Data Fusion Pipeline for UAV-Based Real-Time Night Crowd Counting for Public Safety

Kiat Nern Yeo
Q Team Centre of Expertise
Home Team Science and Technology Agency
Singapore
yeo_kiat_nern@htx.gov.sg

Yan Ling Lau
Q Team Centre of Expertise
Home Team Science and Technology Agency
Singapore
lau_yan_ling@htx.gov.sg

Gee Wah Ng
Q Team Centre of Expertise
Home Team Science and Technology Agency
Singapore
ng_gee_wah@htx.gov.sg

***Abstract*— Performing crowd management in large scale outdoor events at night is a challenging yet essential task for public safety and security purposes. Traditional methods of carrying out crowd counting require the deployment of massive manpower and are unable to provide a reliable count for effective resource and manpower planning. In recent years, deep learning based crowd counting methods trained on static images were introduced. However, there still exist real world challenges of variations in crowd density across the scene, illumination, environmental conditions, and perspective problems which these methods are unable to fully address. This paper attempts to address the problems of varying crowd densities and illumination through a crowd counting pipeline that fuses illumination enhancement processes with crowd density estimation and crowd localisation techniques to achieve improved accuracy for crowd counts from live UAV video feeds. This data fusion pipeline approach has been demonstrated to provide improved count accuracy on both dataset and real-world images compared against standalone state-of-the-art methods.**

Keywords — crowd counting, UAV, low-illumination, public safety, data fusion

I. Introduction

In space-constrained Singapore, with her bustling nightlife, public events where massive crowds gather presents a safety and security challenge for law enforcement officers. These can range from events such as the F1 race, our National Day, busy festive streets to concerts and rave parties. Law enforcement officers from our Home Team are often tasked [1] to perform crowd monitoring and control activities during such events in their utmost effort to prevent crowd related disasters such as stampedes [2] and crowd crush [3] incidents. Traditionally, our officers had manually kept track of crowd sizes through methods such as handheld clickers or simply giving a ballpark estimate from field experience. Other manual methods include aggregating statistics from ticket sales and entrance scanners. These manual methods were inadequate in reflecting the true turnout of the events and are also unable to represent the crowd density and distribution in various subparts of public events.

To support our Home Team officers in crowd size monitoring and management, the following were considered during the design of our pipeline:

1) Scene coverage: Instead of using closed-circuit television (CCTV) setups, which are unable to provide a full coverage of the event grounds, UAVs with gimballed cameras are deployed, allowing complete coverage from a single vantage point.

2) Lighting conditions: Large-scale events at night are often accompanied with inconsistent, dazzling or strobe lighting. This causes illumination within an image to be infeasible even for hand-counting. As such, a robust and adaptable illumination technique is required to preprocess the input video frame before it is fed into the crowd counting models.

3) Crowd density variation: Within each scene, there may be dense and sparse crowds which our pipeline addresses by a fusion of the outputs of the crowd density estimation and crowd localisation methods.

To best account for the effect of these factors, our work employs scene enhancement techniques, prior to the fusion of 2 existing state-of-the-art crowd counting models operating on different techniques, one based on crowd density estimation, and the other on crowd localisation. This effort aims to create an adaptive crowd counting method to support quick on-the-field decision-making to further enhance public safety for large scale events.

The key contributions of this paper are:

1) To demonstrate the application of existing state-of-the-art crowd counting techniques to UAV views with our pipeline that is adaptable to any UAV capable of streaming video without affecting existing operations setups.

2) To demonstrate the value of pre-processing low-illumination images to achieve improved crowd counting accuracy using existing pre-trained state-of-the-art (SOTA) methods, which would otherwise be negatively impacted by the lack of image details in low illumination scenes or would require transfer learning.

To verify the effectiveness of our pipeline, we tested on images that represent the real-world use case of crowd counting at night from UAV perspective. From the existing open-source crowd datasets, relevant images based on the illumination information and camera parameters like altitude, zoom factor and pitch angle, are selected to be tested on.

The rest of the paper is organised as follows. Section II explores the works related to crowd counting, Section III describes our methods and materials, which includes our data fusion pipeline for UAV-based real time crowd counting operations, then into the details of the hardware and software setup with human-in-the-loop decision making. Section III

