Information Fusion 111 (2024) 102537

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Towards zero-shot object counting via deep spatial prior cross-modality fusion

Jinyong Chen [a,1], Qilei Li [a,b,1], Mingliang Gao [a,*], Wenzhe Zhai [a], Gwanggil Jeon [c,*], David Camacho [d]

[a] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom
[c] Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, South Korea
[d] Computer Science Department, Universidad Politécnica de Madrid, 28040, Spain

ARTICLE INFO

Keywords:
Object counting
Cross-modality
Deep Spatial Prior
Grounding DINO
Zero-shot

ABSTRACT

Existing counting models predominantly operate on a specific category of objects, such as crowds and vehicles. The recent emergence of multi-modal foundational models, *e.g.*, Contrastive Language-Image Pre-training (CLIP), has facilitated class-agnostic counting. This involves counting objects of any given class from a single image based on textual instructions. However, CLIP-based class-agnostic counting models face two primary challenges. Firstly, the CLIP model lacks sensitivity to location information. It generally considers global content rather than the fine-grain location of objects. Therefore, adapting the CLIP model directly is suboptimal. Secondly, these models generally freeze pre-trained vision and language encoders, while neglecting the potential misalignment in the constructed hypothesis space. In this paper, we address these two issues in a unified framework termed Deep Spatial Prior Interaction (DSPI) network. The DSPI leverages the spatial-awareness ability of large-scale pre-trained object grounding models, *i.e.,* Grounding DINO, to incorporate spatial location as an additional prior for a specific query class. This enables the network to be more specifically focused on the precise location of the objects. Additionally, to align the feature space across different modalities, we tailor a meta adapter that extracts textual information into an object query. This serves as an instruction for cross-modality matching. These two modules collaboratively ensure the alignment of multi-modal representations while preserving their discriminative nature. Comprehensive experiments conducted on a diverse set of benchmarks verify the superiority of the proposed model. The code is available at https://github.com/jinyongch/DSPI.

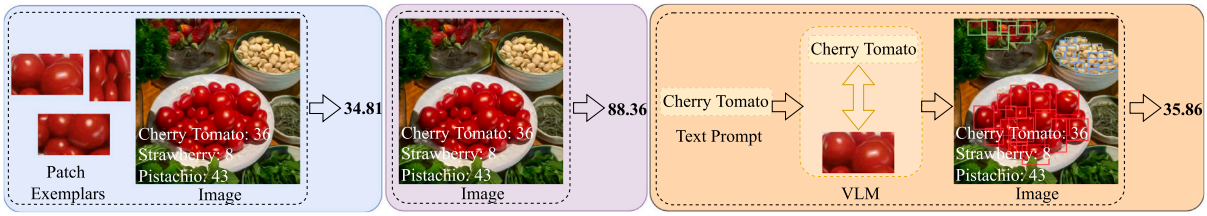
1. Introduction

Over the last decade, object-specific counting has garnered substantial attention [1–3] and significant progress had been achieved, especially for crowd counting and vehicle counting. However, these models face constraints when it comes to counting specific objects, thereby restricting their effectiveness in various real-world applications, especially when dealing with unseen object categories. Consequently, there is a significant demand for a versatile counting model that can seamlessly operate across previously untrained categories and generate corresponding density estimations [4].

This demand has led to the emergence of class-agnostic counting models [5–7]. They are designed to train a unified/shared model to estimate the number of arbitrary objects of interest within a provided image, as depicted in Fig. 1-(a). Through the annotation of select image patches as exemplars and subsequently computing the similarities between these exemplars and various image regions, these models have demonstrated commendable levels of generalizability and counting precision. However, most class-agnostic counting methods are built on the fragile assumption that the precise bounding boxes can be easily obtainable during inference, which is challenging to realize in practical scenarios. Therefore, one of the requirements is to annotate samples of objects to be counted manually, which can present a significant inconvenience for users. Besides, the substantial intra-class variance among query objects may introduce biased object counts [7,8]. To address these issues, reference-less counting methods have been developed, which enable the exploration and enumeration of salient objects during inference [9,10]. Although these methods relieve from manually annotating samples, they fail to specify the category of

* Corresponding authors.
E-mail addresses: mlgao@sdut.edu.cn (M. Gao), gjeon@inu.ac.kr (G. Jeon).
[1] Authors contributed equally.

https://doi.org/10.1016/j.inffus.2024.102537
Received 10 April 2024; Received in revised form 28 May 2024; Accepted 14 June 2024
Available online 18 June 2024



(a). Few-shot Counting (b). Reference-less Counting (c). Spatial Prior Guided Zero-shot counting

Fig. 1. Illustration of various generalized object counting schemes: (a) Few-shot counting: This approach necessitates manually labelled image patches for training and inference. (b) Reference-less counting: This method automatically identifies and counts salient objects. However, it cannot specify objects of interest. (c) Spatial Prior Guided Zero-shot counting: The diagram proposed in this paper. It makes use of the spatial information from objects of interest and provide text instructions to specify the counting object.

the object of interest when multiple categories are presented, as shown in Fig. 1-(b). In general, the limitations of the existing counting models lie in their relative inflexibility. Therefore, it is hard to adapt them directly to real-world applications.

Contrastive Language-Image Pre-training (CLIP) [11] is an efficacious and scalable approach derived from natural language supervision. The CLIP captures semantic correspondences between images and text, enabling robust generalization even without annotations. Recent model CLIP-Count [12] a frozen vision encoder was built to extract visual presentation from the given image. The text representation from the corresponding text input instructs the specific object class to be counted. Unlike the existing reference-less counting methods, it does not require additional samples to warm up the model for the interested object. This can make domain-agnostic counting more practical. Although CLIP provides generic knowledge regarding the objects represented in the given image, directly applying CLIP encoders into the model design as CLIP-Count, is inherently limited by the following two aspects. (1) CLIP is pre-trained by contrastive comparing the vision and language representations, therefore, it intrinsically decides if one object exists in the image, rather than being attentive to the spatial location of the objects. Therefore, simplifying the use of the vision encoder to an exact representation for counting is suboptimal, as counting objects' dominance relies on their spatial locality. (2) The visual inputs for pretraining CLIP are mainly natural images and the containing objects are relatively sparse. While counting the objects, the inputs normally contain many objects. The data distribution shifts regarding the density of the objects, making the text representation not aligned with the visual representation.

In this work, we aim to solve the aforementioned two limitations using frozen CLIP for zero-shot object counting. To make the image representation attentive to the spatial information of the objects, instead of training the auxiliary branch to label the object's bounding box from scratch, we design the spatial prior guided module, as illustrated in Fig. 1-(c). To this aim, we explore deep prior from the cutting-edge multi-modal grounding models, *i.e.*, Grounding DINO [13]. We incorporated Grounding DINO as a training-free module and used it to provide the network with deep prior regarding the location of the specific objects. The spatial prior extractor is frozen, so that it does not introduce any additional parameters for training. Secondly, we address the density shift issue while deploying pre-trained CLIP encoders for object counting. Instead of training the resource-intensive heavy vision and language encoders, we introduce a meta adapter designed to function as a translator. This meta adapter converts the text instruction into a vision query specifically related to the object. The meta adapter is a lightweight module designed to extract the text instruction into an object query and facilitate interaction with the visual information. This enables the visual representation to be more attentive to the specific object. The contributions of the paper are summarized as:

1. We address the spatial location oversight in foundational models by integrating a pre-trained multi-modal grounding model. This model generates spatial priors for guidance and enhances the attentiveness of the vision encoder to specific object regions while suppressing background interference.
2. We tackle the misalignment issue between textual instructions and visual representations using the designed meta adapter. It can extract instructive descriptors from the text and transform them into an object query aligned with the vision representation. Thus, it facilitates the subsequent cross-modality interaction.
3. We validate the effectiveness of the proposed Deep Spatial Prior Interaction (DSPI) model through a comprehensive set of experiments. The results demonstrate that DSPI can extract distinctive representations aligned across various modalities, while also incorporating precise spatial information, to help improve the generalization capability of the model.

The remainder of the paper is structured as follows. In Section 2, the relevant research is reviewed. Section 3 offers a detailed explanation of the proposed method. Comprehensive experimental analyses are provided in Section 4. The conclusion is drawn in Section 5.

