CrowdCLIP: Unsupervised Crowd Counting via Vision-Language Model

Dingkang Liang[*1], Jiahao Xie[*2], Zhikang Zou[3], Xiaoqing Ye[3], Wei Xu[2], Xiang Bai[†1]

[1]Huazhong University of Science and Technology, {dkliang, xbai}@hust.edu.cn
[2]Beijing University of Posts and Telecommunications, {xiejiahao, xuwei2020}@bupt.edu.cn
[3] Baidu Inc., China

Abstract

Supervised crowd counting relies heavily on costly manual labeling, which is difficult and expensive, especially in dense scenes. To alleviate the problem, we propose a novel unsupervised framework for crowd counting, named CrowdCLIP. The core idea is built on two observations: 1) the recent contrastive pre-trained vision-language model (CLIP) has presented impressive performance on various downstream tasks; 2) there is a natural mapping between crowd patches and count text. To the best of our knowledge, CrowdCLIP is the first to investigate the vision-language knowledge to solve the counting problem. Specifically, in the training stage, we exploit the multi-modal ranking loss by constructing ranking text prompts to match the size-sorted crowd patches to guide the image encoder learning. In the testing stage, to deal with the diversity of image patches, we propose a simple yet effective progressive filtering strategy to first select the highly potential crowd patches and then map them into the language space with various counting intervals. Extensive experiments on five challenging datasets demonstrate that the proposed CrowdCLIP achieves superior performance compared to previous unsupervised state-of-the-art counting methods. Notably, CrowdCLIP even surpasses some popular fully-supervised methods under the cross-dataset setting. The source code will be available at `https://github.com/dk-liang/CrowdCLIP`.

1. Introduction

Crowd counting aims to estimate the number of people from images or videos in various crowd scenes, which has received tremendous attention due to its wide applications in public safety and urban management [14, 42]. It is very challenging to accurately reason the count, especially

---
[*]Equal contribution. [†]Corresponding author.
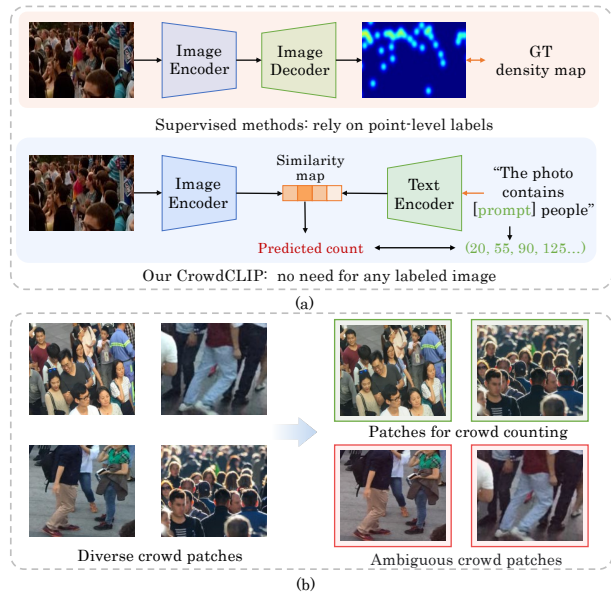Work done when Dingkang Liang was an intern at Baidu.

Figure 1. (a) The supervised methods require point-level annotations, which need heavy manual labor to label a large-scale dataset. The proposed method transfers the vision-language knowledge to perform unsupervised crowd counting without any annotation; (b) Crowd counting aims to calculate the number of human heads, while some crowd patches do not contain human heads, *i.e.*, ambiguous patches.

in dense regions where the crowd gathers.

The recent crowd counting methods [5, 18, 41, 62] attempt to regress a density map (Fig. 1(a)). To train such density-based models, point-level annotations are required, *i.e.*, assigning a point in each human head. However, annotating point-level object annotations is an expensive and laborious process. For example, the NWPU-Crowd [48] dataset, containing 5,109 images, needs 30 annotators and 3,000 human hours for the entire annotation process. To reduce the annotation cost, some weakly-supervised methods [16, 19, 59] and semi-supervised methods [22, 28] are proposed, where the former usually adopts the count-level

annotation as supervision, and the latter uses a small fraction of fully-labeled images and massive unlabeled images for training. However, both weakly and semi-supervised methods still need considerable label costs, especially when annotating dense or blurry images.

