

Multimodal Crowd Counting with Pix2Pix GANs

Muhammad Asif Khan¹^a, Hamid Menouar¹^b and Ridha Hamila²^c

¹*Qatar Mobility Innovations Center, Qatar University, Doha, Qatar*

²*Department of Electrical Engineering, Qatar University, Doha, Qatar*

Keywords: Crowd Counting, CNN, Density Estimation, Multimodal, RGB, Thermal.

Abstract: Most state-of-the-art crowd counting methods use color (RGB) images to learn the density map of the crowd. However, these methods often struggle to achieve higher accuracy in densely crowded scenes with poor illumination. Recently, some studies have reported improvement in the accuracy of crowd counting models using a combination of RGB and thermal images. Although multimodal data can lead to better predictions, multimodal data might not be always available beforehand. In this paper, we propose the use of generative adversarial networks (GANs) to automatically generate thermal infrared (TIR) images from color (RGB) images and use both to train crowd counting models to achieve higher accuracy. We use a Pix2Pix GAN network first to translate RGB images to TIR images. Our experiments on several state-of-the-art crowd counting models and benchmark crowd datasets report significant improvement in accuracy.

1 INTRODUCTION


Crowd counting has gained significant attention in recent years due to its diverse applications across various domains, including crowd management, urban planning, security surveillance, event management, and public safety. The ability to accurately estimate crowd density and count individuals in congested areas holds immense practical significance, enabling better resource allocation, improved crowd control strategies, and enhanced decision-making processes. The advent of deep learning (DL) has led to a paradigm shift in crowd counting techniques, achieving higher accuracy and scalability in diverse real-world applications.


Most state-of-the-art works on crowd counting use the density estimation approach (Khan et al., 2022; ?). In density estimation, a deep learning model (typically a convolution neural network (CNN) or a vision transformer (ViT)) is trained using annotated crowd images to learn the crowd density. The annotations come in the form of sparse localization maps that indicate the head positions of individuals present in each scene or region of interest. The localization maps are converted into continuous density maps for each image by applying a Gaussian kernel centered at each


head location. The intensity of the map thus represents the crowd density at different spatial locations. The density maps serve as the ground truth for the crowd images to train the crowd counting model.

A number of different models have been proposed in the last few years using the density estimation approach. These models apply various model architectural innovations and novel learning functions to improve the accuracy performance. However, most of these works use optical color images mainly with reasonable lighting conditions. However, in many surveillance scenarios images captured using optical cameras will have poor lighting conditions resulting in poor performance of the counting models. To improve the accuracy, thermal infrared (TIR) cameras are used along with the optical camera to capture both color RGB images and thermal images. Then a crowd counting model may use one (monomodal) or both (multimodal) RGB and TIR images to learn the crowd density in low-light conditions.

Multimodal learning using both optical (RGB) and thermal (TIR) images for crowd counting has been proposed in recent works (Wu et al., 2022; Liu et al., 2023; Thißen and Hergenröther, 2023). The simultaneous use of both RGB and TIR images provides more information to extract enriched features from the crowd scene leading to better predictions. Nevertheless, the interest in multimodal crowd counting is increasing, there exist only a few multimodal

^a <https://orcid.org/0000-0003-2925-8841>

^b <https://orcid.org/0000-0002-4854-909X>

^c <https://orcid.org/0000-0002-6920-7371>

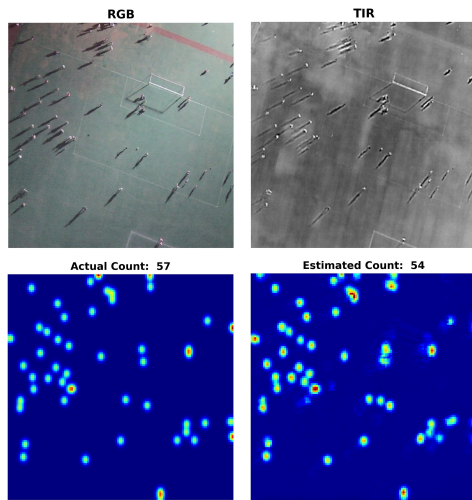


Figure 1: Illustration of counting prediction on a single sample: RGB image (top-left), corresponding TIR image (top-right), ground truth density map (bottom-left), and estimated density map (bottom-right).