2. Related work

2.1. Object grounding

Object grounding refers to connecting or associating natural language descriptions with actual objects in an image. This aids in empowering models to comprehend the relationship between textual descriptions and the content present in images. This field has been tremendously advanced over the last few years and numerous models have been proposed. One of the most representative works is DETR [14]. It has undergone various improvements from different perspectives [15–17]. DAB-DETR [18] incorporates anchor boxes as queries in DETR to enhance the precision of box predictions. Furthermore, the DETR with improved denoising anchor boxes (DINO) [19] is built upon the foundations of DAB-DETR and DN-DETR. It further advances several techniques, including contrastive denoising, and achieving new records on the COCO object detection benchmark.

However, such detectors focus on closed-set detection and may face challenges when generalizing to new classes due to limited predefined categories. Open-set object detection challenges the model to infer unknown object classes. It goes beyond recognizing familiar target classes present in the training set. This is vital in real-world situations, where the model may encounter unfamiliar targets. OV-DETR [20] leverages image and text encodings from the contrastive language-image pre-training (CLIP) [11] model as queries to decode category-specific boxes within the DETR framework. ViLD [21] extracts knowledge from the CLIP teacher model into the R-CNN class detector. DETR and Deformable DETR [22], attempt to formalize object detection as a set prediction problem that can eliminate the post-processing non-maximum suppression (NMS). However, previous efforts are limited to integrating multi-modal information at certain stages. Although these approaches may have yielded relatively good results, they might not be optimal [13]. Grounding DINO is a multi-modal model renowned for extracting detailed representations. It effectively captures the precise spatial positioning of objects and can create bounding boxes for various object categories. Moreover, it fits into current multi-modal designs to

provide meaningful guidance information. In this paper, We employ a pre-trained multi-modal grounding model, *i.e.*, Grounding DINO, to extract a deep spatial prior. This spatial prior guides the CLIP image encoder in being attentive to the location of objects.

2.2. Zero-shot object counting

Few-shot object counting aims to determine the number of objects in an image with limited training samples. It can quickly learn and adapt to new object categories in a relatively short amount of time. Thus, it provides flexibility and efficiency for a broader range of scenarios in practical applications. FamNet [23] employed ROI pooling to predict density maps and introduced a dataset for class-agnostic counting, known as FSC-147 [23]. The further progress can be divided into two main aspects. One method includes utilizing advanced visual backbones, like Vision Transformers (ViT), to improve the extracted feature representations [5,8]. The second approach aims to improve exemplar matching by either explicitly modelling exemplar-image similarity [24,25], or by further utilizing exemplar guidance, as explored in [6,26]. Despite their excellent performance, they are not applicable when some samples are not obtained.

Recently, reference-less counting has become an effective approach for class-agnostic counting without relying on human annotations. RepRPN-Counter [10] introduced a region proposal module specifically designed to extract prominent objects, eliminating the need for sampled inputs. RCC [9] employed pre-trained ViT [27,28] to extract salient objects implicitly and directly regress a scalar for estimating object counts. Several contemporary few-shot counting models [5,6] can also be adapted for reference-less counting. Although these approaches do not depend on samples, they lack an effective method for specifying the object of interest in the presence of multiple object classes. Simultaneously, Xu et al. [7] introduced zero-shot object counting, which requires only the class name during inference. They trained a text-conditional variational autoencoder (VAE) on a known object set and a few-shot object counter with exemplar supervision to generate exemplar prototypes. However, these approaches still depend on patch exemplars. To facilitate end-to-end training without the need for patch-level supervision, Jiang et al. incorporated Contrastive Language-Image Pre-training (CLIP) [11] into the counting network [12]. CLIP endows the model with zero-shot image-text alignment capability. To transfer the robust image-level representations from CLIP to dense tasks such as density estimation, a text-contrastive loss, and a hierarchical patch-text interaction module are devised within the model. In this paper, we focus on zero-shot object counting given its practical application value. Zero-shot counting methods count objects in a class-agnostic manner without the need for additional image patch annotations. However, the previously mentioned models lack the ability to perceive spatial location information. Therefore, while addressing multi-modal alignment, we also emphasize the importance of spatial location awareness.

2.3. Feature attentive learning

The attention mechanism is adopted to enable the network to concentrate on the discriminative part of the input data. It has been widely incorporated in different types of networks, including recurrent neural networks (RNN), convolutional neural networks (CNN), and the Transformer [29] based networks. It has been broadly adopted in applications, such as image segmentation, object detection, and crowd counting [30,31]. The most typical attention mechanisms include spatial attention mechanisms, channel attention mechanisms, and self-attention mechanisms.

Spatial attention methods focus on critical regions within input data and enhance spatial context information. The channel attention mechanism focuses primarily on the channel dimension of input data, and it enhances the essential channel features. Woo et al. [32] proposed a Convolutional Block Attention Module (CBAM), that incorporates both channel and spatial attention. Fu et al. [33] proposed a multiscale feature fusion method, called Dual Attention Network (DANet), which can adaptively integrate local features and their global dependencies. DANet introduced channel and spatial attention branches to improve semantic segmentation performance. The advantage of self-attention over traditional spatial and channel attention lies in its minimal dependence on external information and its superior ability to capture non-local correlations. This quality facilitates the extraction of global information representations in transformer networks without employing traditional RNN or CNN networks. Self-attention and cross-attention are based on the same core mechanism, but their applications and purposes differ. Self-attention is designed to deal with relationships within a single sequence, while cross-attention is designed to deal with relationships between two different sequences. In this paper, we introduce the deep spatial prior that encodes the spatial location of the probe objects as hard-coded attention. This guidance mechanism aims to enhance the model's spatial awareness of the query objects.

2.4. Deep learning for crowd counting

Crowd counting is particularly prominent in all kinds of object counting tasks because of its special significance to social security and development. Zhang et al. [34] proposed a method based on multi-column convolutional neural networks for crowd counting from a single image. The network can learn to capture features at different scales and perspectives by designing multiple columns. Li et al. [35] proposed a congested scene recognition network (CSRNet), which introduced extended convolutional neural networks to understand highly congested scenarios better. By introducing an extended convolution layer, the network effectively captures the information of different density regions. Zhang et al. [36] proposed an adaptive convolutional neural network in which the structure can be adjusted in response to density changes in the input image. This adaptability enables the network to better adapt to various crowded scenarios. Wang et al. [37] developed an automated data acquisition that used domain adaptation from synthetic images to real images to address the limitations of labelled real-world data. Earlier CNN-based approaches utilized multi-column architectures with different receptive fields to learn features at different scales. Li et al. [35] employed an expanded convolution layer to augment the receptive field to address scale variation.

Meanwhile, some works aim to address domain offsets between training and test images in crowd counting. Reddy et al. [38] proposed a meta-learning heuristic strategy to address learning mechanisms in scenarios with limited data shots. Unlike the prior methods that tackled individual distributions, Zhu et al. [39] proposed a domain adaptive approach by employing optimal transport in both the source and target domains. This alignment strategy addresses misalignments caused by domain-agnostic factors. Jiang et al. [12] proposed a CLIP-Count model in which pre-trained vision-language models (VLMS) can directly adopt text-guided object counting tasks in an end-to-end manner. It can detect and count the target by the learned semantic relation without the direct label. These methods are specifically designed for human counting and do not apply to categorize beyond humans. In this paper, we test the generalizability of the model on crowd counting to validate the class-agnostic merit of the DSPI model.

