Zero-shot Object Counting with Vision-Language Prior Guidance Network

Wenzhe Zhai, Xianglei Xing, Mingliang Gao, Qilei Li

***Abstract*—The majority of existing counting models are designed to operate on a singular object category, such as crowds or vehicles. The emergence of multi-modal foundational models, *e.g.*, Contrastive Language-Image Pre-training (CLIP), has paved the way for class-agnostic counting. This approach facilitates the counting of objects across diverse classes within a single image based on textual indications. However, class-agnostic counting models based on CLIP confront two primary challenges. Firstly, the CLIP model exhibits limited sensitivity towards location information, which prioritizes global content over the precise localization of objects. Therefore, directly employing the CLIP model is regarded as suboptimal. Secondly, these models commonly employ frozen pre-trained vision and language encoders while disregarding potential misalignment within the constructed hypothesis space. In this paper, we propose a unified framework, named the Vision-Language Prior Guidance (VLPG) Network, to tackle these two challenges. The VLPG consists of three key components, namely the Grounding DINO module, Spatial Prior Calibration (SPC) module, and Object-Centric Alignment (OCA) module. The Grounding DINO module utilizes the spatial-awareness capability of extensive pre-trained object grounding models to incorporate the spatial position as an additional prior for a particular query class. This adaptation enables the network to concentrate more precisely on the exact location of the objects. Meanwhile, the SPC module is built to extract the long-range dependencies and local regions of the spatial position. Additionally, to align the feature space across different modalities, we design an OCA module that condenses textual information into an object query which serves as an instruction for cross-modality matching. Through the collaborative efforts of these three modules, multimodal representations are aligned while maintaining their discriminative nature. Comprehensive experiments conducted on various benchmarks validate the effectiveness of the proposed model.**

***Keywords*—Zero-shot object Counting, Multi-modal foundational model, Vision-language prior guidance network, Cross-modality.**

I. Introduction

IN the past decades, object-specific counting has played a considerable role in many real-world applications [1]–[3]. Nonetheless, current models frequently encounter difficulties in extending to new object categories not seen during training, which limits their practicality across various real-world contexts [4]–[6]. Therefore, there is an urgent need for a versatile counting model that can adjust to unseen categories and provide corresponding density estimates [7]–[9].

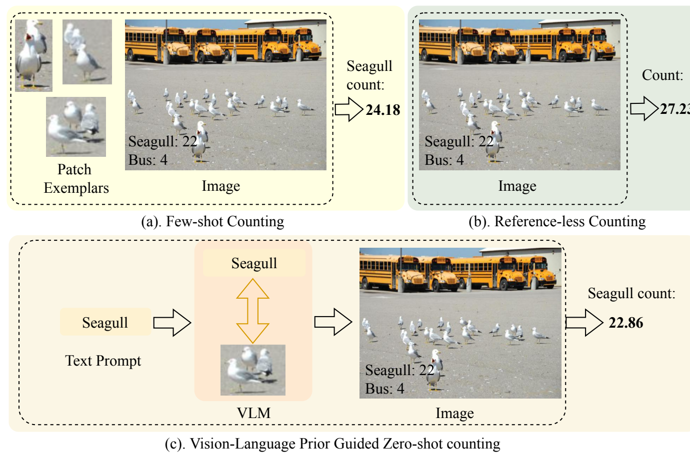


Fig. 1. Schema of few-shot counting, reference-less counting, and Vision-language Prior Guided (VLPG) Zero-shot counting. In contrast to conventional methods, the proposed VLPG model does not require specific image patch labels or counting all salient objects in the image. Instead, it counts objects of any category specified by text prompts. It is worth noting that the numbers on the image represent the actual quantities of all categories of objects, while the output numbers indicate the predicted quantity of a specified category.

This demand has resulted in the emergence of class-agnostic counting models [10]–[12]. These models adopt a unified/shared approach to estimate the quantity and density of objects within a given image, as depicted in Fig. 1-(a). By annotating specific image patches as exemplars and subsequently assessing the similarities between these exemplars and various image regions, these models have demonstrated notable generalization and counting accuracy. However, the majority of class-agnostic counting methods rely on the unrealistic assumption that object bounding boxes are available during inference, which is not realistic in practical application. Consequently, they necessitate users to manually annotate certain object samples for counting, which can be cumbersome and time-consuming. Moreover, the substantial intra-class variability among query objects may lead to biased counts [12], [13]. To tackle these issues, reference-less counting methods have been proposed to detect and count salient objects without annotations during inference [14], [15]. Although these methods alleviate the need for manual annotation, they struggle to specify the object category of interest in the presence of multiple categories, as illustrated in Fig. 1-(b). Overall, existing counting models exhibit relatively limited flexibility and are

The work is supported by the National Natural Science Foundation of China No. 62076078 and the CAAI-Huawei MindSpore Open Fund No. CAAIXSJLJJ-2020-033A. (Corresponding author: Xianglei Xing)

Wenzhe Zhai and Xianglei Xing are with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, 150001, China. (e-mail: wenzhezhai@163.com and xingxl@hrbeu.edu.cn.)

Mingliang Gao, Qilei Li is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China. (e-mail: mlgao@sdut.edu.cn, qilei@ieee.org)

challenging to apply in real-world scenarios.

Contrastive Language-Image Pre-training (CLIP) [16] is an effective and scalable method. It utilizes natural language supervision to learn semantic alignments between images and text, which enables robust generalization of CLIP even in the absence of annotations. Jiang *et al.* [17] proposed a recent variant, namely CLIP-Count, which employs a static vision encoder to extract visual features from input images and a textual encoder to capture the textual representation of the object category intended for counting. Unlike existing referenceless counting methods, it does not require any additional samples for fine-tuning the model for the target object, which makes domain-agnostic counting more feasible. However, the direct application of CLIP encoders to the model architecture, as demonstrated in CLIP-Count [17], has two inherent limitations. (1) CLIP undergoes pre-training through contrastive analysis of visual and language representations, which facilitates object recognition within images while lacking precise spatial localization. Consequently, utilizing the vision encoder for feature extraction in counting tasks is suboptimal, given that object counting primarily depends on spatial distribution. (2) CLIP is pre-trained using natural images characterized by sparse object occurrences. Nevertheless, input images typically exhibit a denser distribution of objects in object counting tasks, leading to a shift in data distribution. Consequently, textual representations may deviate from their corresponding visual representations.

This study aims to tackle the aforementioned limitations by employing frozen CLIP for zero-shot object counting. To focus on spatial information within image representations, we propose the Vision-Language Prior Guidance (VLPG) Network. It leverages textual information for guidance and uses object bounding box annotations as prior information for class-agnostic counting. The proposed schema is illustrated in Fig. 1-(c). Specifically, we incorporate the Grounding DINO [18] as a training-free module to equip the network with extensive prior information concerning the spatial positioning of specific objects. The spatial prior extractor is frozen and does not introduce any further trainable parameters. Secondly, we incorporated a spatial prior calibration (SPC) module to capture both long-range dependencies and local regions associated with spatial positions. Besides, to address the challenge of density shift encountered when employing pre-trained CLIP encoders, we build the object-centric alignment (OCA) module. The OCA module serves as a bridge between textual instructions and visual queries. It is built to distill textual instructions into object queries, thereby promoting interaction with visual information. Consequently, this enhances the attentiveness of visual representations towards specific objects. In a nutshell, the key contributions of the paper are summarized as follows:

- A VLPG Network is proposed for zero-shot object counting. It can extract distinctive representations aligned with multi-modalities while incorporating positional information to suppress background interference and enhance the generalization capability of the network.
- An SPC module is built to enhance the visual representation by correcting deviations in the visual feature space. It can extract the long-range dependencies and local regions within regions of spatial position.
- An OCA module is established to extract instructive descriptors from the text and transform them into an object query aligned with the vision representation. It can tackle the misalignment between textual instructions and visual representations.

