Interim Report: Object Counting using deep learning.

Submitted December 10, 2024, in partial fulfillment of
the conditions for the award of the degree **BSc Hons Computer Science with Artificial Intelligence.**

Yuzhe Wu
20411994

Supervised by Jianfeng Ren

School of Computer Science University of Nottingham Ningbo China

Abstract

Exemplar-Free Counting aims to count objects of interest without intensive annotations of objects or exemplars. To achieve this, we propose Gated Context-Aware Swin-UNet (GCA-SUN) to directly map an input image to the density map of countable objects. Specifically, a Gated Context-Aware Modulation module is designed in the encoder to suppress irrelevant objects or background through a gate mechanism and exploit the attentive support of objects of interest through a self-similarity matrix. The gate strategy is also incorporated into the bottleneck network and the decoder to highlight the features most relevant to objects of interest. By explicitly exploiting the attentive support among countable objects and eliminating irrelevant features through the gate mechanisms, the proposed GCA-SUN focuses on and counts objects of interest without relying on predefined categories or exemplars. Experimental results on the FSC-147 and CARPK datasets demonstrate that GCA-SUN outperforms state-of-the-art methods.

Contents

Abstract **1**

List of Figures **3**

List of Tables **4**

List of Abbreviations **5**

1 Introduction **6**

2 Background & Related Work **8**

3 Design & Specifications **9**
3.1 Overview of Proposed Method 9
3.2 Swin-T Encoder with GCAM 10
3.3 Bottleneck with GEFS 11
3.4 Swin-T Decoder with GAFU 11

4 Implementation **12**
4.1 Datasets 12
4.2 Experiment Setting 12

5 Evaluation **13**
5.1 Comparison with State-of-the-Art Methods 13
5.2 Cross-Domain Evaluation on CARPK Dataset 14
5.3 Visualization of GCAM 14
5.4 Ablation Study 15

6 Summary and Reflections **16**
6.1 Project management 16
6.1.1 Overview of the Project Plan 16
6.1.2 Comparison of Planned and Actual Progress 16
6.1.3 Reflections on the Project Plan 17
6.1.4 Revised Timeline 18
6.2 Conclusion and future work 18

References **20**

List of Figures

1.1 Examples of Object Counting. 7

3.1 Overview of proposed GCA-SUN. It consists of an encoder, bottleneck, and a decoder. The encoder consists of a set of GCAM blocks to highlight the features relevant to countable objects while suppressing others, and Swin transformer to extract features. The GEFS in the bottleneck and the GAFU in the decoder also enhance the features of objects of interest. Finally, a regression head generates a density map for estimating the number of objects. 9

5.1 Visual comparisons to CounTR [1] on the FSC-147 dataset. 14
5.2 Visualization of the effects of GCAM. 15

6.1 Original Plan. 17
6.2 Revised Plan. 18

List of Tables

5.1 Comparison with other methods on the FSC-147 dataset [2], with best results highlighted in bold. 13
5.2 Comparison with other methods on the CARPK dataset [3]. 14
5.3 Ablation study of each component on the FSC-147 dataset [2]. 15

List of Abbreviations

CAC Class Agnostic Counting

EFC Exemplar Free Counting

FSC Few Shot Counting

OC Object Counting

ZSC Zero Shot Counting

Chapter 1

Introduction

Object counting determines the number of instances of a specific object class in an image [2], *e.g.*, vehicles [4], crowd [5], and cells [6]. As 1.1 shown, one of the datasets used for Object Counting tasks, *i.e.*, FSC-147, contains various categories and scenarios from real-world. It can be broadly categorized as: 1) Class-Specific Counting (CSC), counting specific categories like crowd [7], fruits [8] and animals [9]; 2) Class-Agnostic Counting (CAC), counting objects based on visual exemplars [1,2,10] or text prompts [11,12]; 3) Exemplar-Free Counting (EFC), counting objects without exemplars, presenting a significant challenge in discerning countable objects and determining their repetitions [1, 13, 14].

Exemplar-Free Counting shows promise for automated systems such as wildlife monitoring [15], healthcare [16], and anomaly detection [17]. Hobley and Prisacariu directly regressed the image-level features learned by attention modules into a density map [14]. CounTR [1] and LOCA [18] are originally designed for CAC tasks, but can be adapted to EFC tasks by using trainable components to simulate exemplars. RepRPN-Counter identifies exemplars from region proposals by majority voting [13], and DAVE selects valuable objects using a strategy similar to majority voting based on [19].

Despite the advancements, existing models [1,18,19] often explicitly require exemplars to count similar objects. EFC methods such as RepRPN-Counter do not require exemplars but generate them through region proposal [13]. Either explicit or implicit exemplars may induce sample bias as exemplars can't cover the sample distribution. To address the challenge, we propose Gated Context-Aware Swin-UNet (GCA-SUN), which directly maps an input image to the density map of countable objects, without any exemplars. Specifically, the encoder consists of a set of Swin Transformers to extract features, and Gated Context-Aware Modulation (GCAM) blocks to exploit the attentive supports of countable objects. The bottleneck network includes a Gated Enhanced Feature Selector (GEFS) to emphasize the encoded features that are relevant to countable objects. The decoder includes a set of Swin transformers for generating the density map, with the help of Gated Adaptive Fusion Units (GAFUs) to selectively weigh features based on their relevance to countable objects. Finally, a regression head is utilized to derive the density map from the aggregated features.

