GigaCount: Enhancing Crowd Counting by Integrating a Multi-Scale Feature Fusion Model into CLIP-EBC

Hoang Bach Nguyen Phan[1][0009−0007−1540−8622], Minh Nghia Le[1][0009−0006−5226−3699], An Tran Duc[1][0009−0003−1560−4325], and Minh Triet Tran[1][0000−0003−3046−3041]

University of Science, VNUHCM, Vietnam National University, HCMC

Abstract. Crowd counting has emerged as a vital task in computer vision, driving applications ranging from urban planning to public safety. Despite advances, challenges remain in handling diverse crowd scenarios, such as low-light scenes, distorted human figures, and extremely dense crowds. To handle these problems, we propose GigaCount, a multi-scale vision-language model that leverages *Contrastive Language-Image Pre-training with Enhanced Blockwise Classification* to enhance crowd counting performance. Our approach integrates ConvNeXt and its multi-scale feature fusion capabilities into CLIP, further addressing key challenges in crowd analysis. We evaluate the model's effectiveness by analyzing density maps and conducting ablation studies, which reveal patterns in prediction errors and their underlying causes. These findings guide targeted enhancements, including data augmentation to boost robustness across diverse lighting conditions, loss function adjustments to enhance accuracy in dense scenes, and layer removal to minimize model size and computational cost. Achieving a competitive MAE of 103.3, our model introduces a novel, lightweight architecture that integrates multi-scale feature fusion into CLIP's image encoder. Although it does not outperform state-of-the-art methods, this approach highlights the potential of multiscale vision-language models in crowd analysis and lays a foundation for further advancements. The implementation is available at https://github.com/AdamHermes/GigaCount.

Keywords: Crowd Counting · Multiscale Feature Fusion · Vision Language model · CLIP-EBC.

1 Introduction

Crowd counting is the task of estimating the number of individuals in various crowded scenes, such as public events and traffic congestion. In 2020, interest in this problem has increased significantly, particularly following the COVID-19 pandemic and the implementation of social distancing measures. Consequently, numerous high-quality datasets [17] have been developed to advance research in this area.

Traditional approaches in crowd counting primarily relied on convolutional neural network (CNN)-based regression models [1] that generated density maps using Gaussian kernels. Although these methods achieved promising results, they struggled to capture long-range dependencies within complex crowd scenes. To address this limitation, Vision Transformers (ViTs) [4] have been introduced, leveraging self-attention mechanisms to model global relationships across the image. Consequently, ViT-based methods have become the predominant framework in recent crowd counting research.

The emergence of CLIP [9] further advanced the field by aligning visual and textual representations through contrastive learning, enabling zero-shot transfer without task-specific fine-tuning. Ziang et al. [10] extended CLIP for crowd counting by integrating its vision-language representation with density estimation, transforming crowd counting from a regression to a classification problem.

In 2024, Yiming Ma et al. proposed **CLIP-EBC** [11], which incorporates an Enhanced Blockwise Classification framework into CLIP to mitigate ambiguity near count boundaries and improve prediction precision. As a result, this method achieves remarkable performance gains, enhancing classification-based approaches by up to 44.5% MAE score reduction on the UCF-QNRF dataset.

Through comprehensive experiments with CLIP-EBC, we identify several challenges affecting the model's performance. The heavy backbone caused slow training, and the model struggled with dark or low-contrast images, reducing object distinction accuracy. These limitations highlight the need for improvements in the model's architecture to enhance the overall performance.

To address these issues, we propose **GigaCount** that possess several key improvements:

- **Replacing the Backbone:** We replace CLIP image encoder's backbone with a modified version of ConvNeXt [13], which performs strongly in object counting and detection. ConvNeXt, as a fully convolutional neural network, is optimized for spatial feature extraction, making it highly suitable for counting tasks [13]. Most importantly, our modified version of ConvNext includes a feature fusion module to integrate hierarchical features from different levels of the image encoders, thus enhancing spatial resolution and enabling precise multi-scale object localization and counting.
- **Reducing Computational Cost/Training Time:** Compared to CLIP's ViT-B backbone with 86M parameters, our ConvNeXt-Tiny model is a much more lightweight backbone with 29M parameters. On a T4 GPU in Google Colab, CLIP-EBC takes 16 hours for 100 epochs, while our GigaCount model takes 7.5 hours. The lower computational cost also allows for larger batch sizes (ConvNeXt: 24–32 vs ViT : 16), further reducing training time to 4–5 hours.
- **Adaptive Hybrid Loss Function:** We introduce the Adaptive Hybrid Loss Function, which combines the probability Eucledian loss with the standard classification cross entropy loss and the DMCount loss to improve the overall accuracy of the final density map prediction.