datasets with both optical and thermal images. Thus, our aim in this paper partly is to address this challenge. We propose the use of generative adversarial networks (GANs) (Goodfellow et al., 2014) to generate high-quality synthetic thermal images from optical images. Generative Adversarial Networks (GANs) are powerful deep learning models capable to generate realistic and high-quality data. The Pix2Pix (Isola et al., 2016; Khan et al., 2023a) is a type of GAN that translate images from a given source domain to a target domain using paired training data. Pix2Pix has been successfully used in applications such as image-to-image translation, image colorization, style transfer, etc. The generated thermal images using Pix2Pix GAN may not contain all the information that real thermal images would contain, but they certainly provide more enriched information that the model can learn. This paper thus investigates the use of Pix2Pix GANs to populate the existing monomodal RGB crowd datasets with multimodal thermal images to improve the performance of existing crowd models.

The contribution of this paper is as follows:

- We used a Pix2Pix GAN trained on an existing RGBT crowd dataset to generate high-quality realistic thermal images for RGB images. Fig. 1 illustrates the RGB-to-TIR image translation. The efficacy of using synthetic images generated by Pix2Pix GAN in crowd counting is evaluated. For this purpose, various crowd counting models are trained on RGB, real thermal, and synthetic thermal images.
- We design a minimal end-to-end multimodal crowd counting method (MMCount) that takes an

RGB input image, generates a thermal counterpart, and trains on both RGB and thermal images to predict the crowd density map.

- The performance of the proposed method is evaluated on three benchmark datasets using RGB-only, TIR-only, and RGB+TIR inputs to show the efficacy of the multimodal scheme.

2 RELATED WORK

2.1 Crowd Counting

Traditional computer-vision methods were based on detecting people in crowd images using shapes, body parts, or other image information such as texture, and foreground segmentation. However, more accurate counting models today use density estimation using convolution neural networks (CNNs) and vision transformers (ViTs). The first CNN-based crowd-counting model i.e., CrowdCNN (Zhang et al., 2015) was a single-column 6-layer CNN architecture. Afterward, a large number of CNN architectures using the density estimation approach were proposed with progressive improvement in the network architecture, learning functions, and evaluation methods. MCNN (Zhang et al., 2016), CrowdNet (Boominathan et al., 2016), SCNN () proposed multi-column architectures to overcome the scale variations in crowd images. MCNN uses three CNN columns with receptive fields of different sizes and combines the features learned by each column to predict the final density map. CrowdNet (Boominathan et al., 2016) is a 2-layer architecture containing a 5-layer deep network and a 3-layer shallow network with feature fusion for final prediction. Scale-aware architectures replaced simple multi-column networks for learning scale variations using special multi-scale modules for scale-relevant feature extraction. MSCNN (Zeng et al., 2017), a single-column network uses Inception modules (Szegedy et al., 2015) called multi-scale blobs (MSBs) for feature extraction in different layers. A cascaded multi-task learning (CMTL) model (Sindagi and Patel, 2017) is proposed to adapt to the wide variations of density levels in images. CMTL is also a two-column network. The first column is a high-level prior that classifies an input image into groups based on the total count in the image. The features learned by the high-level prior are shared with the second column that estimates the respective density map.

In highly dense crowd images, simple CNN architectures such as multi-column or scale-aware models often gain limited accuracy. In highly congested

scenes, deeper architectures relying on transfer learning provide better performance. Authors in (Li et al., 2018) propose CSRNet (Li et al., 2018) uses VGG16 (Simonyan and Zisserman, 2015) as the front-end to extract features, and a CNN network with dilated convolution. Other models using transfer learning include CANNet (Liu et al., 2019), GSP (Aich and Stavness, 2018), TEDnet (Jiang et al., 2019), Deepcount (Chen et al., 2020), SASNet (Song et al., 2021), M-SFANet (Thanasutives et al., 2021), and SGANet (Wang and Breckon, 2022).