3. Proposed method

3.1. Preliminary: Grounding DINO

The Grounding DINO model [13] adopts a dual-encoder-single-decoder architecture. It comprises five parts, *i.e.,* image backbone, text backbone, feature enhancer, language-guided query selection module, and cross-modality decoder. For each (Image, Text) pair, it first extracts vanilla image and text features using an image and text backbone, respectively. Each pair (Image, Text) is processed through an image

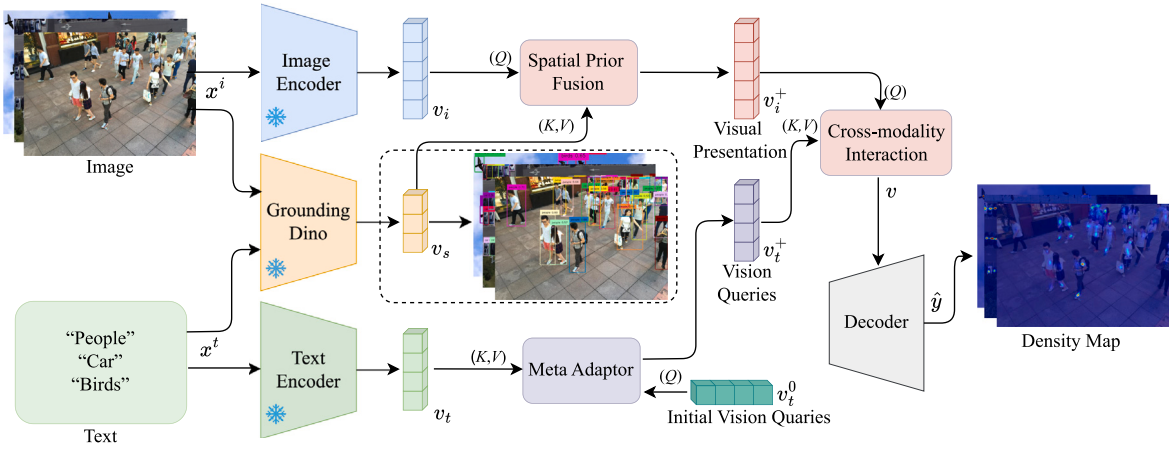


Fig. 2. The framework of the proposed DSPI model.

backbone and a text backbone to extract basic image and text features. These features are fed into a feature enhancer module for cross-modality feature fusion. Following the acquisition of cross-modality text and image features, a language-guided query selection module is employed to identify cross-modality queries based on the extracted image features. The identified cross-modality queries will be input into a cross-modality decoder, which enables the probing of desired features from both modalities and subsequently updating themselves. The output queries from the final decoder layer will be employed to predict object boxes and extract the corresponding phrases.

3.2. Overview

The goal of the DSPI method is to develop a deep neural network, denoted as f_θ, that processes a visual image x^i and a text instruction x^t as inputs. The network is expected to produce a density map, represented as $\hat{y} = f_\theta(x^i, x^t)$, which accurately marks the spatial location of the target object(s) described in the text instruction. To this aim, we design the DSPI model that can interact with the cross-modality representations produced by the frozen CLIP encoders. The overall framework of the proposed Deep Spatial Prior Interaction (DSPI) model is shown in Fig. 2.

Given a batch of samples $\mathcal{X}_{tr} = \{x_1, x_2, \ldots, x_B\}$, and the corresponding text instruction $\mathcal{T}_{tr} = \{t_1, t_2, \ldots, t_B\}$, where H and W are the width and height of the image, and B is the batch size, we use image encoder f_{ci} in CLIP model to extract the visual representation from the input images as $\mathcal{V}_i = f_{ci}(\mathcal{X})$. Similarly, the text representation can be extracted by the CLIP text encoder f_{ct} as $\mathcal{V}_t = f_{ct}(\mathcal{T})$. To make the visual representation attentive to the spatial location of the probe object, we design a prior fusion module to incorporate the deep spatial prior produced by the grounding DINO model into the visual representation. Moreover, we designed a meta adapter to bridge the modality gap between the text probe and visual representation. This adapter translates the text probe into an object query, enabling effective interaction with the visual representation. Finally, given the discriminative multi-modal representation, a decoder is applied to regress a density map that indicates the spatial location of the query object.

3.3. Deep spatial prior attentive injection

The visual representation extracted by the CLIP vision encoder is generally focused on the overall category of objects within the given images, while insensitive to the spatial location of the objects. For counting the objects, it is vital to know the fine-grain location of objects so that a density map can be generated and the number of objects can be counted by summarizing the objects in the generated density map. To make the vision representation spatial-aware, we utilize the spatial prior extracted by the large-scale pre-trained Grounding DINO model to focus on relevant object regions. The illustration of the deep spatial prior extraction is shown in Fig. 3. Grounding DINO is also a multi-modal model, which can label the spatial location of certain objects. We extract the deep spatial prior using the frozen Grounding DINO model. Given the paired input $\{x^i, x^t\}(x^i \in \mathbb{R}^{C\times H\times W}, x^t \in \mathbb{R}^L)$, we employed the image backbone and text backbone to extract the representations in each modality. These representations are fused by a cross-modality decoder to generate the bounding boxes. Instead of directly using the bounding box, we use the intermediate representation produced by the cross-modality decoder as the deep spatial prior. It is produced after the cross-attention layer so that it can encode global information while being locally attentive to the spatial location of the query object.

The spatial prior fusion conducted by the cross-attention layer is applied among the three counterparts, namely query (Q), key (K), and value (V). The attentive interaction process is formulated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where d_k indicates the dimension of K. Softmax$(\cdot)$ represents the normalization function.

Once we obtain the deep spatial prior v_s, it will be fused into the CLIP image representation to emphasize the spatial location of a certain object. To this end, we design the spatial prior fusion module. The pipeline of the spatial prior fusion module is illustrated in Fig. 4-(a). It contains a multi-head cross-attention (MHCA) layer that takes the image representation v_i as the Q, and the spatial prior v_s as K and V. Following the MHCA, we use a Multi-Layer Perceptron (MLP) to refine the extracted representation. The whole process can be denoted as:

$$v_i^+ = \text{MLP}(\text{softmax}(\frac{\text{FC}_Q(v_i) * \text{FC}_K(v_s)^T}{\sqrt{d_k}}) * \text{FC}_V(v_s)), \tag{2}$$

where the $FC_{Q|K|V}(\cdot)$ denotes the project layers for the three counterparts, $MLP(\cdot)$ is the function of the MLP layer, and v_i^+ is the spatially enhanced visual representation.

3.4. Vision queries learning by meta adapter

Considering the intrinsic difference in object density between the input image and the samples used to train the CLIP encoders, there is a major challenge caused by the holistic distribution shift that hinders the alignment between text representation and visual representation. To mitigate this issue, We propose incorporating a meta adapter as a translator to extract informative knowledge from the text probe. This acquired knowledge can be seamlessly integrated with the visual representation, fostering effective cross-modal interaction.

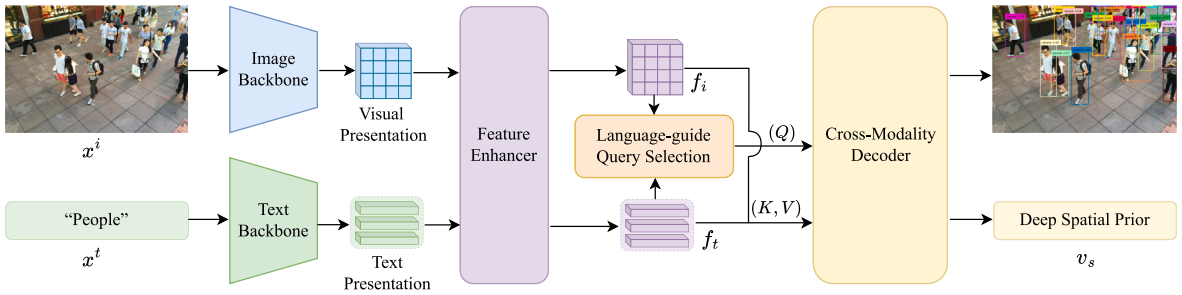


Fig. 3. Illustration of the deep spatial prior.