One key challenge in EFC is to effectively differentiate countable objects from other objects. The GCAM blocks tackle the challenge by first evaluating feature qualities by computing the feature score for each token, and then prioritizing those with informative content. In addition, GCAM computes pairwise similarities between tokens through a self-similarity matrix, exploiting the support of repeating objects in the same scene. Lastly, a gate mechanism is incorporated to highlight the most relevant features while suppressing

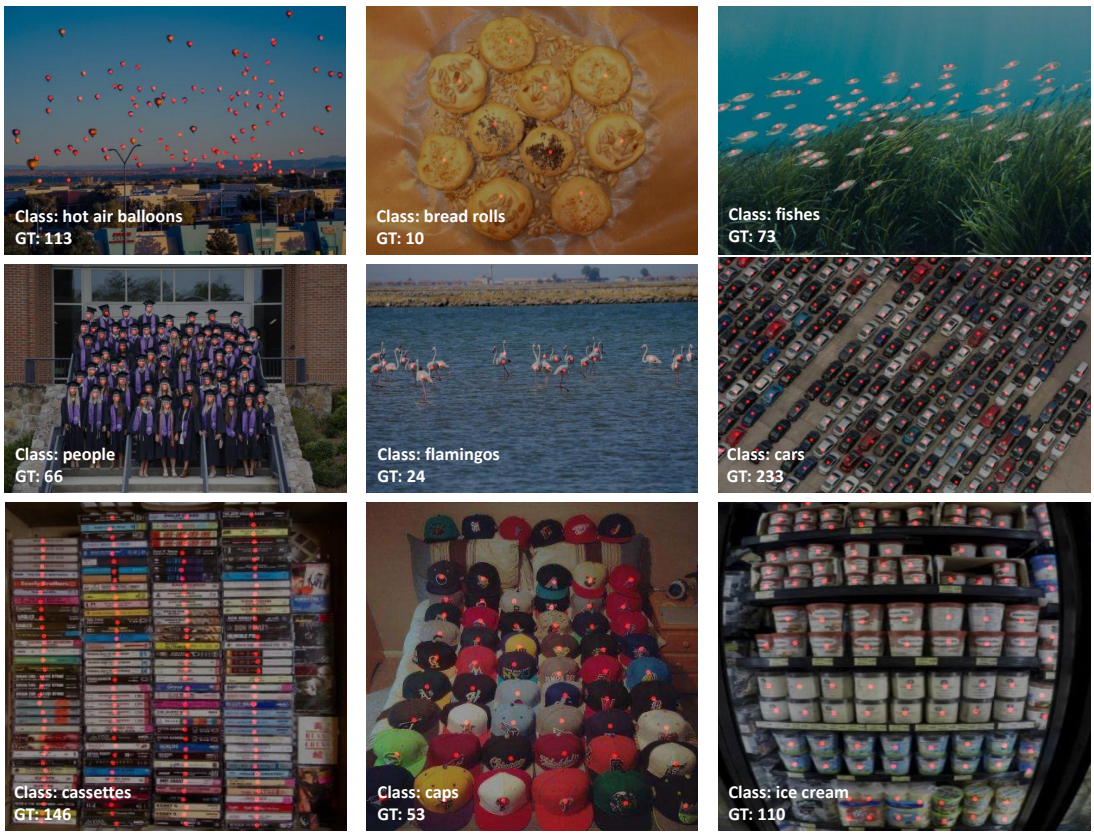

Figure 1.1: Examples of Object Counting.

irrelevant ones.

Another challenge is that foreground objects often share similar low-level features with background content. The skip connections directly fuse low-level features in the encoder with high-level semantics in the decoder, potentially impeding counting performance as the background information could disturb the foreground objects. To tackle this issue, gate mechanisms are incorporated into both GEFS and GAFU to suppress irrelevant low-level features while preserving as much information on objects of interest as possible. The former selectively enhances the compressed features at the bottleneck, and the latter filters the features in the decoder.

Our contributions can be summarized as follows. 1) The proposed GCA-SUN achieves exemplar-free counting through a UNet-like architecture that utilizes Swin transformer blocks for feature encoding and decoding, avoiding the sample bias of exemplar-based approaches [13]. 2) The proposed GCAM exploits attentive support of repetitive objects through the self-similarity matrix, to focus on countable objects. 3) The gate mechanism is integrated into various modules, *e.g.*, GCAM, GEFS and GAFU, which suppresses the features of irrelevant objects or background while highlighting the most relevant features to countable objects. 4) The proposed GCA-SUN is evaluated on the FSC-147 and CARPK datasets. It outperforms state-of-the-art methods for exemplar-free counting.

Chapter 2

Background & Related Work

In the literature, object counting methods are generally categorized into three main approaches: **Class-Specific Counting**, **Class-Agnostic Counting**, and **Exemplar-Free Counting**.

Class-Specific Counting is a detection-based approach that relies on object detectors to localize instances of specific classes within an image [20–22]. While effective, these methods require specialized training datasets for each target class, making them less generalizable to unseen classes during inference. This reliance on class-specific data presents significant challenges, particularly in scenarios where the training set does not contain all potential object categories.

Class-Agnostic Counting adopts regression-based methods [1, 2, 10]. Instead of addressing the complex multi-class training required for open-set detection, these methods focus on learning a mapping from dense image features to a density map. This mapping is guided by visual exemplars [1, 2, 10] or textual prompts [11, 12]. While achieving improved generalization across diverse object classes, these approaches are susceptible to sample bias or misinterpretation introduced by the provided exemplars or text prompts, potentially leading to discrepancies between human-like and machine-based counting.

Exemplar-Free Counting has been introduced as a promising alternative [1, 13, 14] to address these limitations. EFC approaches eliminate the dependency on manually annotated exemplars or textual prompts by learning visual features that autonomously identify human-countable regions within an image. This reduces annotation overhead and mitigates biases introduced by predefined exemplars or text prompts. By use this autonomous feature extraction, EFC methods offer a more robust and generalizable framework for object counting across diverse and unseen scenarios.