II. Related work

A. Prompt-based foundation model

The emergence of extended language models, such as ChatGPT, has revolutionized the field of natural language processing and extended its application to computer vision. These models are referred to as "foundation models" and have shown remarkable generalization capabilities in both zero-shot and few-shot scenarios. In computer vision, Contrastive Language-Image Pre-training (CLIP) [16] is a prominent foundational model that employs contrast learning to train text and image encoders. The CLIP model has emerged as a powerful tool for bridging the gap between text and images. By training on an extensive dataset of images and text, the CLIP model has unlocked the potential for tasks like image-text matching. It can understand images and their associated descriptions, enabling it to perform tasks like finding matching images for given textual queries.

In recent years, numerous object grounding models have been proposed. Carion *et al.* [19] proposed the DEtection TRansformer (DERT) model. It employed a Transformer to predict the class and location of objects within images. Zhang *et al.* [20] introduced the concept of dynamic anchor boxes in DINO. In this approach, each position query is represented as a four-dimensional anchor box, which is dynamically updated at every layer of the decoder. Liu *et al.* [21] utilized dynamic anchor boxes for query formulation in DETR. The box coordinates are directly used as queries for the Transformer decoder and are updated layer by layer. However, previous research only performed well when dealing with a limited label set, but their effectiveness diminished when addressing a broader range of labels. Grounding DINO [18] effectively addresses the challenges of complex label spaces and significantly improves performance under diverse labeling conditions. It effectively captures the precise spatial positioning of objects and can create bounding boxes for various object categories. Moreover, the Grounding DINO fits into current multimodal designs to provide meaningful guidance information. The advent of foundation models has ushered in a transformative era in computer vision. These models can handle diverse data distributions without requiring explicit training on those specific instances.

B. Attention-based method

The attention mechanism enables the network to focus on the discriminative features in the input data. The attention mechanism has been widely applied in diverse network architectures, which encompass Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer-based networks [22]. It has been employed in

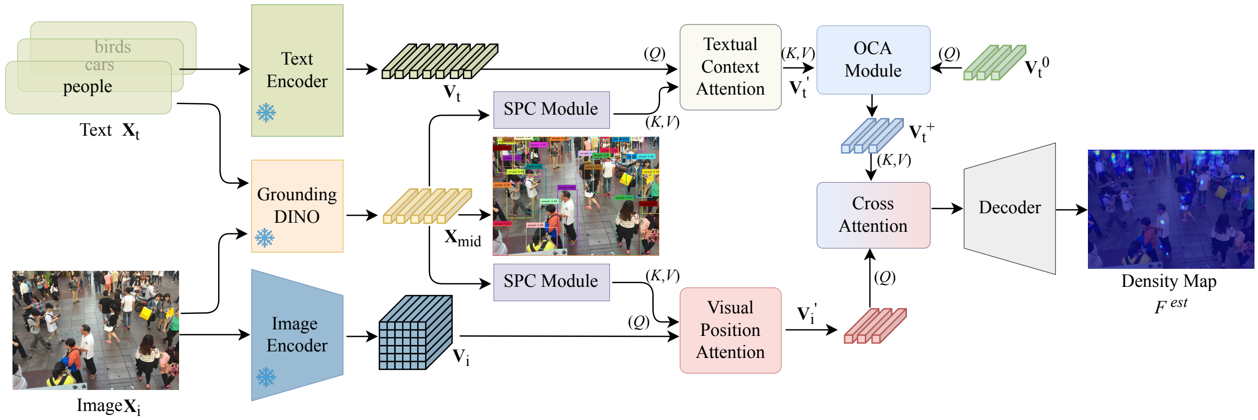


Fig. 2. Framework of proposed VLPG network. It integrates pre-trained image and text encoders from the CLIP model to extract image and text representations, respectively. To incorporate spatial context into the image representation, we utilize the multi-modal object detection model, *i.e.*, Grounding DINO module, to extract deep positional prior into the visual representation. Besides, a spatial prior calibration (SPC) module is utilized to capture both long-range dependencies and local regions within spatial positions. Furthermore, an object-centric alignment (OCA) module is established to translate text representations into visual features for cross-modality fusion. Finally, the density map is generated by the decoder.

diverse domains, such as semantic segmentation, object detection, and crowd counting [23]–[25]. Predominant attention mechanisms encompass spatial attention, channel attention, and self-attention mechanisms. The spatial attention prioritizes crucial regions within the input data and enhances the spatial context information. The channel attention mechanism primarily focuses on the channel dimension of input data, which augments the critical features within the channels. Woo *et al.* [26] introduced the Convolutional Block Attention Module (CBAM), which integrates channel attention and spatial attention. Fu *et al.* [27] presented the Dual Attention Network (DANet) which integrates local features and global dependencies to improve semantic segmentation performance.

The superiority of self-attention over traditional spatial and channel attention methodologies lies in its minimal reliance on external information and its enhanced ability to capture non-local correlations [28]–[30]. This characteristic facilitates the extraction of global information representations in transformer networks without employing traditional RNNs or CNNs. Both self-attention and cross-attention share a common core mechanism, yet their applications and purposes are different [31], [32]. Self-attention is specifically designed to handle relationships within a single sequence, while cross-attention addresses relationships between two distinct sequences. In this paper, we build the spatial positional prior that encodes the spatial position of the probe objects as hard-coded attention. This guidance mechanism aims to enhance the model's spatial awareness of the query objects.

C. Class-agnostic object counting

The class-agnostic object counting is broadly categorized into three groups according to the method of identification, *e.g.*, few-shot counting methods, reference-less counting methods, and zero-shot counting methods. Few-shot object counting involves estimating the object quantity in an image with a restricted number of training samples. This approach enables rapid learning and adaptation to new object categories in a short time, which provides flexibility and efficiency across diverse practical applications. FamNet [33] utilized ROI pooling to predict density maps and introduced a dataset for class-agnostic counting, known as FSC-147 [33]. The further advancement can be divided into two main aspects. One approach involves the utilization of advanced visual backbones, such as Vision Transformers (ViT), to enhance the extracted feature representations [10], [13], [34]. The second approach focuses on refining exemplar matching either by explicitly modeling exemplar-image similarity [35], [36] or by further incorporating exemplar guidance, as explored in [11], [37]. Despite the remarkable performance of these methods, they are not suitable in scenarios where samples are unattainable. Meanwhile, the method of reference-less counting has gained attention as an effective approach for class-agnostic counting that does not rely on human annotations. RepRPN-Counter [15] introduced a region proposal module tailored for extracting prominent objects, which eliminates the need for sampled inputs. RCC [14] used the pre-trained Vision Transformer [38], [39] to extract salient objects implicitly and directly regress a scalar for estimating object counts. Various contemporary few-shot counting models [10], [11] can be adapted for reference-less counting.

Despite their independence from specific samples, these approaches face a challenge in effectively specifying the object of interest, particularly in the presence of multiple object classes. Recently, zero-shot object counting methods have been proposed to facilitate end-to-end training without the need for patch-level supervision. Jiang *et al.*integrated Contrastive Language-Image Pre-training (CLIP) [16] into the counting network [17]. CLIP equips the model with the ability for zero-shot image-text alignment. To transfer 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 incorporated within the model. In

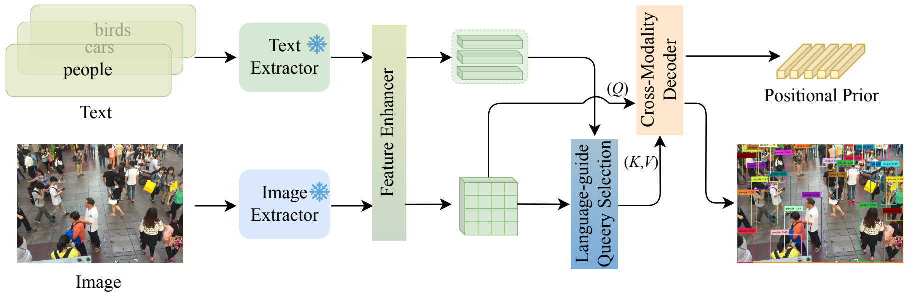


Fig. 3. Illustration of the positional prior. It is taken from the frozen Grounding DINO module. The image and text extractors are first utilized to extract the visual and textual features. Then, the similarity of visual and textual features is calculated by the language-guide query selection. Finally, the cross-modality decoder generates the positional prior.

this paper, we focus on zero-shot object counting given its practical application value.