2 Related Work

2.1 Vision Transformers

In 2022, Vision Transformers (ViTs) [4] emerged as a promising solution for crowd counting. First introduced by Dosovitskiy et al., they were later applied to this task in TransCrowd by D. Liang et al. [5]. Unlike CNNs, ViTs capture long-range spatial dependencies by dividing images into patches and using self-attention to model their relationships [5]. This enables effective analysis of local and global contexts, making ViTs suitable for densely populated complex scenes [6].

Our study suggests that the Vision Transformer approach provides a flexible and robust solution to the crowd counting problem, efficiently utilizing spatial data while requiring less supervision than traditional models. As of this writing, ViT- based hybrid models remain state-of-the-art in terms of performance. However, their complexity comes with drawbacks, particularly high computational demands and large training dataset requirements.

2.2 CLIP based methods

With the advancements in vision-language models, we find that CLIP (Contrastive Image-Language Pretraining)[9] shows outstanding performance when dealing with crowd counting tasks. A natural approach to leverage vision–language knowledge is to discretize the crowd count into a set of intervals, thereby reformulating the counting problem as a classification task rather than regression. In this setup, the similarity between the image embedding produced by the image encoder and the text embedding generated by the text encoder can be directly computed, with the count prediction being the label corresponding to the most similar image–text pair.[10]

CLIP-EBC CLIP-EBC [11] provides a solution that preserves the strengths of CLIP as an image–text encoder while introducing an enhanced blockwise classification (EBC) framework to reduce data bias and improve performance. In essence, EBC divides each image into blocks and assigns them to bins, where each bin corresponds to a specific range of people counts. These bins serve as labels, which are determined by computing the similarity between image–text pairs during training. During optimization, CLIP-EBC applies noise reduction to the images and addresses both the mismatch in count values and the loss in density maps by leveraging a unified loss function called Distance-Aware Cross-Entropy (DACE) loss[11].

We choose CLIP-EBC as the baseline due to its ability to generate density maps. While other CLIP-based methods focus solely on generating the count values for each image, CLIP-EBC produces a detailed density map for crowd analysis and visualization. This allows CLIP-EBC to compute the DMCount loss between the predicted density map and the ground truth one, further enhancing the model's convergence rate.

2.3 Fuss Free Network

Although ResNet has proven to be an effective backbone for CLIP's image encoder, our team believes that it does not retain enough spatial information when downsampled, which limits its ability to capture fine-grained details which are crucial for vision-language tasks. To address this limitation, we introduce the Fuss Free Network (FFN) [16], a lightweight feature fusion network proposed by Lei Chen et al .

Fuss Free Network is inspired by the ConvNeXt [13] architecture and introduces a simple yet effective fusion mechanism that enhances spatial awareness by preserving local and global context features during downsampling. Unlike standard hierarchical backbones that progressively discard resolutions, Fuss Free Network integrates feature maps from different layers using shortcut connections and multiscale fusion blocks, allowing it to maintain spatial granularity while producing semantically rich representations.

Fuss Free Network stands out for its plug-and-play design, requiring minimal computational overhead and training cost, yet delivering performance comparable to much heavier vision backbones. It has demonstrated superior performance in image-text retrieval and zero-shot classification tasks. Moreover, Fuss Free Network achieves strong results without any extensive pretraining or architectural complexity, making it an ideal choice for efficiency and interpretability in vision-language modeling[18].