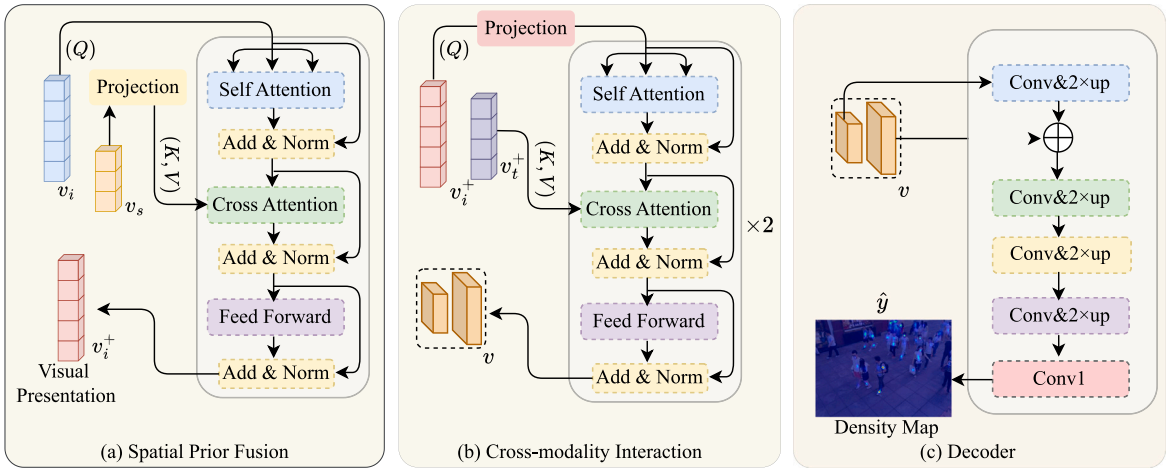


Fig. 4. Illustration of spatial prior fusion module, cross-modality interaction module and Decoder module.

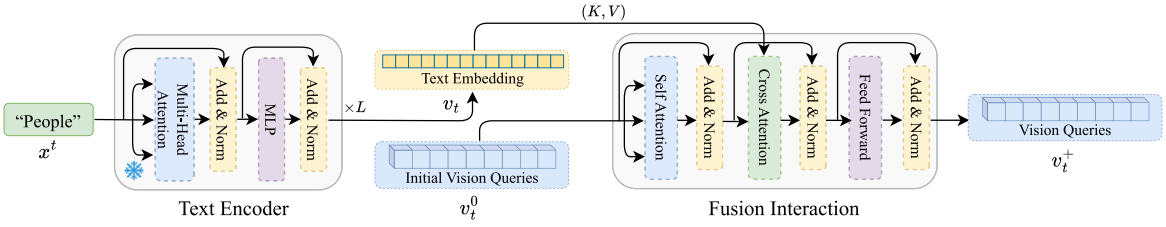


Fig. 5. Illustration of the meta adapter.

The overview of the meta adapter is shown in Fig. 5. Once we extracted the text representation v_t, we will extract the query information regarding the object, and inject it into the initial object query, which is randomly initialized. The fusion module, composed of the conventional multi-head attention module, performs extraction and injection. We use the random initialized query v_t^0 as Q, and the text embedding v_t as V and K. Following the definition in Eq. (1), we can form the object query as:

$$v_t^+ = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{3}$$

where v_t^+ is the enhanced object query that contains the discriminative knowledge from the text embedding v_t.

3.5. Cross-modality interaction and density map regression

Given spatial representation v_i^+ and the object query v_t^+, we can perform cross-modality interaction. The pipeline of cross-modality interaction is shown in Fig. 4-(b). To this aim, we design the multi-modal transformer composed of the multi-head self-attention layer (MHSA). It takes v_i^+ as input, and multi-head cross-attention layer (MHCA) that takes the output of the MSHA layers as a query and v_t^+ as key and value to perform the knowledge transfer and interaction. Then, a two-layer feed-forward network is appended after the MHCA to refine the feature representation. Finally, given the cross-modality fused representation v, we use a conventional CNN-based decoder to generate the density map $\hat{y}$, from which we can calculate the number of interested objects. The process of density map regression is illustrated in Fig. 4-(c).

4. Experimental results and analysis

4.1. Implementation detail

All experiments were conducted using the PyTorch deep learning framework [12], and with an NVIDIA A100 GPU. The Adam optimizer with a weight decay of 5×10^{-2} was employed to optimize the learnable parameters model. The learning rate was set to 10^{-5}. The batch size was set to 32, and the model was trained for 200 epochs to ensure the convergence. We used the Grounding DINO [13] with Swin Transformer to extract the spatial prior with a dimension of 256.

Table 1
Details of the datasets adopted for comparison.

Dataset	# Year	# Categorize	# Images	# Train	# Val	# Test	Average resolution	# Min	# Max	# Avg	# Total
FSC-147 [23]	2021	147	6135	3659	1286	1190	774 * 938	7	56	3701	343,818
ShanghaiTech Part A [40]	2016	1	482	300	–	182	589 * 868	33	3139	501	241,677
ShanghaiTech Part B [40]	2016	1	716	400	–	316	768 * 1024	9	578	123	88,488
CARPK [41]	2017	1	1448	989	–	459	720 * 1280	1	188	62	89,777
UCF_CC_50 [42]	2013	1	50	40	–	10	2101 * 2888	94	4543	1279	63,974
UCF-QNRF [43]	2018	1	1535	1201	–	334	2013 * 2902	49	12,865	815	1,251,642
JHU-Crowd++ [44]	2020	1	4372	2272	500	1600	910 * 1430	0	25,791	346	1,515,005
NWPU-Crowd [45]	2020	1	5109	3109	500	1500	2191 * 3209	0	20,033	418	2,133,375

4.2. Benchmarking datasets

The essential information of benchmarking datasets is shown in Table 1.

FSC-147 [23] is a meticulously annotated collection of images designed specifically for class-agnostic object-counting research. The images within each category are non-overlapping, primarily consisting of kitchen utensils, office supplies, stationery, vehicles, and animals. Each image in the dataset undergoes careful annotation, and it serves as fundamental ground truth data for model evaluation. The annotations offer detailed insights into the spatial distribution of objects within the images. In the experiments, we utilize the class names as textual input, without employing annotations on image patches.

ShanghaiTech [40] is a large-scale crowd-counting dataset comprising 1198 annotated images. The dataset is divided into two subsets: Part A and Part B. Part A images are derived from the internet, featuring densely populated targets. In contrast, Part B images are authentic captures of bustling streets in Shanghai, exhibiting sparser target distributions. The disparate origins of these two segments pose challenges for cross-scene evaluations.

CARPK [41] is an image dataset designed for vehicle counting. It comprises 1148 bird's-eye-view images of parking lots capturing vehicles under various time and weather conditions. The dataset comprises 89,777 cars, with diverse scenarios illustrated in density, occlusion, and scale. All images in the dataset are annotated, providing counting information for vehicles and pedestrians and serving as a benchmark for evaluation.

UCF_CC_50 The UCF_CC_50 dataset [42] includes 50 images, each exhibiting extremely high crowd density. The annotations per image range from 94 to 4543, with an average of 1280 pedestrians per image. In accordance with [46,47], we randomly divided the dataset into five parts for cross-validation.

UCF-QNRF [43] features a large-scale crowd with diverse scenes, multiple perspectives, and variations in lighting. The dataset includes 1535 high-resolution images, each averaging 2013 × 2902 pixels in size.

JHU-Crowd++ [44] is composed of 4250 images which include numerous images featuring weather-based degradations and illumination variations.

NWPU-Crowd [45] includes 5109 images and 2,133,375 head annotations. It is a large-scale crowd counting dataset obtained from the internet and it contains the presence of negative samples.

4.3. Evaluation metrics

Following prior researches [46,48–51], the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were employed as metrics for evaluating. The MAE was adopted to assess the accuracy of the model. It is mathematically denoted as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|, \tag{4}$$

where N represents the total number of images in the test set, y_i denotes the ground truth of the actual number of people in the ith image, and $\hat{y}_i$ corresponds to the total predicted count from the density map for the same image. The advantage of MAE lies in its insensitivity to outliers, as it solely considers absolute differences.

However, due to the nature of absolute values, MAE cannot provide deeper insights into the analysis of squared errors. Conversely, RMSE was utilized to evaluate the robustness of the model. It is formulated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2}. \tag{5}$$

In comparison to MAE, the primary advantage of RMSE is its sensitivity to large errors, thereby revealing inadequacies in the performance of the model on certain samples.