III. Methodology

A. Framework overview

The flowchart of the proposed Vision-Language Prior Guidance (VLPG) Network is illustrated in Fig. 2. Initially, the visual image $\mathbf{X}_{\mathrm{i}}$ and the text instruction $\mathbf{X}_{\mathrm{t}}$ are employed as paired inputs. The VLPG utilizes two separate frozen CLIP encoders to encode both the image and the text, which facilitates interaction with cross-modal representations. First, the Grounding DINO [18] module is utilized to incorporate the spatial positional prior into the visual representations. Afterward, the spatial prior calibration (SPC) module is utilized to extract the long-range dependencies and local regions of the spatial position. Furthermore, the object-centric alignment (OCA) module is introduced to translate the text instruction into an object query, enabling effective cross-modal interaction. Finally, the network produces a density map, represented as $\mathbf{M} = F_\theta(\mathbf{X}_{\mathrm{i}}, \mathbf{X}_{\mathrm{t}})$, which accurately identifies the spatial positions of the target objects specified in the textual instructions.

B. Positional prior attentive injection

The visual depiction obtained through the CLIP vision encoder tends to emphasize the overall object categories in the given images while showing limited regard for the spatial position of objects. For counting the objects, it is essential to model the fine-grained location of the object. Nevertheless, the image encoder only focuses on image global information and is insensitive to the spatial position information of the objects. To improve the spatial perception ability of visual features, we apply the spatial priors extracted from the large-scale pre-trained Grounding DINO [18] model to focus on relevant object regions. The illustration of the positional prior extraction process is depicted in Fig. 3. It comprises five components: an image encoder, a text encoder, a feature enhancer, a text-guided selection querier, and a cross-modal decoder. First, visual and textual features are extracted using the visual encoder and text encoder, respectively. Subsequently, semantic consistency constraints are performed by the feature enhancer to align the visual and textual features. Then, the likelihood of the textual and visual features is calculated using the text-guided query selection to match the parts of the visual information that are related to the textual prompt and guide the model to focus on the object region. Lastly, the matched features are fed into the cross-modal decoder to generate the spatial positional prior $\mathbf{X}_{\mathrm{mid}}$. In particular, the positional prior contains spatial location information of local objects and global information of object distribution. By conducting further text-guided selection on the visual features, it will be transformed as query (Q), and the textual prompt information is transformed to key (K) and value (V), which are fed into the cross-modality decoder for positional prior fusion. It is formulated as follows,

$$\mathbf{X}_{\mathrm{mid}} = \mathbf{S}(\frac{QK^T}{\sqrt{d_k}})V. \tag{1}$$

where $\mathbf{S}(\cdot)$ represents the softmax function. d_k represents the dimension corresponding to each attention head.

C. Spatial prior calibration module

The spatial prior calibration (SPC) module is constructed with two blocks, as shown in Fig. 4. First, the dimension of the feature is reshaped to transport the spatial perception (SP) block and explicit calibration (EC) block. In particular, an SP block is utilized to capture global long-range dependencies and a parallel EC block is employed to capture local key points within regions of spatial position.

The SP block captures the long-range dependencies to identify object location information, which employs the global channel-based MLP operation with the full connection layer. It comprises two residual units: a deep convolutional unit and a channel-based MLP unit. Particularly, the input features are inputted into the deep convolutional unit, which employs the group-normalized depthwise convolution layer. The channel scaling and drop path operations are applied to enhance feature generalization and robustness. Subsequently, a residual connection of $\mathbf{X}_{\mathrm{mid}}$ is introduced. These procedures can be formalized as follows,

$$\tilde{\mathbf{X}}_{\mathrm{mid}} = \mathrm{DP}(\mathrm{CS}(\mathrm{DConv}(\mathrm{GN}(\mathbf{X}_{\mathrm{mid}})))) + \mathbf{X}_{\mathrm{mid}}, \tag{2}$$

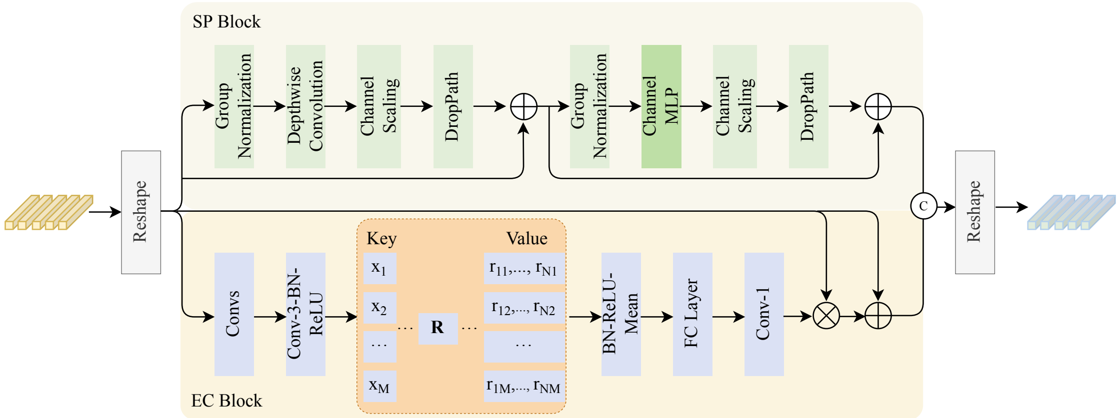


Fig. 4. Illustration of the SPC module. The SPC module consists of a spatial perception (SP) block and an explicit calibration (EC) block. The SP block depends on the global channel MLP with the fully connected layer to capture the long-range dependencies. Besides, the EC block utilizes the different scaling ratio convolution to extract the local feature.

where $\tilde{\mathbf{X}}_{\text{mid}}$ represents the output of the depthwise convolution-based unit. $\text{DP}(\cdot)$ employs the drop path operation and $\text{CS}(\cdot)$ represents the channel scaling operation. $\text{GN}(\cdot)$ represents group normalization, and $\text{DConv}(\cdot)$ denotes a depthwise convolution with a kernel size of 1×1. The middle features $\tilde{\mathbf{X}}_{\text{mid}}$ of the MLP-based unit is the output from the deep convolutional unit. Then, the features are passed through group normalization, followed by the channel MLP operation. Subsequently, the operations of channel scaling, drop path, and a residual connection for $\tilde{\mathbf{X}}_{\text{mid}}$ are applied sequentially. It is expressed as follows,

$$\mathbf{SP}(\mathbf{X}_{\text{mid}}) = \text{DP}(\text{CS}(\text{CMLP}(\text{GN}(\tilde{\mathbf{X}}_{\text{mid}})))) + \tilde{\mathbf{X}}_{\text{mid}}, \quad (3)$$

where $\text{CMLP}(\cdot)$ denotes the channel MLP.

