RGB-T Crowd Counting from Drone: A Benchmark and MMCCN Network

Tao Peng#, Qing Li#, and Pengfei Zhu*

College of Intelligence and Computing, Tianjin University, Tianjin 300072, China
{wspt, liqing, zhupengfei}@tju.edu.cn

Abstract. Crowd counting aims to identify the number of objects and plays an important role in intelligent transportation, city management and security monitoring. The task of crowd counting is much challenging because of scale variations, illumination changes, occlusions and poor imaging conditions, especially in the nighttime and haze conditions. In this paper, we present a drone based RGB-Thermal crowd counting dataset (DroneRGBT) that consists of 3600 pairs of images and covers different attributes, including height, illumination and density. To exploit the complementary information in both visible and thermal infrared modalities, we propose a multi-modal crowd counting network (MMCCN) with a multi-scale feature learning module, a modal alignment module and an adaptive fusion module. Experiments on DroneRGBT demonstrate the effectiveness of the proposed approach. [1]

Keywords: Crowd counting, RGB-T

1 Introduction

Crowd analysis is of great importance because of a great practical demands such as assembly controlling and other security services. However, it is slow and unreliable to count people using any crowd monitoring system that relies on humans. There is a need for an automatic computer vision algorithm that can accurately count the number of people in crowded scenes based on images and videos of the crowds. Therefore, crowd counting has been widely studied and a growing number of network models have been developed to deliver promising solutions for this mission. These methods usually generate the density map according to the input image, and obtain the crowd counting by integrating the predicted density map.

Previous work [1–4] for scene analysis are mostly based on visible data. However, visible data may have drawbacks of illumination changes and poor imaging conditions in the nighttime. The thermal infrared data has been proven to be effective in boosting image analysis [5–8], and allows scene perception in day and night. However, the research of RGB-T crowd-counting is limited by the lack of

[1] # these authors contributed equally to this paper as co-first authors
* corresponding authors

a comprehensive image benchmark. Therefore, we construct a drone based RGB-Thermal crowd counting dataset, named as DroneRGBT, which consists of 3600 pairs of images and covers different attributes, including height, illumination and density. Compared with the existing crowd-counting datasets, the proposed DroneRGBT has the following main characteristics: 1) Different from most of the existing datasets, it is a drone-view datasets with multi-modalities. 2) Its alignment across modalities is highly accurate, and does not require any pre- or post-processing. 3) It is a large-scale dataset and collected in many different scenes, with 175,698 annotated instances.

With the created benchmark, we propose a novel approach for RGB-T crowd-counting. The main goals of our framework are: 1) The pipeline can predict density map according to a single modality only so that it can still work well when any modality data is missing. 2) Two modalities reuse the model as much as possible to reduce the amount of model parameters. 3) The fusion results are better than the results based on single modality. Hence, our pipeline, named Multi-Modal Crowd Counting Network (MMCCN), is based on ResNet-50 with three specific modules, i.e., multi-scale feature learning module, modal alignment module, adaptive fusion module. All the modules are optimized jointly and trained in an end-to-end manner. The pipeline can effectively extract low-level modality-specific feature and high-level modality-aligned semantic feature, and adaptively combine the prediction results to acquire a good fusion estimation. We design some experiments to demonstrate that our proposed pipeline can effectively utilize two modalities, RGB-Thermal, to estimate more accurate density map, resulting in more precise counting. Compared with other baselines in different aspects, we conclude that our model is more efficient than two-stream baseline and more precise than simple average baseline. Additionally, we also propose a special modal transfer method for our MMCCN framework to solve the problem of modal missing or the case of having single modality only. To be specific, we present a DM-CycleGAN, which can effectively generate thermal infrared data through visible data. During the training process of DM-CycleGAN, we introduce a density-map (DM) loss. It can make the generated image and the real image as similar as possible in the space which is used to count the instances. Extensive experiments prove that the transfer performance of DM-CycleGAN is better than that of original CycleGAN in field of crowd-counting. Fig. 1 demonstrates the flowchart of the proposed method.

This paper makes three major contributions for RGB-T crowd-counting.

* We create a new benchmark dataset containing 3600 registered RGB and thermal image pairs with ground truth annotations for evaluating RGB-T crowd counting methods.
* We propose a novel end-to-end pipeline, MMCCN, for RGB-T crowd-counting. Extensive experiments on our benchmark dataset demonstrate the effectiveness of the proposed approach and the importance of each component of the pipeline.
* We prove a useful way to use massive pairs of registered multi-modal images to train a modal transfer model. The proposed model, DM-CycleGAN, can

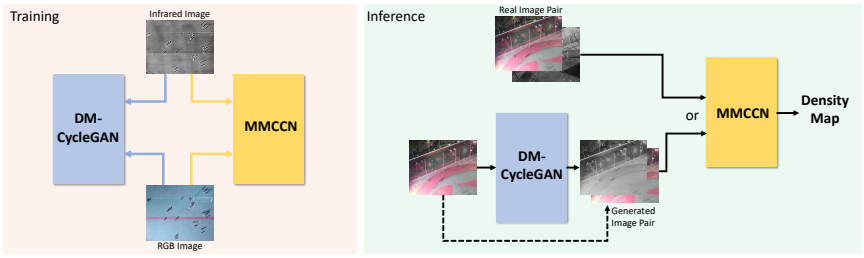

Fig. 1. MMCCN and DM-CycleGAN can be trained by using DroneRGBT benchmark. In the inference process, a pair of registered RGB-T data or data generated by DM-CycleGAN can be used as the input of MMCCN to estimate the density map.

effectively generate thermal infrared data though visible data and improve the counting performance in multi-modal tasks.

