HER-Count: Learning Hyper-Exemplar Representation for Generalized Zero-Shot Object Counting

Jianing Li[1], Xiaobin Liu[2]*, and Ruihan Xu[1]

[1] College of Computer Science, Peking University, Beijing, China
[2] College of Artificial Intelligence, Nankai University, Tianjin, China
ljn-vmc@pku.edu.cn, liuxb@nankai.edu.cn, ruihan.xu@foxmail.com

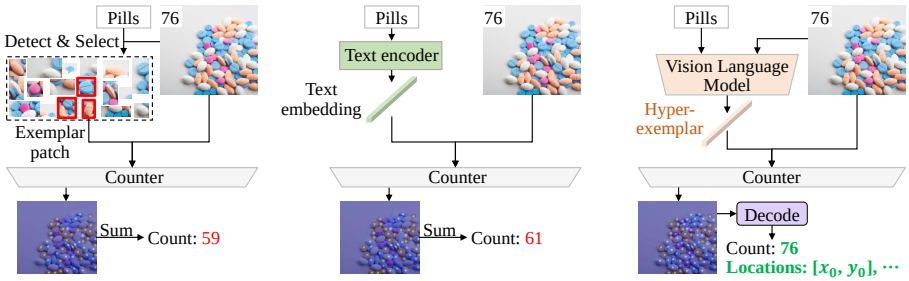
Abstract. Zero-shot object counting aims to count objects in images from given class names. Existing methods typically select representative image patches or extract text embeddings as exemplars to match image regions for counting. However, a limited number of selected patches is insufficient to cover intra-class diversity, and the accumulated errors in exemplar detection and selection substantially degrade counting accuracy. While the text embeddings remain at a generic semantic level, blind to the specific instance diversity of target objects in an image. To handle these issues, this paper proposes the Hyper-Exemplar Representation (HER) counting framework (HER-Count) to represent the objects of a specific category in a given image. Rather than image patches or text embeddings, HERs are synthesized from holistic images and textual class names via an multi-modal large-language model. By jointly integrating all target objects in each image, HERs avoid exemplar generation errors and are more customized representative, thus yielding more accurate counting results. To effectively exploit HERs in counting, HER-Count introduces a hierarchical fusion strategy that injects HERs into multiple intermediate layers of the vision encoder when predicting density maps. This enables the model to progressively refine counting cues across layers, significantly enhancing the performance. A global discrimination enhancement constraint at both image sample and semantic class levels is designed to further improve the capacities for both representation and discrimination of HERs. Extensive experiments on widely used benchmarks demonstrate both the state-of-the-art performance and generalization ability of the proposed HER-Count. Benefiting from the compact end-to-end design, HER-Count also achieves promising inference speed.

Keywords: Object counting · Multi-modal Large Language Model · Zero-shot

1 Introduction

The object counting task aims to infer the number of objects in given images. As an important step in visual monitoring applications, this task has attracted increasing attention. Traditional works focus on counting objects of predefined classes, *e.g.*, crowds [2], vehicles [13], animals [1], and cells [23]. However, these methods are restricted by specific classes, lacking the flexibility to adapt to unseen classes.

* Corresponding author.



(a) Exemplar selection-based method (b) Text embedding-based method (c) Our method with hyper-exemplar

Fig. 1: Comparison between our HER-Count and others. (a) Automated detectors produce incomplete and limited exemplars. (b) The generic text embedding is agnostic to the specific instances. (c) Our multi-modal hyper-exemplar is synthesized based on both the image and class name, thus is more representative and robust to instance variations, leading to more accurate counting result.

Recently, class-agnostic counting [5, 12, 19, 20, 26] has been proposed to count objects of arbitrary categories. Based on provided information, this task can be divided into few-shot and zero-shot settings. In few-shot methods, high quality human-annotated exemplar patches are provided to specify the objects of interest. Researchers commonly formulate this task as a matching problem, *i.e.*, using the annotated exemplar patches to identify and count objects. However, relying on human annotations is inflexible for real-world applications. To pursue more practical counting solutions, recent works introduce the zero-shot counting [6,22,24,28] which uses class names alone, enabling autonomous counting without manual bounding box annotations. Existing methods focus on detecting and selecting high-quality image patches as exemplars, or directly matching the text embedding with image regions, as shown in Fig. 1(a) and (b). The core challenge of of existing methods lies in selecting high-quality image patches as exemplars based on class name, and various algorithms [24, 28] have been designed for exemplar detection and selection.

However, these paradigms still suffer from several drawbacks. Image patches selected by automated detectors often yield incomplete or even incorrect exemplars, failing to represent all objects of the same class due to the large intra-class appearance variation and thus reducing counting accuracy. Furthermore, subsequent counting relies solely on the visual modality, neglecting cues in the textual modality. Moreover, performing detection and counting separately hinders the joint optimization of them, leading to error accumulation across these two stages. Those text embedding-based methods are confined to generic text embeddings of given classes. They neglect intra-class instance appearance variations in specific images, thus failing to capture the specific diversity of target objects and leading to inaccurate counting results.

To address these limitations, this paper generates exemplars from a new perspective, *i.e.*, synthesizing Hyper-Exemplar Representations (HERs) via a Multi-modal Large-Language Model (MLLM) based on both images and class names. As shown in Fig. 1(c), rather than detecting image patches or using generic text embeddings, we employ an MLLM to jointly capture the relationship between image and category, and subse-

quently synthesize corresponding HERs that encode target objects in given image. HERs and the image are then fed into a counter to predict the object number. Compared with the detect-and-select pipeline in previous works, our method directly learns hyper-exemplars from the input image in the latent space under the guidance of text categories, thus avoiding exemplar detection errors caused by automated detectors. By integrating the textual class name and all object visual appearances in the target image, the synthesized HERs exhibit stronger representative capacity. Moreover, our end-to-end pipeline enables the joint training of exemplar generation and object counting. Synthesizing HERs via MLLM also leverages its rich prior knowledge to improve the generalization capacity of HERs on unseen classes during testing.