The EC block is built to capture local features at multiple scales, which utilizes the various scaling ratio convolution layers. It consists of two components: 1) an inherent codespace denoted as $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M\}$, where $M = H \times W$ represents the total spatial number of the input features and H, W denotes the feature map of height and width. 2) a set of scaling ratios $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M\}$ is employed to capture multiscale features. Initially, the middle features from $\mathbf{X}_{\text{mid}}$ are encoded through a series of convolution layers of 1×1, 3×3, and 1×1. The encoded features are then processed by a 3×3 convolutional operation followed by a Batch Normalization (BN) layer and a Rectified Linear Unit (ReLU) activation function. Following the aforementioned steps, the encoded features $\check{\mathbf{x}}_n$ are mapped to the codespace. It involves sequentially applying a set of scaling ratio $\mathbf{r}$ to ensure the correspondence between each encoded feature $\mathbf{x}_{\text{mid}}$ and codespace entry $\mathbf{b}_m$. The information about the m-th intermediate feature can be calculated as follows,

$$\mathbf{e}_n = \sum_{i=1}^{N} \frac{e^{-\mathbf{r}_m \|\check{\mathbf{x}}_n - \mathbf{b}_m\|^2}}{\sum_{j=1}^{M} e^{-\mathbf{s}_m \|\check{\mathbf{x}}_n - \mathbf{b}_m\|^2}} (\check{\mathbf{x}}_n - \mathbf{b}_m), \quad (4)$$

where $\mathbf{r}_m$ represents the m-th scaling ratio, $\check{\mathbf{x}}_n$ represents the n-th pixel point, and $\mathbf{b}_m$ denotes the m-th learnable visual code-word. M denotes the total number of visual centers. $(\check{\mathbf{x}}_n - \mathbf{b}_m)$ indicates the relative position of each pixel with respect to a code word.

Afterwards, the Φ is utilized to combine all $\mathbf{e}_n$. It is formalized as follows,

$$\mathbf{e} = \Phi(\mathbf{e}_n), \quad (5)$$

where $\Phi(\cdot)$ comprises a BN layer with ReLU activation function and mean layer.

The fusion feature $\mathbf{e}$ is further fed into a 1×1 convolutional layer and a fully connected layer. Then, we employ channel-wise multiplication between the input features $\mathbf{X}_{\text{mid}}$ and the scaling ratio factor $\mathbf{Sig}(\cdot)$. It is expressed as follows,

$$\mathbf{E} = \mathbf{X}_{\text{mid}} \otimes (\mathbf{Sig}(\text{Conv}_1(\mathbf{e}))), \quad (6)$$

where $\mathbf{Sig}(\cdot)$ represents the sigmoid function and Conv_1 is the 1×1 convolutional layer. $\otimes$ denotes channel-wise multiplication. Subsequently, we conduct channel-wise addition between the features $\mathbf{X}_{\textbf{mid}}$ output from the middle feature and the features $\mathbf{E}$ of the local region. It is calculated as follows,

$$\mathbf{EC}(\mathbf{X}_{\text{mid}}) = \mathbf{X}_{\text{mid}} \oplus \mathbf{E}, \quad (7)$$

where $\oplus$ denotes the channel-wise addition.

The positional prior $\mathbf{P}$ is generated by averaging the channels between the SP block and the EC block. It is formalized as follows,

$$\mathbf{P}(\mathbf{X}_{\text{mid}}) = \mathbf{SP} \copyright \mathbf{EC}, \quad (8)$$

where $\mathbf{P}$ represents the positional prior information. © denotes the element-wise concatenation. The $\mathbf{P}$ contains the spatial distribution information and scale information of objects.

D. Visual position attention and textual context attention

To accentuate the spatial position of a specific object, the positional prior $\mathbf{P}$ is integrated into the image representation. To this end, a multi-head cross-attention (MHCA) layer is used as a visual position attention module. Especially, the image representation $\mathbf{V}_{\rm i}$ serves as the query (Q), while the spatial prior $\mathbf{P}$ functions as both the key (K) and the value (V). Following the MHCA, an MLP is utilized to fine-tune the extracted representation. It is denoted as follows,

$$\mathbf{V}_{\rm i}^{'} = \mathbf{MLP}(\mathbf{S}(\frac{\mathbf{FC}_Q(\mathbf{V}_{\rm i}) * \mathbf{FC}_K(\mathbf{P})}{\sqrt{d_k}}) * \mathbf{FC}_V(\mathbf{P})), \quad (9)$$

where $\mathbf{FC}_{Q|K|V}(\cdot)$ represents the projection layers for the three counterparts, $\mathbf{MLP}(\cdot)$ denotes the function of the MLP layer, and $\mathbf{V}_{\rm i}^{'}$ is indicative of the spatially enhanced visual representation. Finally, the dimension is reshaped to the input dimension size.

Similarly, a positional prior $\mathbf{P}$ is fed into textual context attention, which integrates textual features into prior information. It also leverages a multi-head cross-attention (MHCA) layer. Here, the textual representation $\mathbf{V}_{\rm t}$ acts as the query (Q), while the prior context $\mathbf{P}$ serves as both the key (K) and the value (V). Following the MHCA, an MLP is applied to refine the textual representation. This process is defined as follows,

$$\mathbf{V}_{\rm t}^{'} = \mathbf{MLP}(\mathbf{S}(\frac{\mathbf{FC}_Q(\mathbf{V}_{\rm t}) * \mathbf{FC}_K(\mathbf{P})}{\sqrt{d_k}}) * \mathbf{FC}_V(\mathbf{P})), \quad (10)$$

where $\mathbf{V}_{\rm t}^{'}$ denotes the enhanced textual representation.

E. Object-centric alignment module

Given the inherent contrast in object density between the input image and the samples employed for CLIP encoder training, a significant challenge arises due to the overall distribution shift, which impedes the alignment between text and visual representations. Inspired by Q-former in BLIP-2 [40], an Object-Centric Alignment (OCA) module is designed to learn text queries that align the feature spaces of visual and textual modalities, as illustrated in Fig. 5. The prior information about object representations is extracted from textual prompts across modal interactions to assist visual features. Upon extracting the text representation $\mathbf{V}_{\rm t}^{'}$, we proceed to distill the query information of the object and inject it into the initially randomized object query. The extraction and injection processes are carried out through the fusion module, which consists of the conventional multi-head attention module. The randomly initialized query $\mathbf{V}_{\rm t}^{0}$ serves as Q, while the textual context attention information $\mathbf{V}_{\rm t}^{+}$ functions as both V and K. The object query can be constructed as follows,

$$\mathbf{V}_{\rm t}^{+} = \mathbf{S}(\frac{QK^T}{\sqrt{d_k}})V, \quad (11)$$

where $\mathbf{V}_{\rm t}^{+}$ represents the augmented object query.

Finally, the Context Interact (CI) unit is employed to encompass discriminative knowledge derived from the text embedding $\mathbf{V}_{\rm t}^{+}$. It is calculated as follows,

$$\mathbf{CI}(\mathbf{V}_{\rm t}^{+}) = \frac{\mathbf{V}_{\rm t}^{+} + \frac{1}{N}\sum_{i=1}^{N}\mathbf{V}_{\rm t}^{+}}{2}, \quad (12)$$

where N stands for N-dimension along the channel direction..

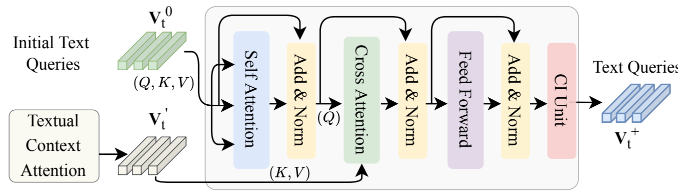


Fig. 5. Illustration of the OCA module. The OCA module extracts prior information on object representation from textual prompts, which enables cross-modal interactions to assist visual features.

