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

Villanelle O'Reilly*§, Jonathan Cox*, Georgios Leontidis†, Marc Hanheide*, Petra Bosilj‡, James M. Brown§
†Interdisciplinary Institute, University of Aberdeen, UK & UiT The Arctic University of Norway
‡Department of Advanced Computing Sciences, Maastricht University, The Netherlands
*L-CAS, University of Lincoln, UK; §AVAIL, University of Lincoln, UK
✉ {voreilly, jcox, mhanheide, jamesbrown}@lincoln.ac.uk ✉ georgios.leontidis@uit.no ✉
✉ petra.bosilj@maastrichtuniversity.nl ✉

***Abstract*—Density map estimation enables accurate object counting in heavily occluded, and densely packed scenes where detection-based counting fails. In multi-class density estimation, class awareness can be introduced by modelling classes non-exclusively, better reflecting crowded and visually ambiguous contexts. However, existing multi-class density estimators often degrade in less-dense scenes, while state-of-the-art detectors still struggle in the most congested settings. To bridge this gap, we propose the first vision-transformer-based approach to multi-class density estimation. Our model combines a Twins-SVT pyramid vision transformer backbone with a multiscale CNN decoder that leverages hierarchical features for robust counting across a wide range of densities. Further to that, the method adds an auxiliary segmentation task with the *Category Focus Module* to suppress inter-category interference at training time. The module improves the density estimation head without the need for constraining assumptions added by the application of the auxiliary task at inference time, as required in previous methods. Training and evaluation on the VisDrone and iSAID benchmarks demonstrates a leap in performance versus the previous state-of-the-art multi-class density estimation methods, attaining a 33%, 43%, and 64% reduction to MAE in testing evaluation. The method outperforms YOLO11 in less busy scenes, exceeding it by an order of magnitude in the most crowded testing samples.**

Code, containers, and trained weights available at https://github.com/LCAS/gnr_mcdest.

***Index Terms*—image recognition, object counting, multi-class density estimation, vision transformers**

I. Introduction

A fundamental computer vision task, object counting produces valuable automated insights for urban planning [1], [2], biodiversity monitoring [3], [4] and in medicine [5]. It is an expensive and tedious task for humans to perform at scale [6], and the high density of objects in occluded scenes pose a significant challenge for automated methods. Computer vision object counting methods include counting by detection [3], direct regression of object counts, without localisation information [2], and density estimation providing ''weak'' localisation, directly regressing a heatmap that integrates into the total counts of objects [7], [8].

Density estimation can be extended to a multi-class paradigm, where one density map is predicted for a fixed set of object classes, and is rarely studied [11]–[14]. While multi-class object detection has seen extensive research [15], the continuous density estimation model yields more accurate counts in crowded and occluded scenes than detection methods, as demonstrated in Gomez *et al.* [16]. *Crowd counting* methods, first applied to estimating the population of congested urban streets [7], can be directly applied for counting singular abstract classes, such as agricultural items (i.e. one of: almonds, apples, wheat kernels, etc.) in Gomez *et al.* [16]. More recent developments see density estimation used in an open-vocabulary textual prompting paradigm, but the state-of-the-art has been found to lack class-awareness in Ciampi *et al.* [17], and such methods lack the single-stage design of multi-class density estimation—as only one class or prompt can be isolated at a time.

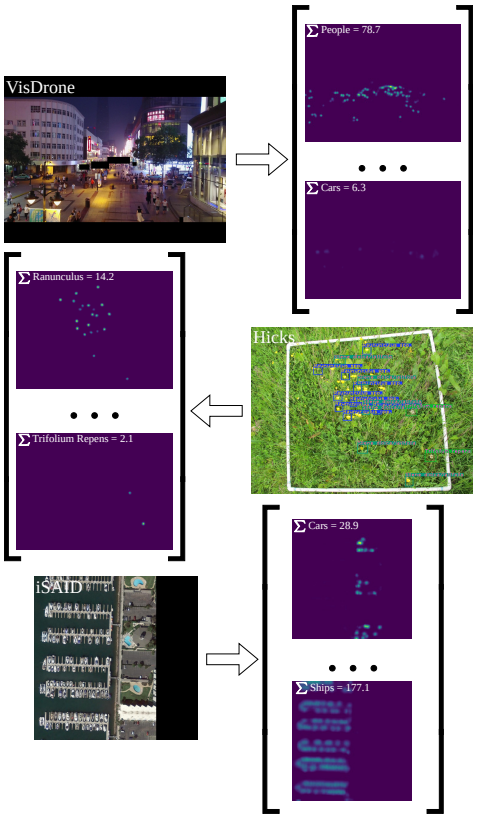

Fig. 1. **Multi-class Density Estimation**. The proposed model's inference on testing samples from the Hicks *et al.* [3], VisDrone-DET [9], and iSAID [10] datasets. The estimated count, from the integration of the heatmap Eq. (2a), is superimposed category-wise.

