occlusions, perspective effects, variations in object shapes, ”ground truth” density map through a Gaussian kernel is imperfect. We use BL algorithm [20] to solve this problem.

During training, we randomly select one of every 5 TIR images in the training set and add it to the val set. The backbone network is VGG19 [25].

A.7. Semi-balanced Sinkhorn with Scale Consistency (S3)

Hui Lin
linhuixjtu@gmail.com

Traditional crowd counting approaches usually use Gaussian assumption to generate pseudo density ground truth, which suffers from problems like an inaccurate estimation of the Gaussian kernel sizes. In this paper, we propose a new measure-based counting approach to regress

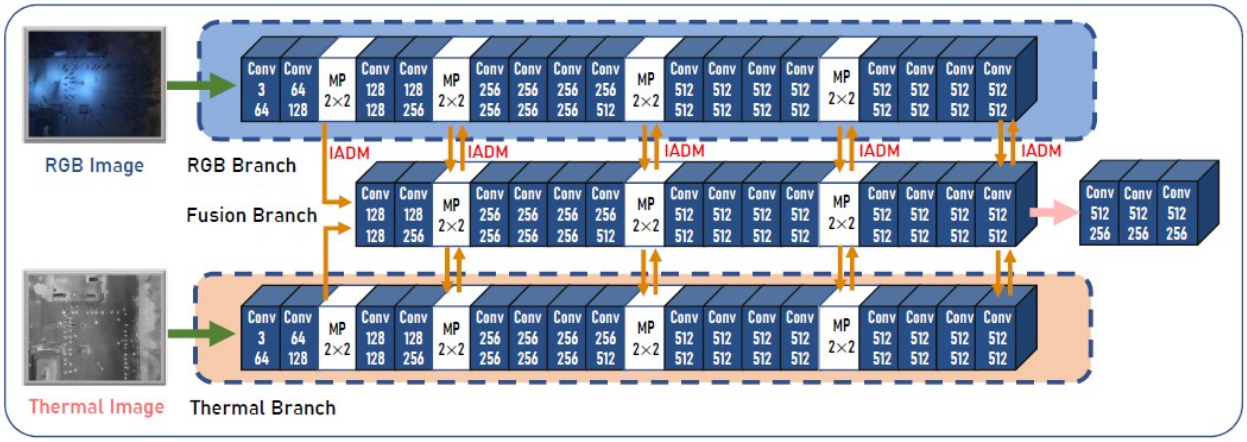


Figure 4. The model architecture of IADNet.

the predicted density maps to the scattered point-annotated ground truth directly. First, crowd counting is formulated as a measure matching problem. Second, we derive a semi-balanced form of Sinkhorn divergence, based on which a Sinkhorn counting loss is designed for measure matching. Third, we propose a self-supervised mechanism by devising a Sinkhorn scale consistency loss to resist scale changes. Finally, an efficient optimization method is provided to minimize the overall loss function. Extensive experiments on four challenging crowd counting datasets namely ShanghaiTech, UCF-QNRF, JHU++, and NWPU have validated the proposed method. The algorithm is detailed in [16].

A.8. MMCCN+Upconv Ver.HD

Binyu Zhang
zhangbinyu@bupt.edu.cn

Pre-process: We modified the training dataset to unify the feature of people. We reverse the false TIR images with a high value on the background and a low value on the people. And we only reserved the images taken with straight down angles. We still using the Gaussian kernel to represent people, but we applied a smaller sigma to separate crowding people with each other. To reduce the contrast in RGB images, we adopted RetinexNet [28] to enhance the low-light area in RGB images.

Data Augmentation: Random horizontally flip. Random rotate. Random brightness on RGB image. Random crop. Mosaic.

Model: MMCCN (the benchmark) with an upsample decoder. We change the backbone from ResNet50 to ResNet18 and using a larger feature map to capture the detailed information.

Loss Function: We adopt MSELoss to supervise the result on the pixel level. We proposed a Block MAELoss to supervise the result on local.

Post-process: We found that the people would have different scales with the drones on different altitudes. So we adopt multi-scale testing and merge the results from the different scales. Also, we adopt the model fusion to get a more stable result.

References

[1] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] Reza Bahmanyar, Elenora Vig, and Peter Reinartz. Mrcnet: Crowd counting and density map estimation in aerial and ground imagery. *arXiv preprint arXiv:1909.12743*, 2019.

[3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[5] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, volume 1, page 3, 2012.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Helia Farhood, Xiangjian He, Wenjing Jia, Michael Blumenstein, and Hanhui Li. Counting people based on linear, weighted, and local random forests. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2017.

[8] Min Fu, Pei Xu, Xudong Li, Qihe Liu, Mao Ye, and Ce Zhu. Fast crowd density estimation with convolutional neural net-

works. *Engineering Applications of Artificial Intelligence*, 43:81–88, 2015.

[9] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. Cnn-based density estimation and crowd counting: A survey. *arXiv preprint arXiv:2003.12783*, 2020.

[10] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely overlapping vehicle counting. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 423–431. Springer, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017.

[13] Jingxian Huang, Guanchen Ding, Yujia Guo, Daiqin Yang, Sihan Wang, Tao Wang, and Yunfei Zhang. Drone-based car counting via density map learning. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 239–242. IEEE, 2020.

[14] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[15] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

[16] Hui Lin, Xiaopeng Hong, Zhiheng Ma, Xing Wei, Yunfeng Qiu, Yaowei Wang, and Yihong Gong. Direct measure matching for crowd counting. *arXiv preprint arXiv:2107.01558*, 2021.

[17] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4823–4833, 2021.

[18] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting objects by blockwise classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3513–3527, 2019.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[20] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[21] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 235–252. Springer, 2020.

[22] Tao Peng, Qing Li, and Pengfei Zhu. Rgb-t crowd counting from drone: A benchmark and mmccn network. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[23] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[24] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. *arXiv preprint arXiv:2009.13077*, 2020.

[27] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020.

[28] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[29] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network. *arXiv preprint arXiv:1912.01811*, 2019.

[30] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] Huicheng Zheng, Zijian Lin, Jiepeng Cen, Zeyu Wu, and Yadan Zhao. Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):787–799, 2018.