4.4. Experiments on FSC-147 dataset

Table 2 presents the objective comparison results of the proposed method DSPI against State-Of-The-Art (SOTA) methods on the FSC-147 [23] dataset. In comparison to the CLIP-Count [12], both MAE and MSE have shown an improvement of 10.6% and 12.5% on the validation set, which indicates superior counting performance over advanced zero-shot counting methods. Due to the incorporation of spatial priors, the DSPI can effectively distinguish between real objects and background noise as shown in Fig. 13. The attention is focused on foreground objects and thus enables a more accurate prediction of object quantity. To comprehensively assess the performance of the counting model, we included comparisons with several few-shot and reference-less counting methods in Table 2. It is observed that the proposed method DSPI achieved a reduction of 30.9% and 24.5% in MAE and RMSE on the validation set, and 23.7% in MAE on the test set, compared to the SOTA few-shot method FamNet [23], which relies on a limited number of labelled samples. When compared to the reference-less counting method CounTR [8], which operates without the need for labelled data, the proposed method DSPI achieves reductions of 7.0% and 25.5% in MAE and RMSE on the validation set, and 5.0% in RMSE on the test set. This further validates the exceptional performance of the proposed method DSPI not only in zero-shot scenarios with high accuracy and robustness but also in handling few-shot and reference-less scenarios.

Fig. 6 visualizes the estimated density map on the FSC-147 dataset. The first row consists of the original input images, and the second row showcases the visualization of spatial prior information obtained through the Grounding DINO pre-trained model. The third row is the corresponding ground truth (GT). The "GT" is obtained by summing the object center point coordinates from the dataset labels. The fourth row and the last row respectively exhibit the predicted density maps overlaid on the original images by Clip-Count and the proposed DSPI. The "Pred" represents the number of people predicted by the network, which is obtained by summing the pixels on the density map. Evidently, the proposed DSPI model maximally leverages spatial and textual prior information, accurately counting different types of objects with guidance from textual prompts. Moreover, the predicted density maps exhibit spatial consistency with the ground truth density distributions.

Table 2
Objective comparison results on the FSC-147 dataset. The best results are highlighted in **bold**.

Scheme	Method	Source	#Shot	Val set		Test set	
				MAE	RMSE	MAE	RMSE
Few-shot	FamNet [23]	CVPR2021	3	24.32	70.94	22.56	101.54
	CFOCNet [52]	WACV2021	3	21.19	61.41	22.10	112.71
	CounTR [8]	BMVC2022	3	13.13	49.83	11.95	91.23
	LOCA [5]	ICCV2023	3	10.24	32.56	10.97	56.97
	FamNet [23]	CVPR2021	1	26.05	77.01	26.76	110.95
Reference-less	FamNet* [23]	CVPR2021	0	32.15	98.75	32.27	131.46
	RepRPN-C [10]	ACCV2022	0	29.24	98.11	26.66	129.11
	CounTR [8]	BMVC2022	0	18.07	71.84	14.71	106.87
	LOCA [5]	ICCV2023	0	17.43	54.96	16.22	103.96
	RCC [9]	arXiv2022	0	17.49	58.81	17.12	104.53
Zero-shot	Xu et al. [7]	CVPR2023	0	26.93	88.63	22.09	115.17
	Clip-Count [12]	MM2023	0	18.79	61.18	17.78	106.62
	DSPI (Ours)	–	0	**16.80**	**53.56**	**17.22**	**101.48**

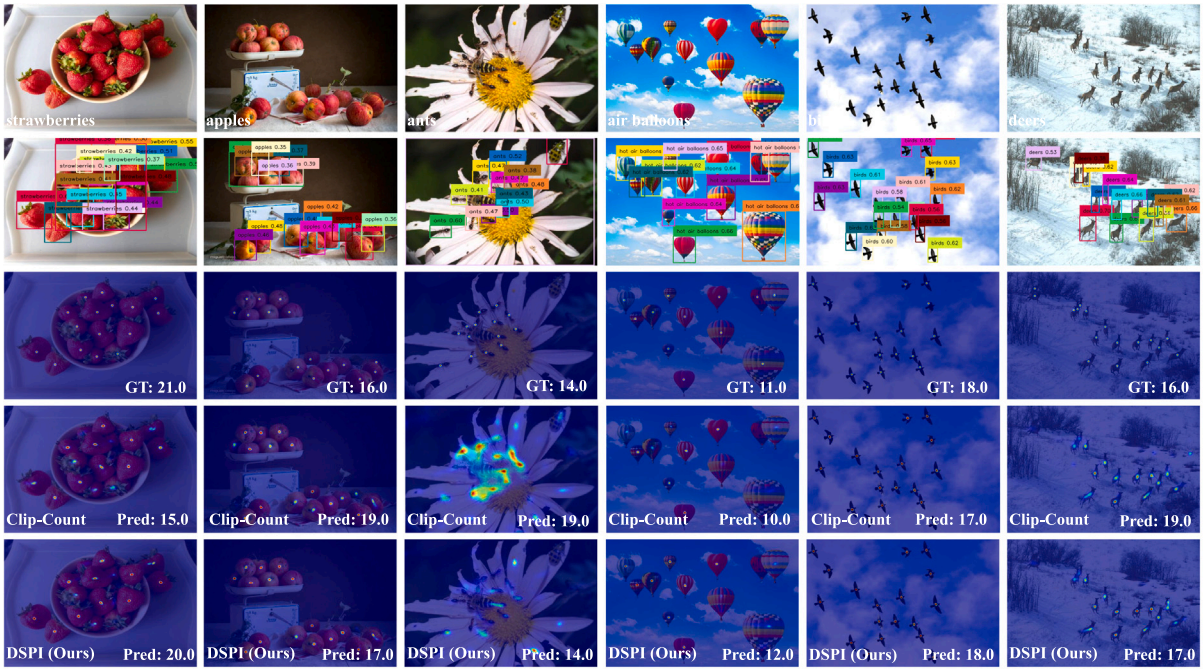


Fig. 6. Visualization of the input image and generated density maps for the samples from the FSC-147 dataset.

4.5. Experiments on ShanghaiTech dataset

We also tested the cross-domain generalization ability of the DSPI model. We directly utilized the model trained on the FSC-147 dataset and tested it on the ShanghaiTech test set. During this process, we only needed to update the input textual prior information to "person" to specify the target crowd to be counted. The objective comparison results are shown in Table 3. It can be observed that the proposed method outperforms the reference-less counting method RCC [9] and the zero-shot counting method CLIP-Count [12]. Specifically, on the Part A dataset, the MAE and RMSE are reduced by 11.2% and 11.0% compared to CLIP-Count [12]. On the Part B dataset, MAE and RMSE are reduced by 6.3% and 8.9% compared to CLIP-Count [12], respectively. This indicates the advanced generalization capability of the proposed method. The subjective results shown in Fig. 7 further verify the effectiveness of the proposed model on ShanghaiTech, especially in cross-dataset scenarios. The experimental results indicate that the density map generated by the proposed model accurately represents crowd distribution.

4.6. Experiments on CARPK dataset

We also tested the cross-domain generalizability of DSPI the model on the CARPK [41] dataset. Similar to the ShanghaiTech [40] dataset, the model was trained on FSC-147 without fine-tuning and directly tested on the CARPK dataset. The input textual prior information was set to "car" to specify the target object to be counted. The objective comparison results are shown in Table 4. Compared with the RCC [9], the proposed method DSPI achieved reductions of 46.2% and 36.8% in MAE and RMSE, respectively. When compared with the few-shot counting method BMNet [24], the proposed method DSPI demonstrated decreases of 20.2% and 32.8% in MAE and RMSE, respectively. These consistent improvements further validate the superiority of the proposed method DSPI in counting tasks. Visualization results on the CARPK dataset are illustrated in Fig. 8. The CARPK scenarios present substantial background clutter. The samples demonstrate that the proposed method can suppress cluttered backgrounds.

Table 3
Cross-dataset evaluation on ShanghaiTech crowd counting dataset. The best results are highlighted in **bold**.

Method	Type	Training ⟶ Testing	MAE	RMSE	Training ⟶ Testing	MAE	RMSE
MCNN [34]	Specific	Part A ⟶ Part B	85.2	142.3	Part B ⟶ Part A	221.4	357.8
CrowdCLIP [53]			69.6	80.7		217.0	322.7
RCC [9]	Generic	FSC147 ⟶ Part B	66.6	104.8	FSC147 ⟶ Part A	240.1	366.9
Clip-Count [12]			45.7	77.4		192.6	308.4
DSPI (Ours)			**42.8**	**70.5**		**171.1**	**274.4**

Fig. 7. Visualization of the input image and generated density maps for the samples from the ShanghaiTech dataset.