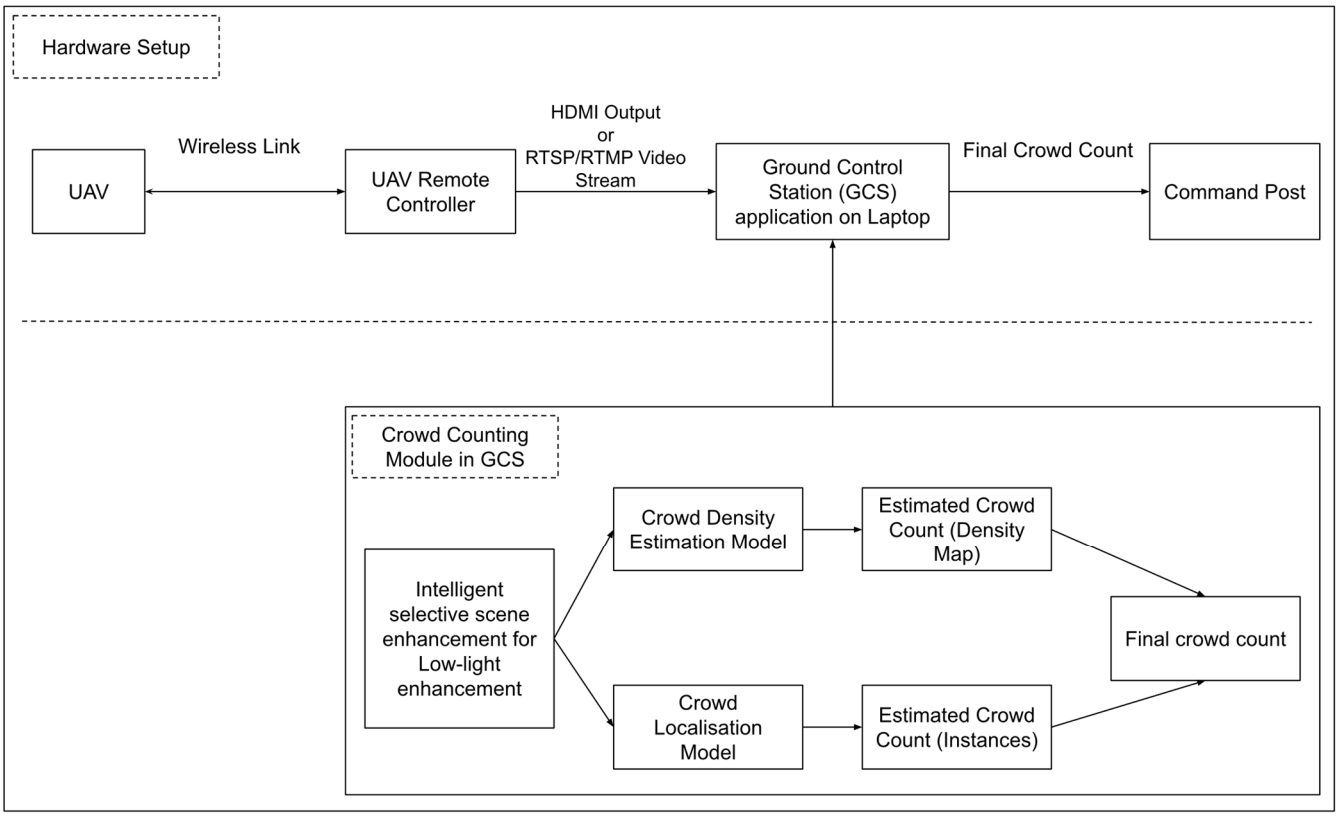


Fig. 1. Hardware setup and crowd counting module in GCS

also describes the dataset selection and filtering process, as well as the figures of merit used for evaluation. Section IV documents our tests results, Section V discusses prospective future works and conclusion.

II. Background and Related Work

Crowd counting has been a longstanding problem that has drawn the attention of many research works. A study conducted by Fiandeiro et al. [4] discusses the wide range of techniques that have been applied to crowd counting, initially ranging from detection-based methods using a classifier or individual and density-based regression methods [5]. These methods encountered problems as variations in crowd density, complexity and occlusions increased [6][7].

There have been extensive studies on crowd density estimation techniques that are based on deep learning methods. Convolution neural networks (CNN) gained immense popularity with their better performance. A prominent example is the scale aware multi-column CNN (MCNN) [8] by Zhang et al. Amongst them, the dual path multiscale fusion network architecture with attention mechanism SFANet [9] by Zhu et al., utilises attention mechanisms to generate density maps which are then summed up to attain the final count. Other works have taken the CNN one step further by fusing it with an encoder-decoder network, such as by the works of Khan et al. [10] and Ding et al. [11] Closer to our UAV-based use case, Drone-SCNet [12] by Elharrouss et al. showed promising results. The model was trained on the VisDrone2020 dataset. The authors succinctly summarised the key difficulties of UAV-based crowd counting, namely object scale variation, fast motion, flying height, complex background, and weather changes. These algorithms were geared towards providing crowd density estimation, where a pixel level quantity of heads is outputted, and the integral represents the final estimated count.

