Getting the Numbers Right—Modelling Multi-Class Object Counting in Dense and Varied Scenes

Villanelle O'Reilly

they/them or **she/her**, in order of preference

Supervised by:
Marc Hanheide & Georgios Leontidis
Further advice from:
James Brown & Petra Bosilj
Contributions from:
Jonathan Cox

June 24th, 2026

Why Count?

- Quantification constitutes a fundamental prerequisite for data-driven decision-making.
- Quantity translates directly into actionable metrics.

Why Count?

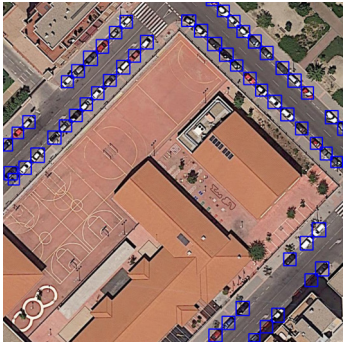
- Agricultural yield estimation: counts of fruit and flowers translate into expected volume.
- Biodiversity monitoring: save expert's time for thinking, and use computers to collect transient ecological data.
- Unconstrained pedestrian counting: precise resource management, transportation engineering, urban design, and advertising Shen et al., 2019

Figure 1: A robot autonomously surveys in a strawberry polytunnel

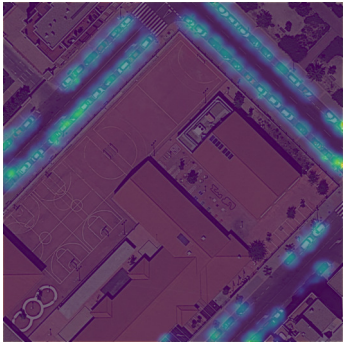
The Impact of Automated Counting

- Numerical extraction transforms qualitative visual scenes into quantitative insights necessary for strategic planning.
- Manual enumeration is impractical at scale. In this work we try to improve the accuracy of counting for dense, occluded, and heterogeneous environments.

How can we count?



(a) Object detection



(b) Density estimation

Figure 2: Demonstration of counting methods in an occlusion-free iSAID sample.

Crowd Counting as a General Counting Problem

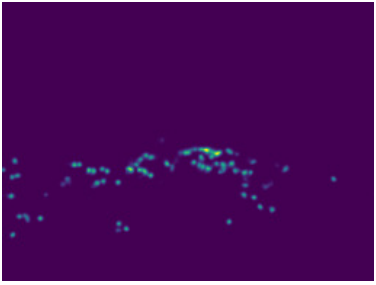
- Crowd counting methods typically focus on regressing counts of people in a scene, but as density estimation and count estimation can be used for arbitrary objects, we use those terms
- Density and count estimation typically outperform discrete counting-by-detection methods in busy scenes, or in scenes where occlusion is a problem

Figure 3: A busy scene from the VisDrone-DET dataset Zhu et al., 2022.

Crowd Counting as a General Counting Problem

Density Estimation

(a) The same VisDrone sample



(b) Output for the "people" class

Figure 3: Fig. 3a is the same VisDrone sample, and Fig. 3b shows our model's output for the people category - the integration of the $\frac{1}{4}$ scale density map produces the count of people in the image, which is 79. The same principle applies to fruit, flowers, or plants.