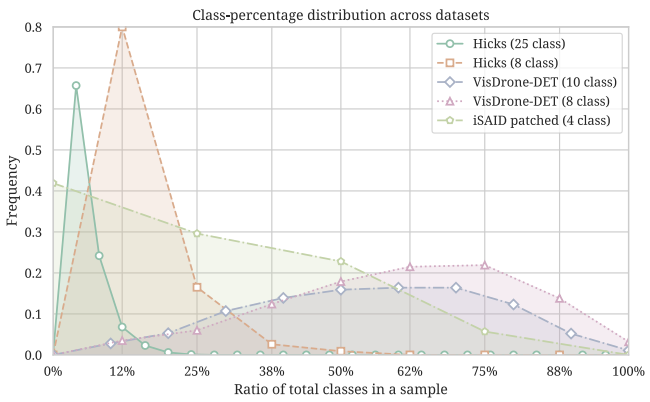

Fig. 2. **Class-Appearance Distribution**. The plot illustrates, across several datasets, the distribution of the proportion of classes present within a single sample. VisDrone-DET and iSAID images typically contain many classes, whereas biodiversity samples from Hicks generally contain no more than one class in a sample.

Multi-class methods simultaneously count several categories of objects in one pass, such as people, cars, trucks in DSACA [11] or additionally cars, trucks, ships and planes from satellite imagery in Michel *et al.* [12]. Previous multi-class density estimation methods [11]–[14] have only been evaluated on urban datasets including VisDrone-DET [9], iSAID [10] and RSOC [18], datasets containing either few classes, or a uniform class-frequency distribution. Visualised in Fig. 2, the biodiversity dataset Hicks *et al.* [3] is extremely skewed, compared to the more uniformly distributed VisDrone [9] and iSAID [10] datasets. It is common for most of the categories to appear in a single sample in the urban datasets, contrasting the rare occurrence of more than one category present in Hicks *et al.* [3], where some species of flowers may never naturally occur next to another. We find our vision-transformer-based method is able to successfully learn in this new domain, beating the previous open-source state-of-the-art DSACA [11], which we train with the dataset.

Contributions: **1)** The first vision-transformer-based multi-class density estimation architecture, outperforming the state of the art with an average 46.7% MAE decrease over the two benchmarks, and outperforming the object detection YOLO11 method by an order of magnitude in the most occupied scenes. We evidence the robustness of the method with the superior counting predictions over YOLO11 in less busy scenes that are historically favourable to object detection. **2)** Fusing a Convolutional Neural Network (CNN) decoding features from the Vision Transformer (ViT) Twins-SVT [19] backbone, the novel multiscale multi-class counting head is built around the smoothly-differentiable and numerically stable softplus activation, guaranteeing directly explainable density estimations modelling positive counts. **3)** The Category Focus Module, used in a lightweight auxiliary masking task, further increases class awareness, and directly leverages the density output at inference time. To the best of our knowledge, this is the first multi-class object counting approach to cast the task entirely as continuous density estimation at inference time, avoiding the discreteness assumptions inherent to detection-augmented methods [12], and the effectively thresholding object-presence-likelihood assumptions implicit in segmentation-augmented approaches [11].

II. Related Work

YOLO: YOLO11 is a recent YOLO architecture [20], [21], and the state-of-the-art in single-stage object detection. We use it as a benchmark for counting-by-detection performance.

DSACA: DSACA [11] was the first method reformulating crowd counting into a multi-class counting and density estimation problem. It extracts multiscale features from a VGG-16 [22] backbone with a Dilated Scale Aware Module (DSAM). They suppress inter-category interference with their Category Attention Module (CAM) masking out *low confidence* regions of the output density map at inference time. DSACA achieved state-of-the-art performance on the multi-class VisDrone-DET [9], and RSOC [18] datasets, benchmarking against various [7], [23]–[26] single-class density estimation methods.

Class-aware Object Counting: Michel *et al.* [12] employ a CNN multitask architecture with a ResNet50 [28] backbone. A region proposal network, and a detection head generate class-aware detections, alongside a parallel multi-class density-estimation branch that learns a separate density map for each category. Therefore, the key difference from DSACA [11] is the use of an object-detection-based auxiliary task rather than a segmentation-based one. The two branches are fused at inference time in a Count-Estimation Network (CEN), which predicts integer object counts by abstracting the continuous density maps, and suppressing inter-category interference by combining density-map and detection information. This design offers another way to reduce inter-category interference via multitask learning, combining a task suited to low-density scenes with a task suited to high-density scenes, and achieving reasonable counting performance across both count ranges. Michel *et al.* [12] evaluate their method against object-detection approaches [29]–[32] and report state-of-the-art results on those datasets.

CCTwins: Dong *et al.* [8] propose a weakly-supervised ViT and CNN based single-class crowd counting method that relies solely on count-level labels rather than dense location annotations. Building on the Twins-SVT backbone [19], it introduces an adaptive scene-consistency attention module to enhance feature extraction in highly uneven crowd distributions, and a multi-level weakly-supervised loss that progressively refines the density map from coarse to fine stages. The method uses a multiscale convolutional decoder repeatedly fusing multi-scale feature representations, supervising each layer against the global count. Rooted in their success concatenating and convolving multiple layers of the pyramid-ViT backbone, we take the Twins-SVT backbone for our multi-class method.