F. Cross-modal fusion and density map regression

Given visual representation $\mathbf{V}_{\rm i}^{'}$ and the textual query $\mathbf{V}_{\rm t}^{'}$, we construct a multi-head attention module for cross-modal interaction and knowledge transfer between visual features and text queries to obtain multi-modal features. Specifically, the model incorporates a multi-head self-attention mechanism, which takes $\mathbf{V}_{\rm i}^{'}$ as input. It further employs a multi-head cross-attention layer that utilizes the output of the multi-head self-attention layers as queries, and $\mathbf{V}_{\rm t}^{'}$ as keys and values to facilitate knowledge transfer and interaction. Subsequently, a two-layer feedforward network follows the multi-head cross-attention to enhance the feature representation. Finally, the CNN-based decoder is used to regress the density map, and the predicted number of objects F^{est} is obtained by integration.

G. Loss function

The Mean Squared Error (MSE) loss is utilized for model optimization during the training stage. The representation of this loss is as follows,

$$L_{\rm MSE} = \frac{1}{N}\sum_{i=1}^{N}\left\|F_i^{est} - F_i^{gt}\right\|_2^2, \quad (13)$$

where N denotes the total headcount. F_i^{est} and F_i^{gt} represent the estimated and the ground-truth count of the i-th image. $\|\cdot\|_2^2$ represents Euclidean norm squared.

IV. Experimental Results and Analysis

A. Implementation detail

All experiments were conducted using the PyTorch deep learning framework [17], and with an NVIDIA RTX3090 GPU. To optimize the learnable parameters model, the Adam optimizer with a weight decay of 5×10^{-2} was employed. 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.

B. Benchmarking datasets

FSC-147 [33] serves as a meticulously annotated image collection specifically crafted for class-agnostic object-counting research. It encompasses a comprehensive assemblage of 7,135 images categorized into 147 distinct classes, and each category features non-overlapping images predominantly depicting items, *e.g.*, kitchen utensils, office supplies, stationery, vehicles, and animals. Each image in the dataset undergoes thorough annotation, which establishes it as a foundational source of ground truth data for the evaluation of counting models. The annotations provide 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 [41] presents a comprehensive crowd-counting dataset with 1,198 annotated images. It is segregated into two subsets, namely Part A and Part B. Images in Part A are obtained from the internet and depict densely populated targets. It includes 482 images, with 300 assigned for training and 182 for testing. In contrast, Part B includes authentic captures of lively streets in Shanghai, and displays relatively sparse target distributions. It includes a total of 716 images, with 400 designated for training and 316 for testing. The distinct origins of these two segments pose challenges for cross-scene evaluations.

CARPK [42] represents an image dataset specifically crafted for the task of vehicle counting. It incorporates 1,148 bird's-eye-view images of parking lots and captures vehicles in varying time and weather conditions. The dataset embodies a total of 89,777 cars and vividly illustrates variations in density, occlusion, and scale. Each image within the dataset is meticulously annotated, which offers comprehensive counting data for both vehicles and pedestrians.

C. Evaluation metrics

Following prior researches [43]–[45], the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were employed as metrics for evaluating. MAE was used to assess the accuracy of the model. It is mathematically formulated as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|, \tag{14}$$

where N represents the total number of images in the test set, y_i denotes the ground truth of the actual number of objects in the $i{-}th$ 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, with the mathematical expression as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|^2}, \tag{15}$$

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.

D. Experiments on FSC-147 dataset

Table I presents the objective comparison results of the proposed method VLPG against State-Of-The-Art (SOTA) methods on the FSC-147 [33] dataset. In comparison to the CLIP-Count [17], which achieves zero-shot object counting by correcting the visual feature space through textual prompts, both MAE and MSE have shown an improvement of 14.58% and 12.57% on the validation set, which indicates superior counting performance over advanced zero-shot counting methods. To comprehensively assess the performance of the counting model, we included comparisons with several few-shot methods and reference-less counting methods in Table I. It is observed that the proposed method VLPG achieved a reduction of 24.26% and 11.27% in MAE and RMSE on the validation set, and 20.36% in MAE on the test set, compared to the SOTA few-shot method CFOCNet [46], which leverages the similarity between query images and reference images to achieve few-shot object counting. The proposed method reduces the reliance on manually annotated samples during the training and testing phases by utilizing textual descriptions. Importantly, it demonstrates its unique strengths when dealing with a wide range of categories and large-scale sample sets. When compared to the reference-less counting method LOCA [10], which achieves zero-shot counting by iteratively blending shape and appearance information with image features, the proposed method VLPG achieves reductions of 7.92% and 25.54% in MAE and RMSE on the validation set, and 6.06% in RMSE on the test set. This further validates the exceptional performance of the proposed method VLPG not only in zero-shot scenarios with high accuracy and robustness but also in handling few-shot and reference-less scenarios.

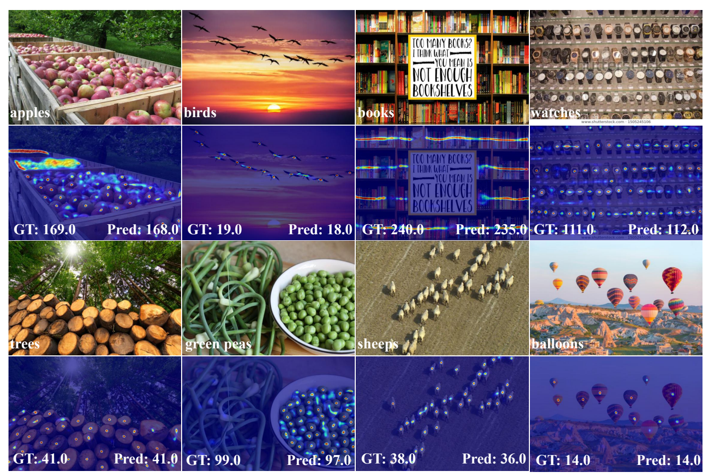


Fig. 6. Visualization of the input image and generated density maps for the samples from the FSC-147 dataset.

The visualization results for the FSC-147 dataset are depicted in Fig. 6. The second and fourth rows display the application of predicted density maps overlaying the original images. It is evident that the proposed VLPG model optimally exploits both spatial and textual prior information, which enables accurate counting of various object types guided by tex-

TABLE I. 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 [33]	CVPR2021	3	24.32	70.94	22.56	101.54
	CFOCNet [46]	WACV2021	3	21.19	61.41	22.10	112.71
	CounTR [13]	BMVC2022	3	13.13	49.83	11.95	91.23
	LOCA [10]	ICCV2023	3	**10.24**	**32.56**	**10.97**	**56.97**
	FamNet [33]	CVPR2021	1	26.05	77.01	26.76	110.95
Reference-less	FamNet* [33]	CVPR2021	0	32.15	98.75	32.27	131.46
	RepRPN-C [15]	ACCV2022	0	29.24	98.11	26.66	129.11
	CounTR [13]	BMVC2022	0	18.07	71.84	**14.71**	106.87
	LOCA [10]	ICCV2023	0	**17.43**	**54.96**	16.22	**103.96**
	RCC [14]	CVPR2023	0	17.49	58.81	17.12	104.53
Zero-shot	ZSC [12]	CVPR2023	0	26.93	88.63	22.09	115.17
	Clip-Count [17]	MM2023	0	18.79	61.18	17.78	106.62
	VLPG (Ours)	This Paper	0	**16.05**	**53.49**	**17.60**	**97.66**

TABLE II. CROSS-DATASET EVALUATION ON SHANGHAITECH CROWD COUNTING DATASET.

Method	Type	Training $\longrightarrow$ Testing	MAE	RMSE	Training $\longrightarrow$ Testing	MAE	RMSE
MCNN [41]	Specific	Part A $\longrightarrow$ Part B	85.2	142.3	Part B $\longrightarrow$ Part A	221.4	357.8
CrowdCLIP [47]			69.6	80.7		217.0	322.7
RCC [14]	Generic	FSC147 $\longrightarrow$ Part B	66.6	104.8	FSC147 $\longrightarrow$ Part A	240.1	366.9
Clip-Count [17]			45.7	77.4		192.6	308.4
VLPG (Ours)			**42.4**	**71.6**		**178.9**	**284.6**

tual prompts. Furthermore, the predicted density maps exhibit spatial consistency with the ground truth density distributions.