Another closely related subtask is crowd localisation, where instances of each person are located. These instances may appear as points as in Independent Instance Maps (IIM) [13] by Gao et. al, or as blobs such in the work done by Laradji et al. [14]. Taking one step further from CNNs, transformers were used as well, for instance by Liu et al. [15]. Further algorithms attempt to address both issues in one sitting, such as a Deep CNN with Composition loss by Idrees et al. [16] while also introducing the UCF-QNRF dataset. The authors noted that counting, crowd density estimation and crowd localisation are 3 closely related tasks. Another algorithm close to our UAV-based use case, STANet [17] by Wen et al. attempts to address both challenges and introduces their DroneCrowd dataset. However, we did not utilise the VisDrone2020 and DroneCrowd datasets in our testing due to them containing only top-down views, which will not be encountered during operations by our Home Team officers as direct overhead flying is prohibited for safety reasons. Thus, we also did not employ the models from Drone-SCNet and STANet as they were trained on these respective datasets.

From our observation, the leading performers are based on CNNs, and the literature demonstrates value in addressing both the challenges of crowd density estimation and crowd localisation to best provide an accurate crowd count. Thus, our selection narrowed to these two niches. The algorithms employed and datasets utilised for testing are discussed further below.

III. Methods and materials

A. Data fusion pipeline for UAV-based real time crowd counting

Fig. 1 demonstrates the concept of operations for our UAV-based crowd counting pipeline. The pipeline with its hardware and software is furthered elaborated in the next section.

B. Hardware Setup

In terms of hardware, the challenge we address is "How might we improve the crowd counting operations while minimising disruptions to the current operating procedure?" For our crowd counting pipeline to function, our pipeline design is simplified such that we only require a video stream input to the computing system equipped with our Ground Control Station (GCS). The video can be streamed into the computing system via a HDMI cable with a video capture card, or through wireless means such as RTMP or RTSP stream. As such, our crowd counting pipeline is adaptable to any UAV capable of video streaming and does not impose additional overheads to the command-and-control system.

C. Crowd Counting Pipeline

Our data fusion pipeline for UAV-based real time crowd counting approach incorporates the use of low-light enhancement on the input video frames, and finally obtaining the final count from the weighted fusion of two crowd counting algorithms. One algorithm is based on crowd density estimation method, and the other is based on crowd localisation method. For each of these steps, the necessary preprocessing performed includes resizing of the input images to a fixed width of 1280px while maintaining the same aspect ratio of the original images.

Scene enhancement – We recognise that the state-of-the-art crowd counting models perform exceptionally well in well-lit conditions; problems with count accuracy only arise with low-illumination scenes such as during night operations. In our work, we demonstrate the application of several existing computer-vision based scene enhancement methods on low-illumination scenes and their respective impacts on the count accuracy, as well as processing time required. These image enhancement methods leverage on the OpenCV python library. They include, histogram equalisation (HE), Contrast Limited Adaptive Histogram Equalisation (CLAHE), a histogram threshold based Automatic Brightness Contrast Adjustment (AutoBC), and other works like Multiscale Retinex [18] with Colour Restoration (MSR-CR) and Colour Preservation (MSR-CP).

Crowd density estimation – To obtain a crowd density estimation, our pipeline employs the algorithm based on the Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting (SFANet) [9] initially presented by Zhu et al., and then further modified by Thanasutives et al. [19]. According to the authors, the SFANet algorithm uses VGG16-bn as the feature map extractor to extract multi-scale features. In our pipeline, we utilise the pre-trained weights and models shared by Thanasutives et al. which were trained on the UCF-QNRF dataset.

Crowd localisation instances – To obtain crowd localisation instances of each person, our pipeline employs the algorithm Independent Instance Maps (IIM) [13] presented by Gao et al. The authors have also kindly shared their python implementation of their algorithm on GitHub, with the implementations utilising different feature extraction backbones. One is using the VGG16 backbone, while another implementation using the HRNet backbone [20].

Considering that the same input image is passed into both models, it is vital that each model employs a different backbone for feature extraction, so as to best provide a more robust feature detection step. We note that the SFANet algorithm adopts the first 13 layers from the VGG16-bn feature detector. Thus, in our pipeline, we have adopted the HRNet backbone instead of VGG16 for the IIM, with pre-trained weights that were trained on the NWPU dataset.

Fusion of model outputs – We observed that the crowd localisation and density estimation models perform better on sparse and dense crowds respectively. Thus, after model inference is completed, it is beneficial to fuse the outputs of the models to achieve counts of a higher accuracy. Several fusion methods were considered.

First, a pixel-level fusion involving summation of the maximum counts per pixel from each model's output. Initial testing revealed that this is conceptually inaccurate, resulting in significant overcount of the involved crowds. Further work is necessary to explore such a method.

Second, a two-stage process by first determining the crowd density of sub-regions of each image, before applying crowd density estimation for dense regions and crowd localisation instances for sparse regions. Initial testing showed that this method added significant processing overheads and was thus unsuitable or real-time operations. This method involves the deployment of a support vector machine (SVM), an additional machine learning model, to determine if a region of the image is dense or sparse before using the count from the respective crowd counting model in the final count.