MRCNet: As a recent publication with state-of-the-art fully-supervised performance, we draw from Yu and Hu [27] as the most modern method for a density estimation multiscale feature extraction decoder. As a modern technique not using vision transformers, they introduce a dilated *Multiscale Aware*

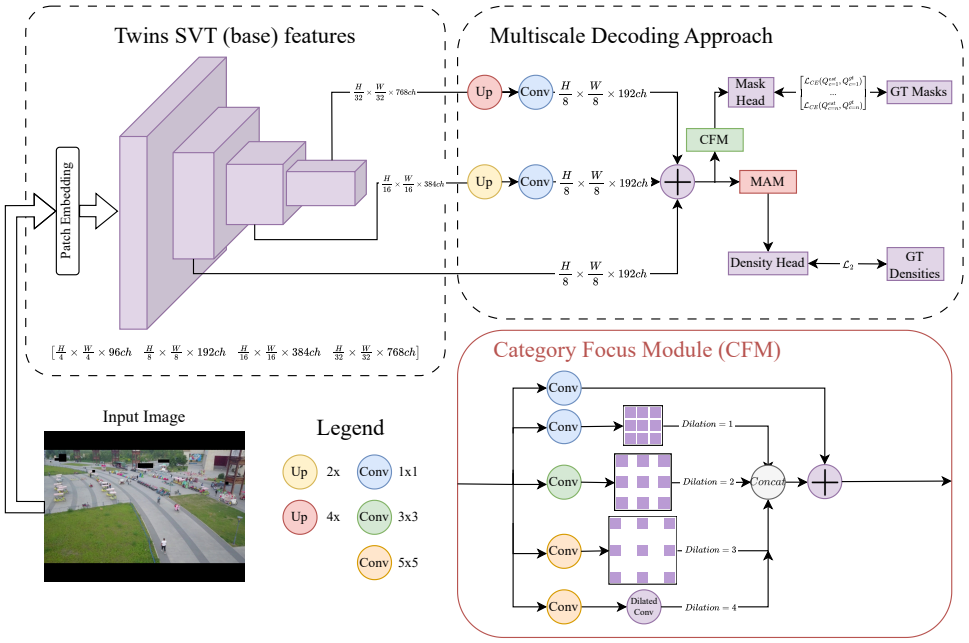

Fig. 3. **The Model Architecture**. In the Category Focus Module (CFM), an extension of the Multiscale Aware Module (MAM) from MRCNet [27], the first column of convolutions is followed by a column of batch norm and ReLU activations. The same batch norm and ReLU column is present in the MAM.

Module increasing the receptive field of their decoder. To increase our architecture's robustness to scale and density variance, we adopt this component of MRCNet.

For our density estimation head, we use a similar multi-scale convolutional decoding scheme, leveraging hierarchical features as with CCTwins [8] and MRCNet [27], for scale-variance-robust feature extraction capabilities from the pyramid ViT Twins-SVT backbone; introducing self-attention to multi-class density estimation. Similar to DSACA [11], we employ a segmentation-based secondary task to reduce inter-category interference. We extend their two-task design by introducing the *Category Focus Module* that extracts features from several layers of the backbone, learning at multiple scales with its *Multiscale Aware Module-style* design. By designing our lightweight masking head, we rely on the unified propagation of the segmentation and density losses as the sole benefit of the masking task, realised at training time, without the need for the masking head at inference. By not applying the assumptions and issues a segmentation [11] or object detection [12] task has in dense and occluded scenes, where the tasks could lead to exclusion of correct density contributions, ours is the first multi-class density estimation method that truly operates with a continuous model designed for such contexts.

III. Method

Following previous work [11], [12], we formulate multi-class density estimation as learning to predict a set of density maps per input image, where each map represents the spatial distribution of a single class in the image. Let D_i^{est} denote the predicted set of density maps estimated from the RGB input image $I_i \in \mathbb{R}^{H\times W\times 3}$. We assume a non-linear mapping $\mathcal{F}$, parameterised by θ, such that:

$$D_i^{\text{est}}(I_i) = \mathcal{F}(I_i;\theta) \tag{1}$$

For a dataset of N images $I = \{I_1, I_2, \ldots, I_N\}$, θ is learned using N ground-truth density maps, each containing C channels representing the set of classes in a dataset. The estimated and ground-truth maps share the same spatial dimensions, discretised at $\frac{1}{4^2}^{\text{th}}$ the resolution of I_i. Thus, the maps have the width $W_d = \frac{W}{4}$, and the height $H_d = \frac{H}{4}$, yielding $D_i^{\text{gt}}, D_i^{\text{est}} \in \mathbb{R}^{H_d\times W_d\times C}$.

We index the classes of D_i^{est} as $\{D_{i,1}^{\text{est}}, \ldots, D_{i,C}^{\text{est}}\}$. For example, $D_{i,1}^{\text{est}}$ might represent the *people* class for VisDrone, or the *planes* class for iSAID.