Chapter 3

Design & Specifications

3.1 Overview of Proposed Method

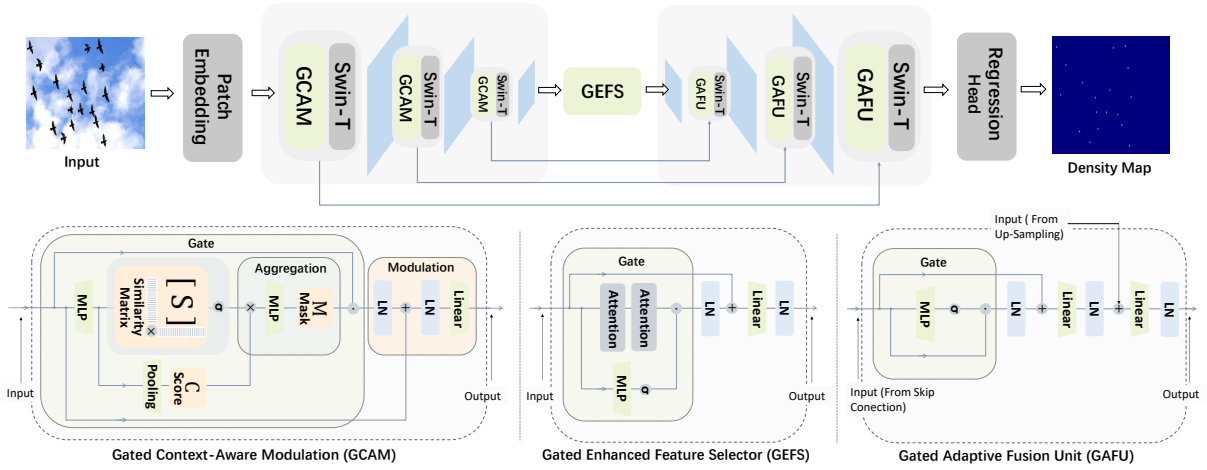

Figure 3.1: Overview of proposed GCA-SUN. It consists of an encoder, bottleneck, and a decoder. The encoder consists of a set of GCAM blocks to highlight the features relevant to countable objects while suppressing others, and Swin transformer to extract features. The GEFS in the bottleneck and the GAFU in the decoder also enhance the features of objects of interest. Finally, a regression head generates a density map for estimating the number of objects.

The proposed Gated Context-Aware Swin-UNet (GCA-SUN) is built upon a Swin-UNet architecture [23], with three new building blocks, GCAM, GEFS and GAFU, to exploit attentive support of countable objects and suppress irrelevant tokens or features, as outlined in Fig. 3.1. It begins with patched image feature $\boldsymbol{F}$, following by feature encoding,

$$\boldsymbol{F}_i^{\mathrm{E}} = \mathcal{F}_i^{\mathrm{Down}}(\mathcal{F}_i^{\text{Swin-T}}(\mathcal{F}_i^{\mathrm{GCAM}}(\boldsymbol{F}_{i-1}^{\mathrm{E}}))), \tag{3.1}$$

where $\mathcal{F}_i^{\mathrm{Down}}$, $\mathcal{F}_i^{\text{Swin-T}}$, $\mathcal{F}_i^{\mathrm{GCAM}}$ denote down-sampling, GCAM, and Swin-T processing, and $\boldsymbol{F}_{i-1}^{\mathrm{E}}$ and $\boldsymbol{F}_i^{\mathrm{E}}$ are the input and output features at the i-th stage, respectively. GCAM enhances the token for countable objects and suppresses others.

At the bottleneck, the features are enhanced through the proposed GEFS, *i.e.*, $\boldsymbol{F}^{\mathrm{BN}} = \mathcal{F}^{\mathrm{GEFS}}(\boldsymbol{F}_K^{\mathrm{E}})$, where $\mathcal{F}^{\mathrm{GEFS}}(\cdot)$ denotes the operation of GEFS, and $\boldsymbol{F}_K^{\mathrm{E}}$ denotes the output features of the encoder of K stages. GEFS selects the features corresponding to the countable object using a gate mechanism.

Subsequently, a set of Swin transformer blocks are utilized as the decoder to derive the density map. Specifically, the features at the j-th stage of the decoder are derived as,

$$\boldsymbol{F}_j^{\mathrm{D}} = \mathcal{F}_j^{\mathrm{Up}}(\mathcal{F}_j^{\text{Swin-T}}(\mathcal{F}_j^{\mathrm{GAFU}}(\boldsymbol{F}_{j-1}^{\mathrm{D}}, \boldsymbol{F}_{K+1-j}^{\mathrm{E}}))), \tag{3.2}$$

where $\mathcal{F}_j^{\mathrm{Up}}$, $\mathcal{F}_j^{\text{Swin-T}}$, and $\mathcal{F}_j^{\mathrm{GAFU}}$ denote the operation of up-sampling, Swin transformer, and GAFU block, respectively. The GAFU enhances features through a gate mechanism, prioritizing crucial information with a dynamic assigned weight.

Finally, these features are processed through a regression head, $\boldsymbol{F}^{\mathrm{head}} = \mathcal{F}^{\mathrm{Head}}(\boldsymbol{F}_K^{\mathrm{D}})$, where $\mathcal{F}^{\mathrm{Head}}$ denotes the regression head consisting of a series of convolutional blocks. The output is a density map that accurately represents the object count.