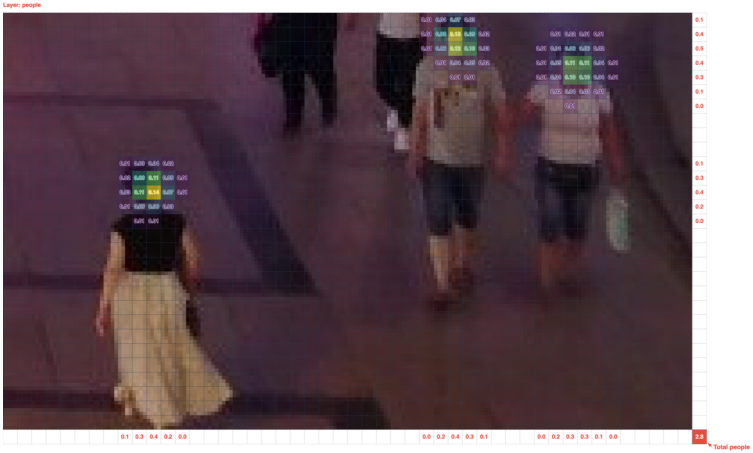
Density Estimation

Density Estimation

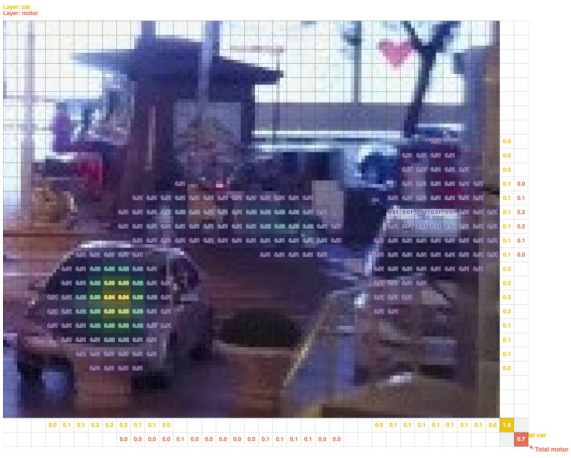
The estimate count for the class c, $\hat{n}_{i,c}$, is the integration of that class in the density map:

$$\hat{n}_{i,c} = \sum_{w=1}^{W_d} \sum_{h=1}^{H_d} D^{\text{est}}_{i,c,w,h}. \tag{1}$$

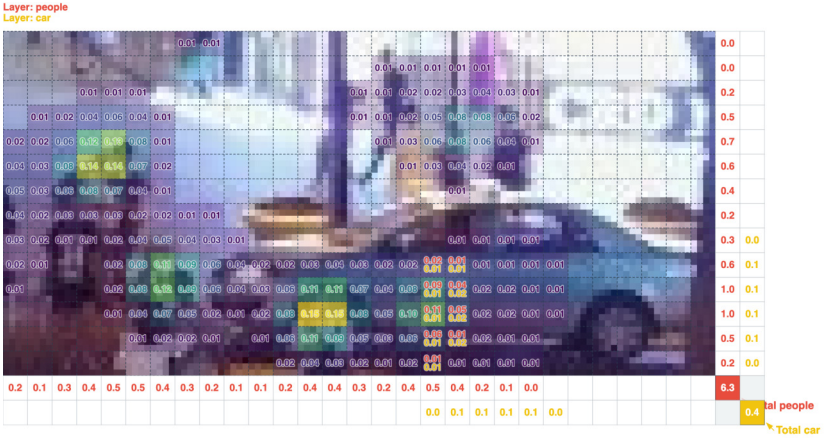
Visualising Density Integration



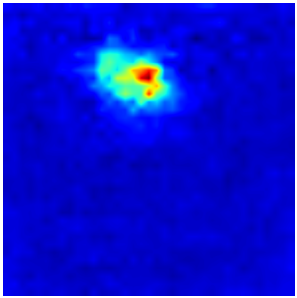
Visualising Density Integration



Visualising Density Integration

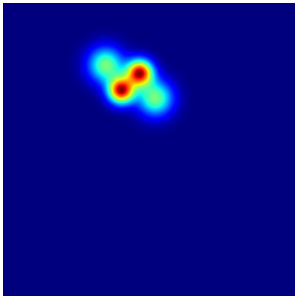


Novel Domains



(a) Model output

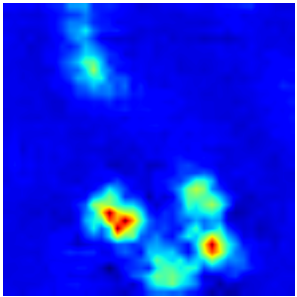
(b) Input RGB



(c) Ground truth

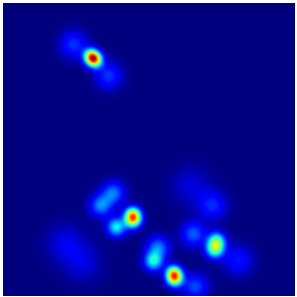
Figure 3: Counting almonds in Gomez et al., 2021 with the ancient CAN Cao et al., 2018

Novel Domains



(a) Model output

(b) Input RGB



(c) Ground truth

Figure 3: Counting almonds in Gomez et al., 2021 with the ancient CAN Cao et al., 2018