2 Related Work

2.1 Crowd Counting Datasets

According to image acquisition methods, the existing crowd counting datasets can be divided into three parts: surveillance-view datasets, free-view datasets and drone-view datasets. **Surveillance-view datasets** are collected by surveillance camera, which usually contain crowd images in specific indoor scenes or small-area outdoor locations. UCSD [9], Mall [10], WorldExpo'10 [11] and ShanghaiTech Part B [12] are typical surveillance-view datasets. **Free-view datasets** contain images collected from the Internet. The attributes of these type datasets vary significantly. There are also many free-view datasets for evaluation criteria, such as UCF_CC_50 [13], UCF-QNRF [14] and ShanghaiTech Part A [12]. Our dataset is a **drone-view based dataset** which is collected by UAV.

2.2 Crowd Counting Methods

In recent years, there are more and more researches on crowd counting and more recent methods used CNNs to tackle crowd counting [12, 15, 1–3]. Zhang et al. [12] propose a classical and lightweight counting model called Multi-Column Convolutional Neural Network (MCNN), which can estimate density map by learning the features for different head sizes by each column CNN. A spatial FCN (SFCN) [3] is designed by Wang et al. to produce the density map. After the spatial encoder, a regression layer is added in SFCN. In this work, we propose a baseline network which is used to predict crowd number on RGB-T datasets.

2.3 Multi-Modal Learning

Multi-Modal learning has drawn more attentions in the computer vision community. In this paper, we focus on integrating RGB and thermal infrared data

Table 1. Comparison of the DroneRGBT dataset with existing datasets.

Dataset	Resolution	Frames	Thermal	View	Max	Min	Ave	Total
UCSD [9]	158×238	2000	-	surveillance	46	11	24.9	49,885
MALL [10]	640×480	2000	-	surveillance	53	13	31.2	62,316
UCF_CC_50 [13]	-	50	-	free	4543	94	1279	63,974
WorldExpo [11]	576×720	3980	-	surveillance	253	1	50.2	199,923
SHT A [12]	-	482	-	free	3139	33	501	241,677
SHT B [12]	768×1024	716	-	surveillance	578	9	123	88,488
UCF-QNRF [14]	-	1535	-	free	12,865	49	815	1,251,642
DroneRGBT	512×640	3600	√	drone	403	1	48.8	175,698

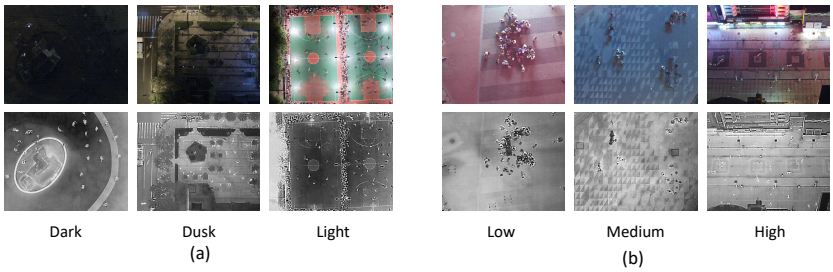

Fig. 2. Some example image pairs in the DroneRGBT dataset.

[16–19]. The typical problems that use these two modalities are as follows. (1) **RGB-T Saliency Detection.** Li et al.[16] propose a novel approach, multitask manifold ranking with cross-modality consistency, for RGB-T saliency detection. (2) **RGB-T tracking.** Li et al. [18] provided a graph-based cross-modal ranking model for RGB-T tracking, in which the soft cross-modality consistency between modalities and the optimal query learning are introduced to improve the robustness. Different from these typical works, our work focus on crowd-counting and it is the first benchmark and baseline for RGB-T crowd-counting.

3 DroneRGBT BENCHMARK

3.1 Data Collection and Annotation

Our DroneRGBT dataset is captured by drone-mounted cameras (DJI Phantom 4, Phantom 4 Pro and Mavic), covering a wide range of scenarios, e.g., campus, street, park, parking lot, playground and plaza. After cleaning the unavailable data, we use the Homography method to register RGB images with infrared image. We label the number of people based on the head count of infrared images. The ground truth annotation file is saved as xml format.

And then, we divide the training set and the test set according to the illumination. We first divide the dataset into three categories: dark, dusk, and light, and then divide each category into two parts, one part for training and

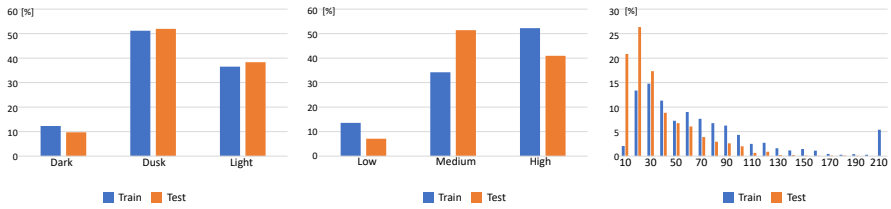

Fig. 3. The distribution of illumination, height, and density attributes in the training set and the testing set from left to right. Bars represent the percentage of this property in the training and testing sets.

the other for testing, while ensuring that the training set and testing set have different scenes to reduce the chances of overfitting to particular scenes.