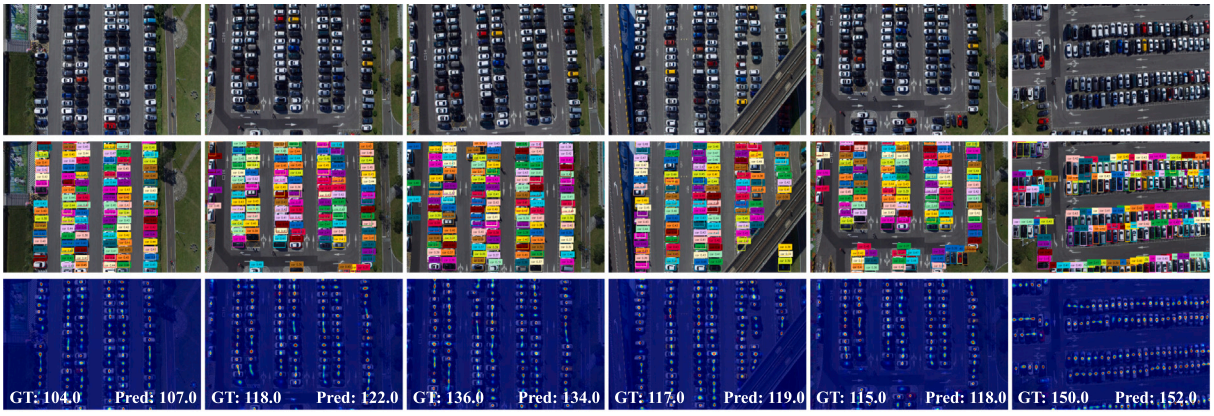


Fig. 8. Visualization of the input image and generated density maps for the samples from the CARPK dataset.

Table 4
Cross-dataset evaluation on CARPK dataset. The best results are highlighted in **bold**.

Method	#Shot	MAE	RMSE
FamNet [23]	3	28.84	44.47
BMNet [24]	3	14.41	24.60
BMNet+ [24]	3	**10.44**	**13.77**
RCC [9]	0	21.38	26.15
Clip-Count [12]	0	11.96	16.61
DSPI (Ours)	0	**11.50**	**15.52**

4.7. Experiments on other dense crowd datasets

To further validate the generalization ability of the proposed model, we perform cross-domain analysis on three other dense crowd datasets, namely UCF_CC_50 [42], UCF-QNRF [43], JHU-Crowd++ [44], NWPU-Crowd [45]. The UCF_CC_50 [42] dataset comprises 50 images with extremely high crowd density, while the UCF-QNRF [43] dataset consists of 1535 high-dense images. The JHU-Crowd++ [44] dataset includes a considerably large number of samples (*i.e.*, 4372 images). Additionally, the NWPU-Crowd dataset [45] contains 5109 images. This dataset presents several challenges, including negative samples, high resolution, and significant appearance variations. The CLIP-Count [12] is adopted as the competitor, and it employs multimodel information to count the crowds. Comparative results are reported in Table 5. It proves that the DSPI outperforms CLIP-Count [12] in terms of MAE and RMSE, which verifies the generalization ability of the proposed method on dense crowd datasets. Subjective results in Fig. 9 illustrate a visual comparison between DSPI and CLIP-Count [12] on the dense crowd dataset. It demonstrates the proposed method in accurately predicting the number while precisely crowd density distribution.

4.8. Component analysis

Ablation study on pivotal components To validate the effectiveness of the adapter module and Prior module in the proposed DSPI model, extensive ablation experiments were conducted on the FSC-147 dataset, and the objective comparison results are presented in Fig. 10. Considering the complexity of the network and the computational resource requirements, we detailed the learnable parameters and computation costs of pivotal modules in Table 6. The input image size is 384 × 384.

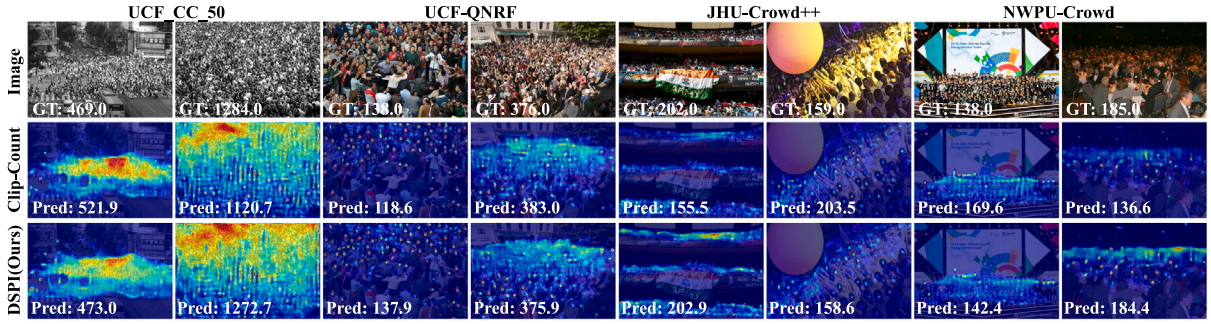


Fig. 9. Visualization of the input image and generated density maps for the samples from the dense crowd datasets.

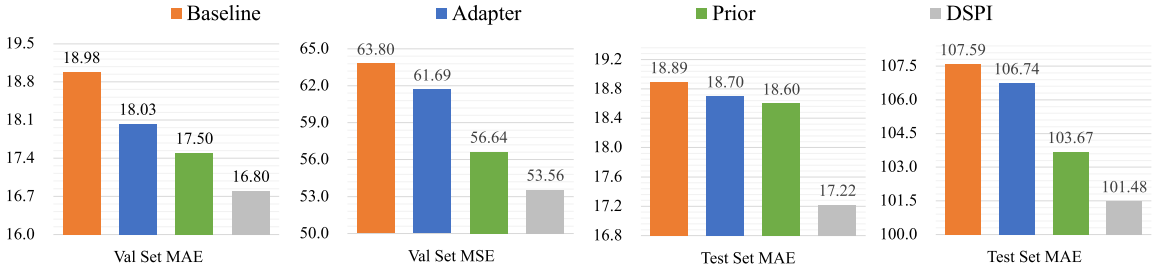


Fig. 10. Performance analysis comparison of the Adapter and Prior modules in the proposed DSPI model.

Table 5
Cross-dataset evaluation on dense crowd counting datasets. The best results are highlighted in **bold**.

Method	Type	Training ⟶ Testing	MAE	RMSE
Clip-Count [12]	Generic	FSC147 ⟶ UCF_CC_50	739.1	992.2
DSPI (Ours)			**676.5**	**915.2**
Clip-Count [12]	Generic	FSC147 ⟶ UCF-QNRF	452.9	765.5
DSPI (Ours)			**414.2**	**705.9**
Clip-Count [12]	Generic	FSC147 ⟶ JHU-Crowd++	201.9	682.4
DSPI (Ours)			**187.6**	**656.7**
Clip-Count [12]	Generic	FSC147 ⟶ NWPU-Crowd	316.2	962.7
DSPI (Ours)			**302.2**	**941.4**

Table 6
Comparison results of the pivotal components on the model complexity.

Methods	Params (M)	GFLOPs
Baseline	16.36	123.06
Baseline + Prior	64.46	124.69
Baseline + Adapter	20.58	123.13
DSPI (Ours)	68.67	124.76

1. **Baseline** represents the baseline model without the Adapter and Prior modules. It can be observed that the performance is not optimal.
2. **Adapter** introduces the Adapter module on top of the Baseline. As shown in Fig. 10, MAE increases by 1.01%, and RMSE decreases by 0.79% compared with the baseline model on the test set. This indicates that the Adapter module enhances the modal alignment between text and targets in certain image samples, making matching text and targets more accurate by incorporating visual information.
3. **Prior** incorporates the Prior module on top of the Baseline. It is evident from Fig. 10 that by adding spatial prior position information of targets, the model achieves a reduction of 1.54% in MAE and 3.64% in RMSE compared with the baseline model on the test set. This validates the effectiveness of the Prior module.
4. **DSPI** simultaneously introduces the Adapter and Prior modules on top of the Baseline. Compared to introducing only the Prior module on the test set, MAE and RMSE decreased by 8.84% and 5.68%, respectively. This implies that the Adapter module further improves counting accuracy and robustness on the foundation of the Prior module. Therefore, the decreasing trend of errors in Fig. 10 demonstrates the effectiveness of the Adapter and Prior modules in the DSPI model.