Multi-class Imbalance Problems

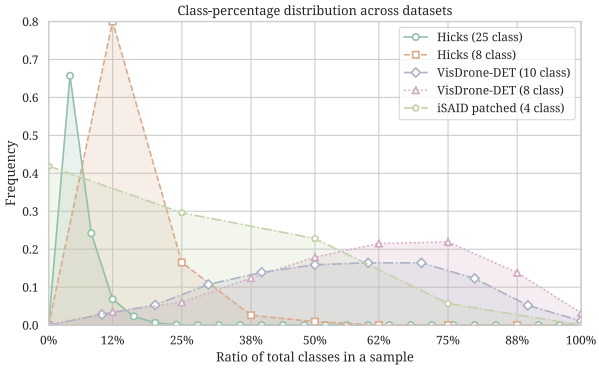


Figure 4: Class appearance density

- Density estimation can be applied to more than just people and cars, such as biodiversity monitoring with the Hicks et al., 2021 dataset
- Existing methods fail where few classes appear in a given sample

Multi-class Problem

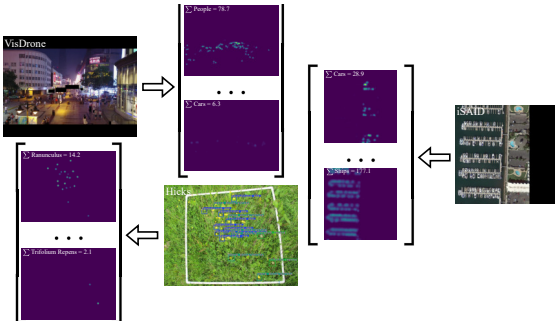
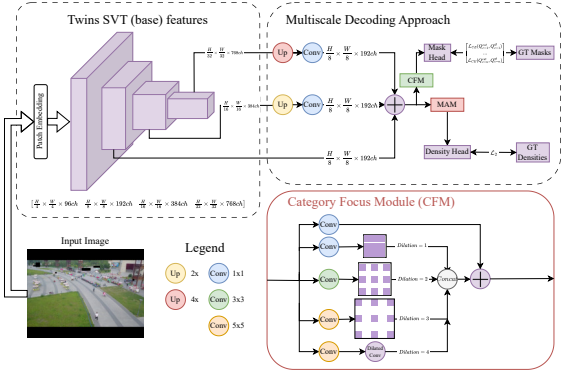


Figure 5: Multi-class outputs from our model trained on VisDrone-DET Zhu et al., 2022, iSAID Waqas Zamir et al., 2019 and the Hicks et al., 2021 datasets

- Density estimation is expressed as C unique counting problems to solve for multi-class
- Outputting C density estimations in one matrix brings inter-category interference

Our Approach



- The interference is addressed by solving auxiliary tasks - past work Xu et al., 2021 has used segmentation as we do or an auxiliary detection task in Michel et al., 2022.
- We learn from segmentation at training time, without needing it at inference time

Ablation Studies: Category Focus Module (CFM)

- **CFM Impact:** The module significantly improves performance, especially at smaller count ranges.
- **Scale Independence:** The multiscale CFM handles larger ranges of counts better than the single-scale variant.

VisDrone-DET (8-category) CFM Ablation

Count Range		YOLO 11x	No CFM	Single-scale	Ours (Base)
No. Params		56.97m	58.20m	60.95m	60.95m
0–1000	MAE	2.65	2.49	2.28	**2.29**$^{+8.0\%}$
	RMSE	10.06	**8.00**	8.16	**8.28**$^{-3.5\%}$
0–10	MAE	1.51	0.63	0.53	**0.46**$^{+27.0\%}$
	RMSE	2.44	1.04	0.99	**1.08**$^{-3.8\%}$
11–50	MAE	7.49	1.62	1.41	**1.39**$^{+14.2\%}$
	RMSE	10.46	3.10	**2.78**	**2.79**$^{+10.0\%}$
51–100	MAE	22.21	3.31	**3.06**	**3.09**$^{+6.6\%}$
	RMSE	27.29	6.59	6.32	**6.49**$^{+1.5\%}$
101–1000	MAE	80.63	8.55	**8.49**	**8.61**$^{-0.7\%}$
	RMSE	105.17	**24.64**	25.78	**26.12**$^{-6.0\%}$

Ablation Studies: Backbone Scale

- Tested across Twins-SVT-small, base, and large backbones.
- The **small** model (26.23m params) provides high-quality predictions.