3.2 Data Characteristic

The DroneRGBT dataset is the first drone-view crowd counting dataset with both RGB and thermal infrared data and it contains images pairs taken at different locations with large variations in scale, viewpoint and background clutters. Tab. 1 compares the basic information of DroneRGBT and existing datasets. In addition to the above properties, DroneRGBT is more diverse than other datasets. Specifically, the following main aspects are considered in creating the DroneRGBT dataset.

* *Illumination*. The image pairs are captured under different light conditions, such as *dark*, *dusk* and *light*. Under different conditions, the difference in illumination is obvious, which can be distinguished by experience..
* *Scale*. Like most of surveillance-view and free-view based benchmarks which usually include instances with different scales, different object scales are also taken into account for our dataset. So our dataset are collected in different altitudes which significantly affects the scales of object. We delineate 30-50 meter as low altitude, 50-80 meter as medium, and 80-100 meter as high.
* *Density*. Density means the number of objects in each image pair. In our dataset, the density varies from 1 to 403. The distribution of our dataset based on these attributes is shown in Fig 3.

Some typical sample image pairs in different attributes from our DroneRGBT dataset are shown in Fig. 2. It shows the diversity of our datasets.

3.3 Evaluation Metrics

Following previous works for crowd counting, we use the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to evaluate the performance of

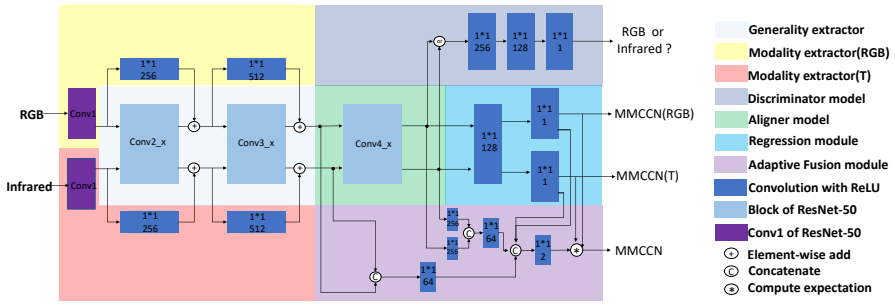

Fig. 4. The architecture of our multi-modal crowd counting network. The pipeline is color-filled to represent the different modules as shown in the color legend on the right side of the figure. Numbers in blue rectangles stand for the kernel size of convolution and numbers of kernel. The block is corresponding to the Tab. 1 in the paper [20].

our proposed method. The MAE and RMSE can be computed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|C_i - \hat{C}_i\right|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(C_i - \hat{C}_i\right)^2} \tag{1}$$

where n is the number of images, C_i is the counting label of people and $\hat{C}_i$ is the estimated value for the i-th test image.

4 Proposed Approach

4.1 Multi-Modal Crowd Counting Network

The pipeline of Multi-Modal Crowd Counting Network (MMCCN) is shown in Fig. 4. Our network is based on ResNet-50 [20] with three specific modules, i.e., multi-scale feature learning module, modal alignment module, adaptive fusion module. The feature learning module is used to extract both generality feature and modality-special feature of the input data. The pair of extracted features are separately fed into the modal alignment module to further extract the high-level semantic feature and each pair of semantic feature are aligned to same feature space at the same time. After using the high-level semantic feature to regress the crowd number, the pipeline fuse the prediction output based on visible spectrum and thermal infrared data by adaptive fusion module to obtain the final result.

Multi-Scale Feature Learning Module Due to different principles of imaging physics, the distributions between visible spectrum and thermal infrared data are different. One intuitive thought will be to extract their discriminative features respectively. However, it will increase the parameters of network and might degrade efficiency. Besides, two-stream also ignore modality-shared feature learning. To reduce computational burden, we use generality extractor to obtain the

common information and modality extractor to extract modality-special feature. The generality extractor is the first two blocks of ResNet-50. Modality extractor, where consists of a convolution layer, is used to extract modality-special feature representations with a little computational burden.

In multi-scale feature learning module, each block has a modality extractor. The outputs of each generality extractor and modality extractor is element-wise added together. For example, the feature of thermal infrared data f^T is computed by Equ. 2.

$$f^T = F_g(T; \sigma_g) + F_m^T(T; \sigma_m^T) \tag{2}$$

where F_g and F_m^T stand for generality extractor and modality extractor of thermal modality, respectively. And σ_g and σ_m^T are the parameters of the corresponding extractors. So, the feature of thermal infrared data depends on both σ_g and σ_m^T. Although the modality extractor only has less parameters, the module also can effectively extract discriminative feature of visible spectrum and thermal infrared data respectively. The reason is as follows. Firstly, Equ. 2 can be simplified when we denote the transfer function as a convolution operation. And then we can merge the matrix as follows.

$$f^T = W_g * T + W_m^T * T = (W_g + W_m^T) * T = M^T * T \tag{3}$$

where W_g represents the parameter of generality extractor and W_m^T stands for that of modality extractor. Convolution operation is denoted as $*$. As a result, we can find a new weight matrix M^T which can focus on modality-specific feature.