Considering the above issues, a crowd counting model that can be trained without any labeled data is worth exploring. So far, there is only one approach called CSS-CCNN [3] for pure unsupervised crowd counting. Based on the idea that natural crowds follow a power law distribution, CSS-CCNN ensures the distribution of predictions is matched to the prior. Though the performance of CSS-CCNN is better than the random paradigm, there is a significant performance gap compared to the popular fully supervised methods [4, 62]. Recently, Contrastive Language-Image Pre-Training (CLIP) as a new paradigm has drawn increasing attention due to its powerful transfer ability. By using large-scale noisy image-text pairs to learn visual representation, CLIP has achieved promising performance on various downstream vision tasks (*e.g.*, object detection [38], semantic segmentation [56], generation [10]). Whereas, how to apply such a language-driven model to crowd counting has not been explored. Obviously, CLIP cannot be directly applied to the counting task since there is no such count supervision during the contrastive pre-training of CLIP.

A natural way to exploit the vision-language knowledge is to discretize the crowd number into a set of intervals, which transfers the crowd counting to a classification instead of a regression task. Then one can directly calculate the similarity between the image embedding from the image encoder and the text embedding from the text encoder and choose the most similar image-text pair as the prediction count (called zero-shot CLIP). However, we reveal that the zero-shot CLIP reports unsatisfactory performance, attributed to two crucial reasons: 1) The zero-shot CLIP can not well understand crowd semantics since the original CLIP is mainly trained to recognize single-object images [38]; 2) Due to the non-uniform distribution of the crowd, the image patches are of high diversity while counting aims to calculate the number of human heads within each patch. Some crowd patches that do not contain human heads may cause ambiguity to CLIP, as shown in Fig. 1(b).

To relieve the above problems, in this paper, we propose CrowdCLIP, which adapts CLIP's strong vision-category correspondence capability to crowd counting in an unsupervised manner, as shown in Fig. 1(a). Specifically, first, we construct ranking text prompts to describe a set of size-sorted image patches during the training phase. As a result, the image encoder can be fine-tuned to better capture the crowd semantics through the multi-modal ranking loss. Second, during the testing phase, we propose a simple yet effective progressive filtering strategy consisting of three stages to choose high-related crowd patches. In particular, the first two stages aim to choose the high-related crowd patches with a coarse-to-fine classification paradigm, and the latest stage is utilized to map the corresponding crowd patches into an appropriate count. Thanks to such a progressive inference strategy, we can effectively reduce the impact of ambiguous crowd patches.

Extensive experiments conducted on five challenging datasets in various data settings demonstrate the effectiveness of our method. In particular, our CrowdCLIP significantly outperforms the current unsupervised state-of-the-art method CSS-CCNN [3] by 35.2% on the challenging UCF-QNRF dataset in terms of the MAE metric. Under cross-dataset validation, our method even surpasses some popular fully-supervised works [40, 62].

Our major contributions can be summarized as follows: 1) In this paper, we propose a novel unsupervised crowd counting method named CrowdCLIP, which innovatively views crowd counting as an image-text matching problem. To the best of our knowledge, this is the first work to transfer vision-language knowledge to crowd counting. 2) We introduce a ranking-based contrastive fine-tuning strategy to make the image encoder better mine potential crowd semantics. In addition, a progressive filtering strategy is proposed to choose the high-related crowd patches for mapping to an appropriate count interval during the testing phase.

2. Related Works

2.1. Fully-Supervised Crowd Counting

The mainstream idea of supervised methods [18, 25, 30, 41, 55, 64, 65] is to regress a density map, which is generated from an elaborately labeled point map. In general, the labeled points are hard to reflect the size of the head, meaning the density-based paradigm easily meets the huge variation issue. To tackle the scale variations, various methods make many attempts. Specifically, some works [41, 62] adopt multi-column networks to learn multi-scale feature information. Some methods propose to utilize the scaling mechanism [13, 35, 54, 55] or scale selection [45] to relieve the scale variations. The attention mechanism is also a valuable tool to improve the feature representation, such as self-attention [23], spatial attention [39, 57], and other customized attention blocks [6, 32, 51]. Different from regressing density maps, some methods [47, 52, 53] leverage supervised-classifier to classify the crowd into different intervals, achieving appealing performance.