3 Methods

3.1 Overview

In this work, we propose GigaCount, a multi-scale vision-language model for crowd density estimation, leveraging a modified version of ConvNeXt and a Feature Fusion module to enhance feature extraction and representation learning. An overview of our method is demonstrated in Fig 1

Our method builds upon CLIP by incorporating the Multi-Scale Feature Fusion techniques of Fuss Free Network with the Enhanced Blockwise Classification (EBC) strategy of CLIP-EBC. The key breakthrough of our approach is the integration of a modified version of ConvNeXt, a multi-scale feature fusion CNN model, into CLIP's image encoder.

This section is organized as follows: First, we describe the core CLIP's encoder blocks and its relevance to our approach. In this part, we describe the ModifiedConvNeXt image encoder along with the multiscale feature fusion module, which replaces the traditional ResNet and Vision Transformer architecture within CLIP. Then, we will explain how the Enhanced Blockwise Classification (EBC) strategy enhances the stability of our training process and improves overall model performance.

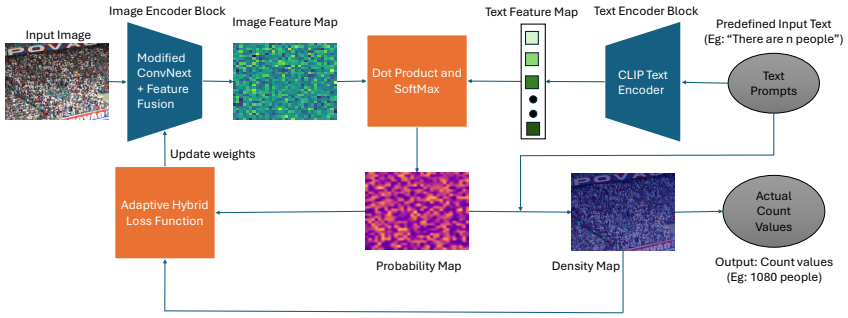

Fig. 1. Overview of GigaCount, our Modified ConvNeXt-based framework for crowd density estimation with text supervision. GigaCount integrates a feature fusion mechanism within a Modified ConvNeXt backbone to enhance multi-scale visual representation. The extracted image features are projected and combined with text feature maps obtained from CLIP's text encoder through a dot-product similarity operation to generate probability maps. These probability maps are subsequently transformed into density maps, which are further refined using an adaptive hybrid loss function. Finally, the refined density maps are aggregated to produce the estimated crowd count.

3.2 GigaCount - MultiScale Feature Fusion Model with CLIP-EBC

Image Encoder Block Our image encoder builds upon a modified ConvNeXt architecture, in which the standard ConvNeXt blocks are replaced with ODConv blocks, and the Stage 4 layer is omitted to improve spatial adaptability and reduce computational complexity.[13]. The image encoder first extracts multi-scale feature maps, which are subsequently refined using a composite attention pooling module. This module comprises Channel Attention, Filter Attention, Spatial Attention, and Kernel Attention [13], each designed to selectively emphasize informative features while suppressing less relevant ones.

To ensure effective aggregation of spatial and semantic information, we incorporate a Feature Fusion module, which integrates hierarchical features from different levels of the image encoder to enhance crowd localization and density prediction. The overall architecture of our ModifiedConvNext and the Feature Fusion module is illustrated in Fig 2

Specifically, given an input crowd image, the encoder applies multiple ODConv and batch normalization layers to extract feature representations. These features pass through three sequential downsampling stages, each producing a feature map that progressively deeper representations-shallow, mid-level, and deep features. The resulting multi-scale features are then upsampled and concatenated through a feature fusion module to generate the final image feature map. These image feature maps are later used by GigaCount to compute similarity with the text feature maps provided by CLIP Text Encoder.

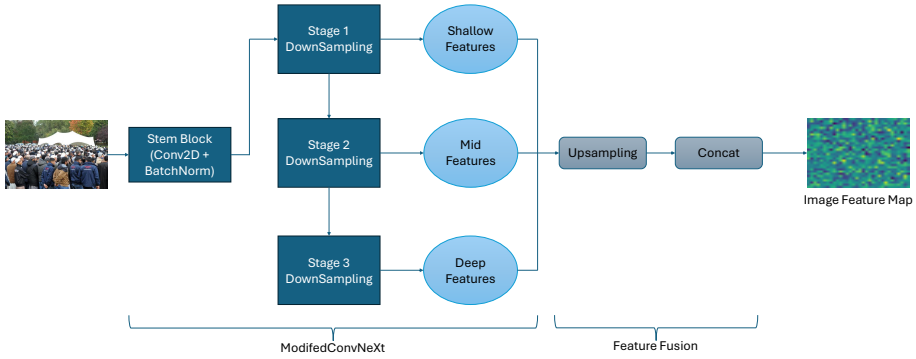

Fig. 2. Overview Archtecture of the Modified ConvNeXt and Multi-Scale Feature Fusion Module