Modal Alignment Module Our hypothesis is that the distribution changes between bi-modality are low-level characteristics rather than high-level. The high-level semantic information between visible spectrum and thermal infrared data is similar, because these pairs of images are shot in the same place with registration. Therefore, we try to reuse the latter network and attend to map the RGB input to features which are aligned with thermal feature space. In the spirit of adversarial training in GAN [21, 22], the modal alignment module is trained by a minimax game. It consists of a aligner model and a discriminator model. The aligner learn to align the feature maps between RGB and thermal infrared data, and the discriminator differentiate the feature distributions. These two models are alternatively optimized and compete with each other. Specifically, the backbone of aligner, which aims to align feature, is the $Conv4\_x$ of ResNet-50 with 1 stride for all convolution. In order to avoid gradient unstable, we explore the least squares loss rather than the sigmoid cross entropy loss function to optimize the our model, where the least squares loss function can relieves the problem of vanishing gradients. Therefore, the loss for learning the aligner is:

$$\min_A J(A) = \min_A \frac{1}{2} E_{f^V \sim P_{f^V}} [D(A(f^V)) - c)]^2 \tag{4}$$

where A stands for the aligner and D means the discriminator. The low-level modality-special feature of visible spectrum data extracted by multi-scale feature

learning module is denoted as f^V. The discriminator would differentiate the complicated feature space, which has 3 convolution layers with 1×1 kernel size. And it is optimized via:

$$\min_D J(D) = \min_D \frac{1}{2} E_{f^T \sim P_{f^T}}[D(A(f^T)) - a]^2 + \frac{1}{2} E_{f^V \sim P_{f^V}}[D(A(f^V)) - b]^2 \quad (5)$$

The definitions of symbols are same as the earlier ones. To make A align the modal features as close as possible, we set $c = b$. So, by using the 0-1 binary coding scheme, the parameters is set by $a = c = 1$ and $b = 0$ in this model. By alternative updating of D and A, decision boundary of the least squares loss function can force the aligner to generate feature of both modality toward decision boundary. Note that the discriminator model is only used in training processing to provide the supervised signal for the align model. During the inference process, only the align model is used to obtain aligned high-level features.

Adaptive Fusion Module By using a regression module, we can obtain the density map predicted through visible spectrum and thermal infrared input, respectively. M^T denote the density map predicted by our pipeline when the input is thermal infrared data. M^V is density map when the input is visible spectrum data. In ensemble learning, the result of multiple-model fusion usually can achieve better than the direct result of single model. Therefore, the prediction result of our model is a expectation, which combines M^T and M^V.

$$E(M) = M^T \times p(M^T) + M^V \times p(M^V) \quad (6)$$

where $p(M^T)$ is the probability of M^T. Similarly, the probability of M^V is denoted as $p(M^V)$. The probability here means the confidence of corresponding output. Given that the confidence depends on the pair of input, we use a additional network to regress the probability based on multi-scale feature. The details of this module are shown in Fig. 4. Therefore, $p(M) = \widetilde{p}(M) \times \hat{p}(M)$, where $\widetilde{p}$ is the prior confidence and $\hat{p}$ is the confidence predicted by network. Using Equ 6, each modality density map multiplies corresponding confidence map to get the final result. All the parameters of the pipeline are learned by minimizing the loss function $J(M)$.

$$\begin{aligned} J(M) = \frac{1}{MN} \sum_{j=1}^{M} \sum_{k=1}^{N} [(E(M)_{i,j} - M_{i,j}^{GT})^2 \\ + \lambda_T (M_{i,j}^T - M_{i,j}^{GT})^2 + \lambda_V (M_{i,j}^V - M_{i,j}^{GT})^2] \end{aligned} \quad (7)$$

where M^{GT} is the ground-truth of the density map. λ_T and λ_V are weighting factors of loss.

Loss Function Our final objective for network becomes

$$J = J(M) + \lambda_a (J(D) + J(A)) \quad (8)$$

where λ_a is the weight for the align loss. $J(D)$ is just used to optimize the discriminator model.

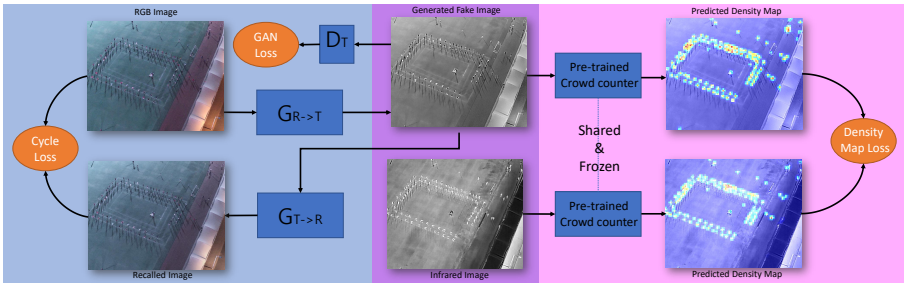

Fig. 5. The framework of the DM-CycleGAN. The pipeline in blue color area is the original CycleGAN, and the extra branch of our DM-CycleGAN are shown in red color area. Pre-trained model is shared and frozen.