Third, a late fusion by taking the mean of the total count of the models' output. This was found to be the better balance between count accuracy and processing time, and was the chosen method applied in this pipeline.

D. Human-in-the-loop

The crowd count is then presented in real-time as a numerical figure alongside the live video stream to the officers at the command post. From there, they can utilise this crowd count number of a particular sector of the operation grounds for decision making on the fly, allowing them to respond to unforeseen circumstances, allocate manpower and resources for crowd management.

E. Benchmark datasets

5 publicly available crowd counting datasets were used for testing our crowd counting pipeline. The variety ensures that a diverse range of scenes and a realistic range of crowd sizes are accounted for. To focus on crowd counting during low-illumination or night conditions, we selected a subset of images from each dataset based on criterion described in the next section on Dataset Filtering and Image Selection. Some sample images from these datasets are shown in Fig. 2. An overview of the images shortlisted from various datasets is shown in Table I. Finally, for the purpose of this paper, we will also perform the pipeline on images obtained from actual operations. Some photos obtained from these operations are shown in Fig. 3.

Fig. 2. Sample night images with respective ground truth from 5 publicly avalable datasets

Fig. 3. Sample photos from HTX Singapore Dataset

F. Dataset filtering and image selection

1) **GCC** [21] dataset features synthetic images generated using the game, Grand Theft Auto V, described by the time of day, weather, camera location(x, y, z) and attitudes (roll, pitch, yaw). The selection criteria for time is set such that all images tagged with time information less than or equal to 6 and greater than or equal to 19 (to represent before 0659H and after 1900H) were chosen to describe low-illumination scenes. To reflect the higher viewpoint of UAVs, a pitch angle of minimum 20 degrees and z-coordinate of at least 20 was used.

2) **MOT20Det** [22] is traditionally used for Multi-Object Tracking (MOT). For our work, we have selected the sequences of MOT20-03 and MOT20-05 as these represent low-illumination night time scenes with elevated viewpoints.The number of bounding boxes were treated as the ground truth headcount.

3) **JHU-Crowd++** [23] dataset features image level annotations. However, these annotations do not contain illumination details. Thus, we employed a criteria of mean pixel intensity values less than 85 for selecting low-illumination scenes.

4) **NWPU** [24] dataset features image level annotations, one of which describes the illumination level of the image. We selected all images from the "train" and "val" distributions with illumination level 0. We did not select any from the "train" distribution as there were no ground truth values.

5) **UCF-QNRF** [16] dataset does not contain image level annotations. Thus, we employed the same criteria of mean pixel intensity values less than 85 for selecting low-illumination scenes.

6) **HTX Singapore Dataset** contains 24 images that reflects our real world operations and actual problems faced during crowd counting for large scale public events at night in the urban spaces of Singapore. These images were taken from UAVs flown at varying altitudes, at a safe distance away from the crowds, and each head has been manually hand-labelled to ensure accuracy.

The descriptors for the overview of the dataset table shown in Table I are as follows:

- **Source dataset**: Indicates the original dataset from which the smaller subset of images were filtered and selected;
- **Selection criteria**: Describes the numerical filtering criteria, where possible.
- **Selected sequences**: describes the subsets that were chosen to qualitatively represent the desired conditions when no numerical filtering criteria was applied.
- **Image number**: represents the quantity of images after the filtering and selection process.
- **Min count**: represents the lowest number of headcounts in the selection.
- **Max count**: represents the highest number of headcounts in the selection;
- **Avg count**: represents the averaged headcount in the selection, rounded to the nearest whole number;
- **Total count**: represents the total number of labelled objects.

TABLE I. OVERVIEW OF DATASETS

Source Dataset	Image Number	Min Count	Max Count	Avg Count	Total Count
GCC	331	0	3995	620	205358
MOT20Det	5720	84	248	187	1073155
JHU-Crowd++	1103	0	21859	332	400520
NWPU	214	0	7636	317	67879
UCF-QNRF	427	49	5520	628	268235
HTX Singapore	24	48	1977	611	14652

G. Figures of merit

As we are interested in the overall accuracy of the count, we will use the metrics of Mean Absolute Error (MAE) and Mean Squared Error (MSE) to compare the performances of each enhancement. The average processing and inference time (in seconds) is also recorded to quantify the effect of the scene enhancements on the overall inference time. The equations for MAE and MSE are given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i - \hat{C}_i|, \quad (1)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(C_i - \hat{C}_i)^2, \quad (2)$$

where N is the total number of test images, C_i is the ground truth count for the i-th image and $\hat{C}_i$ is the estimated count.