Our new zero-shot object counting framework with proposed HERs is termed as HER-Count. Given an image and a target object name, HERs are synthesized via a powerful MLLM and then used to guide the counter to generate the density map for object counting. To better exploit HER, we introduce a deep fusion strategy in HER-Count that injects HERs into multiple intermediate layers of the counter. This hierarchical integration enables the model to progressively refine counting cues across layers, significantly boosting the counting performance. We further introduce a global discrimination enhancement constraint on HERs that simultaneously maximizes the intra-class similarities and ensures the inter-class separability, effectively enhancing the discriminative power of HERs against diverse instance appearances.

We evaluate our HER-Count on widely used zero-shot counting datasets include FSC147 [19] and CARPK [4]. Experiments show that HER-Count achieves superior performance compared with recent SOTA methods. Benefiting from the compact end-to-end design, our method also shows favorable inference speed. The contributions of this work can be summarized as follows.

1) We propose HER-Count, a novel framework that learns representative multi-modal hyper-exemplars based on both images and textual classes for zero-shot counting.

2) A hierarchical fusion strategy is proposed to inject HERs into multiple intermediate layers of vision backbone, significantly boosting the counting performance. We also design a global discrimination enhancement constraint to further enhance the representative capacity of HERs.

3) To the best of our knowledge, HER-Count is the first MLLM-based end-to-end method to explicitly integrate multi-modal cues into exemplar learning for zero-shot counting. Extensive experiments show that HER-Count achieves superior performance on both standard counting benchmarks and cross dataset generalization.

2 Related Work

2.1 Class-Specific Object Counting

Early works focus on class-specific object counting, which counts objects of predefined categories, such as human crowds [2], vehicles [13], animals [1], and cells [23]. They mainly leverage object detection or density estimation to count, and achieve promising performance on trained categories. However, these methods cannot count objects of arbitrary categories at test, limiting their applicability in real scenarios.

2.2 Class-Agnostic Object Counting

Class-agnostic counting aims to count objects of arbitrary classes. It can be categorized into few-shot counting and zero-shot counting based on used annotations.

Few-shot counting requires manually annotated exemplar patches. For example, GMN [12] feeds features of both test image and exemplar patches into a matching module to regress the counting result. FamNet [19] adopts ROI pooling to extract exemplar features and performs correlation matching. CounTR [9] utilizes more advanced vision transformers as visual backbones to enhance the extracted feature representations. BMNet [20] designs a bilinear matching network to explicitly model the exemplar-image similarity. Despite their promising performance, these methods require human-annotated bounding boxes, making them impractical in real-world scenes.

Zero-shot counting aims to count objects only using given class names, hence is more practical and challenging. ZSC [24] extracts textual prototypes via a conditional VAE and selects exemplars by the distance of image patches and prototypes. PseCo [5] designs a framework for both few-shot and zero-shot object counting based on SAM and CLIP. VA-Count [28] designs an exemplar selection algorithm to discover high-quality exemplars for counting. DAVE [16] proposes a detect-and-verify paradigm to merge the benefits of both density and detection-based counting methods. However, a limited number of exemplars fail to adequately represent all target instances in the given image, and detection errors will further degrade the counting performance. CLIP-Count [6] employs CLIP to encode text and images separately, and designs a patch-text loss to align visual and textual embeddings. YOLO-Count [27] introduces a light-weight model with special-designed cardinality map. T2ICount [17] proposes a diffusion-based framework that consists of a hierarchical semantic correction module to refine the text embedding. However, the text embedding in these methods only describes generic semantic cues, failing to capture the intra-class appearance variations.

Difference from recent methods: The key difference between our HER and existing methods lies in the modal cues used for learning counting representations. As illustrated in Fig. 1, existing approaches either: 1) rely solely on textual cues [6,22], thereby overlooking valuable visual information; or 2) employ additional detection models (e.g., GroundingDINO) [16,27,28] to extract only 1$\sim$3 image patches, which may introduce visual bias due to the limited and potentially inaccurate patches, while also suffering from error accumulation in multi-stage pipelines. In contrast, our HER leverages both textual and visual cues within a unified model to represent the target category, providing richer and more accurate multimodal information for counting. Moreover, our end-to-end framework effectively avoids the error accumulation issue inherent in multi-stage methods.

2.3 Multi-modal Large Language Model

Multi-modal Large Language Model (MLLM) has demonstrated advanced zero-shot image-text understanding ability. For example, CLIP [18] aligns image and text embeddings in a shared space, making it an ideal zero-shot classifier for many downstream vision tasks. BLIP-2 [7] proposes the Q-former that uses learnable parameters

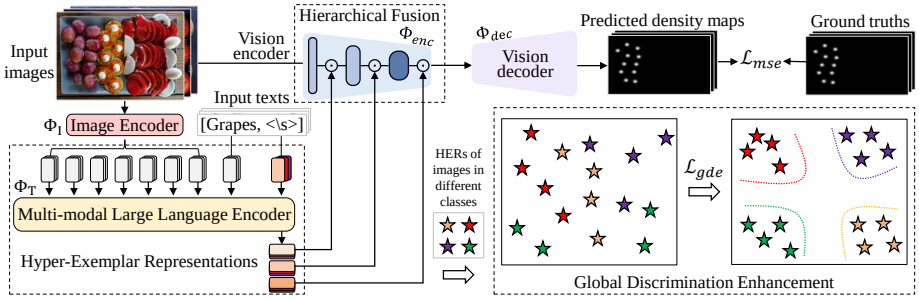


Fig. 2: Overview of the proposed HER-Count framework with hyper-exemplar representations.

to zip visual cues and enhances multi-modal understanding capabilities by bootstrapping. Grounding DINO [11] enables open set detection by incorporating language into transformer-based detector. LLaVA [10] only fine-tunes a single projection layer in MLLM to align features of different modalities. Mini-gemini [8] introduces an extra fine-grained vision encoder and unlocks the multi-modal output capacity of MLLM. Currently, the multi-modal comprehension capacity of MLLMs effectively facilitates the development of downstream tasks. However, their capacity on the zero-shot counting task has not been explored.