4.2 DM-CycleGAN

In some cases, due to lack of the infrared acquisition equipment, the thermal infrared data is not acquired. So, we consider whether we can utilize the visible data to generate infrared data and use generated infrared data and visible data as the input of our MMCCN network for crowd-counting. The original idea was to use CycleGAN [23] to handle this modality transfer problem, which translates visible data into infrared data. However, given that the loss function of CycleGAN do not constrain local details, original CycleGAN can not focus on local patterns and texture features. So, we propose a modality translator, called DM-CycleGAN, to generate meaningful infrared image.

Framework For making generated image meaningful, we assume that the distance between the generated image and the real image in the space which is used to count the instances needs to as close as possible. In this pipeline, a pre-trained crowd counter is viewed as the spatial mapping converter, which can transfer the image from image space into density-map space. Therefore, the fake image generated by the original CycleGAN and real infrared image are transferred into density-map space by this converter, respectively. And a Density Map Mean Square Loss is introduced to force generated image to become close to real image in density-map space. This extra loss can force GAN to focus on person in the image. The specific description of the DM-CycleGAN framework is shown in Fig. 5.

Loss Function Firstly, the symbol definitions are the same as CycleGAN. Generator and discriminator are defined as G and D, respectively. $\mathcal{R}$ and $\mathcal{T}$ stand for visible data and thermal infrared data. Therefore, the Density Map Mean Square(DM) Loss is defined as:

$$\mathcal{L}_{DM}\left(G_{\mathcal{R}\to\mathcal{T}}, G_{\mathcal{T}\to\mathcal{R}}, \mathcal{R}, \mathcal{T}\right) = \mathbb{E}_{(i_{\mathcal{T}}, i_{\mathcal{R}})\sim I_{\mathcal{T},\mathcal{R}}}[MSE(\mathcal{C}(G_{\mathcal{R}\to\mathcal{T}}(i_{\mathcal{R}})), \mathcal{C}(i_{\mathcal{T}})) + MSE(\mathcal{C}(G_{\mathcal{R}\to\mathcal{T}}(G_{\mathcal{T}\to\mathcal{R}}(i_{\mathcal{T}})), \mathcal{C}(i_{\mathcal{T}})))] \tag{9}$$

where $\mathcal{C}$ stands for the space mapping by pre-trained crowd-counter, MSE represents the mean square error between the generated image and real image in density-map space. Finally, the final objective of DM-CycleGAN is defined as:

$$\begin{aligned}&\mathcal{L}_{\text{final}}\left(G_{\mathcal{R}\to\mathcal{T}}, G_{\mathcal{T}\to\mathcal{R}}, D_{\mathcal{R}}, D_{\mathcal{T}}, \mathcal{R}, \mathcal{T}\right)\\&\quad=\mathcal{L}_{GAN}\left(G_{\mathcal{R}\to\mathcal{T}}, D_{\mathcal{T}}, \mathcal{R}, \mathcal{T}\right)\\&\quad+\mathcal{L}_{GAN}\left(G_{\mathcal{T}\to\mathcal{R}}, D_{\mathcal{R}}, \mathcal{T}, \mathcal{R}\right)\\&\quad+\lambda\mathcal{L}_{\text{cycle}}\left(G_{\mathcal{R}\to\mathcal{T}}, G_{\mathcal{T}\to\mathcal{R}}, \mathcal{R}, \mathcal{T}\right)\\&\quad+\mu\mathcal{L}_{DM}\left(G_{\mathcal{R}\to\mathcal{T}}, G_{\mathcal{T}\to\mathcal{R}}, \mathcal{R}, \mathcal{T}\right)\end{aligned} \tag{10}$$

where the definition of $\mathcal{L}_{GAN}$ and $\mathcal{L}_{\text{cycle}}$ are the same as original CycleGAN,

$$\begin{aligned}&\mathcal{L}_{GAN}\left(G_{\mathcal{R}\to\mathcal{T}}, D_{\mathcal{T}}, \mathcal{R}, \mathcal{T}\right)\\&\quad=\mathbb{E}_{i_{\mathcal{T}}\sim I_{\mathcal{T}}}\left[\log\left(D_{\mathcal{T}}\left(i_{\mathcal{T}}\right)\right]+\mathbb{E}_{i_{\mathcal{R}}\sim I_{\mathcal{R}}}\left[\log\left(1-D_{\mathcal{T}}\left(G_{\mathcal{R}\to\mathcal{T}}\left(i_{\mathcal{R}}\right)\right)\right)\right]\end{aligned} \tag{11}$$

$$\begin{aligned}&\mathcal{L}_{\text{cycle}}\left(G_{\mathcal{R}\to\mathcal{T}}, G_{\mathcal{T}\to\mathcal{R}}, \mathcal{R}, \mathcal{T}\right)\\&\quad=\mathbb{E}_{i_{\mathcal{R}}\sim I_{\mathcal{R}}}\left[\left\|G_{\mathcal{T}\to\mathcal{R}}\left(G_{\mathcal{R}\to T}\left(i_{\mathcal{R}}\right)\right)-i_{\mathcal{R}}\right\|_1\right]+\mathbb{E}_{i_{\mathcal{T}}\sim I_{\mathcal{T}}}\left[\left\|G_{\mathcal{R}\to\mathcal{T}}\left(G_{\mathcal{T}\to\mathcal{R}}\left(i_{\mathcal{T}}\right)\right)-i_{\mathcal{T}}\right\|_1\right]\end{aligned} \tag{12}$$