VisDrone-DET (8-category) Backbone Ablation

Count Range		Small	Base	Large
No. Params		26.23m	60.95m	107.95m
0–1000	MAE	2.38	**2.29** $^{+3.8\%}$	**2.27**
	RMSE	8.06	**8.28** $^{-2.7\%}$	8.18
0–10	MAE	0.59	**0.46** $^{+22.0\%}$	**0.42**
	RMSE	0.98	**1.08** $^{-10.2\%}$	**0.83**
11–50	MAE	1.51	**1.39** $^{+7.9\%}$	**1.37**
	RMSE	2.80	**2.79** $^{+0.4\%}$	2.81
51–100	MAE	3.12	**3.09** $^{+1.0\%}$	3.09
	RMSE	**6.22**	**6.49** $^{-4.3\%}$	6.49
101–1000	MAE	8.63	**8.61** $^{+0.2\%}$	8.61
	RMSE	25.45	**26.12** $^{-2.6\%}$	25.73

Ablation Studies: Final Activation Function

- **ReLU**: Truncates negative predictions *before* the calculation & propagation of loss.
- **Leaky ReLU**: Unbounded, causing metrics to be impacted by negative count predictions.
- **Softplus**: Guarantees numerically stable, positive outputs without sacrificing gradients.

VisDrone-DET (8-category) Activation Ablation

Count Range		Softplus (Base)	ReLU	Leaky ReLU
0–1000	MAE	**2.29** $^{+60.7\%}$	5.83	5.67
	RMSE	**8.28** $^{+49.2\%}$	16.30	10.38
0–10	MAE	**0.46** $^{+45.9\%}$	0.85	3.97
	RMSE	**1.08** $^{+45.7\%}$	1.99	4.36
11–50	MAE	**1.39** $^{+61.6\%}$	3.62	4.79
	RMSE	**2.79** $^{+62.0\%}$	7.35	5.86
51–100	MAE	**3.09** $^{+63.2\%}$	8.40	6.39
	RMSE	**6.49** $^{+60.6\%}$	16.46	9.32
101–1000	MAE	**8.61** $^{+57.4\%}$	20.23	12.06
	RMSE	**26.12** $^{+43.6\%}$	46.31	28.69

Counting Results

Hicks et al. (8-category)

Count Range (samples in range)		DSACA	Ours
0-1000 (420)	MAE	0.92	**0.38** $^{+58.7\%}$
	RMSE	2.10	**1.71** $^{+18.6\%}$
0-5 (252)	MAE	0.61	**0.17** $^{+72.1\%}$
	RMSE	0.91	**0.82** $^{+9.9\%}$
6-10 (71)	MAE	0.93	**0.31** $^{+66.7\%}$
	RMSE	1.54	**0.97** $^{+37.0\%}$
11-25 (71)	MAE	1.39	**0.69** $^{+50.4\%}$
	RMSE	2.78	**2.01** $^{+27.7\%}$
26-1000 (26)	MAE	2.64	**1.76** $^{+33.3\%}$
	RMSE	5.96	**5.22** $^{+12.4\%}$

iSAID (4-category)

Count Range (samples in range)		YOLO 11x	Michel et al.	Ours
0-10000 (4056)	MAE	10.95	7.85	**2.84**
	RMSE	68.20	31.64	**29.12**
0-10 (2862)	MAE	1.67	2.5	**1.04**
	RMSE	13.49	**10.26**	16.27
11-50 (784)	MAE	6.84	6.51	**3.32**
	RMSE	19.61	12.18	28.07
51-100 (201)	MAE	20.15	17.86	**6.54**
	RMSE	36.05	26.38	**24.43**
101-10000 (209)	MAE	144.53	50.31	**22.11**
	RMSE	291.71	**95.57**	96.44

Counting Results

VisDrone-DET (8-category)

Count Range (samples in range)		YOLO 11l	YOLO 11x	Ours (small)	Ours (base)	Ours (large)
No. Params		25.37m	56.97m	26.23m	60.95m	107.95m
0-1000 (1610)	**MAE**	2.75	2.65	2.38	2.29	**2.27**
	RMSE	10.49	10.06	8.06	8.28	8.18
0-10 (126)	**MAE**	1.85	1.51	0.59	0.46	**0.42**
	RMSE	2.67	2.44	0.98	1.08	**0.83**
11-50 (980)	**MAE**	7.81	7.49	1.51	1.39	**1.37**
	RMSE	10.83	10.46	2.80	2.79	2.81
51-100 (377)	**MAE**	23.47	22.21	3.12	3.09	3.09
	RMSE	28.35	27.29	**6.22**	6.49	6.49
101-1000 (127)	**MAE**	84.94	80.63	8.63	8.61	8.61
	RMSE	109.41	105.17	**25.45**	26.12	25.73

Open-Vocabulary Counting

- The new direction is looking at open-vocabulary counting
- We want a multi-class vision-language counting model that accepts natural language prompts such as “the unripe and rotten strawberries” or “yellow and white flowers”
- We’re not quite there yet...