3 Method

3.1 Formulation

Given an image $\mathcal{I}$ and a textual class name $\mathcal{T}$, the zero-shot object counting task aims to count objects in $\mathcal{I}$ belonging to $\mathcal{T}$. As a zero-shot task, there is no overlap among the categories in the training set, validation set, and testing set. Existing methods [6,24,28] formulate this task as the density map estimation. To establish the connection between the image and the textual class name, they generates the exemplar $\mathcal{E}$ by either detecting and selecting representative image patches from $\mathcal{I}$ based on off-the-shelf automated detectors, or encoding $\mathcal{T}$ with a text encoder. A density map is then estimated by a matching module Match based on $\mathcal{I}$ and $\mathcal{E}$ as:

$$\mathcal{D} = \mathrm{Match}(\mathcal{I}, \mathcal{E}), \tag{1}$$

where $\mathcal{D} \in \mathbb{R}^{1\times H\times W}$ denotes the predicted density map that represents the existence of objects at each location. H and W are the height and width of $\mathcal{I}$, respectively. The counting result N is obtained by summing $\mathcal{D}$ in previous works as: $N = \mathrm{sum}(\mathcal{D})$. The training objective is commonly formulated as minimizing the Mean Squared Error (MSE) loss $\mathcal{L}_{mse}$ between $\mathcal{D}$ and its corresponding ground truth $\mathcal{D}^*$ as:

$$\mathcal{L}_{mse}(\mathcal{D}^*, \mathcal{D}) = \frac{1}{HW}\sum \|\mathcal{D}^* - \mathcal{D}\|_2^2 . \tag{2}$$

However, since objects of the same class often exhibit substantial appearance variations, a limited numbers of image patches or one single generic textual name embedding cannot adequately represent all target instances in a image. And incorrect patches by automated detectors will mislead the matching module. These issues limit the counting accuracy of current methods, and also restrict their generalizability and scalability in real-world applications.

To pursue a robust representation of target objects, we propose the new HER-Count framework with Hyper-Exemplar Representations (HERs), as shown in Fig. 2. A set of multi-modal HERs $\mathbf{F}$ is directly synthesized based on both $\mathcal{I}$ and $\mathcal{T}$ to comprehensively represent target object appearance via an MLLM, which can be formulated as:

$$\mathbf{F} = \mathrm{MLLM}(\mathcal{I}, \mathcal{T}), \tag{3}$$

$\mathbf{F}$ is then used for density map estimation, and Eqn. (1) can hence be rewritten as:

$$\mathcal{D} = \mathrm{Match}(\mathcal{I}, \mathbf{F}). \tag{4}$$

Details of synthesizing $\mathbf{F}$ will be given in Sec. 3.2.

Directly integrating $\mathbf{F}$ with $\mathcal{I}$ leads to insufficient fusion due to the deep structure of Match. To adequately use HERs to guide the generation of $\mathcal{D}$, we introduce a Hierarchical Fusion Strategy (HFS) that integrates $\mathbf{F}$ into multiple intermediate layers of the transformers in Match. Details of HFS will be given in Sec. 3.3.

To further handle the significant intra-class visual variations and enhance the discriminative ability of $\mathbf{F}$ against diverse object appearance, we propose the Global Discrimination Enhancement (GDE) constraint on $\mathbf{F}$ at both image sample and semantic class levels to enhance the intra-class invariance and the inter-class discrimination. The final training objective for the overall HER-Count framework can be formulated as:

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_{gde}\mathcal{L}_{gde}, \tag{5}$$

where $\mathcal{L}_{gde}$ denotes the objective function of GDE and λ_{gde} denotes its loss weight. Details of $\mathcal{L}_{gde}$ will be given in Sec. 3.4.

Existing methods obtain counting results by simply summing $\mathcal{D}$. However, this fails to provide accurate integer counts and also loses the object locations. Moreover, the summation procedure may introduce substantial errors with uncontrollable background noise, limiting its applications in real-world scenarios. To address these issues, HER-Count adopts the Local-Maxima Detection (LMD) strategy, which is widely used in keypoint localization works [3, 14], to count the objects in $\mathcal{D}$. Compared with summation, LMD is more robust to background noise and additionally able to provide object locations. The final counting results obtained by LMD can be formulated as:

$$N = \mathrm{LMD}(\mathcal{D}). \tag{6}$$

It is worth noting that, LMD is previously used in many localization works and is not the core of our HER-Count framework. Befitted by discriminative HERs that are trained by GDE and fused into $\mathcal{D}$ by HFS, HER-Count is able to predict high quality $\mathcal{D}$ with clear object boundaries, as shown in Fig. 4.

The overview of our HER-Count framework is illustrated in Fig. 2. As shown in the figure, $\mathbf{F}$ is generated via an MLLM based on input $\mathcal{I}$ and $\mathcal{T}$. A vision encoder is used to encode both $\mathcal{I}$ and $\mathbf{F}$ into intermediate features with HFS. A vision decoder decodes these intermediate features to $\mathcal{D}$, which is finally decoded into counting result via LMD. The following parts introduce the details of each component in HER-Count.