3.2 Swin-T Encoder with GCAM

The encoder consists of a set of Swin transformers to extract features relevant to countable objects. The GCAM employs a dynamic token modulation process to simultaneously exploit the attentive support of tokens relevant to countable objects and suppress features of irrelevant objects. This process facilitates self-probing among objects and precise capture of objects of the same category for exemplar-free counting. We first compress token features $\boldsymbol{F}_i^{\mathrm{E}}$ using an MLP, $\boldsymbol{F}_i^{\mathrm{proj}} = \mathcal{F}^{\mathrm{MLP}}\left(\boldsymbol{F}_i^{\mathrm{E}}\right)$. To identify the objects of interest, we resort to two key observations: 1) The objects should be salient enough to step out from the background; 2) Similar objects could support each other to boost the saliency. The former is exploited by computing the average feature score $\boldsymbol{C}_i$ for each token through average pooling $\mathcal{F}^{\mathrm{AVG}}$ as, $\boldsymbol{C}_i = \mathcal{F}^{\mathrm{AVG}}\left(\boldsymbol{F}_i^{\mathrm{proj}}\right)$. The score reflects the importance of tokens, prioritizing those with rich content. Tokens that frequently appear in similar contexts are more likely to be related to the target object of interest. To identify them, we employ a similarity matrix $\boldsymbol{S}_i = \sigma\left(\boldsymbol{F}_i^{\mathrm{proj}}\boldsymbol{F}_i^{\mathrm{proj}^T}\right)$, where σ is a softmax function to normalize similarities across rows. $\boldsymbol{S}_i$ captures the semantic similarity of tokens in a spatial context to emphasize tokens that repeatedly share similar features, thereby emphasizing potential countable objects. A mask $\boldsymbol{M}_i$ is derived by aggregating $\boldsymbol{S}_i$ and $\boldsymbol{C}_i$ as,

$$\boldsymbol{M}_i = \sigma\left(\mathcal{F}^{\mathrm{MLP}}\left(\boldsymbol{S}_i, \boldsymbol{C}_i\right)\right). \tag{3.3}$$

$\boldsymbol{C}_i$ encodes the token importance when considering the token alone, while $\boldsymbol{S}_i$ encodes the token importance after interacting with other tokens. The tokens are then filtered by the mask as,

$$\boldsymbol{F}_i^{\mathrm{GCAM}} = \mathcal{F}^{\mathrm{Linear}}(\mathcal{F}^{\mathrm{LN}}(\mathcal{F}^{\mathrm{LN}}(\boldsymbol{F}_i^{\mathrm{E}} \odot \boldsymbol{M}_i) + \boldsymbol{F}_i^{\mathrm{proj}}))), \tag{3.4}$$

where $\odot$, $\mathcal{F}^{\mathrm{LN}}$ and $\mathcal{F}^{\mathrm{Linear}}$ denote element-wise product, layer normalization and linear layer, respectively. The GCAM applies the mask $\boldsymbol{M}_i$ to $\boldsymbol{F}_i^{\mathrm{E}}$, filtering out less relevant features and reinforcing those critical ones for countable objects.

The proposed GCAM selectively amplifies the importance of tokens related to significant object features through pairwise similarities. It is significantly different from LOCA [18] and DAVE [19] which depends on predefined prototypes to predict object densities. In contrast, our GCAM leverages the self-similarity matrix for more dynamic and precise

modulation of features. It is also different from RCC [14] which relies on global feature comparisons, and CounTR [1] which uses attention-driven similarity matrices. The GCAM emphasizes a clear distinction between relevant and irrelevant tokens.

3.3 Bottleneck with GEFS

The proposed Gated Enhanced Feature Selector selectively filters out features in the bottleneck that are semantically irrelevant to the object of interest, but allows critical compressed features to pass through. The GEFS is implemented by first deriving the local token weights as $\boldsymbol{W}^{\mathrm{GEFS}} = \sigma(\mathcal{F}^{\mathrm{MLP}}(\boldsymbol{F}_K^{\mathrm{E}}))$, and then applying them on features as,

$$\boldsymbol{F}_0^{\mathrm{D}} = \boldsymbol{F}_K^{\mathrm{E}} + (\boldsymbol{W}^{\mathrm{GEFS}} \odot (\mathcal{F}^{\mathrm{Attn.}}(\mathcal{F}^{\mathrm{Attn.}}(\boldsymbol{F}_K^{\mathrm{E}})))). \tag{3.5}$$

The GEFS is positioned at the bottleneck where features transit from the down-sampling to the up-sampling pathways. As a vital bottleneck, GEFS compresses and filters essential object-related features, ensuring that only the most relevant information of countable objects is advanced into the up-sampling path. Specifically, the attention blocks within GEFS refine the model's ability to extract high-level semantic representations, leading to more accurate feature representation. Furthermore, the gate mechanism that is incorporated into GEFS selectively prioritizes specific aspects of this condensed representation, effectively filtering out less relevant semantics. This process not only refines features by strengthening relevant inter-dependencies, but also lays a solid foundation for comprehensive reconstruction of the up-sampling pathway.

3.4 Swin-T Decoder with GAFU

The decoder contains a set of Swin transformers to articulate the density map and a set of Gated Adaptive Fusion Units (GAFUs) to integrate low-level encoder features from skip connections with abstract features from the up-sampling pathway. In each GAFU, we employ a gate mechanism to determine the token weights as, $\boldsymbol{W}^{\mathrm{GAFU}} = \sigma(\mathcal{F}^{\mathrm{MLP}}(\boldsymbol{F}_i^{\mathrm{E}}))$, and then apply them to modulate the features as,

$$\boldsymbol{F}_i^{\mathrm{G}} = \boldsymbol{F}_i^{\mathrm{E}} + (\boldsymbol{W}^{\mathrm{GAFU}} \odot \boldsymbol{F}_i^{\mathrm{E}}). \tag{3.6}$$

Subsequently, these features are fused with the decoder features as $\boldsymbol{F}_j^{\mathrm{GAFU}} = \mathcal{F}^{\mathrm{Linear}}([\boldsymbol{F}_j^{\mathrm{D}}, \boldsymbol{F}_i^{\mathrm{G}}])$. By weighing the features during the fusion process, the GAFU effectively concentrates on semantic information pertinent to countable objects, minimizing interference from irrelevant details.