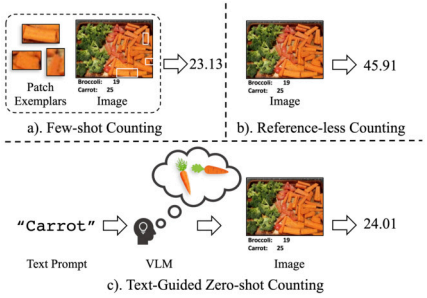
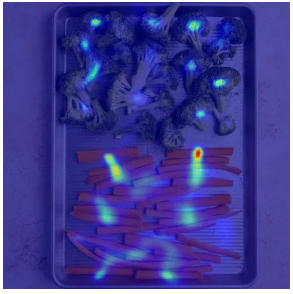


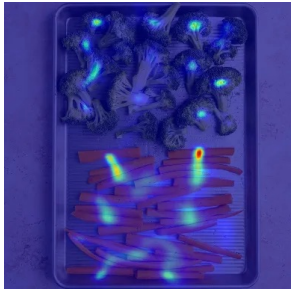
Figure 6: CLIP-Count

CLIP-Count Preliminary Results

- Results show CLIP-Count method fails to respond to text modality



(a) Prompt: “carrots”



(b) Prompt: “broccoli”

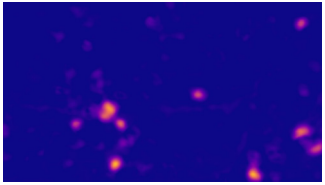
Figure 7: CLIP-Count not paying attention

CLIP-Count

(a) CLIP-Count found 15.3 “unripe strawberries”

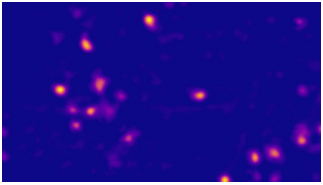
(b) CLIP-Count found 16.6 “ripe strawberries”

DSACA with Strawberries



(a) Ripe strawberries map

(b) Input RGB



(c) Unripe strawberries map

Figure 9: Qualitative results counting strawberries with DSACA Xu et al., 2021

Mind the Prompt: A Novel Benchmark for Prompt-Based Class-Agnostic Counting (WACV, Ciampi et al., 2025)

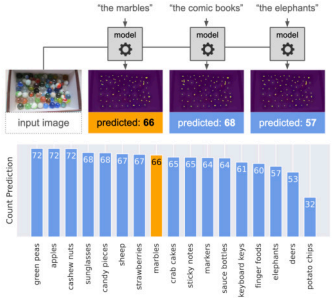


Figure 1. Prompt-based counting models – CounTX [2] in this example – exhibit difficulties in accurately interpreting user-provided texts that specify object classes to be counted. The confusion occurs even between classes that are semantically very distinct – like *marbles* and *elephants*. In some cases, the count of classes not present in the image is even higher than that for the ground-truth object category (highlighted in orange).

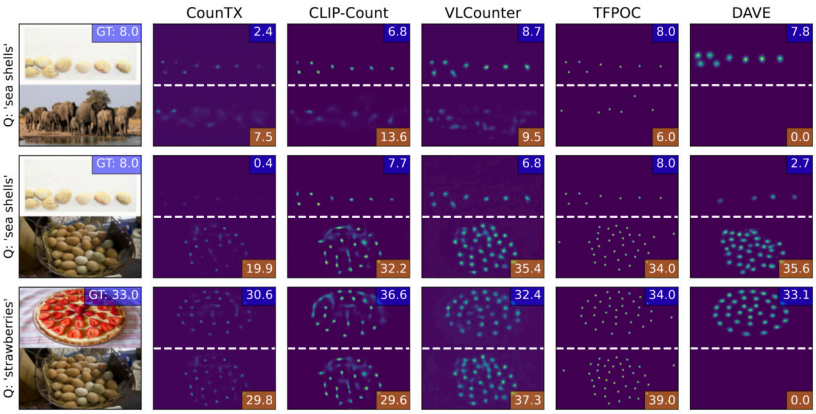


Figure 5. This figure shows, for each model, the output density maps for three different *(mosaic, input prompt)* pairs. The count reported in the blue box is c_{ij}^{pos}, while the count reported in the dark orange box corresponds to c_{ij}^{neg}. We can notice how the models often misidentify instances from the negative image in the mosaic, though most accurately estimate the positive instances in the upper part.