To determine the count of objects for the class c present in the input image I_i, the density estimation for all classes is predicted from the function $\mathcal{F}$ Eq. (1), from which the corresponding channel c in the density prediction is isolated as $D_{i,c}^{\text{est}}$. The estimated count of objects for the class c, defined as $\hat{n}_{i,c}$, is the integration of $D_{i,c}^{\text{est}}$, and the ground-truth count, $n_{i,c}$, is the integration of $D_{i,c}^{\text{gt}}$ such that:

$$\hat{n}_{i,c} = \sum_{w=1}^{W_d}\sum_{h=1}^{H_d} D_{i,c,w,h}^{\text{est}}, \tag{2a}$$

$$n_{i,c} = \sum_{w=1}^{W_d}\sum_{h=1}^{H_d} D_{i,c,w,h}^{\text{gt}}. \tag{2b}$$

Our method implements $\mathcal{F}$ using a Pyramid Vision Transformer (PVT) backbone coupled with a Convolutional Neural Network (CNN) decoder (Section III-A). The parameters θ

are optimised iteratively using the loss functions defined in Section III-B.

A. Backbone and Decoding Approach

Many density estimation methods [11], [13], [27], [33], use the CNN VGG16 [22] backbone, or the ResNet50 [12] backbone. A number of density and count-estimation methods using vision transformer [2], [33], or pyramid vision transformer backbones [8], [34] have emerged with greater single-class performance than methods using a CNN backbone. Therefore, we use the Twins-SVT [19] backbone, combined with a multiscale CNN-based decoder. By using a pyramid vision transformer with spatial attention, as opposed to a flat transformer encoder (e.g. [2]), we are still able to extract multi-scale features enabling the convolutional decoder to learn scale variances expected in problems suited to density estimation.

As we are interested in the mask's cross-entropy loss improving performance of the density head, we present a minimal masking head design in Fig. 4, so that the cross-entropy segmentation loss has a greater impact on the density head. Our *Category Focus Module* (CFM), combines the approaches of DSACA [11] and CCTrans [34], passing fused features from layers of the encoder through dilated convolutions to improve robustness to scale variation by increasing the receptive field. The masking head outputs $2C$ channels of logits, so that for each class there is a segmentation task not mutually exclusive of other classes, predicting a positive and background mask for each class. This corresponds to the assumption that any given pixel could be representative of more than one class. The *Multiscale Aware Module* similarly processes features from the backbone at different scales with dilated convolutions to increase the receptive field of the density head, to the end of scale-robust density predictions.

Fig. 4 shows the class-aware density head enabling the method to simultaneously predict several classes of density maps. The approach uses the smoothly differentiable softplus activation:

$$\text{softplus}(x) = \frac{\log(1 + \exp(\beta \times x))}{\beta}. \tag{3}$$

Softplus provides a bounded output, as with ReLU (used in [11], [12] etc.), constraining the predictions of the model to positive numbers. However, where ReLU would truncate negative predictions before the calculation of a loss, the softplus function guarantees a numerically stable positive output, applying our counting constraint without sacrificing gradients at training time.

B. Loss Functions

We use the $\mathcal{L}_2$ loss for optimising the density estimation task. This computes the class-pixel-wise squared error, therefore incorporating spatial and class-localisation error, as well as count error into the optimisation problem:

$$\mathcal{L}_2(D_i^{\text{gt}}, D_i^{\text{est}}) = \sum_{c=1}^{C} \sum_{h=1}^{H_d} \sum_{w=1}^{W_d} \left(D_{i,c,w,h}^{\text{gt}} - D_{i,c,w,h}^{\text{est}}\right)^2. \tag{4}$$

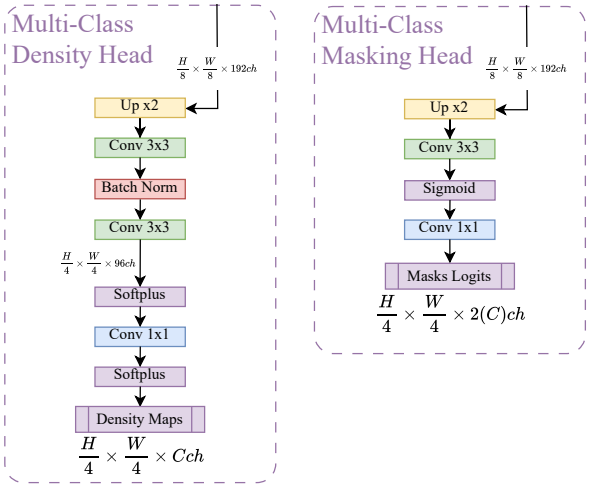

Fig. 4. **Model Heads**. The two output heads of the model.