Chapter 4

Implementation

We utilize two benchmark datasets for evaluation. Following [1, 18, 19], we employ the Mean Average Error (MAE) and Root Mean Squared Error (RMSE) as evaluation metrics. MAE is calculated as

$$\frac{1}{N}\sum_{i=1}^{N}\left|\mathcal{C}_i - \mathcal{C}_i^{\mathrm{GT}}\right|,$$

where $\mathcal{C}_i$ and $\mathcal{C}_i^{\mathrm{GT}}$ represent the predicted and ground truth counts of objects for the i-th image, respectively. Similarly, $RMSE$ is computed using

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathcal{C}_i - \mathcal{C}_i^{\mathrm{GT}})^2},$$

where N is the number of query images.

4.1 Datasets

FSC-147 [2] consists of 6,135 images of 147 categories, mainly composed of foods, animals, kitchen utensils, and vehicles. It is officially split into 3,659, 1,286 and 1,190 images for training, validation and testing, respectively.

CARPK [3] comprises 1,448 images taken from four parking lots using a bird's-eye view. It is primarily intended for object counting and vehicle localization tasks, and it is officially split into 989 training images and 459 testing images.

4.2 Experiment Setting

The Swin-T blocks are pre-trained on ImageNet-22k [24], and other modules are randomly initialized. AdamW optimizer [25] is employed for training, with an initial learning rate of 0.003, a decay rate of 0.95 and a batch size of 16. The model is trained with a warm-up period of 50 epochs. The input image size is 384×384. Data augmentation [1] is employed to facilitate efficient training. Experiments are conducted using two NVIDIA RTX A5000 GPUs.

Chapter 5

Evaluation

5.1 Comparison with State-of-the-Art Methods

Comparison experiments are conducted on the FSC-147 dataset. The results are summarized in Table 5.1. Following the practice in [1, 14], we report the errors on the test and validation sets. We have the following observations. 1) The proposed GCA-SUN outperforms all the compared methods regarding test errors, while performing slightly poorer than DAVE [19] regarding validation errors. The superior results demonstrate its effectiveness for EFC tasks. It outperforms not only the dedicated methods for solving EFC tasks such as RepRPN-C [13] and RCC [14], but also state-of-the-art models for CAC tasks. Compared to the second-best method, CounTR [1], the performance gain is 0.71 for of MAE and 14.68 for RMSE. 2) Although the GCA-SUN performs slightly poorer than DAVE [19] in terms of validation errors, it significantly outperforms DAVE in terms of test errors, *i.e.*, a performance gain of 1.14 for MAE and 11.30 for RMSE. DAVE tends to overfit to the validation set, while generalizing poorly on the test set. In contrast, the GCA-SUN generalizes well on the novel test set with minimal errors. We

Method	Test Set		Val Set	
	MAE	RMSE	MAE	RMSE
FamNet [2] CVPR'21	32.27	131.46	32.15	98.75
LOCA [18] ICCV'23	16.22	103.96	17.43	54.96
DAVE [19] CVPR'24	15.14	103.49	**15.54**	**52.67**
CounTR [1] BMVC'22	14.71	106.87	18.07	71.84
RepRPN-C [13] ACCV'22	26.66	129.11	29.24	98.11
RCC [14] CVPR'23	17.12	104.53	17.49	58.81
Proposed GCA-SUN	**14.00**	**92.19**	16.06	53.04

Table 5.1: Comparison with other methods on the FSC-147 dataset [2], with best results highlighted in bold.

visually compare the density maps on the FSC-147 dataset. As shown in Fig. 5.1, our method can capture fine-grained details of objects. CounTR sometimes generates density maps that do not accurately distinguish between individual objects in the map, *e.g.*, in the fourth column, CounTR can't identify the far-way small fruit, while our method can.

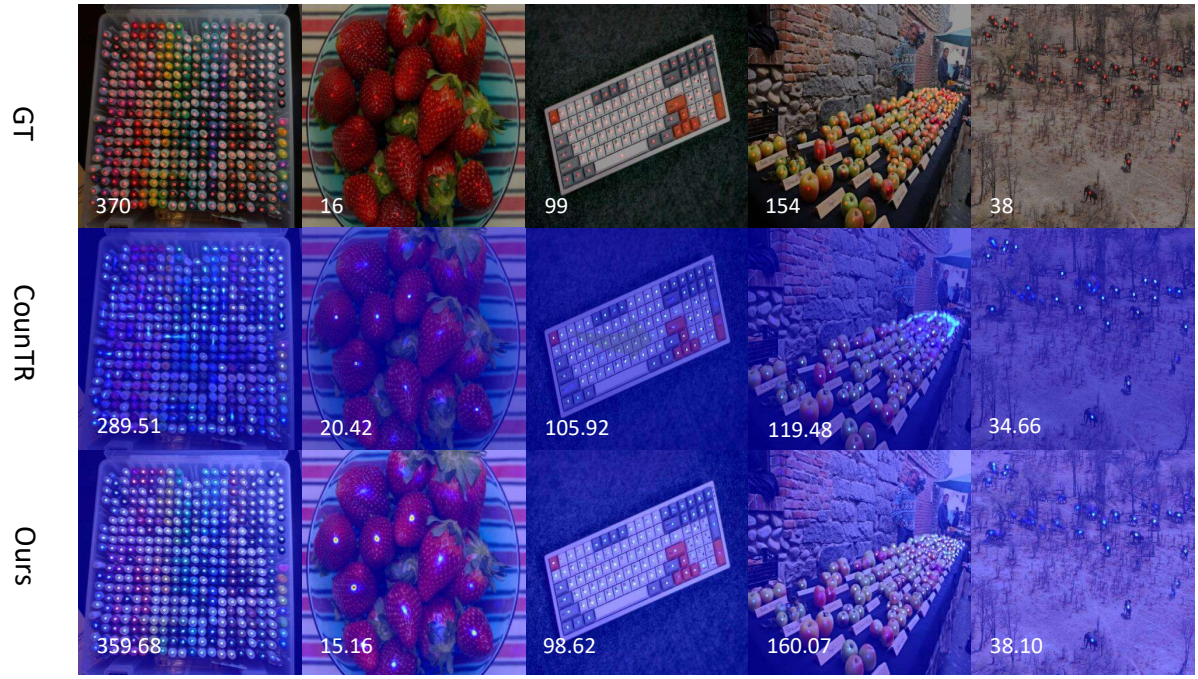

Figure 5.1: Visual comparisons to CounTR [1] on the FSC-147 dataset.