Ablation study on the Adapter To investigate the effect of the Adapter on textual features, we conducted ablation experiments on top of the Baseline with the added Prior Module. The comparative results are illustrated in Fig. 11. Within the meta adapter, we randomly initialized learnable visual queries with dimensions of 2^n to extract instructive knowledge from the text probe. Through these ablation experiments, we verified the role of the Adapter module and analysed the impact of different dimensions of initial visual queries on its counting performance. As depicted in Fig. 11, appropriately setting the dimensions of initial visual queries maximizes the effectiveness of the Adapter. Additionally, we observed that when the initial visual query dimension is 16, the adapter could effectively extract and reinforce crucial information from textual features, thereby reducing noise and redundancy within them. This enhanced the ability of the model to capture semantic relationships between text and visual information, and thus improve the overall performance.

Visualization of the Prior and Adapter To further analyse the impact of Prior and Adapter modules on counting performance, we visualized the intermediate visual features of each module in Fig. 13. The second column reveals that incorporating the Prior module reduces the focus on irrelevant background information. The third column introduces the Adapter module, which further integrates text features and concentrates attention more directly on the counting objects corresponding to textual cues. It can be observed that the DSPI aligns the visual features of the text by integrating information from the Prior and Adapter modules, and thus effectively suppress background interference and enhance the generalization ability. The density map accurately reflects the distribution of objects, to demonstrate the effectiveness of the Prior and Adapter modules.

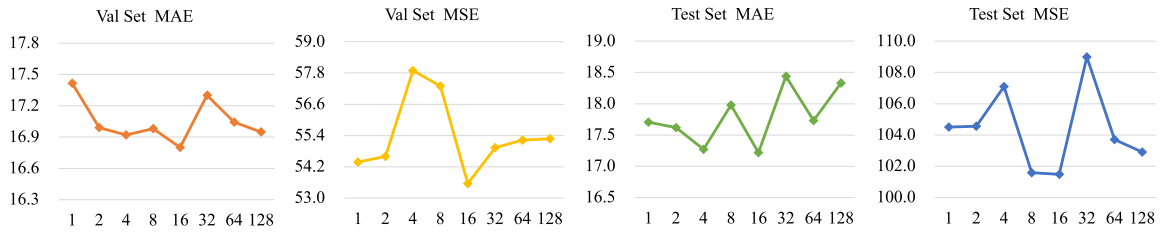


Fig. 11. The ablation analysis of the Adapter module on the FSC-147 dataset. The horizontal axis represents the dimension sizes of learnable initial visual queries 2^n. The vertical axis of the two left subplots represents MAE or MSE on the validation set, while the vertical axis of the right two subplots represents MAE or MSE on the test set.

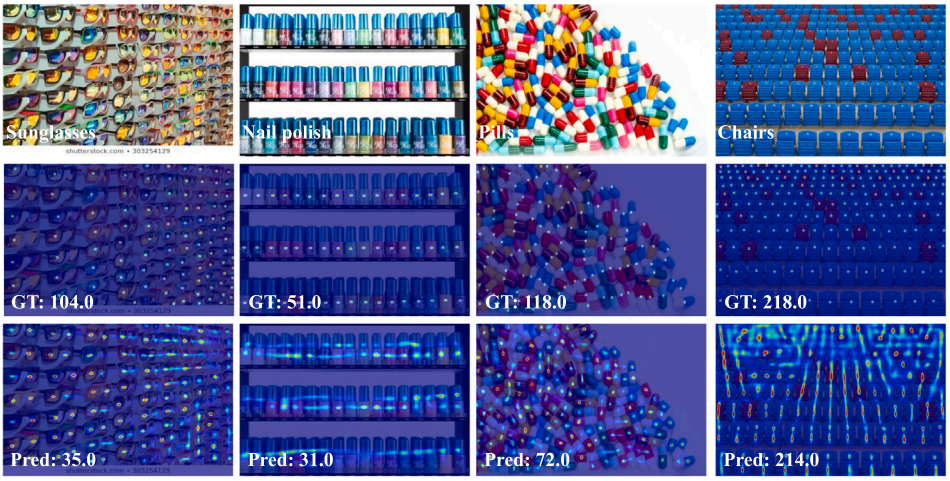


Fig. 12. Visualization of failure cases.

4.9. Failure cases

The proposed DSPI may encounter limitations in certain scenarios due to ambiguous text guidance, as shown in Fig. 12. In some cases, if an object has two identifiable components, it may be counted as two separate objects. This limitation primarily arises from insufficient utilization of textual information to comprehend the integrity of targets. In future work, we tend to introduce additional contextual information based on textual prompts and design region-merging strategies to help the model better understand the holistic nature of objects and mitigate the ambiguity of query objects.

5. Conclusion

In this paper, we identified the limitation of the existing class-agnostic counting model, include the lack of sensitivity to location information and potential misalignment in the hypothesis space. To solve these two problems, we proposed Deep Spatial Prior Interaction (DSPI). The DSPI leverages the spatial-awareness ability of pre-trained object grounding model and incorporates spatial location as an additional prior for a specific query class, enabling a more precise focus on the object's exact location. Additionally, we designed a meta adapter to align feature spaces across different modalities. The proposed model demonstrated superior performance through extensive experiments on diverse benchmarks. It showcased the effectiveness in addressing the identified challenges and advancing class-agnostic counting in a multi-modal context.

CRediT authorship contribution statement

Jinyong Chen: Data curation, Conceptualization. **Qilei Li:** Funding acquisition, Formal analysis. **Mingliang Gao:** Methodology, Investigation. **Wenzhe Zhai:** Validation, Software. **Gwanggil Jeon:** Supervision, Resources, Data curation. **David Camacho:** Software, Resources, Project administration.

Declaration of competing interest

None Declared.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work has been funded by Grants: PLEC2021-007681 (XAI-DisInfodemics), PID2020-117263GB-100 (FightDIS), and PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"; by Calouste Gulbenkian Foundation, United Kingdom, under the project MuseAI - Detecting and matching suspicious claims with AI, and by "Convenio Plurianual with the Universidad Polit'ecnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario".