Another trend is based on localization [2, 7, 21, 36, 46, 49, 50], which can be divided into three categories: predict pseudo-bounding boxes [29, 36, 49] or customize special localization-based maps [2, 21, 54], the other methods [20, 44] directly regress the point coordinates, removing the need for pre-processing or post-processing.

2.2. Weakly-/Semi-/Unsupervised Crowd Counting

The fully-supervised methods need expensive costs to label points for each head. To this end, weakly or semi-supervised methods are proposed to reduce the annotations burden. The weakly-supervised methods [15, 16, 19, 59] suggest using count-level instead of point-level annotation as the supervision. The semi-supervised methods [28, 31, 58] leverage small-label data to train a model and further use massive unlabeled data to improve the performance. Method in [37] optimizes almost 99% of the model parameters with unlabeled data. However, all the above methods still require some annotated data. Once the labeled data is left, these models can not be trained.

So far, only one method, CSS-CCNN [3], focuses on pure unsupervised settings, *i.e.*, training a model without any label. CSS-CCNN [3] argues that natural crowds follow a power law distribution, which could be leveraged to yield error signals for back-propagation. We empirically find that there is a significant performance gap between CSS-CCNN and some popular fully-supervised methods [60, 62]. In this paper, we propose a novel method named CrowdCLIP, which transfers the counting into an image-text matching problem, to boost the performance for unsupervised crowd counting by a large margin.

2.3. Vision-Language Contrastive Learning

Recently, vision-language pre-training (CLIP [34]) using massive image-text pairs from the Internet has attracted more and more attention. Several methods [10,17,33,38,56] transfer the vision-language correspondence to the downstream tasks, such as object detection [38], semantic segmentation [56], and generation [10]. Benefiting from the strong zero-shot ability of CLIP, these unsupervised methods achieve promising performance. In this paper, we study how to transfer the vision-language knowledge to the unsupervised crowd counting task.

3. Preliminary

In this section, we revisit CLIP [34] and then introduce the setting of unsupervised crowd counting.

3.1. A Revisit of CLIP

CLIP [34], a representative pre-trained vision-language model, focuses on how to build the connection between visual concepts and language concepts. CLIP contains two encoders used for encoding the image feature and text feature, respectively. Given an image-text pair, the CLIP aims to compute the semantic similarity between the encoding image feature and the encoding text feature. The pre-trained CLIP can be easily extended to the zero-shot/open-vocabulary image classification. Specifically, one can utilize a series of class names (*e.g.*, 'cat,' 'dog') for replac-ing the pre-defined text prompt template, *e.g.*, "a photo of [CLASS]". Then the text is fed into the text encoder to generate the class embeddings used to compute the similarity with image embedding for classification.

In this paper, we are the first to study how to extend the strong correspondence between the image and text of CLIP to the crowd counting task.

3.2. Unsupervised Crowd Counting

The goal of crowd counting is to estimate the pedestrian number given crowd images. Following the unsupervised definition from previous methods [3, 8], *i.e.*, during the training, the model does not need a single annotated image while allowing to use of the supervision provided by the original data. The manually annotated validation or test set is only used for evaluation. Note that using CLIP [34] to the downstream tasks without the task-related training label is in an accepted unsupervised manner [1, 38, 63].

4. Our Method

The overview of our method is shown in Fig. 2. During the training phase, we fine-tune the image encoder by introducing ranking text prompts while the parameters of the text encoder are frozen. During the testing phase, we propose a progressive filtering strategy that progressively queries the highly potential crowd patches and maps the filtered patches into the specific crowd intervals.

4.1. Ranking-based Contrastive Fine-tuning

In this section, we introduce how to fine-tune the CLIP [34] to improve the ability to extract crowd semantics. Note that the original CLIP needs image-text pairs to complete the contrastive pre-trained process. However, no label is provided in the unsupervised setting, *i.e.*, the fine-tuning lacks the corresponding text modality as supervision. To this end, we construct ranking prompts using texts of counting intervals to describe the size-ordered input images, as shown in Fig. 2(a). As a result, the image encoder can be fine-tuned through multi-modal ranking loss.