3.2 Hyper-Exemplar Representation

As shown in Fig. 2, the MLLM used for HER generation consists of an image encoder Φ_{I} and a multi-modal large language encoder Φ_{T}. Given $\mathcal{I}$ and $\mathcal{T}$, Φ_{I} is first employed to encode $\mathcal{I}$ into a sequence of image tokens. To synthesize $\mathbf{F}$ that covers all instance appearances in $\mathcal{I}$, a special token `</s>` is appended after $\mathcal{T}$ to encode the overall multi-modal cues of both image and text tokens. Thus, the input text is: [$\mathcal{T}$,`</s>`]. The image tokens and the text tokens are concatenated and fed into Φ_{T} together to obtain the embeddings of the last n transformer layer in Φ_{T} as:

$$\{\mathbf{e}_i | i = L - n + 1, ..., L\} = \Phi_{\mathrm{T}}(\Phi_{\mathrm{I}}(\mathcal{I}), [\mathcal{T}, \texttt{</s>}]), \tag{7}$$

where L denotes the depth of Φ_T.

Then, the embedding $\mathbf{e}_i$[`</s>`] corresponding to the token `</s>` is mapped by a Multi-Layer Perceptron (MLP) to generate the HER embedding $\mathbf{f}_i$ as:

$$\mathbf{f}_i = \mathrm{MLP}(\mathbf{e}_i[\texttt{</s>}]). \tag{8}$$

In this paper, n HER embeddings from the last n transformer layers in Φ_T are selected to compose the set of HERs, *i.e.*, $\mathbf{F}$, which can be denoted as:

$$\mathbf{F} = \{\mathbf{f}_i | i = L - n + 1, ..., L\}. \tag{9}$$

Different from previous methods that detect and select patches as exemplars, the proposed HERs are directly generated based on the holistic image and the textual class name, thus avoiding detection errors and being able to integrate cues from all objects in $\mathcal{I}$. This makes $\mathbf{F}$ a comprehensive capture of intra-class variance in the specific image, thus enabling more accurate counting results. Subsequently, $\mathbf{F}$ is used to guide the vision encoder Φ_{enc} by the proposed HFS to encode $\mathcal{I}$ into intermediate features $\mathcal{H}$, which can be formulated as:

$$\mathcal{H} = \Phi_{enc}(\mathbf{F}, \mathcal{I}). \tag{10}$$

The vision decoder Φ_{dec} is then used to decode $\mathcal{H}$ to $\mathcal{D}$ as:

$$\mathcal{D} = \Phi_{dec}(\mathcal{H}). \tag{11}$$

$\mathcal{D}$ is finally used to compute the counting result via LMD as shown in Eqn. (6).

Unlike sparse raw patches that amplify instance variations and hurt accuracy, $\mathbf{F}$ aggregates all target instances into the latent embedding corresponding to token `</s>`. The image tokens, class name text token, and `</s>` are concatenated together and fed into MLLM as shown in Eqn. (7). Cross-modal attentions in MLLM guide $\mathbf{e}_i$[`</s>`] to encode holistic visual cues of all target instances with the text token specifying semantic

class and image tokens providing visual cues. In MLLM, rich visual cues enable fine-grained discrimination of both intra-class variations and inter-class distinctions, while text-guided cross-modal attentions actively aggregate aligned visual cues to enhance the class-specific representative capacity of $\mathbf{F}$. The proposed GDE in Sec. 3.4 will further boost the discriminative power of HERs.

3.3 Hierarchical Fusion Strategy

After synthesizing $\mathbf{F}$, the key challenge is to inject its rich cues of instance appearance variations into Φ_{enc} to guide the feature extraction for counting as formulated in Eqn. (10), while keeping the decoding processing effective. A straightforward method is to directly fuse $\mathbf{F}$ with the output features of Φ_{enc} as: $\mathcal{H} = \mathbf{F} \odot \Phi_{enc}(\mathcal{I})$. $\odot$ denotes the Hadamard product. However, this vanilla method injects the specific instance appearance cues relevant to counting only at the late stage of feature extraction, preventing earlier layers of Φ_{enc} from dynamically extracting features according to the counting objective, thus limiting the counting accuracy.

In order to integrate the abundant multi-modal cues embedded in $\mathbf{F}$ into the extraction of $\mathcal{H}$, we propose the HFS to enhance the feature extraction by hierarchically fusing multiple HERs into different layers of Φ_{enc}. Since HERs in $\mathbf{F}$ are extracted layer-wise from different layers in Φ_{T}, they convey hierarchical cues of target objects that range from shallow texture to deep semantics. Thus, each HER embedding is able to provide specific visual and semantic cues for different vision encoder layers.

As shown in Fig. 2, given the stacked layers in Φ_{enc}, multi-modal cues embedded in HERs are integrated layer-wise into those intermediate layers. Specifically, the output of the i-th layer in Φ_{enc} can be formulated as:

$$\mathbf{h}_i = \mathcal{P}_i(\mathbf{h}_{i-1}) \odot \mathbf{f}_i, \tag{12}$$

where $\mathcal{P}_i$ denotes the i-th layer in Φ_{enc}, and $\mathbf{h}_i$ denotes the intermediate feature of $\mathcal{P}_i$. Via Eqn. (12), the multi-modal HER embedding $\mathbf{f}_i$ provides specific guidance on $\mathbf{h}_i$ to focus on both appearance and semantic cues that are most relevant to the counting objective, yielding a specific $\mathbf{h}_i$ for the given $\mathcal{I}$ and $\mathcal{T}$. HFS layer-wise integrates HERs into the intermediate layers of Φ_{enc}, thus the extracted features are better aligned with the accurate object counting task.