5.2 Cross-Domain Evaluation on CARPK Dataset

Following [26], we conduct a cross-domain evaluation, training the model on FSC-147 [2] and directly evaluating on CARPK [3], with results summarized in Table 5.2. The results for all compared methods are reproduced from [1, 14, 18] under the same settings. Our model has shown superior cross-domain performance compared with other methods, achieving a performance gain of 0.56 on MAE and 0.61 on RMSE compared to the previous best-performing method CounTR. Compared to the earlier EFC model [14], the gains are even more significant, highlighting GCA-SUN's superior generalization over all compared methods.

Methods	MAE	RMSE
LOCA [18] ICCV'23	16.84	19.72
CounTR [1] BMVC'22	11.52	14.56
RCC [14] CVPR'23	21.38	26.61
Proposed GCA-SUN	**10.96**	**13.95**

Table 5.2: Comparison with other methods on the CARPK dataset [3].

5.3 Visualization of GCAM

We visualize the effects of the proposed GCAM in Fig. 5.2. Sub-figure (b) and (c) are obtained by projecting the density maps into two two-dimensional spaces, followed by kernel density estimation [27] to calculate the density distribution. The resulting images

are then normalized and visualized to illustrate the model's focus areas. Clearly after applying GCAM, a general decrease of density values in the background areas (sky) can be observed, with the foreground objects (birds) becoming more prominent. This indicates that the GCAM module effectively enhances the representation of foreground tokens while suppressing irrelevant ones.

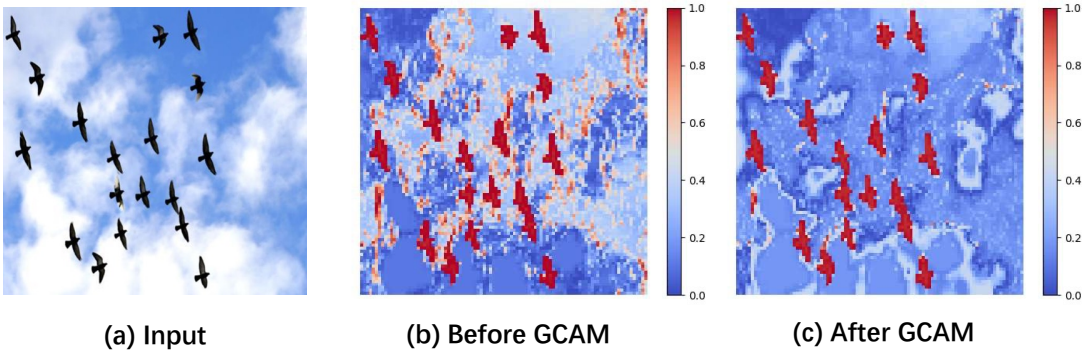

Figure 5.2: Visualization of the effects of GCAM.

5.4 Ablation Study

We conduct a set of comprehensive ablation studies on the three major modules of proposed method on the FSC-147 dataset [2]. The results are summarized in Table 5.4. The GCAM module alone significantly decreases the MAE on the test set by 1.77 and on the validation set by 2.33, highlighting its capability to enhance feature selectivity crucial for complex scenes. Similarly, utilizing GEFS alone or GAFU alone also greatly reduces the errors in both test set and validation set, demonstrating the importance of the gate mechanism in highlighting the relevant features while suppressing irrelevant ones. The full integration of all three components produces the most substantial enhancement, reducing MAE by 2.83 on the test set and by 3.34 on the validation set. This underscores the effectiveness of their synergistic interaction and affirms the component's design.

$\mathcal{F}^{\text{GCAM}}$	$\mathcal{F}^{\text{GEFS}}$	$\mathcal{F}^{\text{GAFU}}$	Test Set		Val Set	
			MAE	RMSE	MAE	RMSE
✗	✗	✗	16.83	88.41	19.40	71.58
✓	✗	✗	15.06	91.68	17.07	53.22
✗	✓	✗	15.78	88.21	18.26	57.30
✗	✗	✓	15.82	84.64	17.92	59.61
✓	✗	✓	14.47	89.16	16.40	54.18
✗	✓	✓	14.86	86.85	17.72	57.27
✓	✓	✗	14.66	94.47	17.59	57.85
✓	✓	✓	14.00	92.19	16.06	53.04

Table 5.3: Ablation study of each component on the FSC-147 dataset [2].

Chapter 6

Summary and Reflections

6.1 Project management

6.1.1 Overview of the Project Plan

At the beginning of the project, a detailed plan was developed, as shown in Fig 6.1. The plan consisted of several key phases, including preliminary work, literature review, design, experimentation, and report writing. Each phase was subdivided into specific tasks with estimated start and finish dates. The estimated duration of each task was based on its complexity, with more time allocated to critical tasks, such as the design and development of methods and conducting experiments.

The project timeline was designed to ensure steady progress toward the final goal while allowing flexibility for adjustments based on task complexity and unforeseen challenges. The initial plan emphasized the importance of foundational tasks and performing literature reviews to identify relevant methodologies.