E. Experiments on ShanghaiTech dataset

Table II presents the objective comparison results of the proposed method VLPG against State-Of-The-Art (SOTA) methods on the ShanghaiTech dataset [41] dataset. We assessed the model's cross-domain generalization capability by conducting tests on the ShanghaiTech dataset using the model trained directly on the FSC-147 dataset. Throughout this process, we only needed to update the input textual prior information to "person" to specify the target population for counting. It can be observed that, even in this scenario, the proposed method outperforms other counting methods listed in Table II. Specifically, MAE and RMSE were reduced by 7.11% and 7.72% in the Part A dataset and 7.22% and 7.49% in the Part B dataset compared to CLIP-Count [17]. The experimental results demonstrate that the proposed method reduces interference among objects, which enhances long-distance dependencies to improve counting accuracy. Subjective results in Fig. 7 provide additional confirmation of the effectiveness of our method on ShanghaiTech, particularly in cross-dataset scenarios. Visualizations further indicate that the VLPG can extract the long-range dependencies to suppress the background and capture the local region to address the scale variation. The proposed method can enhance counting precision in regions with high density.

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

F. Experiments on CARPK dataset

We also tested the cross-domain generalizability of VLPG model on the CARPK [42] dataset. Similar to the ShanghaiTech [41] 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 III. Compared with the Shi *et al.* [49], which incorporates the Segment Anything Model into the

TABLE III. Cross-dataset evaluation on CARPK dataset.

Method	#Shot	MAE	RMSE
FamNet [33]	3	28.84	44.47
BMNet [35]	3	14.41	24.60
BMNet+ [35]	3	**10.44**	**13.77**
RCC [14]	0	21.38	26.15
Clip-Count [17]	0	11.96	16.61
DSPI [48]	0	11.50	15.52
Shi *et al.* [49]	0	10.97	14.24
VLPG (Ours)	0	**10.14**	**13.79**

counting network to achieve zero-shot object counting, the proposed method VLPG achieved reductions of 7.57% and 3.16% in MAE and RMSE, respectively. The objective results indicate that the introduction of spatial location priors can effectively enhance the precision of object identification within images, thereby improving the accuracy of object counting. When compared with the few-shot counting method BMNet [35], which jointly learns representation and similarity measurement to achieve zero-shot counting, the proposed method VLPG demonstrated decreases of 29.63% and 43.94% in MAE and RMSE, respectively. These consistent improvements further validate the superiority of the proposed method VLPG in counting tasks. Visualization results on the CARPK dataset are illustrated in Fig. 8. Subjective observations reveal that the integration of spatial information substantially aids in distinguishing between targets and backgrounds, which highlights the distinct advantage of combining textual descriptions with spatial priors.

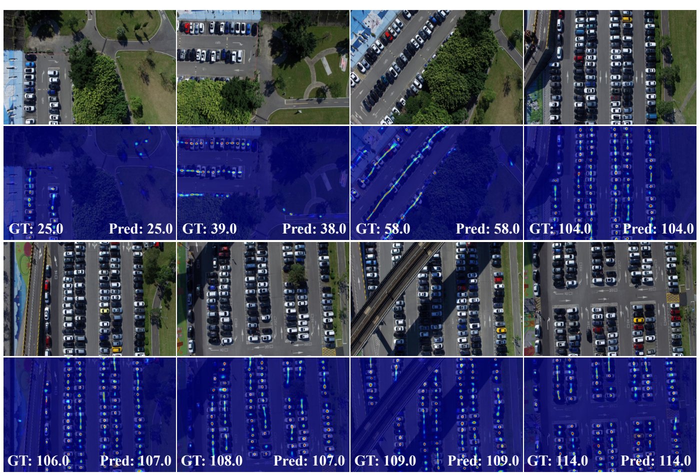


Fig. 8. Visualization of the input image and generated density maps for the samples from the CARPK dataset.

G. *Efficiency comparison*

To assess the efficiency of the proposed method, we conducted a series of comparative experiments on the CAPRK dataset using two different GPUs (*i.e.*, RTX 3090 and RTX 3060). The input size was set to 384×384. Four evaluation metrics, namely parameters, FLOPs, inference time, and frames per second (FPS), were utilized to assess the efficiency of different methods. The comparative results are illustrated in Table IV. On the CAPRK dataset, the proposed VLPG scores 10.14 and 13.79 in MAE and RMSE, which outperform other methods in terms of counting accuracy. Nevertheless, in terms of parameters and processing time, the VLPG is slightly less efficient than other methods. Specifically, the proposed method has 90.11M parameters, which is higher than DSPI (68.67M). The VLPG has 127.37G FLOPs, which is comparable to other methods. Regarding processing time and frame rate, the proposed method takes 14.40ms and 24.00ms for each image on RTX 3090 and RTX 3060 GPUs, namely achieving FPS of 69.47 and 41.66. It indicates that the VLPG can process in real-time (30FPS) in video surveillance and security scenarios. In the future, we will explore more efficient model architectures, which aim to reduce parameter count and computational complexity while maintaining or even improving the accuracy of the model.

H. *Ablation studies*

Component analysis To investigate the individual contributions of different components in the VLPG model and assess its effectiveness, ablation experiments were extensively conducted on the FSC-147 dataset, with the objective comparison results shown in Table V. Additionally, we performed intermediate feature visualizations for various combinations, as shown in Fig. 9.

1) **Scheme-a** represents the baseline model without the Grounding DINO (Prior), SPC, and OCA modules.
2) **Scheme-b** indicates the addition of the OCA module to the baseline model. The results show that MAE and RMSE decreased by 5.43 and 1.89, respectively. Additionally, one can see from Fig. 9 that the model with the OCA module pays more attention to the foreground object areas compared with the baseline model. This indicates that the optimized textual features can provide a stronger alignment capability.
3) **Scheme-c** incorporates the Prior module on the baseline to offer spatial prior positional information for target objects. As depicted in Table V, compared with the baseline model, it reduces the MAE and RMSE by 9.84% and 10.27% on the validation set. This verifies the effectiveness of the deep spatial prior. Besides, the visual representation of the positional prior reduces attention to irrelevant background information, as shown in Fig. 9.
4) **Scheme-d** introduces the SPC module on the Baseline for capturing both global long-range dependence and local key points within spatial regions. As shown in Table V, compared to adding only the Prior module, MAE and RMSE decreased by 0.71% and 4.23% on the test set, respectively. Fig. 9 indicates that the SPC module assists the model in obtaining a more comprehensive context at both global and local levels, which enhances its understanding and representation of the input.