Image to patches. Given an input crowd image, we first crop a set of square patches $\{\mathcal{O}_M\}$, where M is the predefined number of patches. The cropped patches obey the following rules:

- For any two cropped image patches ($\mathcal{O}_i$ and $\mathcal{O}_j$, $0 \leq i < j \leq M - 1$), the size of patch $\mathcal{O}_i$ is smaller than the size of patch $\mathcal{O}_j$.

- The patches from the same image share the same image center. During the training, all patches are resized to the same size and fed into the image encoder to generate the image rank embeddings $\mathbf{I} = [\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_{M-1}]$, where $\mathbf{I} \in \mathbb{R}^{M \times C}$.

(a) Training phase

(b) Inference phase

Figure 2. Overview of our framework. (a) During the training phase, we fine-tune the image encoder by introducing ranking prompts while the parameters of the text encoder are frozen. (b) During the testing phase, we propose a progressive filtering strategy that progressively queries the most likely crowd patches and maps the filtered patches into the specific quantitive count.

Prompt design. The original CLIP [34] is not designed for the counting task. How to customize feasible text prompts needs to be studied. As mentioned above, we have successfully collected a series of patches whose sizes are ordinal. Obviously, the number of human heads from different patches is ordinality, and the larger patches correspond to the more or equal number of human heads. Thus, we design ranking text prompts to describe the ordinal relationship of image patches. Specifically, we propose to learn the rank embeddings to preserve the order of the image patches in the language latent space. The text prompt is defined as "There are [class] persons in the crowd", where [class] represents a set of base rank number $R = [R_0, R_0+K, ..., R_0+(N-1)K]$, where R_0, K and N denote the basic reference count, counting interval and number of class, respectively. The text prompts will be fed into the text encoder to obtain the output text rank embeddings $\mathbf{R}' = [\mathbf{R}'_0, \mathbf{R}'_1, ..., \mathbf{R}'_{N-1}]$, where $\mathbf{R}' \in \mathbb{R}^{N \times C}$.

Image encoder optimization. Suppose we have a set of image rank embeddings $\mathbf{I} \in \mathbb{R}^{M \times C}$ and text rank embeddings $\mathbf{R}' \in \mathbb{R}^{N \times C}$, where $\mathbf{I}$ and $\mathbf{R}'$ are obtained from the image encoder and text encoder, respectively. For the

image-language matching pipeline, we calculate the similarity scores between $\mathbf{I}$ and $\mathbf{R}'$ via inner product and obtain the similarity matrix $\mathbf{S} = [\mathbf{s}_{izj}]$, where $\mathbf{S} \in \mathbb{R}^{M \times N}$:

$$\mathbf{s}_{i,j} = \mathbf{I}_i \cdot \mathbf{R}'^{T}_{j}, \tag{1}$$

where $\mathbf{I}_i \in \mathbb{R}^{1 \times C}$, $\mathbf{R}'_j \in \mathbb{R}^{1 \times C}$, $0 \leq i \leq M - 1$ and $0 \leq j \leq N - 1$. Due to the inheritance of ranking relationships from the images and text prompts, we hope the similarity matrix $\mathbf{S}$ is a specific ordinal matrix (Fig. 2(a)):

$$\mathbf{s}_{i',i} \leq \mathbf{s}_{i,i}, \tag{2}$$

where $0 \leq i' \leq i \leq M - 1$. To preserve the order of the image-text pair in the latent space, we propose to optimize the image encoder through the multi-modal ranking loss. Specifically, we use the principal diagonal of the similarity matrix as the base and calculate the ranking loss from the bottom up:

$$L_r = \max(0, \mathbf{s}_{i',i} - \mathbf{s}_{i,i}). \tag{3}$$

We set $M = N$ in practice to guarantee that $\mathbf{S}$ is a square matrix. Eq.2 and Eq.3 build intrinsic correspondence between size-sorted patches and ranking text prompts, resulting in the similarity matrix (optimization goal, similar to

multi-modal similarity matrix like PointCLIP [61], Audio-CLIP [9]). During the fine-tuning, the weights of the text encoder are frozen, i.e., L_r aims to align the image embedding into the fixed ranking language space. In this way, the text embeddings are constrained in the well-learned language latent space, leading to robust generalization.