Since $\mathbf{h}_i$ is extracted by $\mathcal{P}_i$ upon $\mathbf{h}_{i-1}$, the intermediate feature of the last layer in Φ_{enc} already contains hierarchical cues extracted from both $\mathcal{I}$ and $\mathbf{F}$. Therefore, we chose $\mathbf{h}_L$ of the last layer in Φ_{enc} as the input of Φ_{dec}, *i.e.*, $\mathcal{H} = \mathbf{h}_L$. Then, the generation of $\mathcal{D}$ in Eqn. (11) can be rewritten as:

$$\mathcal{D} = \Phi_{dec}(\mathbf{h}_L). \tag{13}$$

HFS effectively enhances the image feature map with multi-modal category cues, which facilitates following zero-shot counting.

3.4 Global Discrimination Enhancement

HERs integrate customized cues into the intermediate feature extraction via the proposed HFS as formulated in Eqn. (12). Therefore, HERs should be both robust against

intra-class diversity and discriminative to inter-class similarity, in order to avoid missed instances and false positives in $\mathcal{D}$, respectively. However, the commonly used MSE loss on $\mathcal{D}$ only provides indirect constraint when learning HERs.

To directly enhance the discriminability of HERs, we propose the GDE to optimize them, as shown in Fig. 2. GDE contains constraints at both image sample and semantic class levels. Specifically, HERs of each image and average HERs of each class serve as the optimization anchors at the image sample level and the semantic class level, respectively. GDE then encourages each HER embedding to simultaneously approach all its positive anchors and move away from all its negative anchors at these two levels. By this global distance optimization, GDE enhances the intra-class variation robustness and the inter-class similarity discrimination of HERs, thus enabling accurate density maps for the counting task.

Since $\mathbf{f}_L$ is learned based on preceding HERs and its gradient can also update them, we only use $\mathbf{f}_L$ in the last layer of Φ_{T} for computing GDE. For the computation of GDE, we propose that GDE should satisfy three conditions. 1) GDE should simultaneously constrain relationships of each $\mathbf{f}_L$ to its positive and negative anchors. 2) Rather than absolute distance, GDE should optimize relative similarities between each $\mathbf{f}_L$ and anchors. 3) GDE should be aware of the hardness and produce larger gradients for harder HERs, which are far away to its positive anchors or close to its negative anchors. Based on these motivations, we formulate the objective function of GDE with respect to $\mathbf{f}_L$ as:

$$\mathcal{L}_{gde} = \underbrace{\overline{\log(1+\frac{S_n^I}{\mathrm{sim}(\mathbf{f}_L, A^I[p^I])})}}_{Image\ Level:\ \mathcal{L}_{gde}^I} + \underbrace{\overline{\log(1+\frac{S_n^C}{\mathrm{sim}(\mathbf{f}_L, A^C[p^C])}}}_{Class\ Level:\ \mathcal{L}_{gde}^C}, \tag{14}$$

where overlines indicate averaging. $\mathcal{L}_{gde}^I$ and $\mathcal{L}_{gde}^C$ denote constraints at the image level and the class level, respectively. A^I and A^C denote anchors at the two levels. p^I and p^C denote indexes of positive anchors for $\mathbf{f}_L$ in A^I and A^C. $\mathrm{sim}(\mathbf{f}_L, A^I[p^I]) = \exp(\mathbf{f}_L)^\top \cdot A^I[p^I]/\tau)$ and $\mathrm{sim}(\mathbf{f}_L, A^C[p^C]) = \exp(\mathbf{f}_L)^\top \cdot A^C[p^C]/\tau)$ denote similarities with positive anchors. $\tau = 0.05$ is the temperature coefficient. S_n^I and S_n^C denote the sum of similarities between $\mathbf{f}_L$ and its negative anchors in A^I and A^C, respectively.

$\mathcal{L}_{gde}$ in Eqn. (14) dynamically focuses on hard samples without any sampling strategy. For example, let $\mathrm{sim}[p^I]$ denote $\mathrm{sim}(\mathbf{f}_L, A^I[p^I])$, then $\frac{\partial \mathcal{L}_{gde}}{\mathrm{sim}[p^I]} = -\frac{S_n^I}{(\mathrm{sim}[p^I]+S_n^I)\cdot \mathrm{sim}[p^I]}$, whose scale increases with $\mathrm{sim}[p^I]$ decreasing. Therefore, $\mathcal{L}_{gde}^I$ focuses on hard boundary images and provides fine-grained supervision. Moreover, $\mathcal{L}_{gde}^C$ focuses on hard categories and provides global regularization. Thus, $\mathcal{L}_{gde}^I$ and $\mathcal{L}_{gde}^C$ are complementary to each other, and $\mathcal{L}_{gde}$ combines them to effectively optimize HERs.

$\mathcal{L}_{gde}$ optimizes HERs and enhances the counting accuracy along with $\mathcal{L}_{mse}$. We therefore train the framework with both $\mathcal{L}_{mse}$ and $\mathcal{L}_{gde}$, as formulated in Eqn. (5). Compared with $\mathbf{f}_L$, the intermediate features in Φ_{enc} embed more specific object cues such as object locations and numbers. Therefore, it is not reasonable to additionally constrain the distance relationship of $\mathbf{h}_L$. Moreover, since $\mathbf{h}_L$ is extracted based on $\mathbf{f}_L$, $\mathcal{L}_{gde}$ on $\mathbf{f}_L$ can also improve the accuracy of $\mathbf{h}_L$, thereby boosting the counting performance.

4 Experiment

4.1 Datasets

To evaluate our methods, we conduct experiments on three widely used object counting datasets, *i.e.*, FSC147 [19] and CARPK [4].

FSC147 consists of 6,135 images across 147 object classes, varying from animals, kitchen utensils, to vehicles. The classes in the training, val, and test sets do not overlap. For each image, FSC147 provides a class name and dot annotations of objects.