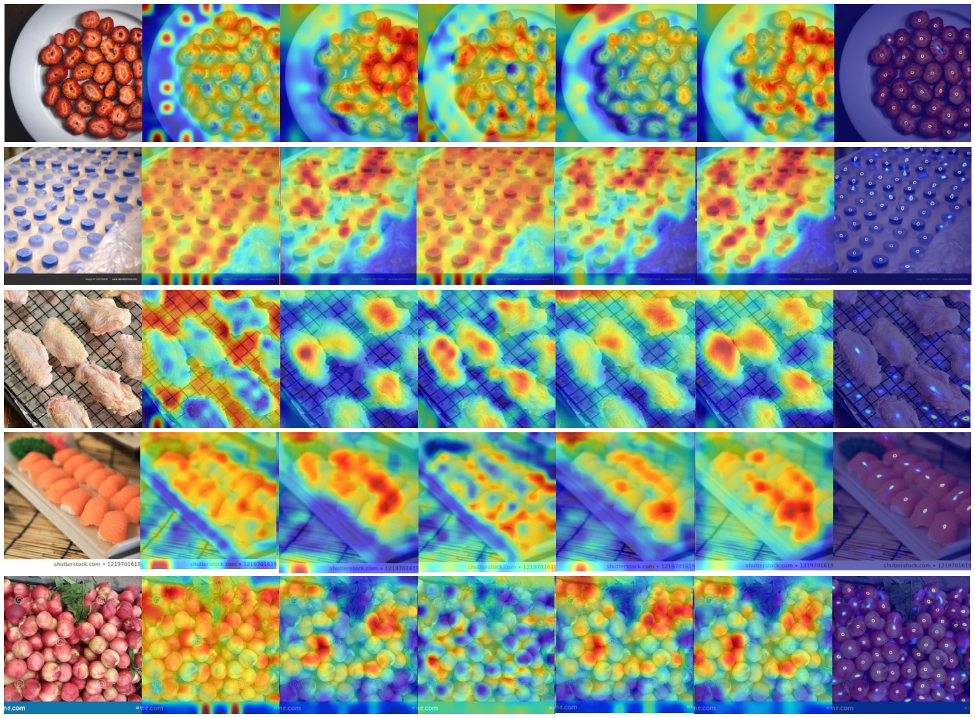


Fig. 9. Visualization of the baseline with different components.

5) **Scheme-e** simultaneously incorporates Prior, SPC, and OCA modules into the baseline. Compared to the model that only included Prior and SPC modules, the MAE and MSE on the validation set decreased by 7.44% and 3.31%, respectively. This shows that the OCA module improves counting accuracy and robustness by matching text and image information on top of the existing foundation. Although the MAE on the test set is not the best, with only a 0.39 difference from the optimal result, Fig. 9 shows that the scheme is more focused on the object area. Additionally, its FLOPs do not differ significantly compared with other schemes, as shown in Table V. Therefore, we select this formula as our final scheme, termed VLPG.

Ablation analysis on the SPC module To validate the impact of different combinations of the global block SP and the local block EC in the SPC module on counting performance, we conducted an ablation study on the FSC-147 dataset, as shown in Fig. 10 and Fig. 11.

1) **SP.** When only the SP block is adopted, the MAE on the test set is 19.11, and the MSE is 104.90. The intermediate feature visualizations are shown in Fig. 11. Particularly, as shown in the third column of the third row, the model utilizes the SP block to suppress the background area in the lower right corner of the image. Furthermore, due to the scale variation in objects, the SP block can extract position information from the target (apple) across different distances from near to far. It indicates that the SP block can capture long-range dependencies between different locations in the image and it enables the model to perceive the connections and information between distant locations of various targets within the image.
2) **EC.** When only the EC block is used, the MAE on the test set is 19.66, and the MSE is 107.55. This result is slightly worse than the performance of the SP block. This is due to the fact that the EC block focuses on extracting local features and lacks global

TABLE IV. COMPARISON RESULTS OF THE MODEL COMPLEXITY ON CARPK DATASET, THE INPUT IMAGE SIZE IS 384×384.

Methods	MAE	RMSE	Params (M)	FLOPs	RTX 3090		RTX 3060	
					Time (ms)	FPS	Time (ms)	FPS
ClipCount [17]	11.96	16.61	16.36	123.06	11.04	90.56	17.61	56.79
DSPI [48]	11.50	15.52	68.67	124.76	12.76	78.40	21.74	46.00
VLPG (Ours)	**10.14**	**13.79**	90.11	127.37	14.40	69.47	24.00	41.66

TABLE V. COMPONENTS ANALYSIS. THE PROPOSED COMPONENTS WERE PROGRESSIVELY INCORPORATED INTO THE BASELINE TO STUDY THE INDIVIDUAL CONTRIBUTION.

Scheme	Components			Val Set		Test Set		Params (M)	FLOPs
	Prior	SPC	OCA	MAE	RMSE	MAE	RMSE		
a)	✗	✗	✗	19.30	66.12	18.52	105.36	16.36	123.06
b)	✗	✗	✓	18.59	60.73	**17.53**	103.37	20.57	123.09
c)	✓	✗	✗	17.40	59.33	17.73	103.89	85.90	124.69
d)	✓	✓	✗	17.34	55.32	17.60	99.50	85.90	127.33
e)	✓	✓	✓	**16.05**	**53.49**	17.60	**97.66**	90.11	127.37

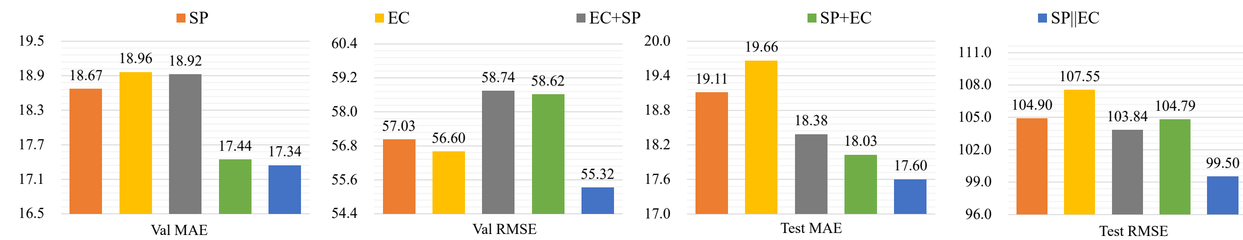


Fig. 10. Quantitative comparisons of different SPC module variations.

information processing, which leads to poorer counting performance compared to the SP block. As shown in the fourth column of the first row of Fig. 11, the EC block effectively extracts the features of individual objects.

3) **EC+SP.** When the EC block is equipped before the SP block, the scores of MAE and MSE on the test dataset are 18.38 and 103.84, respectively. This combination performs better than using the SP or EC block alone. The reason is that the extraction of global features is enhanced by incorporating local features, which combines local details with global information to improve counting accuracy.
4) **SP+EC.** When the SP block is placed before the EC block, the MAE and MSE score 18.03 and 104.79 on the test set, respectively. This configuration performs better than the "EC+SP" combination on the validation set, because "SP+EC" allows the model to better capture both overall information and details.
5) **SP||EC.** When the SP and EC blocks are combined in parallel, they achieve the best performance, with an MAE of 17.60 and an MSE of 99.50 on the test set. Additionally, it can be observed that these intermediate features focus more on the object area compared to other combinations in Fig. 11. This indicates that the parallel combination can effectively utilize both global and local features, thus providing a more comprehensive feature representation.

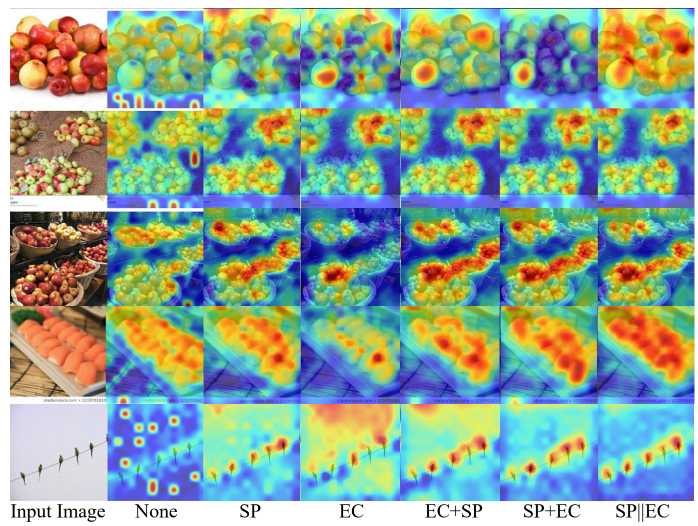


Fig. 11. Qualitative visualization of feature maps obtained from different SPC module variations.