2.2 Generative Adversarial Networks

The first vanilla GAN architecture was proposed in (Goodfellow et al., 2014). The Vanilla GAN is a two-stage architecture comprised of a generator and discriminator engaged in adversarial training. The generator creates data to fool the discriminator, while the discriminator aims to distinguish between real and generated data. During the training, the generator improves its ability to produce realistic samples. The conditional GANs (cGAN) (Mirza and Osindero, 2014) extend the Vanilla GAN by allowing the user to generate specific types of data by providing specific conditioning information. DCGANs (Radford et al., 2016) employs deep networks in both generator and discriminator to generate more realistic images. CycleGANs (Zhu et al., 2017) are proposed for image-to-image translation without the need for paired training data and are used in applications such as style transfer, domain adaptation, etc. Pix2Pix GANs (Isola et al., 2016) are based on cGANs (Mirza and Osindero, 2014) to generate a target image conditional on a given input image. StyleGAN (Karras et al., 2021) allows a style-based generation of more realistic and diverse images. The Pix2Pix GAN (Isola et al., 2016) provides a generic framework for image-to-image translation. Pix2Pix GANs are highly capable of translating images from one domain to another and are used in various tasks such as image colorization, style transfer, etc. Recently, Pix2Pix GANs have been used to generate synthetic near-infrared images of crops (de Lima et al., 2022), lung CT scan images (Toda et al., 2022), MRA images of brain (Aljohani and Alharbe, 2022), etc. (Khan et al., 2023a) propose Pix2Pix GAN based image denoising architecture for fine-grained density estimation of crowd.

3 THE PROPOSED METHOD

We propose a crowd density estimation framework MMCount. There are two essential parts; a Pix2Pix

GAN, and a multimodal crowd counting network. The Pix2Pix GAN generates thermal infrared (TIR) images from optical RGB images of the crowd scene and the crowd model uses both RGB and TIR images to predict the crowd density map. The architecture of MMCount is shown in Fig. 2 and is explained as follows:

3.1 Pix2Pix GANs

A Pix2Pix GAN consists of two main components: a generator and a discriminator. The generator takes an input RGB image and aims to transform it into a target TIR image. The generator uses an encoder-decoder architecture. The encoder encodes the RGB image into a compact representation, and the decoder decodes this representation to generate the TIR image.

The discriminator is a CNN architecture that takes the original RGB image and the generated TIR image (by generator) and attempts to classify it as real or fake. It is trained to distinguish between the real TIR images and the TIR images generated by the generator.

During the training, the generator aims to minimize the pixel-wise L_1 loss to generate more realistic TIR images. The generator also aims to minimize the binary cross-entropy loss (called adversarial loss or GAN loss) to fool the discriminator. The discriminator aims to maximize the adversarial loss to correctly classify real and generated TIR images. The pixel-wise L_1 loss measures the absolute pixel-wise difference between the generated output and the ground truth target and is given in Eq. 1:

$$\mathcal{L}_{L1}(G) = \frac{1}{N} \sum_{i=1}^N |G(x_i) - y_i| \quad (1)$$

where G is the generator network, N is number of samples, x_i is the input sample, and y_i is the ground truth value.

The BCE loss or the adversarial loss measures the similarity between the discriminator's predictions for the generated output and the ground truth target and is given in Eq. 2:

$$\mathcal{L}_{BCE}(D, G) = -\frac{1}{N} \sum_{i=1}^N \left(y_i \log D(x_i) + (1 - y_i) \log(1 - D(G(x_i))) \right) \quad (2)$$

where D is the discriminator network, N is the total number of samples, $D_{(x_i)}$ is discriminator's prediction for input x_i , and $D(G_{(x_i)})$ is discriminator's prediction for the generated output $G_{(x_i)}$.