For our auxiliary function, the masking task employs a per-category cross-entropy loss, where Q^{gt} and Q^{est} refer to the ground-truth and estimated mask data:

$$\mathcal{L}_{mask} = \frac{1}{C} \sum_{c=1}^{C} \mathcal{L}_{CE}(Q_c^{\text{est}}, Q_c^{\text{gt}}). \tag{5}$$

The losses are combined with the weighting factor ω. This enables the parameters θ, representing the lightweight masking task, the density head, and the parameters shared between the tasks to be optimised in one back-propagation operation, creating a dynamic relationship between the tasks in the shared parameters:

$$\mathcal{L} = (\omega \times \mathcal{L}_{mask}) + \mathcal{L}_2. \tag{6}$$

C. Ground-truth Generation

As our datasets provide bounding box annotations, we take centroid points of the bounding boxes and apply an empirically-derived series of Gaussian kernels to generate smooth density maps, via the same procedure as DSACA [11]. As the radius of the kernel can exceed the bounds of the image, which would fractionally subtract from the represented object count, the density maps are scaled class-wise to ensure that the integral of the density map produces the same integer count of objects as the bounding box data provides.

IV. Experiments

We assess our method's performance with the evaluation on the benchmark datasets VisDrone [9] and iSAID [10], [35]. We further apply our method to the biodiversity monitoring problem with data from Hicks *et al.* [3]. A sample for each of the datasets is visualised in Fig. 1, along with some classes of our model's predicted density maps.

A. Datasets

VisDrone-DET [9]: Collected over 14 different Chinese cities, the VisDrone *Object Detection in Images* dataset contains 10,209 images of varying density, with 10 annotated classes including 8 types of vehicle, and people. Michel *et al.* [12] evaluate against the full 10-class VisDrone-DET dataset.

VisDrone-DET (8-category)										
Count Range (samples in range)		**YOLO 11l [20]**	**YOLO 11x [20]**	**Ours (no CFM)**	**Ours (single-scale CFM)**	**Ours (CFM, Twins-SVT-small)**	**Ours (CFM, Twins-SVT-base)**	**Ours (CFM, Twins-SVT-large)**	**Ours (CFM, ReLU)**	**Ours (CFM, leaky ReLU)**
No. Params		25.37m	56.97m	58.20m	60.95m	26.23m	60.95m	107.95m	60.95m	60.95m
0–1000 (1610)	**MAE**	2.75	2.65	2.49	2.28	2.38	2.29	**2.27**	5.83	5.67
	RMSE	10.49	10.06	**8.00**	8.16	8.06	8.28	8.18	16.30	10.38
0–10 (126)	**MAE**	1.85	1.51	0.63	0.53	0.59	0.46	**0.42**	0.85	3.97
	RMSE	2.67	2.44	1.04	0.99	0.98	1.08	**0.83**	1.99	4.36
11–50 (980)	**MAE**	7.81	7.49	1.62	1.41	1.51	1.39	**1.37**	3.62	4.79
	RMSE	10.83	10.46	3.10	**2.78**	2.80	2.79	2.81	7.35	5.86
51–100 (377)	**MAE**	23.47	22.21	3.31	**3.06**	3.12	3.09	3.09	8.40	6.39
	RMSE	28.35	27.29	6.59	6.32	**6.22**	6.49	6.49	16.46	9.32
101–1000 (127)	**MAE**	84.94	80.63	8.55	**8.49**	8.63	8.61	8.61	20.23	12.06
	RMSE	109.41	105.17	**24.64**	25.78	25.45	26.12	25.73	46.31	28.69

TABLE I. Test-set counting results on the merged 8-category VisDrone-DET [9] dataset, ablating the backbone, Category Focus Module (CFM), and softplus activation. Best and second-best results per row are in **bold** and underlined, respectively (ranked using unrounded metrics). For comparison, the VGG-16 backbone used in DSACA [11], and by Michel *et al.* [12] has 138m parameters, making those baselines larger than any of our variations. Unless noted otherwise, models use the Twins-SVT-base backbone (60.95m parameters) and softplus activation. Metrics are grouped by total per-image object counts: '0–1000' represents the full VisDrone-DET test set, while '11–50' includes samples where the total ground-truth count $n_{i,c}$ is $11 \leq \sum_{c=1}^{C} n_{i,c} \leq 50$.

As DSACA [11] merge the *people* and *pedestrian* classes into a unified people class, and the *tricycle* and *awning-tricycle* classes into a unified tricycle class, their VisDrone-DET sub-dataset has 8, instead of the full 10 classes used in Michel *et al.* [12]. Therefore, we evaluate our model against both the 8-class and 10-class VisDrone-DET datasets. Before dataset ground-truth density generation, we resize the images to be 1024 pixels wide, proportionally scaling the height.

iSAID (4-class) [10]: As described in Michel *et al.* [12], patch the iSAID dataset into 800×800 pixel tiles. In the absence of a 4-class or crowd-counting-compatible test challenge, we combine, shuffle, and split the public validation and training datasets 70–10-20 to training-validation-testing. We split the dataset before patching, so that no single image can be patched across the subsets, which could inflate the testing results. Without a provided 4-class iSAID test set, there will be a margin for error comparing our method to Michel *et al.* [12], although it's expected to be minimal because of the large size of the dataset at 20,906 patches, and as we copy their dataset preparation procedure. The YOLO11 model is trained and evaluated on the same split of the dataset we produce for density estimation, before the generation of the ground-truth density maps.