IV. RESULTS AND DISCUSSION

To select the best performing combination of method(s) and enhancement algorithm, we evaluated the performance of the different pipelines on the 5 datasets and compare that with the performance on the HTX Singapore dataset. The results shown in Table II are measured by the MAE and MSE.

Throughout the 5 open-source datasets, the CLAHE enhancement technique performs the best consistently. It is worth noting that IIM with CLAHE enhancement also showed promising results. When weighted fusion, displayed as "AVERAGE" in Table II, is applied to obtain the final count from the estimated counts from the two methods of SFANet and IIM after CLAHE enhancement, counting accuracy was increased as compared to using the individual methods independently without image enhancement.

Thereafter, the pipelines were applied on 24 images obtained from 3 night-time events and the results shown in Table III demonstrated that CLAHE enhancement and weighted fusion is the third best performing pipeline. Fig. 4 shows the results from one of the images in the HTX Singapore dataset where CLAHE enhancement aided the SFANet and IIM methods to count more accurately. That said, it can be seen from these images that SFANet is suitable for dense crowd, but it missed out headcounts from the sparse crowd at the left side of the image. The opposite happens for IIM. Put together, the 2 models compensate for each other's misses.

From the average inference time recorded, we observed that the image enhancement methods of HE, CLAHE and AutoBC do not seem to increase inference times significantly. However, Retinex methods MSR-CR and MSR-CP takes significantly more time. In particular, we noticed in general, the CLAHE enhancement method provides better results in both crowd count accuracy and inference time.

Currently, the weighted fusion of the two methods is achieved by computing the mean count from the outputs. However, more factors can be considered for the weighted fusion component of the crowd counting pipeline to obtain a higher count accuracy. Such factors may include multi-modal methods such as utilising the fusion of RGB images and their thermal equivalent. Another possibility is developing a unified Bayesian neural network to combine illumination enhancement, crowd localisation, and density estimation. As processing time is a concern, a simple application of Bayesian probability on the model's prediction confidence and the final count could be explored as well.

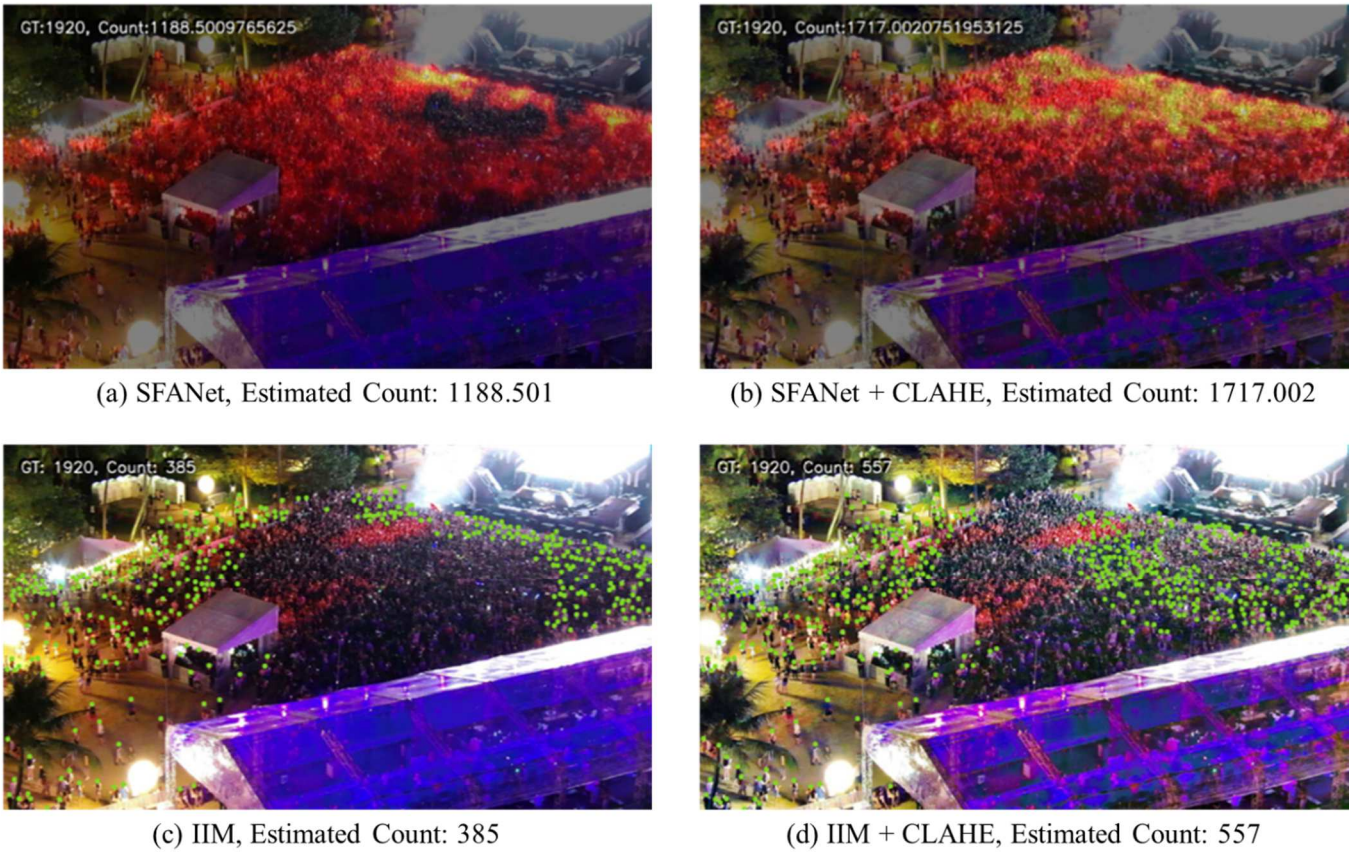
The work presented here demonstrates that there is value to be gained through thorough sensemaking on the input of a single sensor.