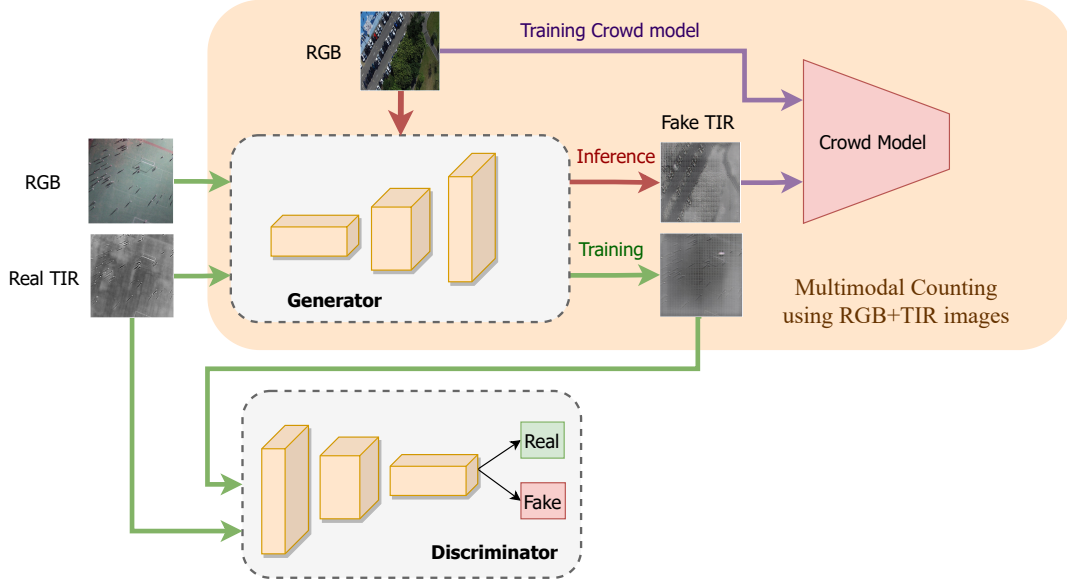


Figure 2: The proposed method for multimodal crowd counting using RGB+TIR images. The TIR images are generated by a Pix2Pix GAN trained earlier on RGB+TIR paired datasets.

3.2 The Multimodal Counting Network

The counting network called "MMCount" is a CNN architecture with two branches, a RGB branch, and a TIR branch. Both branches have similar structures and consist of four convolution layers (conv), each having 16, 32, 64, and 128 filters of (3×3) size, respectively. Each conv layer is also followed by a Relu activation and a (2×2) pooling layer with $stride = 2$. The outputs of both RGB and TIR layers are concatenated and fused in the fusion layer containing 256 filters of size 3×3 . Lastly, a 1×1 conv layer is used to generate the density map. The output density map generated by the MMCount is $1/8$ of the original input images (both RGB and TIR images are of the same size). The complete architecture is presented in Table 1

Table 1: The architecture of MMCount. The vector in Conv layers represents [input channels, output channels, kernel size, padding, stride]. Each vector in Pool layers represents [kernel size, stride].

Type	RGB layer	TIR layer
Conv	[3,16,3,1,1]	[1,16,3,1,1]
Pool	[2,2]	[2,2]
Conv	[16,32,3,1,1]	[16,32,3,1,1]
Pool	[2,2]	[2,2]
Conv	[32,64,3,1,1]	[32,64,3,1,1]
Conv	[64,128,3,1,1]	[64,128,3,1,1]
Conv (Fusion)	[256,256,3,1,1]	
Conv (Regressor)	[256,1,1,1,1]	

To train the MMCount model, we use the original annotations for the RGB images which are in the form of head positions. The head positions are used to generate sparse localization maps i.e., binary matrices of pixel values (same size as the image) in which a value of 1 denotes the head position whereas a 0 represents no head. Such a dot map is used to create density maps that serve as the ground truth for the images to train the model. A density map is generated by convolving a delta function $\delta(x - x_i)$ with a Gaussian kernel G_σ , where x_i are pixel values containing the head positions.

$$D = \sum_{i=1}^N \delta(x - x_i) * G_\sigma \quad (3)$$

where, N denotes the total number of dot points with value 1 in the dot map (i.e., total headcount in the input image). The integral of the density map D is equal to the total head count in the image. Visually, this operation creates a blurring of each head annotation using the scale parameter σ . The value of σ can be fixed (Cao et al., 2018) or adaptive (Idrees et al., 2018; Zeng et al., 2017). The typical loss function used in most crowd density estimation works is the l_2 loss (euclidean distance) between the target and the predicted density maps and is given in Eq. 4.

$$L(\Theta) = \frac{1}{N} \sum_1^N \|D(X_i; \Theta) - D_i^{gt}\|_2^2 \quad (4)$$

where N is the total number of samples in training data, X_i is the input image, D_i^{gt} is the ground truth density map, and $D(X_i; \Theta)$ is the predicted density map.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

4.1.1 DroneRGBT

The dataset contains 3600 pairs of RGB and thermal images. All images have a fixed resolution of 512×640 pixels. The images cover different scenes e.g., campus, streets, parks, parking lots, playgrounds, and plazas. The dataset is divided into a training set (1807 samples) and a test set (912 samples).