Text Encoder Block For textual supervision, we adopt OpenAI's CLIP Text Encoder with frozen weights. A set of prompts such as "There are n people in the image" is tokenized and transformed into text feature matrices. These serve as semantic cues that provide count-aware guidance to the image encoder backbone.

Vision–Language Alignment and Density Map Generation GigaCount combines the image feature maps from the Image Encoder with the text feature maps from the Text Encoder to perform vision–language alignment in the CLIP embedding space. The text embeddings act as semantic guidance, allowing the model to associate visual regions of the image with discrete count levels. This is achieved by computing dot products between image features and text embeddings, followed by softmax normalization to obtain probability distributions over count bins. Aggregating these probabilities across the image produces a density map that reflects both spatial distribution and overall crowd count values. Finally, the Adaptive Hybrid Loss refines the predictions by jointly penalizing discrepancies in both count mismatches and density map quality.

EBC Strategy To demonstrate how EBC enhances training, we summarize how CLIP-EBC addresses the limitations of standard blockwise classification.

CLIP-EBC extends Standard Blockwise Classification (SBC) [14] through one important improvement: refined binning strategy.

Binning and Probabilistic Alignment: In the approach, image features are aligned with text embeddings via a dot product, followed by softmax normalization to yield a probability distribution over count bins. Each bin represents a discrete count range of people, and the binning scheme determines how the model allocates probability across count levels. Fine bins provide higher resolution but lower confidence due to sparsity, while coarse bins yield more stable but less precise estimates. Poor bin design can also worsen class imbalance, biasing the

model toward frequently occurring bins. To address this, CLIP-EBC employs a dynamic log-space binning scheme that balances resolution, sample density, and normalizes probability over count bins for more reliable aggregation.

Bin Strategy Crowd counting inherently suffers from highly imbalanced density distributions. Most regions of an image are background where they contain few to no people, while other regions may include up to hundreds of people. The EBC strategy addreses this by introducing three binning schemes: fine, coarse, and dynamic. In fine binning, each bin denotes a single count (one person) and serves as the class label for a given image region. Coarse binning groups counts in pairs to increase per-bin sample sizes; however, both approaches can introduce estimation bias.

Dynamic binning balances these trade-offs by assigning small, frequently occurring counts to individual bins, while grouping larger, less frequent counts into shared bins to reduce data imbalance. To further mitigate long-tailed distributions of count values, bins are normally divided in logarithmic space, producing finer resolution in dense regions and coarser resolution in sparse regions.

3.3 Adaptive Hybrid Loss Function

CLIP-EBC introduced the Distance-Aware Cross-Entropy (DACE) loss, which integrates DMCount loss [12] (to penalize discrepancies in count values) with the standard cross-entropy loss. This joint formulation ensures both classification consistency and accurate count prediction.

Our team acknowledges the power of DACE loss but through extensive experiments, we find there are still some limitations to the loss function. Therefore, we take inspiration from the DACE loss and derive another version of it to handle the training process.

The Adaptive Hybrid Loss function integrates three complementary components: a classification loss (CrossEntropy), a probability-matching loss, and the standard DMCount loss to optimize density map prediction. Specifically, the CrossEntropy term enforces alignment between predicted logits and the ground-truth bin assignments, the probability loss penalizes discrepancies between predicted softmax distributions and one-hot targets to encourage well-calibrated probabilities, and the DMCount term enforces consistency in the overall crowd count. The final loss is defined as:

$$\mathcal{L}_{\text{Hybrid}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Prob}} + \lambda_1 \mathcal{L}_{\text{Count}} \tag{1}$$

where:

- $\mathcal{L}_{\text{CE}}$: Cross-entropy loss between predicted class logits and ground-truth bin indices
- $\mathcal{L}_{\text{Prob}}$: Euclidean distance between predicted softmax probability distributions and one-hot ground-truth distributions, ensuring probability calibration and reducing overconfidence.
- $\mathcal{L}_{\text{Count}}$: DMCount loss for global count consistency [12].
- λ_1: Weighting factor for the count loss (normally selected as 1).