6.1.2 Comparison of Planned and Actual Progress

While the overall structure of the project remained consistent with the original plan, several deviations occurred due to the dynamic nature of research and technical challenges encountered during the project.

- **Preliminary Work**: The preliminary work phase was initially scheduled to span from October 1, 2024, to October 21, 2024. However, due to the efficient execution of the project proposal preparation, this phase was completed ahead of schedule, finishing on October 15, 2024.
- **Literature Review**: The literature review, originally planned for the period between October 21, 2024, and November 15, 2024, was completed by November 1, 2024. This accelerated progress was driven by the additional effort invested in overcoming challenges related to the identification of relevant methods and the selection of appropriate literature.
- **Design and Development**: The design phase, which was initially scheduled from November 15, 2024, to November 30, 2024, began earlier on November 2, 2024, following the early completion of the literature review.

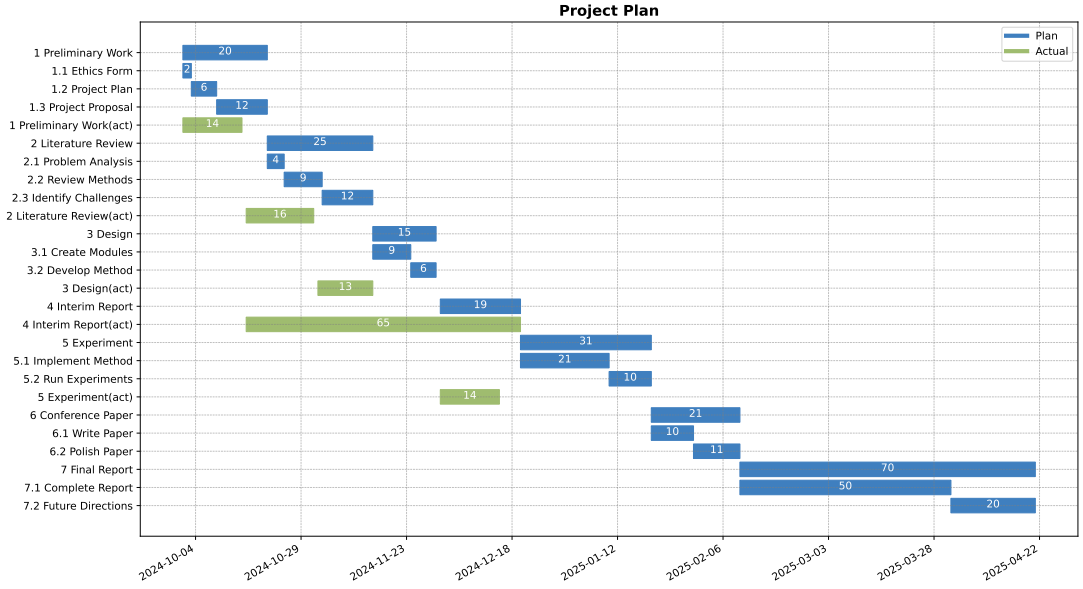

Figure 6.1: Original Plan.

- **Experimentation**: The experimentation phase, initially planned for December 20, 2024, to January 20, 2025, commenced earlier on December 1, 2024. During this phase, baseline results were generated, and improvements were implemented to enhance model performance.
- **Report Writing**: We began writing the interim report ahead of schedule, realizing that the literature review and model design stages provided an opportune time to start drafting.

6.1.3 Reflections on the Project Plan

Reflecting on the deviations from the original plan, several insights were gained that can inform future project management:

- **Flexibility and Adaptation**: The adjustments made during the experimentation phase underscore the importance of flexibility in research. The decision to prioritize writing the interim report enabled the project to stay aligned with its objectives, while the early completion of Stage 1 allowed for a deeper exploration of other tasks in Object Counting (OC).
- **Task Dependencies**: The accelerated progress in preliminary work and the literature review positively influenced subsequent tasks, such as design and experimentation. This highlights the significance of efficiently completing foundational tasks to maintain project momentum.

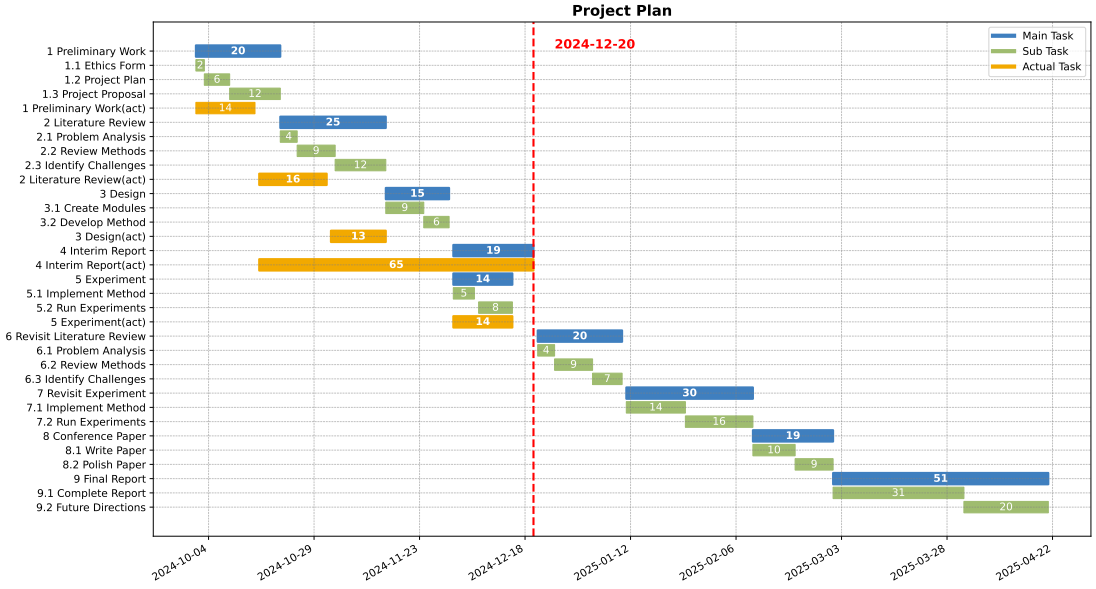

Figure 6.2: Revised Plan.