CARPK offers a bird's-eye view of 89,777 cars in 1,448 parking lot images. Following previous works [6, 28], we use CARPK for the cross-dataset validation.

Following previous works [6,16,28], we report the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) as evaluation metrics in our experiments.

4.2 Implementation Details

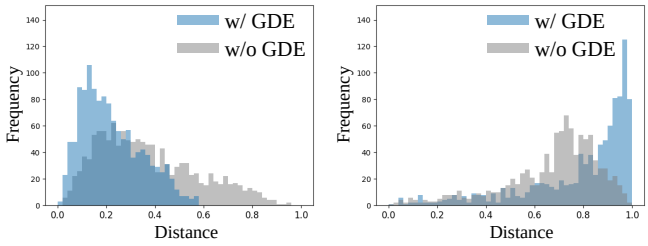
We use Qwen3-VL-2B [25] as our MLLM for synthesizing HERs with given pairs of images and class names. We use HRNet-W32 [21] as Φ_{enc} and 4 convolution layers with 256 channels as vision decoder Φ_{dec}. The input image size is set as 384×384. n in Eqn. (7) is set to 4. The AdamW optimizer is used for training. The learning rate is initialized as 0.03 and decays to 0 with cosine schedule. The framework is totally trained for 100 epochs on FSC147 dataset, taking around 5 hours on 4 NVIDIA 4090 GPU. The batch size is set as 16 and weight decay is 0.05. λ_{gde} is set as 0.1. During testing, we resize the shorter side of input image to 384 and keep the aspect ratio between height and width. In LMD, the threshold is 0.05 and the filter size is 3×3.

4.3 Ablation Study

Effectiveness of components: The evaluation of components in our method is summarized in Table 1, where "HER" denotes only integrating $\mathbf{f}_L$ in the last layer of Φ_{enc} and training the model without GDE. "GDE^I" denotes only using $\mathcal{L}^I_{gde}$ at the image level in Eqn. (14). "HER-Count" equips both HFS and GDE. It is clear that simply using the proposed HER already achieves promising counting performance, *e.g.*, MAE=16.85 on FSC147-Val, outperforming recent method VA-Count [28]. Training with HFS substantially boosts the performance. This shows that HFS can effectively aggregate cues in $\mathbf{F}$ into Φ_{enc} and thus extract features aligned with the counting task. Table 1 also shows that GDE can reduce the counting error by enhancing the discriminability of HERs. These experiments show that each component in HER-Count is important for the performance enhancement, and their combination achieves the best performance.

Table 1: Component evaluation on FSC147.

Method	FSC147-Val		FSC147-Test	
	MAE	RMSE	MAE	RMSE
HER	16.85	67.96	17.17	107.65
HER + HFS	15.53	59.71	15.71	96.12
HER + GDE^I	15.18	60.35	16.25	97.97
HER + GDE	14.33	58.11	15.56	96.79
HER-Count	**13.56**	**55.52**	**14.23**	**95.31**



(a) intra-class distance (b) inter-class distance

Fig. 3: Statistics of the intra-class and inter-class distance between HERs on FSC147 val dataset.

To further validate the effectiveness of GDE, we calculate the distances of intra-class and inter-class HERs after L2-normalization with (w/) and without (w/o) GDE on FSC147 val dataset, and the statistics are illustrated in Fig. 3. It is clear that GDE substantially reduces the intra-class distance and increases the inter-class discriminability. This implies that the model trained with GDE is more robust to intra-class variations and more discriminating to inter-class similarities.

Synthesizing HERs with different inputs: This section evaluates different inputs for synthesizing HERs on FSC147. Results are shown in Table 2. “Image-only" only encodes the image into HERs, while “Text-only" only encodes the class name. “Multi-modal" encodes both the image and class name. As shown, both “Image-only" and “Text-only” perform worse than “Multi-modal”. This is because objects of the same class can exhibit significant appearance variations across images, and the multi-modal hyper-exemplar can capture the variation in a specific image, offering image-specific customization. On the other hand, the text-only hyper-exemplar degrades the method to reference-less counting. For images with multiple object categories, the model struggles to accurately identify the correct category, leading to inferior performance. This experiment highlights the effectiveness of multi-modal learning for hyper-exemplar generation. It is worth noting that our GDE and HFS make “Text-only” a strong baseline, already surpassing recent SOTA method T2ICount [17] in RMSE on FSC147 val. Thus, the improvement of “Multi-modal” over “Text-only” is moderate.

Table 2: Evaluation of different inputs for synthesizing HERs on FSC147.

Input	Val set		Test set	
	MAE	RMSE	MAE	RMSE
Image-only	16.55	60.56	16.83	101.76
Text-only	15.77	58.73	15.13	99.56
Multi-modal	**13.56**	**55.52**	**14.23**	**95.31**

Analysis of LMD threshold: This section evaluates the influence of LMD threshold and the result on FSC147-val with different thresholds is summarised in Table 3. As shown in the table, the performance

Table 3: Analysis of LMD threshold.

thre.	0.01	0.03	0.05	0.09	0.15	0.2
MAE	13.89	13.68	13.56	13.71	13.84	13.92

of our method is robust to LMD threshold because of the high quality of heat maps predicted by HER framework.

Table 4: Comparison with recent methods on FSC147 dataset.