And the λ and μ are the weights of cycle-consistent loss and density-map loss, respectively.

5 Experiments

5.1 Experiments on DroneRGBT Dataset

Training Details The training dataset of DroneRGBT consists of 1800 pairs of registered images and corresponding ground truth annotation files. The annotation is converted into a binary map with a Gaussian filter of standard deviation 5. In addition, data augmentations like rotation, random crop are used to avoid overfitting. The optimizer we use is Adam [24] with the following hyper parameters: learning rate 10^{-5} with stepped decay rate 0.995, $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size=8. And λ_T, λ_V, λ_a are set by 1, 1 and 0.005. We alternatively optimized the aligner model and discriminator model of modal alignment module with the adversarial loss for aligning two domain. In adversarial learning, we utilized the Adam optimizer with a learning rate of 10^{-5} and a stepped decay rate of 0.98 every 100 joint updates, with weight clipping for the discriminator being 0.03. Prior confidence $\widetilde{p}(M^T)$ and $\widetilde{p}(M^V)$ are set as 1 and 0, respectively.

The backbone of our network is ResNet-50. What we should pay special attention to is that there are several Batch-Normalization layers in the ResNet. However, the calculation procedure of Batch-Normalization layer is different between training process and inference process [25]. In the training process, 'mean' and 'variance' are computed by the samples in the mini-batch. 'mean' and 'variance' used in the inference process are the 'moving average mean' and 'moving average variance' counted by training data. In this pipeline, two domains (visible data and thermal infrared data) are share the backbone network. So, 'moving average mean' and 'moving average variance' may be computed by both visible

Table 2. The performance of state-of-art methods with different modalities on DroneRGBT. MAE and RMSE are shown.

Method	Journal/Venue & Year	Thermal		RGB	
		MAE	RMSE	MAE	RMSE
MCNN [12]	CVPR 2016	13.64	19.77	31.13	40.87
CMTL [26]	AVSS 2017	19.35	27.05	19.14	28.46
MSCNN [27]	ICIP 2017	14.89	20.41	23.38	28.40
ACSCP [28]	CVPR 2018	13.06	20.29	18.87	28.31
SANET [29]	ECCV 2018	12.13	17.52	14.91	21.66
StackPooling [30]	CoRR 2018	9.45	14.63	14.72	20.90
DA-NET [31]	Access 2018	9.41	14.10	13.92	20.31
CSRNet [15]	CVPR 2018	8.91	13.80	13.06	19.06
SCAR [32]	NeuCom 2019	8.21	13.12	11.72	18.60
CANNET [33]	CVPR 2019	7.78	12.31	10.87	17.58
BL [34]	CVPR 2019	7.41	11.56	10.90	16.80

data and thermal infrared data so that the distribution between train data and test data becomes more and more different, and as a result it leads to a bad inference result. In our experiment, when the Batch-Normalization layers is frozen during training, the network is easier to overfitting. So, we use different routes to calculate statistics of dataset, respectively.

Compared with Baseline Firsly, we try to prove that multi-modal fusion can achieve better results than single mode. Therefore, we test the performance of state-of-art models on the single modality of our benchmark dataset. We select several advanced RGB crowd counter for evaluations, including MCNN [12], CMTL [26], MSCNN [27], ACSCP [28], SANET [29], StackPooling [30], DA-NET [31], CSRNet [15], SCAR [32], CANNET [33], BL [34]. Specifically, these models are trained by a single modal data and tested in corresponding modal test dataset, respectively. The experiment results is shown in Tab. 2.

In addition, in order to evaluate that our proposed model can more effectively and efficiently integrate the two modal features, we compare our method with other baseline models. Baseline #1 This pipeline is a two-stream network which each stream is the first three blocks of ResNet-50 and is used to extract modal-specific feature. And then the output of each stream are concatenated in channel dimension. After reducing the dimension by a convolution layer with 1*1 kernel size, the high-level fusion feature is learned by the backend network. The regression module of this pipeline is same as our proposed pipeline. Baseline #2 Apart from the first convolution layer which is used to extract modal-specific feature, the whole pipeline is share the same weight like siamese network [35]. The final prediction result is the average of the prediction by each branch. Comparison results are shown in Tab. 3.

It can be seen that our MMCCN performs favorably against the state-of-the-art methods. Our MMCCN obtains 7.27 MAE score and 11.45 RMSE score, but the most competitor BL [34] gets 7.41 MAE score and 11.56 RMSE score. The result shows that modality fusion can further improve the counting result. Compared with other two baselines, although the accuracy of our model is slightly

Table 3. Comparison of our approach with other proposed baseline on DroneRGBT dataset.