Hicks* et al. *[3] (8-class): A biodiversity monitoring dataset, the Hicks dataset of annotates 25,352 tags between 25 species of flowers in natural environments. We reduce the biodiversity monitoring problem to the 8 most common species of flowers. The images in the dataset are resized to have a maximum width of 1024 pixels, and where scaled to reduce the width the height is scaled proportionally. The validation and training sets are combined and split 70–10-20 to training-validation-testing, as the 50-image testing set of the dataset does not contain annotations.

B. Evaluation Metrics

As in DSACA [11] and Michel *et al.* [12], we compare methods on the macro-MAE and macro-RMSE metrics, based on testing data. Best model weights are chosen from macro-MAE score over the validation data:

$$\text{macro-MAE}_i = \frac{1}{C} \sum_{c=1}^{C} |\hat{n}_{i,c} - n_{i,c}|, \tag{7a}$$

VisDrone-DET (10-category)			
Count Range (samples in range)		**Michel *et al.* [12]**	**Ours**
0–1000 (1610)	**MAE**	3.76	**1.99**
	RMSE	9.56	**6.41**
0–10 (126)	**MAE**	2.07	**0.52**
	RMSE	3.25	**0.96**
11–50 (980)	**MAE**	10.49	**1.26**
	RMSE	12.52	**2.52**
51–100 (377)	**MAE**	13.84	**2.68**
	RMSE	25.07	**5.48**
101–1000 (127)	**MAE**	23.55	**7.03**
	RMSE	51.11	**19.54**

TABLE II. Counting (testing) results on the full 10-category VisDrone-DET[9] dataset. The best result on each row is marked in **bold**.

VisDrone-DET (8-category)					
Category		**DOPNet [36]**	**DSACA [11]**	**YOLO 11x [20]**	**Ours**
All	**MAE**	3.48	3.43	2.65	**2.29**
	RMSE	5.60	**5.36**	10.06	8.28
People	**MAE**	8.63	**5.04**	10.00	7.51
	RMSE	13.05	**7.65**	26.29	21.56
Bicycle	**MAE**	2.34	2.35	0.72	**0.70**
	RMSE	4.34	4.33	2.12	**1.94**
Motor	**MAE**	4.48	8.90	**2.18**	2.19
	RMSE	6.55	12.23	5.45	**4.96**
Tricycle	**MAE**	2.54	2.88	**0.43**	0.51
	RMSE	4.21	4.61	1.27	**1.17**
Car	**MAE**	5.49	3.98	3.77	**3.07**
	RMSE	8.65	6.02	7.53	**5.37**
Van	**MAE**	2.57	2.54	2.28	**2.12**
	RMSE	4.57	4.51	4.04	**3.72**
Truck	**MAE**	1.36	1.32	**0.98**	1.18
	RMSE	2.25	2.59	**2.11**	2.42
Bus	**MAE**	0.45	**0.42**	0.80	1.01
	RMSE	1.14	**0.97**	2.03	2.21

TABLE III. Counting (testing) results on the merged 8-category VisDrone-DET [9] dataset. A class-wise breakdown of performance is provided for each model, as DSACA [11] does not provide range-wise breakdowns. The best results in each row are in **bold**, with the second-best results underlined. The best metric is determined before rounding.

$$\text{macro-RMSE}_i = \sqrt{\frac{1}{C}\sum_{i=1}^{c}(\hat{n}_{i,c} - n_{i,c})^2}. \tag{7b}$$

Assuming C is the number of categories in a dataset, where the estimate count $\hat{n}_c$ is defined in Eq. (2a), and the ground-truth count n_c in Eq. (2b). Whilst DSACA [11] report a ''Mean Squared Error (MSE)'' metric, their definition of MSE is equivalent to macro-RMSE, and we refer to their metrics as such. In our tables we refer to the mean macro-MAE and macro-RMSE metrics over the whole testing set, as in previous work. With a testing set of N_{testing} samples, these can be defined as:

$$\text{MAE} = \frac{1}{N_{\text{testing}}}\sum_{i=1}^{N_{\text{testing}}} \text{macro-MAE}_i, \tag{8a}$$

$$\text{RMSE} = \frac{1}{N_{\text{testing}}}\sum_{i=1}^{N_{\text{testing}}} \text{macro-RMSE}_i. \tag{8b}$$

V. Results and Discussion

We achieve a significant improvement in MAE across the 8 and 10 category VisDrone benchmarks (Tables II and III), and an even greater jump on the iSAID benchmark Table IV, attaining state-of-the-art RMSE metrics. To address the large gap in multi-class density estimation research, as DSACA [11] was published in 2021, and Michel *et al.* [12] published their results in 2022, and to address other advances in object counting since then, we compare our results to the centroid prediction method DOPNet [36], and against the largest size of YOLO11. Tables I and IV demonstrate that YOLO11 performs poorly for extremely dense scenes (i.e. scenes with counts ranging 50–10000), and moderately in the ranges 0–10 and 10–50, evidencing the hypothesis that density estimation methods are better suited to dense counting applications than object-detection-based methods.