Method	Source	Input size	Val set		Test set	
			MAE	RMSE	MAE	RMSE
ZSC [24]	CVPR2023	384	26.93	88.63	22.09	115.17
CLIP-Count [6]	ACMMM2023	224	18.79	61.18	17.78	106.62
PseCo [5]	CVPR2024	1024	23.90	100.33	16.58	129.77
VA-Count [28]	ECCV2024	384	17.87	73.22	17.88	129.31
DAVE$_{prm}$ [16]	CVPR2024	512	15.48	52.57	14.90	103.42
DAVE$_{0-shot}$ [16]	CVPR2024	512	15.54	52.67	15.14	103.49
GeCo [15]	NeurIP2024	1024	14.81	64.95	13.30	108.72
YOLO-Count [27]	ICCV2025	640	15.43	58.36	14.80	96.14
T2ICount [17]	CVPR2025	384	13.78	58.78	**11.76**	97.86
HER-Count	Ours	384	**13.56**	**55.52**	14.23	**95.31**

Table 5: Cross-dataset evaluation on CARPK. "C" denotes CARPK and "F" denotes FSC147.

Method	Source	#Shot	C → C		F → C	
			MAE	RMSE	MAE	RMSE
GMN	CVPR22	3	7.48	9.90	-	-
BMNet+	CVPR22	3	5.76	7.83	10.44	13.77
CounTR	BMVC22	3	5.75	7.45	-	-
RCC	CVPR23	0	9.21	11.33	21.38	26.61
CLIP-Count	MM23	0	-	-	11.96	16.61
VA-Count	ECCV24	0	8.75	10.30	10.63	13.20
HER-Count	Ours	0	**5.19**	**6.77**	**5.56**	**7.14**

4.4 Comparisons

Comparison on FSC147: We first compare our method with SOTA zero-shot methods on the FSC147 dataset, as summarized in Table 4. For methods that report multiple performances under different settings, we only select their results under the zero-shot setting, *i.e.*, only using textual names without annotated patches. For DAVE, two variants

are reported, *i.e.*, DAVE$_{prm}$ only uses textual names and DAVE$_{0-shot}$ counts salient objects without names or patches. As shown in Table 4, our method outperforms all compared methods in accuracy. The comparison on FSC147 shows the strong comprehension and generalization ability of our method.

Comparison on CARPK: We further conduct a challenging cross-dataset validation on CARPK [4] to verify the generalization ability of HER-Count. Experimental results are shown in Table 5. "C→C" denotes the model is trained and tested on CARPK, while "F→C" denotes the model is trained on FSC147 and tested on CARPK. As shown in Table 5, HER-Count outperforms all compared methods under the cross-domain setting by a clear margin, demonstrating its superior generalization ability.

4.5 Model Efficiency

We compare the inference speeds and parameter numbers between HER-Count and recent SOTA methods on FSC147-Val in Table 6. FPS is tested on one 4090 GPU with author-released codes. All inference speeds are tested with batch size $= 1$. As shown, HER-Count achieves superior inference speed among the compared methods.

Table 6: Model efficiency comparison on FSC147-Val.

Method	Source	MAE	RMSE	# of param.	FPS
VA-Count	ECCV24	17.87	73.22	0.95B	6.4
GeCo	NeurIPS24	14.81	64.95	1.2B	1.1
T2ICount	CVPR25	13.78	58.78	2.0B	1.4
HER-Count (CLIP)	Ours	14.73	59.64	**0.44B**	**27.1**
HER-Count	Ours	**13.56**	**55.52**	2.1B	14.1

We also test our HER-Count with smaller model size. We use CLIP (ViT-L/14) as a smaller variant with only 437M parameters. The outputs of CLIP vision and text encoders are jointly fed into a self-attention layer to generate HERs. Our CLIP variant, denoted as HER-Count (CLIP) in Table 6, achieves comparable performance with T2ICount which has 2B parameters. This indicates that our HER-Count is also effective with smaller model size and the performance gains come from our new framework to synthesize HERs, rather than the model size of MLLM.

4.6 Visualization

To further validate the effectiveness of our HER-Count framework, we visualize the predicted density maps and counting results of HER-Count, an image patch exemplar-based methods VA-Count [28], and a text embedding based method T2ICount [17] on FSC147, as shown in Fig. 4. It is clear that, both VA-Count and T2ICount generate inaccurate density maps. For examples, density maps by VA-Count exhibit widespread high responses in the background (row 2, column 3), and density maps by T2ICount form contiguous response regions (row 3, column 4). These causes inaccurate counting results, and also loses object locations. The visualization of selected exemplar patches highlighted by red bounding boxes further explains the drawback of the exemplar selection-based methods. For example, in the first row, VA-Count only selects bottle caps in specific color as exemplars, which causes the missed detections of the bottle caps in other colors. Moreover, in the third row, both VA-Count and T2ICount fail to correctly understand the concept of polka dots, leading to incorrect density maps.

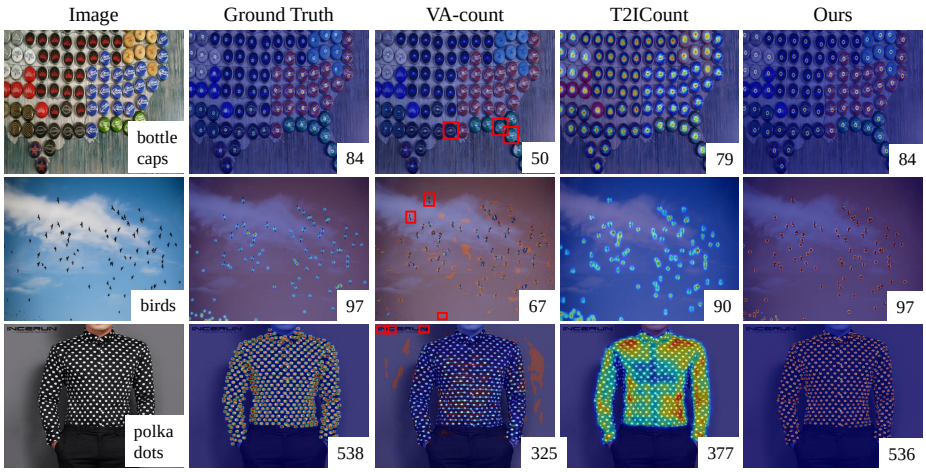


Fig. 4: Illustration of the predicted density maps by our HER-Count, image patch exemplar-based methods VA-Count [28], and text embedding based method T2ICount [17] on FSC147-Val. The selected image patches by VA-Count are highlighted with red bounding boxes. It is clear that HER can better identify the target objects, and its predicted density maps have clearer boundaries.