Method	Precision		Model Size(M)	Speed (fps)	GFLOPs
	MAE	RMSE			
Baseline #1	7.18	11.43	20.72	22	16.54
Baseline #2	11.07	17.15	9.39	24	15.00
MMCCN	7.27	11.45	10.47	17	16.55

Table 4. The performance of three heads on several break-down subsets.

Method	Overall		dark		dusky		light	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MMCCN(RGB)	11.53	16.75	19.76	24.52	11.14	16.07	10.03	15.23
MMCCN(T)	7.49	11.91	**11.35**	**15.95**	7.09	11.19	7.16	11.69
MMCCN	**7.27**	**11.45**	12.02	15.96	**6.92**	**11.00**	**6.83**	**10.69**

lower than that of the baseline #1, our model has less parameters, and each modality is decoupled in MMCCN so that it can work well even if the input is only a single modality. Because there is no module coupling between MMCCN(T) and MMCCN(RGB), they can work independently and still achieve competitive prediction results. Further more, the performance of three heads on several break-down subsets is shown in Tab. 4.

Ablation Study To analyse the importance of each component of our proposed MMCCN model, we additionally construct some variants and evaluate them on the Drone-RGBT dataset. MMCCN(sBN) means that the Batch-Normalization layer of model only has one route as the original ResNet. MMCCN(fBN) stands for the Batch-Normalization of the pipeline is frozen when training the network. MMCCN(w/o mam) indicates the model that removes discriminator model of Modal Alignment Module. And the network is only trained by loss $J(M)$. MMCCN(w/o me) denotes the method that further remove two modality extractors of Multi-scale Feature Learning Module.

All variants are trained on the training set and tested on the testing set. The training steps and other parameters are identical, and meanwhile the evaluation protocol in different experiments are same too. From Tab. 5, it shows that our MMCCN achieves better results than its variants.

Qualitative Results From the quantitative results, we find that fusion based on confidence can improve the prediction results. To test the difference between the density map predicted by MMCCN and MMCCN(T), we visualize the density map respectively shown in Fig. 6. In addition, we also visualize the confidence map predicted by our Adaptive Fusion Module. From the qualitative results, we find that due to the noise of infrared data, result of MMCCN(T) may contain some false positives. By utilizing both RGB and Infrared feature,

Table 5. Comparison of our approach with its variants to prove the importance of each component.

Method	Overall	
	MAE	RMSE
MMCCN(sBN)	17.56 (↓ 10.29)	22.34 (↓ 10.89)
MMCCN(fBN)	11.46 (↓ 4.19)	16.41 (↓ 4.96)
MMCCN(w/o me)	7.34 (↓ 0.07)	11.74 (↓ 0.29)
MMCCN(w/o mam)	7.28 (↓ 0.01)	11.76 (↓ 0.31)
MMCCN	7.27	11.45

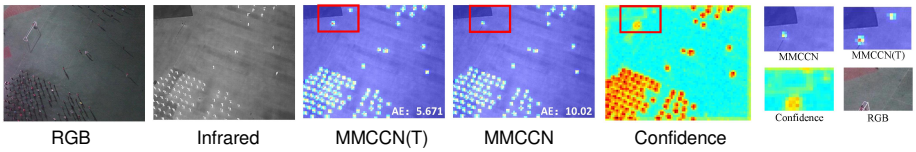

Fig. 6. Qualitative results of MMCCN.

model will reduce the confidence of false positive. As the result, the prediction results improved.

5.2 Experiments on Single Modality

Training Details DM-CycleGAN is trained on the training set of DroneRGBT. During the training phase, the λ and μ are set as 10 and 100, respectively. We use the Adam solver with a batch size of 1. All networks were trained from scratch with a learning rate of 0.0002. Data augmentations like rotation, random crop are used to avoid overfitting.

Results on MMCCN without re-trained We generate fake infrared images by using the visible data in our testing set of DroneRGBT. To test the meaningful of the generated image, we take the generated images as the input of the MMCCN to test whether it can improve the prediction results compared with the single modality. The MMCCN(RGB) method corresponds to the MMCCN method only accepting RGB image. The DM-CycleGAN* and CycleGAN* method receive both RGB image and infrared image. But the infrared image in DM-CycleGAN* is generated from DM-CycleGAN, and CycleGAN* method generates infrared image through CycleGAN. Results in Tab. 6 shows that DM-CycleGAN* performs inferior than MMCCN(RGB) (*i.e.*, 10.92 MAE score vs. 11.53 MAE score). It proves that multi-modality can improve the results though the information content of fake infrared image is based on visible image. At the same time, the Structural Similarity Index (SSIM) between the generated image and real image are also shown in this table. This experiment proves DM-CycleGAN and MMCCN can be used together to boost the result when we only have one modal data.

Table 6. The performance of generated image on pre-trained MMCCN.