A. Ablation Studies

In order to verify the significance of individual contributions, we ablate the CFM, the size of the backbone, and final-layer activation function in Table I. Table I demonstrates that the CFM improves overall model performance, especially at smaller count ranges. The effect of propagating the cross-entropy loss through the CFM is seen when we directly take the final layer of the backbone as the input to the CFM, in the single scale experiment. We see a change in the metrics compared to the full CFM, even when in all cases the output of the mask is discarded, and has no direct effect on the metrics at inference time. Whilst the *0–1000* results of the single-scale experiment are very marginally improved over the base model, the model loses its scale independence, only predicting larger ranges of counts better than the base model with the multiscale CFM, and dropping in performance when predicting lower ranges of counts.

The final-layer activation function is found to have a significant impact on the performance. Selecting leaky ReLU as a philosophical middle ground between the ReLU and softplus activations, we notice a sharp drop in low count ranges with the leaky variant. As the leaky ReLU is unbounded, we notice metrics being impacted from negative count predictions.

iSAID (4-category)				
Count Range (samples in range)		**YOLO 11x [20]**	**Michel *et al.* [20]**	**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 [*sic*]	**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

TABLE IV. Counting (testing) results on the reduced 4-category iSAID dataset [10], as described in Michel *et al.* [12]. The best result on each row is marked in **bold**.

B. Environmental Impact

This work required 2,439.00 GPU-hours (mix of NVIDIA V100, A100, RTX A6000, and RTX 3070), largely during

methodological exploration, and mostly on the University of Lincoln (*Novel*), and University of Aberdeen (*Maxwell*) HPC clusters. These computations consumed 494.095 kWh of GPU energy, generating approximately 96.6 kg of CO_2e emissions in England and Scotland [37].

Hicks *et al.* (8-category)			
Count Range (samples in range)		**DSACA [11]**	**Ours**
0–1000 (420)	**MAE**	0.92	**0.38**
	RMSE	2.10	**1.71**
0–5 (252)	**MAE**	0.61	**0.17**
	RMSE	0.91	**0.82**
6–10 (71)	**MAE**	0.93	**0.31**
	RMSE	1.54	**0.97**
11–25 (71)	**MAE**	1.39	**0.69**
	RMSE	2.78	**2.01**
26–1000 (26)	**MAE**	2.64	**1.76**
	RMSE	5.96	**5.22**

TABLE V. Counting (testing) results on the reduced 8-category Hicks *et al.* [3] dataset. The best result on each row is marked in **bold**.

VI. Conclusion

In this work, we introduce a novel framework for multi-class density estimation, addressing the challenges of object counting in dense, occluded, and heterogeneous environments. By leveraging the Twins-SVT vision transformer backbone and a multiscale convolutional decoder, our method effectively captures both global context and fine-grained spatial details. The proposed Category Focus Module significantly reduces inter-class interference without the need for the application of secondary tasks at inference time, required in previous methods. Extensive experiments on VisDrone, iSAID, and the biodiversity-focused Hicks dataset demonstrate that the method consistently outperforms existing multi-class counting methods and object detection baselines, achieving up to 64% decrease in MAE. Furthermore, our method's applicability to ecological monitoring tasks highlights its potential for real-world applications beyond traditional urban scenarios. We hope this work encourages further exploration into scalable, class-aware density estimation methods and fosters cross-domain innovation in automated counting systems.

Acknowledgment

This work was supported by the UKRI AI Centre for Doctoral Training in Sustainable Understandable agri-food Systems Transformed by Artificial INtelligence (SUSTAIN) [grant reference: EP/Y03063X/1].

References

[1] J. Shen, X. Xiong, Z. Xue, and Y. Bian, ''A convolutional neural-network-based pedestrian counting model for various crowded scenes,'' *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 10, pp. 897–914, 2019. DOI: 10.1111/mice.12454.

[2] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, ''Transcrowd: Weakly-supervised crowd counting with transformers,'' *Science China Information Sciences*, vol. 65, no. 6, p. 160 104, Apr. 2022, ISSN: 1869-1919. DOI: 10.1007/s11432-021-3445-y.

[3] D. Hicks, M. Baude, C. Kratz, P. Ouvrard, and G. Stone, ''Deep learning object detection to estimate the nectar sugar mass of flowering vegetation,'' *Ecological Solutions and Evidence*, vol. 2, no. 3, e12099, 2021. DOI: 10.1002/2688-8319.12099.

[4] N. E. Ocer, G. Kaplan, F. Erdem, D. Kucuk Matci, and U. Avdan, ''Tree extraction from multi-scale uav images using mask r-cnn with fpn,'' *Remote sensing letters*, vol. 11, no. 9, pp. 847–856, 2020. DOI: 10.1080/2150704X.2020.1784491.

[5] F. Lavitt, D. J. Rijlaarsdam, D. van der Linden, E. Weglarz-Tomczak, and J. M. Tomczak, ''Deep learning and transfer learning for automatic cell counting in microscope images of human cancer cell lines,'' *Applied Sciences*, vol. 11, no. 11, 2021, ISSN: 2076-3417. DOI: 10.3390/app11114912.