V. CONCLUSION

In this paper, we recognize limitations within the existing class-agnostic counting model, specifically its insensitivity to position information and potential misalignment within the hypothesis space. To tackle these limitations, we proposed the Vision-Language Prior Guidance (VLPG) Network. The VLPG consists of three critical modules, *i.e.*, Grounding DINO, spatial prior calibration (SPC), and object-centric alignment (OCA) module. The VLPG employs a pre-trained object grounding model integrated to obtain spatial location as an additional prior for a given query class, which facilitates more precise localization of the object. Meanwhile, the SPC module is built for the extraction of long-range dependencies and local regions within spatial position regions. Moreover, the OCA module is designed to harmonize feature spaces across multiple modalities. Through extensive experimentation on various benchmarks, the proposed model showcased superior performance over the SOTA competitors. It contributes to the advancement of class-agnostic counting in a multi-modal context.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] T. Han, L. Bai, J. Gao, Q. Wang, and 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, and 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, T. Lei, B. Ying, M. Xue, and W. Zhao, "A crowd counting network based on multi-scale pyramid transformer," *CAAI Transactions on Intelligent Systems*, vol. 19, no. 2, pp. 67–78, 2024.

[4] X. Wang, Y. Zhan, Y. Zhao, T. Yang, and Q. Ruan, "Semi-supervised crowd counting with spatial temporal consistency and pseudo-label filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4190–4203, 2023.

[5] W. Zhai, M. Gao, Q. Li, G. Jeon, and M. Anisetti, "Fpanet: feature pyramid attention network for crowd counting," *Applied Intelligence*, pp. 1–18, 2023.

[6] S. Jiang, Q. Wang, F. Cheng, Y. Qi, and Q. Liu, "A unified object counting network with object occupation prior," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1147–1158, 2024.

[7] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: comparisons of density maps for crowd analysis tasks—counting, detection, and tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1408–1422, 2018.

[8] W. Zhai, M. Gao, M. Anisetti, Q. Li, S. Jeon, and J. Pan, "Group-split attention network for crowd counting," *Journal of Electronic Imaging*, vol. 31, no. 4, p. 041214, 2022.

[9] S. Jiang, X. Lu, Y. Lei, and L. Liu, "Mask-aware networks for crowd counting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3119–3129, 2019.

[10] N. Djukic, A. Lukezic, V. Zavrtanik, and M. Kristan, "A low-shot object counting network with iterative prototype adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 872–18 881.

[11] M. Wang, Y. Li, J. Zhou, G. W. Taylor, and M. Gong, "Gcnet: Probing self-similarity learning for generalized counting network," *Pattern Recognition*, vol. 153, p. 110513, 2024.

[12] J. Xu, H. Le, V. Nguyen, V. Ranjan, and D. Samaras, "Zero-shot object counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 548–15 557.

[13] L. Chang, Z. Yujie, Z. Andrew, and X. Weidi, "Countr: Transformer-based generalised visual counting," in *Proceedings of the British Machine Vision Conference*, 2022.

[14] M. Hobley and V. Prisacariu, "Learning to count anything: Reference-less class-agnostic counting with weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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

[16] 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 *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

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

[18] 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," in *Proceedings of the European Conference on Computer Vision*, 2024.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 213–229.

[20] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *Proceedings of the International Conference on Learning Representations*, 2023.

[21] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," in *Proceedings of the International Conference on Learning Representations*, 2022.

[22] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[23] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognition*, vol. 50, pp. 107–117, 2016.

[24] W. Zhai, Q. Li, Y. Zhou, X. Li, J. Pan, G. Zou, and M. Gao, "Da2net: a dual attention-aware network for robust crowd counting," *Multimedia Systems*, vol. 29, no. 5, pp. 3027–3040, 2023.

[25] X. Xing, R. Gao, T. Han, S.-C. Zhu, and Y. N. Wu, "Deformable generator networks: Unsupervised disentanglement of appearance and geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1162–1179, 2022.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and 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.

[28] Y. Chen, J. Yang, B. Chen, and S. Du, "Counting varying density crowds through density guided adaptive selection cnn and transformer estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1055–1068, 2022.

[29] W. Zhai, M. Gao, A. Souri, Q. Li, X. Guo, J. Shang, and G. Zou, "An attentive hierarchy convnet for crowd counting in smart city," *Cluster Computing*, vol. 26, no. 2, pp. 1099–1111, 2023.

[30] X. Guo, M. Gao, W. Zhai, Q. Li, and G. Jeon, "Scale region recognition network for object counting in intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[31] W. Wang, X. Yang, and J. Tang, "Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4452–4461, 2023.

[32] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Qtn: Quaternion transformer network for hyperspectral image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7370–7384, 2023.

[33] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, "Learning to count everything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3394–3403.

[34] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4376–4389, 2021.

[35] M. Shi, H. Lu, C. Feng, C. Liu, and 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.

[36] Z. You, K. Yang, W. Luo, X. Lu, L. Cui, and 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.

[37] W. Lin, K. Yang, X. Ma, J. Gao, L. Liu, S. Liu, J. Hou, S. Yi, and A. B. Chan, "Scale-prior deformable convolution for exemplar-guided class-agnostic counting." in *Proceedings of the British Machine Vision Conference*, 2022, p. 313.

[38] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[39] 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," *arXiv preprint arXiv:2010.11929*, 2020.

[40] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 19 730–19 742.

[41] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[42] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 4165–4173.

[43] W. Zhai, M. Gao, X. Guo, Q. Li, and G. Jeon, "Scale-context perceptive network for crowd counting and localization in smart city system," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18 930–18 940, 2023.

[44] Y. Meng, Y. Zhang, and W. Zhou, "Crowd counting method based on proportion fusion and multilayer scale-aware," *CAAI Transactions on Intelligent Systems*, vol. 19, no. 2, pp. 307–315, 2024.

[45] W. Zhai, X. Xing, and G. Jeon, "Region-aware quantum network for crowd counting," *IEEE Transactions on Consumer Electronics*, 2024.

[46] S.-D. Yang, H.-T. Su, W. H. Hsu, and 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.

[47] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu, and 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.

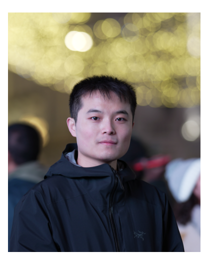
[48] J. Chen, Q. Li, M. Gao, W. Zhai, G. Jeon, and D. Camacho, "Towards zero-shot object counting via deep spatial prior cross-modality fusion," *Information Fusion*, p. 102537, 2024.

[49] Z. Shi, Y. Sun, and M. Zhang, "Training-free object counting with prompts," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 323–331.

Wenzhe Zhai is pursuing a Ph.D. degree at the College of Intelligent Science System and Engineering, Harbin Engineering University, Harbin, China. His research interests include smart city systems, information fusion, crowd analysis, and deep learning.

Xianglei Xing received the M.S. and Ph.D. degrees from the School of Electronic Science and Engineering, Nanjing University, China, in 2006 and 2013, respectively. He is currently a professor with the College of Intelligent System Science and Engineering, Harbin Engineering University. During the years 2017-2019, he was a visiting researcher at UCLA. His research interests include computer vision, statistical modeling and learning, with a focus on representation learning, deep generative models, sparse and structure learning.

Mingliang Gao received his Ph.D. in Communication and Information Systems from Sichuan University. He is now an associate professor at the Shandong University of Technology. He was a visiting lecturer at the University of British Columbia during 2018-2019. His research interests include computer vision, machine learning, and intelligent optimal control. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley.



Qilei Li received the B.S. degree in electronic information engineering from the Shandong University of Technology, and the M.S. degree with the College of Electronics and Information Engineering, Sichuan University, Chengdu, China. His research interests are image processing and deep learning.