DAVE

Figure 10: DAVE failing epically to do what the 2018 model did with ease

- DAVE Pelhan et al., 2024 is praised by Ciampi et al., 2025 as the most class-aware model
- DAVE still falls short when dealing with new scales, so there is work to be done for open vocabulary counting

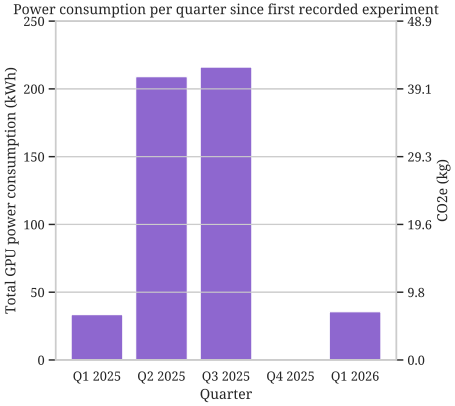
Try DAVE Yourself!

Figure 11: https://gtnr.vnet.tel/

- No data retention
- Try counting things with your phone, and let me know!

Environmental Impact

- 494.099 kWh recorded GPU energy usage $\approx$ 96.611 kg of CO2e emissions from generation and transmission. Conversion factors from UK Gov't Department for Energy Security and Net Zero, 2025
- Every experiment since 20th Jan 2025 tracks cumulative GPU energy usage



Conclusion

- Counting things is useful in agriculture for automated decision support, forecasting and precision farming
- Our multi-class density estimation method provides robust, class-aware counting capabilities
- Open-vocabulary density estimation is progressing, but class-awareness and zero-shot robustness is a challenge
- Thanks for listening

Figure 12: https://gtnr.vnet.tel/

References I

Cao, X., Wang, Z., Zhao, Y., & Su, F. (2018).Scale aggregation network for accurate and efficient crowd counting. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ciampi, L., Messina, N., Pierucci, M., Amato, G., Avvenuti, M., & Falchi, F. (2025).Mind the prompt: A novel benchmark for prompt-based class-agnostic counting. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7970–7979. https://doi.org/10.1109/WACV61041.2025.00774

Gomez, A. S., Aptoula, E., Parsons, S., & Bosilj, P. (2021).Deep regression versus detection for counting in robotic phenotyping. *IEEE Robotics and Automation Letters*, *6*(2), 2902–2907. https://doi.org/10.1109/LRA.2021.3062586

Hicks, D., Baude, M., Kratz, C., Ouvrard, P., & Stone, G. (2021).Deep learning object detection to estimate the nectar sugar mass of flowering vegetation. *Ecological Solutions and Evidence*, *2*(3), e12099. https://doi.org/10.1002/2688-8319.12099

Michel, A., Gross, W., Schenkel, F., & Middelmann, W. (2022).Class-aware object counting. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 469–478. https://doi.org/10.1109/WACVW54805.2022.00053

References II

Pelhan, J., Lukežic, A., Zavrtanik, V., & Kristan, M. (2024).Dave - a detect-and-verify paradigm for low-shot counting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23293–23302.

Shen, J., Xiong, X., Xue, Z., & Bian, Y. (2019).A convolutional neural-network-based pedestrian counting model for various crowded scenes. *Computer-Aided Civil and Infrastructure Engineering*, *34*(10), 897–914. https://doi.org/10.1111/mice.12454

UK Gov't Department for Energy Security and Net Zero. (2025). Greenhouse gas reporting: conversion factors 2025 [[Online; accessed 07-September-2025]]. https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2025

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., & Bai, X. (2019).Isaid: A large-scale dataset for instance segmentation in aerial images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 28–37. https://captain-whu.github.io/iSAID/

Xu, W., Liang, D., Zheng, Y., Xie, J., & Ma, Z. (2021).Dilated-scale-aware category-attention convnet for multi-class object counting. *IEEE Signal Processing Letters*, *28*, 1570–1574. https://doi.org/10.1109/LSP.2021.3096119

References III

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2022).Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7380–7399. https://doi.org/10.1109/TPAMI.2021.3119563