Method	MMCCN		SSIM
	MAE	RMSE	
MMCCN(RGB)	11.53	16.75	-
CycleGAN*	13.45	18.99	0.44
DM-CycleGAN*	10.92	16.19	0.49

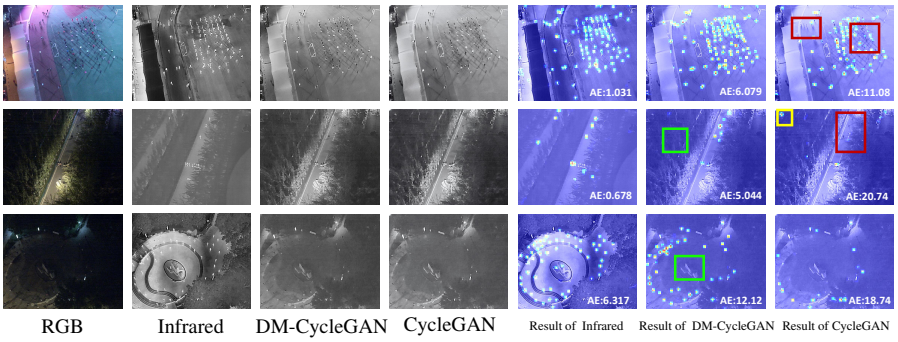

Fig. 7. Qualitative results of CycleGAN and DM-CycleGAN.

Qualitative Results Some generated images and corresponding real thermal infrared images are shown in Fig. 7. Besides, we visualize the density map predicted by thermal infrared head of our model–MMCCN(T). From the results, we find that the infrared images generated by CycleGAN will miss some local information so that the MMCCN can not detect the person (red rectangle). At the same time, without the constraint of DM loss, false positive will appear (yellow rectangle). Our DM-CycleGAN can focus on the local details of people so that it will be predicted by MMCCN. However, it will also miss some information compared with real data due to poor visibility of visible data (green rectangle).

6 Conclusions

In this paper, we presented a benchmark for RGB-T crowd counting. This is a drone-view dataset with different attributes. With the benchmark, we proposed a Multi-Modal Crowd Counting Network for RGBT crowd-counting. DM-CycleGAN is proposed for generating the infrared data for MMCCN when we only have visible data. Through analyzing the quantitative and qualitative results, we demonstrated the effectiveness of the proposed approach.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 61876127 and Grant 61732011, Natural Science Foundation of Tianjin under Grant 17JCZDJC30800 and The Applied Basic Research Program of Qinghai under Grants 2019-ZJ-7017.

References

1. Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Where are the blobs: Counting by localization with point supervision. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 547–562
2. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 4031–4039
3. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2019) 8198–8207
4. Ranjan, V., Le, H., Hoai, M.: Iterative crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 270–285
5. Li, C., Wu, X., Zhao, N., Cao, X., Tang, J.: Fusing two-stream convolutional neural networks for rgb-t object tracking. Neurocomputing **281** (2018) 78–85
6. López-Fernández, L., Lagüela, S., Fernández, J., González-Aguilera, D.: Automatic evaluation of photovoltaic power stations from high-density rgb-t 3d point clouds. Remote Sensing **9** (2017) 631
7. Zhai, S., Shao, P., Liang, X., Wang, X.: Fast rgb-t tracking via cross-modal correlation filters. Neurocomputing **334** (2019) 172–181
8. Zhang, X., Ye, P., Peng, S., Liu, J., Xiao, G.: Dsiammft: An rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion. Signal Processing: Image Communication (2020) 115756
9. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2008) 1–7
10. Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localised crowd counting. In: BMVC. Volume 1. (2012) 3
11. Zhang, C., Kang, K., Li, H., Wang, X., Xie, R., Yang, X.: Data-driven crowd understanding: A baseline for a large-scale crowd dataset. IEEE Transactions on Multimedia **18** (2016) 1048–1061
12. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 589–597
13. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2013) 2547–2554
14. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 532–546
15. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 1091–1100
16. Li, C., Wang, G., Ma, Y., Zheng, A., Luo, B., Tang, J.: A unified rgb-t saliency detection benchmark: dataset, baselines, analysis and a novel approach. arXiv preprint arXiv:1701.02829 (2017)
17. Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J.: Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia **22** (2019) 160–173

18. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: benchmark and baseline. Pattern Recognition **96** (2019) 106977
19. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25** (2016) 5743–5756
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
21. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
22. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2794–2802
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
26. Sindagi, V.A., Patel, V.M.: Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE (2017) 1–6
27. Zeng, L., Xu, X., Cai, B., Qiu, S., Zhang, T.: Multi-scale convolutional neural networks for crowd counting. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017) 465–469
28. Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., Yang, X.: Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 5245–5254
29. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 734–750
30. Huang, S., Li, X., Cheng, Z.Q., Zhang, Z., Hauptmann, A.: Stacked pooling: Improving crowd counting by boosting scale invariance. arXiv preprint arXiv:1808.07456 (2018)
31. Zou, Z., Su, X., Qu, X., Zhou, P.: Da-net: Learning the fine-grained density distribution with deformation aggregation network. IEEE Access **6** (2018) 60745–60756
32. Gao, J., Wang, Q., Yuan, Y.: Scar: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing **363** (2019) 1–8
33. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5099–5108
34. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 6142–6151
35. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Volume 1., IEEE (2005) 539–546