4.1.2 ShanghaiTech Part-B

The dataset is a large-scale crowd-counting dataset used in many studies. The dataset is split into train and test subsets consisting of 400 and 316 images, respectively. All images are of fixed size (1024×768).

4.1.3 CARPK

This dataset contains images of cars from 4 different parking lots captured using a drone (Phantom 3 Professional) at approximately 40-meter altitude. The dataset contains 1448 images split into train and test sets of sizes 989 and 459 images, respectively. The dataset contains a total of 90,000 car annotations and has been used in several object counting and object detection studies.

4.2 Evaluation Metrics

The standard and commonly used metric to evaluate the performance of crowd counting models is the mean absolute error (MAE) calculated using the following Eq. 5.

$$MAE = \frac{1}{N} \sum_1^N \|e_n - \hat{g}_n\| \quad (5)$$

where, N is the size of the dataset, g_n is the target or label (actual count) and e_n is the prediction (estimated count) in the n^{th} crowd image. MAE provides per-image counting and does not take into account the incorrect density estimations in the same image. Grid Average Mean absolute Error (GAME) (Guerrero-Gómez-Olmedo et al., 2015) can overcome these errors by computing patch-wise error i.e., an image is divided into 4^L non-overlapping patches and MAE is computed over each patch. GAME is thus a more robust and accurate metric for crowd-counting. It is defined in Eq. 6.

$$GAME = \frac{1}{N} \sum_{n=1}^N \left(\sum_{l=1}^{4^L} |e_n^l - g_n^l| \right) \quad (6)$$

By setting $L = 0$, GAME(0) becomes equivalent to MAE. GAME(1) means the image is divided into 4 patches and GAME(2) means 16 patches.

4.3 Baselines

First we evaluate the quality of generated TIR images by measuring their performance in monomodal crowd counting. For this purpose, four different models i.e., MCNN (Zhang et al., 2016), CMTL (Sindagi and Patel, 2017), CSRNet (Li et al., 2018), SANet (Cao et al., 2018), and LCDnet (Khan et al., 2023b) are used. Then, to measure the performance of MM-Count in crowd density estimation using multimodal crowd counting using RGB+TIR images is evaluated. We also evaluate the counting performance of multimodal learning in MCNN and DroneNet models by feeding one CNN column with TIR images.

4.4 Settings

In our experiments, we used a fixed value of σ to generate the ground truth density maps. To be more specific, a value of 7, 10 and 15 are used for DroneRGBT, CARPK, and ShanghaiTech Part-B datasets, respectively. We used the *Adam* optimizer with a learning rate of 1×10^{-3} . We used pixel-wise L_1 loss function and BCE loss for training the Pix2Pix GAN and L_2 or MSE loss function to train all crowd counting models. The crowd counting model training terminates when the MAE error does not reduce after 10 epochs. All the models are trained on two RTX-8000 GPUs using the PyTorch framework.

4.5 Pix2Pix GAN Results

In the first set of experiments on Pix2Pix GANs, the synthetic TIR images are generated and stored for each dataset. The Pix2Pix model is trained on the paired RGB and TIR images provided by the DroneRGBT dataset and the trained model is then used to generate the TIR images for other datasets. Fig. 3 shows samples TIR images generated for the three datasets.

Interestingly, the Pix2Pix GAN architecture have been effective to generate high quality TIR images of RGB images from different datasets, despite large variations in the crowd scenes and camera settings.

4.6 Crowd Counting Results

To evaluate the efficacy of the generated TIR images, we trained five crowd counting models over three different inputs i.e., (i) using RGB images, (ii) using