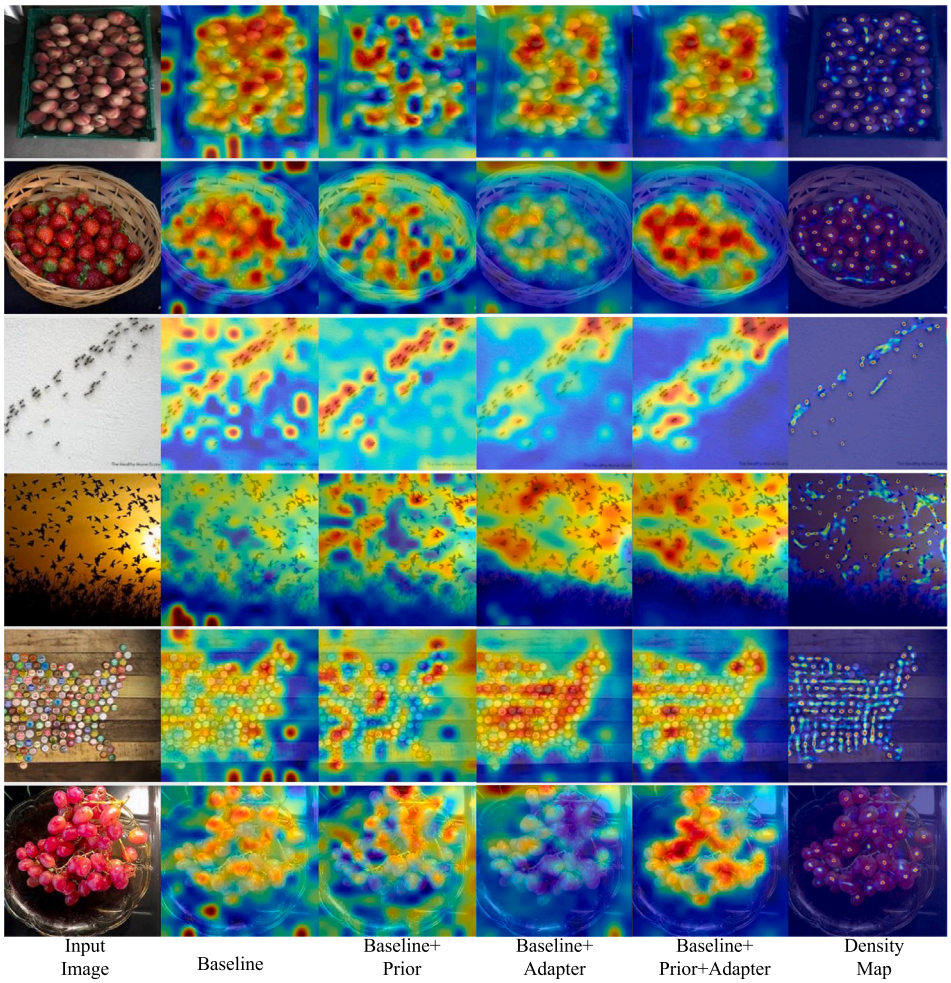


Fig. 13. Visualization of intermediate features. The first column represents the input sample images. The second column depicts the intermediate visual features of the baseline model. The third column illustrates the addition of the Prior module on top of the baseline model. The fourth column illustrates the addition of the Adapter module on top of the baseline model. The fifth column displays the intermediate visual features of the proposed DSPI. The last column showcases the density map.

References

[1] T. Han, L. Bai, J. Gao, Q. Wang, W. Ouyang, Dr. vic: Decomposition and reasoning for video individual counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3083–3092.

[2] L. Liu, J. Chen, H. Wu, G. Li, C. Li, L. Lin, Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4823–4833.

[3] S. Zhang, G. Wu, J.P. Costeira, J.M. Moura, Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3667–3676.

[4] S. Dehaene, The Number Sense: How the Mind Creates Mathematics, OUP USA, 2011.

[5] N. Djukic, A. Lukezic, V. Zavrtanik, M. Kristan, A low-shot object counting network with iterative prototype adaptation, in: Proceedings of the International Conference on Computer Vision, ICCV, 2023, pp. 18872–18881.

[6] M. Wang, Y. Li, J. Zhou, G.W. Taylor, M. Gong, Gcnet: Probing self-similarity learning for generalized counting network, Pattern Recognit. (2024) 110513.

[7] J. Xu, H. Le, V. Nguyen, V. Ranjan, D. Samaras, Zero-shot object counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15548–15557.

[8] L. Chang, Z. Yujie, Z. Andrew, X. Weidi, Countr: Transformer-based generalised visual counting, in: Proceedings of the British Machine Vision Conference, BMVC, 2022, pp. 1–15.

[9] M. Hobley, V. Prisacariu, Learning to count anything: Reference-less class-agnostic counting with weak supervision, arXiv preprint arXiv:2205.10203.

[10] V. Ranjan, M.H. Nguyen, Exemplar free class agnostic counting, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 3121–3137.

[11] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[12] R. Jiang, L. Liu, C. Chen, Clip-count: Towards text-guided zero-shot object counting, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4535–4545.

[13] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, arXiv preprint arXiv:2303.05499.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.

[15] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, J. Wang, Group detr: Fast detr training with group-wise one-to-many assignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6633–6642.

[16] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, L. Zhang, Dynamic head: Unifying object detection heads with attentions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7373–7382.

[17] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, H. Hu, Detrs with hybrid matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19702–19712.

[18] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, L. Zhang, Dab-detr: Dynamic anchor boxes are better queries for detr, in: International Conference on Learning Representations, 2022, pp. 1–15.

[19] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, H.-Y. Shum, Dino: Detr with improved denoising anchor boxes for end-to-end object detection, in: The Eleventh International Conference on Learning Representations, 2022, pp. 1–15.

[20] A. Zareian, K.D. Rosa, D.H. Hu, S.-F. Chang, Open-vocabulary object detection using captions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14393–14402.

[21] X. Gu, T.-Y. Lin, W. Kuo, Y. Cui, Open-vocabulary object detection via vision and language knowledge distillation, in: International Conference on Learning Representations, 2021, pp. 1–15.
[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2020, pp. 1–15.
[23] V. Ranjan, U. Sharma, T. Nguyen, M. Hoai, Learning to count everything, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3394–3403.
[24] M. Shi, H. Lu, C. Feng, C. Liu, Z. Cao, Represent, compare, and learn: A similarity-aware framework for class-agnostic counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9529–9538.
[25] Z. You, K. Yang, W. Luo, X. Lu, L. Cui, X. Le, Few-shot object counting with similarity-aware feature enhancement, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6315–6324.
[26] W. Lin, K. Yang, X. Ma, J. Gao, L. Liu, S. Liu, J. Hou, S. Yi, A.B. Chan, Scale-prior deformable convolution for exemplar-guided class-agnostic counting, in: British Machine Vision Conference, 2022, pp. 0–15.
[27] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020, pp. 1–15.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30.
[30] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Neural Inf. Process. Syst. (2017) 0–15.
[31] W. Zhai, Q. Li, Y. Zhou, X. Li, J. Pan, G. Zou, M. Gao, Da2net: a dual attention-aware network for robust crowd counting, Multimedia Syst. 29 (5) (2023) 3027–3040.
[32] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[34] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 589–597.
[35] Y. Li, X. Zhang, 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, 2018, pp. 1091–1100.
[36] Y. Zhang, D. Zhao, J. Sun, G. Zou, W. Li, Adaptive convolutional neural network and its application in face recognition, Neural Process. Lett. 43 (2016) 389–399.
[37] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8198–8207.
[38] M.K.K. Reddy, M. Hossain, M. Rochan, Y. Wang, Few-shot scene adaptive crowd counting using meta-learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2814–2823.
[39] H. Zhu, J. Yuan, X. Zhong, Z. Yang, Z. Wang, S. He, Daot: Domain-agnostically aligned optimal transport for domain-adaptive crowd counting, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4319–4329.
[40] Y. Zhang, D. Zhou, S. Chen, S. Gao, 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, 2016, pp. 589–597.
[41] M.-R. Hsieh, Y.-L. Lin, W.H. Hsu, Drone-based object counting by spatially regularized regional proposal network, in: Proceedings of the International Conference on Computer Vision, ICCV, ICCV, 2017, pp. 4165–4173.
[42] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2013, pp. 2547–2554.
[43] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, M. Shah, Composition loss for counting, density map estimation and localization in dense crowds, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 532–546.
[44] V.A. Sindagi, R. Yasarla, V.M. Patel, Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2594–2609.
[45] Q. Wang, J. Gao, W. Lin, X. Li, Nwpu-crowd: A large-scale benchmark for crowd counting and localization, IEEE Trans. Pattern Anal. Mach. Intell. 43 (6) (2020) 2141–2149.
[46] W. Zhai, M. Gao, X. Guo, Q. Li, G. Jeon, Scale-context perceptive network for crowd counting and localization in smart city system, IEEE Internet Things J. 10 (21) (2023) 18930–18940.
[47] W. Zhai, M. Gao, Q. Li, G. Jeon, M. Anisetti, Fpanet: feature pyramid attention network for crowd counting, Appl. Intell. (2023) 1–18.
[48] X. Guo, M. Gao, W. Zhai, Q. Li, G. Jeon, Scale region recognition network for object counting in intelligent transportation system, IEEE Trans. Intell. Transp. Syst. 24 (12) (2023) 15920–15929.
[49] X. Guo, M. Gao, G. Zou, A. Bruno, A. Chehri, G. Jeon, Object counting via group and graph attention network, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–12.
[50] J. Chen, M. Gao, X. Guo, W. Zhai, Q. Li, G. Jeon, Object counting in remote sensing via selective spatial-frequency pyramid network, Softw. - Pract. Exp. (2023) 1–15.
[51] J. Chen, M. Gao, Q. Li, X. Guo, J. Wang, X. Xing, et al., Privacy-aware crowd counting by decentralized learning with parallel transformers, Internet Things 26 (2024) 101167.
[52] S.-D. Yang, H.-T. Su, W.H. Hsu, W.-C. Chen, Class-agnostic few-shot object counting, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 870–878.
[53] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu, X. Bai, Crowdclip: Unsupervised crowd counting via vision-language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2893–2903.