TABLE II. COMPARISON OF PIPELINES ON 5 PUBLICLY AVAILABLE DATASETS. THE BEST PERFORMANCE FOR EACH DATASET IS SHOWN IN **BOLD** AND THE SECOND BEST IS UNDERLINED.

Pipeline (Method + Enhancement)	GCC		MOT20Det		JHU-Crowd++		NWPU		UCF-QNRF		Overall Average Inference Time
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	
SFANet	440.9	657795.1	21.2	693.0	145.6	675524.5	120.3	205355.4	146.0	100820.7	0.160
SFANet + HE	**380.2**	520931.3	30.2	1205.1	145.4	633719.3	136.1	173511.4	138.1	95682.8	0.160
SFANet + CLAHE	397.0	553235.8	**15.4**	383.8	**143.4**	621708.5	**111.7**	147407.1	**134.8**	87586.8	0.159
SFANet + AutoBC	456.0	675397.4	57.7	3716.7	151.4	688813.2	136.8	204299.9	188.0	160660.0	**0.155**
SFANet + MSR-CP	400.1	548186.0	73.6	5981.6	184.9	788168.3	162.8	241176.9	248.8	280444.7	2.106
SFANet + MSR-CR	400.1	548151.1	73.6	5982.3	184.9	788161.7	162.8	240818.3	248.8	280532.9	2.064
IIM	472.5	781208.8	42.5	5643446.4	192.3	885215.9	165.1	357007.3	254.5	293649.1	0.329
IIM + HE	425.7	642101.6	59.8	19055685.5	206.2	873112.7	177.3	401540.5	275.5	304517.6	0.322
IIM + CLAHE	434.7	695819.1	46.3	8094803.4	195.2	879642.6	158.7	354062.7	247.6	282602.4	0.319
IIM + AutoBC	434.7	695819.1	85.6	7714.8	210.8	912326.1	194.9	403551.3	294.3	372363.8	0.318
IIM + MSR-CP	416.1	630321.3	59.2	3879.1	204.6	864642.6	176.4	357403.6	300.7	326270.7	1.185
IIM + MSR-CR	392.8	579288.9	80.9	7010.6	202.0	850099.1	173.2	349186.6	287.1	295301.0	2.290
AVERAGE	456.4	713516.1	31.1	1153.6	153.8	753655.1	136.1	269837.8	196.9	177300.5	0.489
AVERAGE + HE	399.9	576792.3	44.9	2271.9	153.5	719587.6	144.1	265610.1	191.3	170870.6	0.482
AVERAGE + CLAHE	414.8	618066.8	29.4	1051.8	148.0	715170.6	125.2	233236.7	183.2	161864.4	0.479
AVERAGE + AutoBC	444.8	679470.1	71.7	5458.1	166.2	772371.3	158.3	290355.1	236.9	241484.7	0.474
AVERAGE + MSR-CP	407.2	585458.5	66.4	4772.6	184.5	810041.8	165.3	283498.3	269.7	286527.4	3.292
AVERAGE + MSR-CR	395.8	560429.8	89.6	8175.0	186.2	807314.6	164.8	283234.9	263.8	274355.9	4.360

TABLE III. PERFORMANCE OF PIPELINES ON HTX SINGAPORE DATASET. THE BEST PERFORMING PIPELINE IS SHOWN IN **BOLD** AND THE SECOND BEST IS UNDERLINED.