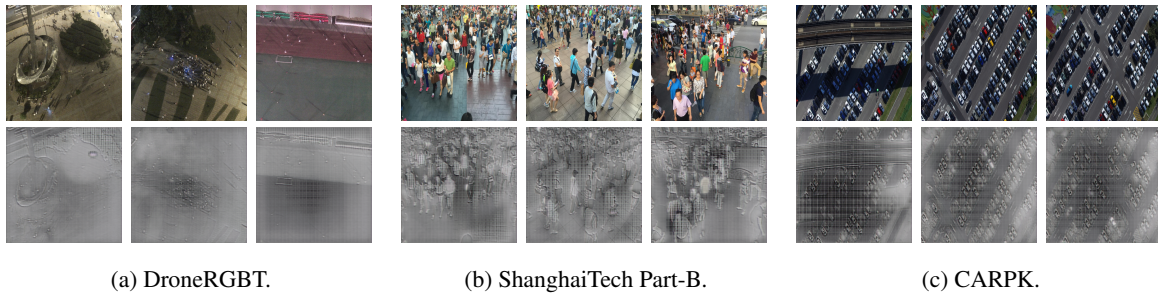


Figure 3: Thermal infrared (TIR) images generated using Pix2Pix GAN.

real TIR images, and (iii) using generated TIR images. The purpose is two fold: first, to understand the ability of crowd models from TIR images and second, the efficacy of generated TIR images to produce comparable results to the real TIR images. Table 2 presents the results of the first set of experiments. It can be observed that RGB images produce

Table 2: Accuracy comparison of crowd counting models over DroneRGBT dataset trained using (i) RGB images only versus (ii) TIR images only versus (iii) Generated TIR images only.

Model	Mean Absolute Error (MAE)		
	RGB	TIR	Generated TIR
MCNN	17.9	20.2	22.5
CMTL	18.1	19.6	21.4
CSRNet	7.6	10.8	13.7
SANet	16.2	18.3	20.8
LCDnet	21.4	23.2	24.4

higher counting accuracy (lower MAE values) due to more enriched information. The interesting information is the MAE values crowd models using generated TIR images which are only slightly lower than those of real TIR images. This strengthens our motivation to use Pix2Pix GAN to generate TIR images for a multimodal counting framework. In the second set of experiments, we trained the proposed multimodal counting network using monomodal data (RGB-only and TIR-only images) and multimodal data (using RGB+TIR images). Furthermore, due to the multi-column structure of MCNN and DroneNet, we extend our experiments to train these models by feeding one of three columns using the TIR images. The TIR images in both monomodal and multimodal settings in this experiment are generated by the Pix2Pix GAN. The performance is measured using the GAME met-

ric with a value $L = 0, 1, 2$ corresponding to the whole image, four patches per image, and 16 patches per image. The results are compared in Table 3.

5 CONCLUSIONS

This paper addresses the challenge of the limited availability of training data in crowd counting scenarios and proposes the use of generative adversarial networks to instantly generate scene-specific multimodal data. A Pix2Pix GAN is used to generate thermal infrared images for the available color RGB images. The multimodal data (RGB+TIR) is then used to train or fine-tune a crowd counting model. Experiments are conducted using three publicly available datasets. The results show improvements in model performance when trained using actual RGB and synthetic TIR images. The TIR images are generated using a Pix2Pix GAN trained on cross-scene drone images and applied to new and unseen RGB images. We believe the results can be further improved by training the GAN model on images of more correlated scenarios. As a future work, we plan to develop novel lightweight GAN architectures for real-time performance.

ACKNOWLEDGEMENT

This publication was made possible by the PDRA award PDRA7-0606-21012 from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Table 3: Accuracy comparison of crowd counting models over several datasets (trained using RGB-only versus TIR-only versus RGB+TIR images). All TIR images are generated using Pix2Pix GAN. GAME(0) is equivalent to MAE.

Input	Model	DroneRGBT			ShanghaiTech Pat-B			CARPK		
		GAME0	GAME1	GAME2	GAME0	GAME1	GAME2	GAME0	GAME1	GAME2
MCNN	RGB	17.9	24.2	42.0	26.4	34.4	55.2	10.1	21.2	43.4
	TIR	22.5	27.5	51.3	35.2	38.2	66.7	16.8	28.0	49.0
	RGB+TIR	16.2	21.0	35.1	23.2	31.5	48.5	8.9	19.6	36.1
DroneNet	RGB	11.3	22.1	32.7	22.4	30.2	41.9	9.0	20.5	40.1
	TIR	18.6	25.2	40.3	29.2	33.4	52.8	15.3	27.6	46.2
	RGB+TIR	10.1	18.8	28.4	20.0	29.7	38.3	8.1	18.5	26.8
MMCount	RGB	10.8	21.1	32.4	21.6	28.2	40.1	8.2	18.3	37.5
	TIR	16.0	23.3	40.6	27.7	33.6	49.9	15.0	25.6	42.8
	RGB+TIR	9.2	18.0	26.0	18.2	28.0	36.4	7.8	15.2	25.0