Classification Loss. Pixels are discretized into bins according to density thresholds:

$$C_{i,j} = \arg\max_k \mathbf{1}_{\rho_k \le D_{i,j} \le \rho_{k+1}} \tag{2}$$

The cross-entropy objective is:

$$\mathcal{L}_{\text{CE}} = -\sum_{i,j} C_{i,j} \log P_{i,j} \tag{3}$$

Probability Matching Loss. To align distributions, we apply a Euclidean penalty between predicted softmax probabilities $\hat{P}_{i,j}$ and the one-hot target distribution $\mathbf{1}_{C_{i,j}}$:

$$\mathcal{L}_{\text{Prob}} = \frac{1}{N} \sum_{i,j} \left\| \hat{P}_{i,j} - \mathbf{1}_{C_{i,j}} \right\|_2^2 \tag{4}$$

This formulation ensures the model not only predicts the correct bin index but also produces well-calibrated probability maps, while the DMCount term enforces global count consistency. Together, these losses yield more stable training and improved convergence compared to existing DACE loss methods.

4 Experiments

4.1 Experiments with GigaCount

We use the mean absolute error (MAE) and mean square error (MSE) to evaluate the crowd counting performance. These metrics are used to measure the mismatch in count values.

After training for 600 epochs, our model stabilized and achieved an MAE of 103.3 and an RMSE of 198.0. These results are summarized and compared with other methods in Table 1.

The visualization of our GigaCount model is illustrated in Fig 3.

Table 1. Comparison of GigaCount with baseline and other models

Method	Venue	QNRF	
		MAE	RMSE
CSS-CCNN-Random [3]	ECCV'22	718.7	1036.3
CSS-CCNN [3]	ECCV'22	437.0	722.3
CrowdCLIP [10]	CVPR'23	283.3	488.7
MCNN [1]	CVPR'16	277.0	515.0
TransCrowd [5]	2021	97.2	168.5
LoViTCrowd [6]	BMVC'22	87.0	141.9
DMCount [12]	NeurIPS'20	85.6	148.3
DMCount-EBC [11]	2024	77.2	130.4
CLIP-EBC (ResNet50, 2600 epochs) [11]	2024	80.5	136.6
CLIP-EBC (ViT/B-16, 2600 epochs) [11]	2024	87.7	159.3
GigaCount (Ours, 600 epochs)	–	**103.3**	**198.0**

Fig. 3. Performance of GigaCount on UCF-QNRF dataset. Input (left), Ground truth count (middle), Prediction count (right).

While our method does not yet surpass the top-performing approaches, it achieves competitive accuracy with a markedly more efficient and lightweight design. Unlike the CLIP-EBC baseline (ViT-B/16, 86M parameters, 2600 training epochs), our GigaCount model uses only 29M parameters and 600 epochs, resulting in substantially lower computational demands. On a T4 GPU for training, CLIP-EBC requires 16 hours for 100 epochs, whereas GigaCount completes the same in 7.5 hours. With larger batch sizes (24–32 vs. 16), training time can be further reduced to 4–5 hours, while inference remains comparable (5–20 s per image). Runtime comparisons are summarized in Table 2.

4.2 Ablation Studies for GigaCount

We perform simple ablation studies on our GigaCount to evaluate how reducing certain layers, applying learning rate scheduling, and introducing data augmentation affect overall model performance.

Table 2. Training and inference time comparison between CLIP-EBC and GigaCount on a T4 GPU.