Pipeline (Model + Enhancement)	MAE	MSE	Average Inference Time (s)
SFANet	193.2	144010.1	0.678
SFANet + HE	264.0	282958.7	0.750
SFANet + CLAHE	**139.9**	52444.8	0.637
SFANet + AutoBC	260.3	228011.0	0.700
SFANet + MSR-CP	441.9	553812.5	4.209
SFANet + MSR-CR	442.3	553986.5	3.868
IIM	345.0	350378.0	1.390
IIM + HE	284.8	254685.1	0.935
IIM + CLAHE	290.1	253754.1	0.824
IIM + AutoBC	411.2	500334.8	0.882
IIM + MSR-CP	258.0	183612.0	1.812
IIM + MSR-CR	322.8	268294.0	4.320
AVERAGE	265.5	223351.0	2.068
AVERAGE + HE	267.1	254405.5	1.685
AVERAGE + CLAHE	206.5	123731.1	1.461
AVERAGE + AutoBC	332.0	339836.5	1.582
AVERAGE + MSR-CP	348.1	333663.2	6.021
AVERAGE + MSR-CR	379.7	392650.9	8.187



(a) SFANet, Estimated Count: 1188.501

(b) SFANet + CLAHE, Estimated Count: 1717.002

(c) IIM, Estimated Count: 385

(d) IIM + CLAHE, Estimated Count: 557

Fig. 4. Sample results with visualisations of density maps in (a) and (b), and points in (c) and (d) for an image with ground truth of 1920.

V. Future Work and Conclusion

Crowd counting from UAVs during low-illumination conditions or during the night remains a challenging task. While our work may have achieved improved results on low-illumination images, other low-illumination conditions still pose a significant challenge. For instance, in scenes where there may be inconsistent or strobe lighting, there may be images with both overexposed and underexposed portions. Thus, individual images may not contain sufficient visual detail for accurate crowd counting.

We believe crowd counting across consecutive frames or in videos may be able to address this issue. There exists temporal correlation between frames that algorithms may tap on. Recent works such as the Bidirectional ConvLSTM [25] by Hanson et. al. attempts to leverage on such spatial-temporal information. However, some other works, for instance by Bai et. al. [26] have noted that a Long Short Term Memory (LSTM) based framework is difficult to train or generalise. Nonetheless, this challenging task represents an opportunity for further work through the fusion of different techniques.

Another important task is fusion of crowd counting across multiple views of the same scene. Given that these are large-scale events, it is impossible for one static camera to cover the entire scene of operations. UAV-based approaches, such as ours, represent the intermediate step which provides flexibility and mobility in operations spanning a large area. However, it is important to account that multiple cameras having overlapping views may result in a drastic overcount. Research work on crowd counting across multiple views exists, such as in works by Zhang et. al. [27][28]. This will allow crowd count information across an entire operation scene to be available in an integrated format, thereby simplifying information flow and decision-making process. We believe this is another important step in the future of crowd counting.

To summarise, our work has demonstrated the application of existing crowd counting techniques to UAV perspectives through our pipeline that is adaptable to any UAV capable of video streaming without affecting operations setups. In our testing process, our work has also demonstrated the value of pre-processing low-illumination images to achieve improved accuracy for crowd counts. There is still much work to be done for this exciting and unique application of crowd counting during low-illumination or night conditions from UAV perspectives. We hope our work will provide new insights and promote further work in ensuring public safety.

Acknowledgment

We would like to take this opportunity to thank the interns from HTX who had contributed to this project. We would also like to thank the Home Team UAV Command officers who helped to operate the UAVs during the public safety operations.

References

[1] C. Lim, “Police life: Keeping the countdown safe,” Singapore Police Force, https://www.police.gov.sg/Media-Room/Police-Life/2023/01/Keeping-the-Countdown-Safe (accessed Feb. 24, 2024).

[2] “A history of hajj tragedies,” The Guardian, https://www.theguardian.com/world/2006/jan/13/saudiarabia (accessed Feb. 24, 2024).

[3] J. A. Baker, “Seoul crowd crush: How it may have happened and what to do in such a situation,” CNA,

https://www.channelnewsasia.com/asia/seoul-crowd-crush-what-happened-what-do-3034321 (accessed Feb. 24, 2024).

[4] Miguel Fiandeiro, Thanh Thi Nguyen, Hanting Wong, and Edbert B. Hsu, "Modernized Crowd Counting Strategies for Mass Gatherings-A Review," Journal of Acute Medicine, vol. 13, no. 1, Mar. 2023, doi: 10.6705/j.jacme.202303_13(1).0002.

[5] B. Li, H. Huang, A. Zhang, P. Liu, and C. Liu, "Approaches on crowd counting and density estimation: a review," Pattern Analysis and Applications, vol. 24, no. 3. Springer Science and Business Media LLC, pp. 853–874, Feb. 20, 2021. doi: 10.1007/s10044-021-00959-z.

[6] H. Alnabulsi and J. Drury, "Social identification moderates the effect of crowd density on safety at the Hajj," Proceedings of the National Academy of Sciences, vol. 111, no. 25. Proceedings of the National Academy of Sciences, pp. 9091–9096, Jun. 09, 2014. doi: 10.1073/pnas.1404953111.