Different from previous works, HER-Count directly generates the multi-modal hyper-exemplar representation via a MLLM from both image and class name, avoiding detection errors and enhancing discriminative capacity. As shown in Fig. 4, the high quality density maps by HER-Count exhibit clear boundaries at object centers, thus enabling accurate counting and object location decoding. These advantages further highlight the novelty of our HER-Count framework for the zero-shot object counting task.

5 Conclusion

This paper introduces the novel HER-Count framework that learns representative Hyper-Exemplar Representations (HERs) for zero-shot counting. Different from previous works that select image patches or use text embeddings as exemplars to guide the counting, HER-Count explicitly extracts the representative multi-modal HERs from both image and class name via an MLLM. A hierarchical fusion strategy and global discrimination enhancement are designed to improve the representative and discriminative capacities of HERs. Extensive experiments on three widely used benchmarks demonstrate the superiority of HER-Count over recent methods in both accuracy and generalization abilities. To the best of our knowledge, HER-Count is the first counting framework that explicitly exploits MLLM for hyper-exemplar learning. We hope this work could inspire future research on zero-shot object counting.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62401294, 62473208, and 92367105, in part by the National Key Research and Development Program of China under Grant 2024YFB4708900, in part by the China Postdoctoral Science Foundation - Tianjin Joint Support Program under Grant Number 2025T020TJ, in part by the China Postdoctoral Science Foundation under Grant Number 2025M771529, in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20240753, and in part by the Fundamental Research Funds for the Central Universities under Grant Number 63253238.

Finally, Jianing Li and Xiaobin Liu would like to thank Yuning Zhou and Xixi Wang, respectively, for their invaluable support over the years.

References

1. Arteta, C., Lempitsky, V., Zisserman, A.: Counting in the wild. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 483–498. Springer (2016) 1, 3
2. Babu Sam, D., Agarwalla, A., Joseph, J., Sindagi, V.A., Babu, R.V., Patel, V.M.: Completely self-supervised crowd counting via distribution matching. In: European Conference on Computer Vision. pp. 186–204. Springer (2022) 1, 3
3. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020) 6
4. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the IEEE international conference on computer vision. pp. 4145–4153 (2017) 3, 10, 13
5. Huang, Z., Dai, M., Zhang, Y., Zhang, J., Shan, H.: Point segment and count: A generalized framework for object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17067–17076 (2024) 2, 4, 12
6. Jiang, R., Liu, L., Chen, C.: Clip-count: Towards text-guided zero-shot object counting. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4535–4545 (2023) 2, 4, 5, 10, 12
7. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 4
8. Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814 (2024) 5
9. Liu, C., Zhong, Y., Zisserman, A., Xie, W.: Countr: Transformer-based generalised visual counting. arXiv preprint arXiv:2208.13721 (2022) 4
10. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 5
11. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024) 5
12. Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 669–684. Springer (2019) 2, 4

13. Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K.: A large contextual dataset for classification, detection and counting of cars with deep learning. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 785–800. Springer (2016) 1, 3
14. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. Advances in neural information processing systems **30** (2017) 6
15. Pelhan, J., Lukezic, A., Zavrtanik, V., Kristan, M.: A novel unified architecture for low-shot counting by detection and segmentation. NeurIPS (2024) 12
16. Pelhan, J., Zavrtanik, V., Kristan, M., et al.: Dave-a detect-and-verify paradigm for low-shot counting. In: CVPR. pp. 23293–23302 (2024) 4, 10, 12
17. Qian, Y., Guo, Z., Deng, B., Lei, C.T., Zhao, S., Lau, C.P., Hong, X., Pound, M.P.: T2icount: Enhancing cross-modal understanding for zero-shot counting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25336–25345 (2025) 4, 11, 12, 13, 14
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
19. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3394–3403 (2021) 2, 3, 4, 10
20. Shi, M., Lu, H., Feng, C., Liu, C., Cao, Z.: 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. pp. 9529–9538 (2022) 2, 4
21. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020) 10
22. Wang, Z., Pan, Z., Peng, Z., Cheng, J., Xiao, L., Jiang, W., Cao, Z.: Exploring contextual attribute density in referring expression counting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19587–19596 (2025) 2, 4
23. Xie, W., Noble, J.A., Zisserman, A.: Microscopy cell counting and detection with fully convolutional regression networks. Computer methods in biomechanics and biomedical engineering: Imaging & Visualization **6**(3), 283–292 (2018) 1, 3
24. Xu, J., Le, H., Nguyen, V., Ranjan, V., Samaras, D.: Zero-shot object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15548–15557 (2023) 2, 4, 5, 12
25. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 10
26. Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C.: Class-agnostic few-shot object counting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 870–878 (2021) 2
27. Zeng, G., Zhang, X., Wang, Z., Xu, H., Chen, Z., Li, B., Tu, Z.: Yolo-count: Differentiable object counting for text-to-image generation. arXiv preprint arXiv:2508.00728 (2025) 4, 12
28. Zhu, H., Yuan, J., Yang, Z., Guo, Y., Wang, Z., Zhong, X., He, S.: Zero-shot object counting with good exemplars. In: European Conference on Computer Vision. pp. 368–385. Springer (2024) 2, 4, 5, 10, 12, 13, 14