REFERENCES

- Aich, S. and Stavness, I. (2018). Global sum pooling: A generalization trick for object counting with small datasets of large images. *arXiv preprint arXiv:1805.11123*.
- Aljohani, A. A. and Alharbe, N. R. (2022). Generating synthetic images for healthcare with novel deep pix2pix gan. *Electronics*.
- Boominathan, L., Kruthiventi, S. S. S., and Babu, R. V. (2016). Crowdnet: A deep convolutional network for dense crowd counting. *Proceedings of the 24th ACM international conference on Multimedia*.
- Cao, X., Wang, Z., Zhao, Y., and Su, F. (2018). Scale aggregation network for accurate and efficient crowd counting. In *ECCV*.
- Chen, Z., Cheng, J., Yuan, Y., Liao, D., Li, Y., and Lv, J. (2020). Deep density-aware count regressor. In *ECAI*.
- de Lima, D. C., Saqui, D., Mpinda, S. A. T., and Saito, J. H. (2022). Pix2pix network to estimate agricultural near infrared images from rgb data. *Canadian Journal of Remote Sensing*, 48:299 – 315.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.
- Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R. J., Maldonado-Bascón, S., and Oñoro-Rubio, D. (2015). Extremely overlapping vehicle counting. In *IbPRIA*.
- Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S. A., Rajpoot, N. M., and Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. *ArXiv*, abs/1808.01050.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.
- Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D. S., and Shao, L. (2019). Crowd counting and density estimation by trellis encoder-decoder networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6126–6135.
- Karras, T., Laine, S., and Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12):4217–4228.
- Khan, M. A., Menouar, H., and Hamila, R. (2022). Revisiting crowd counting: State-of-the-art, trends, and future perspectives. *ArXiv*, abs/2209.07271.
- Khan, M. A., Menouar, H., and Hamila, R. (2023a). Crowd counting in harsh weather using image denoising with pix2pix gans. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE.
- Khan, M. A., Menouar, H., and Hamila, R. (2023b). Lcd-net: A lightweight crowd density estimation model for real-time video surveillance. *J. Real-Time Image Process.*, 20(2).
- Khan, M. A., Menouar, H., and Hamila, R. (2023c). Visual crowd analysis: Open research problems. *AI Magazine*, 44(3):296–311.
- Li, Y., Zhang, X., and Chen, D. (2018). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100.
- Liu, W., Salzmann, M., and Fua, P. V. (2019). Context-aware crowd counting. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5094–5103.
- Liu, Z., Wu, W., Tan, Y., and Zhang, G. (2023). Rgb-t multi-modal crowd counting based on transformer.
- Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *ArXiv*, abs/1411.1784.

- Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Sindagi, V. A. and Patel, V. M. (2017). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6.
- Song, Q., Wang, C., Wang, Y., Tai, Y., Wang, C., Li, J., Wu, J., and Ma, J. (2021). To choose or to fuse? scale selection for crowd counting. In *AAAI*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.
- Thanasutives, P., ichi Fukui, K., Numao, M., and Kijirikul, B. (2021). Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2382–2389.
- Thißen, M. and Hergenröther, E. (2023). Why existing multimodal crowd counting datasets can lead to unfulfilled expectations in real-world applications.
- Toda, R., Teramoto, A., Kondo, M., Imaizumi, K., Saito, K., and Fujita, H. (2022). Lung cancer ct image generation from a free-form sketch using style-based pix2pix for data augmentation. *Scientific Reports*, 12.
- Wang, Q. and Breckon, T. (2022). Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Transactions on Intelligent Transportation Systems*.
- Wu, Z., Liu, L., Zhang, Y., Mao, M., Lin, L., and Li, G. (2022). Multimodal crowd counting with mutual attention transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Zeng, L., Xu, X., Cai, B., Qiu, S., and Zhang, T. (2017). Multi-scale convolutional neural networks for crowd counting. *2017 IEEE International Conference on Image Processing (ICIP)*, pages 465–469.
- Zhang, C., Li, H., Wang, X., and Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841.
- Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.