[7] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," Pattern Recognition Letters, vol. 107. Elsevier BV, pp. 3–16, May 2018. doi: 10.1016/j.patrec.2017.07.007.

[8] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2016. doi: 10.1109/cvpr.2016.70.

[9] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting." arXiv, 2019. doi: 10.48550/ARXIV.1902.01115.

[10] S. D. Khan, Y. Salih, B. Zafar, and A. Noorwali, "Correction to: A Deep-Fusion Network for Crowd Counting in High-Density Crowded Scenes," International Journal of Computational Intelligence Systems, vol. 14, no. 1. Springer Science and Business Media LLC, Oct. 15, 2021. doi: 10.1007/s44196-021-00035-8.

[11] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, and Y. Huang, "Crowd Density Estimation Using Fusion of Multi-Layer Features," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 8. Institute of Electrical and Electronics Engineers (IEEE), pp. 4776–4787, Aug. 2021. doi: 10.1109/tits.2020.2983475.

[12] O. Elharrouss et al., "Drone-SCNet: Scaled Cascade Network for Crowd Counting on Drone Images," IEEE Transactions on Aerospace and Electronic Systems, vol. 57, no. 6. Institute of Electrical and Electronics Engineers (IEEE), pp. 3988–4001, Dec. 2021. doi: 10.1109/taes.2021.3087821.

[13] J. Gao, T. Han, Q. Wang, Y. Yuan, and X. Li, "Learning Independent Instance Maps for Crowd Localization." arXiv, 2020. doi: 10.48550/ARXIV.2012.04164.

[14] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Where Are the Blobs: Counting by Localization with Point Supervision," Computer Vision – ECCV 2018. Springer International Publishing, pp. 560–576, 2018. doi: 10.1007/978-3-030-01216-8_34.

[15] C. Liu, H. Lu, Z. Cao, and T. Liu, "Point-Query Quadtree for Crowd Counting, Localization, and More," 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Oct. 01, 2023. doi: 10.1109/iccv51070.2023.00161.

[16] H. Idrees et al., "Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds," Computer Vision – ECCV 2018. Springer International Publishing, pp. 544–559, 2018. doi: 10.1007/978-3-030-01216-8_33.

[17] L. Wen et al., "Drone-based Joint Density Map Estimation, Localization and Tracking with Space-Time Multi-Scale Attention Network." arXiv, 2019. doi: 10.48550/ARXIV.1912.01811.

[18] A. B. Petro, C. Sbert, and J.-M. Morel, "Multiscale Retinex," Image Processing On Line, vol. 4. Image Processing On Line, pp. 71–88, Apr. 16, 2014. doi: 10.5201/ipol.2014.107.

[19] P. Thanasutives, K. Fukui, M. Numao, and B. Kijsirikul, "Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting," arXiv, 2020, doi: 10.48550/ARXIV.2003.05586.

[20] J. Wang et al., "Deep High-Resolution Representation Learning for Visual Recognition." arXiv, 2019. doi: 10.48550/ARXIV.1908.07919.

[21] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning From Synthetic Data for Crowd Counting in the Wild," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Jun. 2019. doi: 10.1109/cvpr.2019.00839.

[22] P. Dendorfer et al., "MOT20: A benchmark for multi object tracking in crowded scenes." arXiv, 2020. doi: 10.48550/ARXIV.2003.09003.

[23] V. Sindagi, R. Yasarla, and V. M. M. Patel, "JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method," IEEE Transactions on Pattern Analysis and Machine Intelligence. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–1, 2020. doi: 10.1109/tpami.2020.3035969.

[24] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 6. Institute of Electrical and Electronics Engineers (IEEE), pp. 2141–2149, Jun. 01, 2021. doi: 10.1109/tpami.2020.3013269.

[25] A. Hanson, K. PNVR, S. Krishnagopal, and L. Davis, "Bidirectional Convolutional LSTM for the Detection of Violence in Videos," Lecture Notes in Computer Science. Springer International Publishing, pp. 280–295, 2019. doi: 10.1007/978-3-030-11012-3_24.

[26] H. Bai, J. Mao, and S.-H. G. Chan, "A Survey on Deep Learning-based Single Image Crowd Counting: Network Design, Loss Function and Supervisory Signal." arXiv, 2020. doi: 10.48550/ARXIV.2012.15685.

[27] Q. Zhang and A. B. Chan, "3D Crowd Counting via Multi-View Fusion with 3D Gaussian Kernels." arXiv, 2020. doi: 10.48550/ARXIV.2003.08162.

[28] Q. Zhang, W. Lin, and A. B. Chan, "Cross-View Cross-Scene Multi-View Crowd Counting." arXiv, 2022. doi: 10.48550/ARXIV.2205.01551.