[6] T. D. Breeze *et al.*, ''Pollinator monitoring more than pays for itself,'' *Journal of Applied Ecology*, vol. 58, no. 1, pp. 44–57, 2021. DOI: 10.1111/1365-2664.13755.

[7] W. Liu, M. Salzmann, and P. Fua, ''Context-aware crowd counting,'' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[8] L. Dong, H. Zhang, D. Zhou, J. Shi, and J. Ma, ''Cctwins: A weakly supervised transformer-based crowd counting method with adaptive scene consistency attention,'' *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 22–35, 2024. DOI: 10.1109/TCE.2023.3274651.

[9] P. Zhu *et al.*, ''Detection and tracking meet drones challenge,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2022. DOI: 10.1109/TPAMI.2021.3119563.

[10] S. Waqas Zamir *et al.*, ''Isaid: A large-scale dataset for instance segmentation in aerial images,'' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 28–37. [Online]. Available: https://captain-whu.github.io/iSAID/.

[11] W. Xu, D. Liang, Y. Zheng, J. Xie, and Z. Ma, ''Dilated-scale-aware category-attention convnet for multi-class object counting,'' *IEEE Signal Processing Letters*, vol. 28, pp. 1570–1574, 2021. DOI: 10.1109/LSP.2021.3096119.

[12] A. Michel, W. Gross, F. Schenkel, and W. Middelmann, ''Class-aware object counting,'' in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, Jan. 2022, pp. 469–478. DOI: 10.1109/WACVW54805.2022.00053.

[13] Q. Fu, W. Min, W. Sheng, and C. Peng, ''Counting dense object of multiple types based on feature enhance-

ment,'' *Frontiers in Neurorobotics*, vol. 18, p. 1 383 943, 2024. DOI: 10.3389/fnbot.2024.1383943.

[14] J. Gao, L. Zhao, and X. Li, ''Nwpu-moc: A benchmark for fine-grained multicategory object counting in aerial images,'' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024. DOI: 10.1109/TGRS.2024.3356492.

[15] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, ''Object detection in 20 years: A survey,'' *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023. DOI: 10.1109/JPROC.2023.3238524.

[16] A. S. Gomez, E. Aptoula, S. Parsons, and P. Bosilj, ''Deep regression versus detection for counting in robotic phenotyping,'' *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2902–2907, 2021. DOI: 10.1109/LRA.2021.3062586.

[17] L. Ciampi, N. Messina, M. Pierucci, G. Amato, M. Avvenuti, and F. Falchi, ''Mind the prompt: A novel benchmark for prompt-based class-agnostic counting,'' in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 7970–7979. DOI: 10.1109/WACV61041.2025.00774.

[18] G. Gao, Q. Liu, and Y. Wang, ''Counting from sky: A large-scale data set for remote sensing object counting and a benchmark method,'' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 3642–3655, 2021. DOI: 10.1109/TGRS.2020.3020555.

[19] X. Chu *et al.*, ''Twins: Revisiting the design of spatial attention in vision transformers,'' *Advances in neural information processing systems*, vol. 34, pp. 9355–9366, 2021.

[20] G. Jocher and J. Qiu, *Ultralytics yolo11*, version 11.0.0, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics.

[21] R. Khanam and M. Hussain, *Yolov11: An overview of the key architectural enhancements*, 2024. arXiv: 2410.17725 [cs.CV].

[22] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. DOI: 10.48550/arXiv.1409.1556.

[23] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, ''Single-image crowd counting via multi-column convolutional neural network,'' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

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

[25] Y. Li, X. Zhang, and D. Chen, ''Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes,'' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[26] Z. Ma, X. Wei, X. Hong, and Y. Gong, ''Bayesian loss for crowd count estimation with point supervision,'' in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[27] J. Yu and H. Hu, ''Multiscale regional calibration network for crowd counting,'' *Scientific Reports*, vol. 15, no. 1, p. 2866, 2025. DOI: 10.1038/s41598-025-86247-w.

[28] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[29] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster r-cnn: Towards real-time object detection with region proposal networks,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2577031.

[30] S. Qiao, L.-C. Chen, and A. Yuille, ''Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution,'' in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 208–10 219. DOI: 10.1109/CVPR46437.2021.01008.

[31] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, ''Centernet: Keypoint triplets for object detection,'' in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[32] K. He, G. Gkioxari, P. Dollar, and R. Girshick, ''Mask r-cnn,'' in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[33] H. Lin *et al.*, ''Semi-supervised counting via pixel-by-pixel density distribution modeling,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 5, pp. 3625–3638, 2025. DOI: 10.1109/TPAMI.2025.3532512.

[34] Y. Tian, X. Chu, and H. Wang, *Cctrans: Simplifying and improving crowd counting with transformer*, 2021. arXiv: 2109.14483 [cs.CV].

[35] G.-S. Xia *et al.*, ''Dota: A large-scale dataset for object detection in aerial images,'' in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018. DOI: 10.21227/wwrj-3d46.

[36] M. Cui, G. Ding, D. Yang, and Z. Chen, ''Dopnet: Dense object prediction network for multiclass object counting and localization in remote sensing images,'' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024. DOI: 10.1109/TGRS.2024.3349702.

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