- **Time Allocation**: The unforeseen complexity of tasks like method development and experimentation emphasizes the need for buffer time in project planning. Future projects should account for additional time in tasks involving significant innovation and iterative refinement.
- **Focus on Objectives**: The project benefited from maintaining a clear focus on tasks directly contributing to its core objectives, ensuring efficient progress.

6.1.4 Revised Timeline

The updated Gantt chart (Figure 6.2) reflects the changes made to the original timeline. Key adjustments include the earlier completion of preliminary work and literature review, as well as the shift in focus from reproducing prior results to producing baseline results. Additional time was allocated for reviewing literature and refining experiments for CAC tasks.

6.2 Conclusion and future work

The proposed GCA-SUN effectively tackles the problems of exemplar-free counting by using a Swin-UNet architecture to directly map the input image to the density map of countable objects. The proposed GCAM exploits the attention information among the tokens of repetitive objects through the self-similarity matrix, and suppresses the features

of irrelevant objects through a gate mechanism. The gate mechanism is also incorporated into the GEFS module and the GAFU module, which highlight the features most relevant to countable objects while suppressing irrelevant ones. Our experiments on the FSC-147 and CARPK datasets demonstrate that GCA-SUN outperforms state-of-the-art methods, achieving superior performance in both intra-domain and cross-domain scenarios.

Despite its effectiveness, the EFC approach has certain limitations. To address this, we have turned to explore CAC, aiming to develop more robust models for object counting.

References

[1] Liu Chang, Zhong Yujie, Zisserman Andrew, and Xie Weidi. CounTR: Transformer-based Generalised Visual Counting. In *Brit. Mach. Vis. Conf.*, 2022.

[2] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3394–3403, 2021.

[3] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Int. Conf. Comput. Vis.*, pages 4145–4153, 2017.

[4] Ersin Kilic and Serkan Ozturk. An accurate car counting in aerial images based on convolutional neural networks. *J. Ambient Intell. Humanized Comput.*, pages 1–10, 2023.

[5] Mingliang Dai, Zhizhong Huang, Jiaqi Gao, Hongming Shan, and Junping Zhang. Cross-head supervision for crowd counting with noisy annotations. In *IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023.

[6] Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Comput. Methods Biomech. Biomed. Eng. Imag. Vis.*, 6:283–292, 2018.

[7] Zhuojun Chen, Junhao Cheng, Yuchen Yuan, Dongping Liao, Yizhou Li, and Jiancheng Lv. Deep density-aware count regressor. In *ECAI 2020*, pages 2856–2863. IOS Press, 2020.

[8] Enrico Bellocchio, Thomas A Ciarfuglia, Gabriele Costante, and Paolo Valigi. Weakly supervised fruit counting for yield estimation using spatial consistency. *IEEE Robot. Autom. Lett.*, 4:2348–2355, 2019.

[9] Jayme Garcia Arnal Barbedo, Luciano Vieira Koenigkan, Patricia Menezes Santos, and Andrea Roberto Bueno Ribeiro. Counting cattle in UAV images—dealing with clustered animals and animal/background contrast changes. *Sensors*, 20:2126, 2020.

[10] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9529–9538, 2022.

[11] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15548–15557, 2023.

[12] Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. Vlcounter: Text-aware visual representation for zero-shot object counting. In *Proc. AAAI Conf. Artif. Intell.*, volume 38, pages 2714–2722, 2024.

[13] Viresh Ranjan and Minh Hoai Nguyen. Exemplar free class agnostic counting. In *Proc. Asian Conf. Comput. Vis.*, pages 3121–3137, 2022.

[14] Michael Hobley and Victor Prisacariu. Learning to Count Anything: Reference-less Class-agnostic Counting with Weak Supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[15] Hüseyin Gökhan Akçay, Bekir Kabasakal, Duygugül Aksu, Nusret Demir, Melih Öz, and Ali Erdoğan. Automated bird counting with deep learning for regional bird distribution mapping. *Animals*, 10:1207, 2020.

[16] Vitjan Zavrtanik, Martin Vodopivec, and Matej Kristan. A segmentation-based approach for polyp counting in the wild. *Eng. Appl. Artif. Intell.*, 88:103399, 2020.

[17] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20402–20411, 2023.

[18] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18872–18881, 2023.

[19] Jer Pelhan, Vitjan Zavrtanik, Matej Kristan, et al. DAVE-A Detect-and-Verify Paradigm for Low-Shot Counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23293–23302, 2024.

[20] Olga Barinova, Victor Lempitsky, and Pushmeet Kholi. On detection of multiple object instances using hough transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1773–1784, 2012.

[21] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95:1–12, 2011.

[22] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Int. Conf. Comput. Vis.*, pages 4145–4153, 2017.

[23] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In *Eur. Conf. Comput. Vis. Worksh.*, 2022.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021.

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proc. Int. Conf. Learn. Represent.*, 2019.

[26] Zhicheng Wang, Liwen Xiao, Zhiguo Cao, and Hao Lu. Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting. In *Proc. AAAI Conf. Artif. Intell.*, volume 38, pages 5832–5840, 2024.

[27] Stanisław Węglarczyk. Kernel density estimation and its application. In *Proc. ITM Web Conf.,*, volume 23, page 00037, 2018.