Metric	CLIP-EBC	GigaCount
Training Time (100 epochs)	16 h	7.5 h (4–5 h with larger batch)
Inference Time (per image)	5–20 s	5–20 s

First, the original ConvNeXt-Tiny model includes a deep feature extraction component referred to as Stage 4, which typically captures high-level abstract features. To investigate its contribution, we removed this stage from the backbone. Surprisingly, excluding Stage 4 not only reduced the model's computational complexity but also decreased the overall MAE of the model. Second,

Table 3. Ablation Studies on GigaCount

Variants of GigaCount	**MAE**	**RMSE**
No Learning Rate Scheduler (30 Epochs)	456.75	739.20
With Learning Rate Scheduler (30 Epochs)	196.97	331.67
With Stage 4 (150 Epochs)	135.05	246.41
No Stage 4 (150 Epochs)	116.93	217.17
No Data Augmentation (200 Epochs)	121.76	227.49
With Data Augmentation (200 Epochs)	**113.74**	**214.18**

we experimented with different learning rate scheduling strategies to test model stability. Empirical results show that reducing the learning rate by half every 100 epochs and applying a cosine learning rate scheduler yields better performance compared to using a linear learning rate decay. Finally, although data augmentation slightly worsened MAE during early training, it proved beneficial in stabilizing the model and improving final performance in later epochs.

All experiment results from the ablation study on our GigaCount model are shown in Table 3. These results demonstrate that removing Stage 4 and applying data augmentation significantly improve the overall performance. In addition, setting up a proper learning rate scheduler also plays a critical role in optimization stability.

5 Conclusion

We present GigaCount, an efficient CLIP-based model for crowd counting, that integrates a simplified ConvNeXt backbone and a feature fusion module into the CLIP image encoder. This results in a lightweight architecture which reduces training time by approximately 2× while maintaining competitive accuracy.

We also modify the original CLIP-EBC loss by incorporating probability Euclidean loss between predicted probabilities and one-hot ground truths.

GigaCount's strengths lie in three components: (1) the ConvNeXt backbone effectively preserves spatial information critical for dense prediction tasks; (2) the feature fusion module enhances multi-scale representation learning; and (3) the Adaptive Hybrid loss promotes tighter alignment between predicted and actual class distributions, improving consistency and precision.

1) Impact Statement: Under experiments with UCF-QNRF, GigaCount demonstrates that efficient feature fusion technique within a lightweight modified ConvNeXt architecture can achieve competitive accuracy with significantly less training time than existing state-of-the-art methods.

2) Limitations: As stated before, the limited training time leads to some instability in our model predictions. While GigaCount demonstrates strong potential, further refinements and extended training are necessary to fully enhance its capabilities and minimize prediction errors.

Acknowledgments. This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

1. Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column CNN," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.
2. V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017.
3. V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
4. A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.
5. D. Liang *et al.*, "TransCrowd: Weakly-supervised crowd counting with transformers," *arXiv:2104.09116*, 2021.
6. N. H. Tran *et al.*, "Improving local features with relevant spatial information by Vision Transformer for crowd counting," *British Machine Vision Conf. (BMVC)*, 2022.
7. G. Sun *et al.*, "Boosting crowd counting with transformers," *IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, 2021.
8. H. Lin *et al.*, "Boosting crowd counting via multifaceted attention," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.
9. A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *Proc. Mach. Learn. Res. (PMLR)*, 2021.
10. D. Liang *et al.*, "CrowdCLIP: Unsupervised crowd counting via vision-language model," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
11. Y. Ma, V. Sanchez, and T. Guha, "CLIP-EBC: CLIP can count accurately through enhanced blockwise classification," *arXiv:2403.09281*, 2024.
12. B. Wang, H. Liu, D. Samaras, and M. Hoai, "Distribution matching for crowd counting," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
13. Z. Liu *et al.*, "A ConvNet for the 2020s," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.
14. L. Liu *et al.*, "Counting objects by blockwise classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3125–3139, 2019.
15. H. Xiong *et al.*, "From open set to closed set: Counting objects by spatial divide-and-conquer," *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.
16. L. Chen *et al.*, "The effectiveness of a simplified model structure for crowd counting," *arXiv:2404.07847*, 2024.
17. H. Idrees *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," *Eur. Conf. Comput. Vis. (ECCV)*, 2018.
18. D. Kong *et al.*, "A fuss-free approach to zero-shot image classification with CLIP," *arXiv:2303.08128*, 2023.