Note that the ranking loss we used is different from the previous methods [26, 27]. First, they calculate the ranking loss from one modality (i.e., only images), while ours is designed for the multi-modal (image and text). Second, they still demand labeled data since the ranking loss they used just can judge the order of given image patches, which cannot directly predict the number range of patches. In contrast, our approach does not need any labeled crowd images.

4.2. Progressive Filtering Strategy

In this part, we introduce the detail of the proposed progress filtering strategy consisting of three stages, used to select the real crowd patches and map them into appropriate count intervals at the inference stage, as shown in Fig. 2(b). For simplicity, we name the original image encoder and fine-tuned image encoder as $\mathcal{E}_o$ and $\mathcal{E}_f$, respectively. $\mathcal{E}_o$ is used to choose the high-confidence crowd patches, and $\mathcal{E}_f$ is used for the final counting.

Given an input image, we first divide it into a grid of $P \times P$ patches, then the patches and corresponding text prompts will be respectively fed into $\mathcal{E}_o$ and the first text encoder $\mathcal{T}_0$ to generate similarity scores for coarse classification. The text prompt of $\mathcal{T}_0$ is set to "The object is [class]", which aims to classify the patches into different categories with clear distinction (e.g., 'crowd', 'tree', 'car').

The selected crowd patches from the first stage may contain different components of the human, such as human heads, bodies, and legs. However, the crowd counting task aims to estimate the number of human heads instead of other components since only the heads are not easily obscured compared with other components. Thus, in the second stage, we adopt fine-grain text prompts and feed them into the second text encoder $\mathcal{T}_1$ to further filter the patches. The fine-grain text prompts are defined as "The objects are [class]," where [class] are some fine-grained categories (e.g., 'human heads,' 'human bodies'). As a result, high-confidence crowd patches with human heads are obtained. Note that in this stage, we still use the non-fine-tuned image encoder $\mathcal{E}_o$ for calculating the similarity.

In the third stage, we adopt the fine-tuned $\mathcal{E}_f$ as the image encoder, and the text prompts of the third $\mathcal{T}_2$ are the same as the fine-tuning phase, i.e., the ranking text prompts are defined as "There are [class] persons in the crowd," where [class] is the pre-defined ranking number R. The final count can be obtained by choosing the most similar image-text pair based on the image embedding and class embeddings from $\mathcal{E}_f$ and $\mathcal{T}_2$.

Note that in practice, $\mathcal{T}_0$, $\mathcal{T}_1$ and $\mathcal{T}_2$ share the same parameters, and all images share the same text prompts. In other words, we can get the text embeddings in advance instead of processing the text prompts in each inference phase, thus keeping efficiency.

5. Experiments

5.1. Dataset and Evaluation Metric

UCF-QNRF [12] is a dense counting dataset. There are $1,535$ images with crowd numbers varying from 49 to $12,865$. The images are split into the training set with $1,201$ images and the testing set with 334 images.

JHU-Crowd++ [43] is one of the largest counting datasets. It contains $4,372$ images with $1,515,005$ annotations. Specifically, $2,272,500$, and $1,600$ images are divided into the training, validation, and testing sets. There is a large percentage of images captured on rainy and foggy days. These degraded images increase challenges for crowd counting.

ShanghaiTech [62] includes Part A and Part B. Part A contains 482 images with $241,677$ annotations, the training and testing sets consisting of 300 and 182 images, respectively. Part B contains 716 images and a total of $88,488$ annotated head centers, where 400 images are used for training and the rest 316 images for testing.

UCF-CC50 [11] is a challenging counting dataset. It contains only 50 images but has $63,075$ annotated individuals, where the crowd numbers vary from 94 to $4,543$.

Evaluation metric. We use the mean absolute error (MAE) and mean square error (MSE) to evaluate the crowd counting performance. The two counting metrics are defined as: $MAE = \frac{1}{N_c}\sum_{i=1}^{N_c}|E_i - C_i|$, $MSE = \sqrt{\frac{1}{N_c}\sum_{i=1}^{N_c}|E_i - C_i|^2}$, where N_c is the number of images in the test set. E_i and C_i represent the estimated count and the ground truth of the i-th image, respectively.

5.2. Implement Details

All experiments are conducted on an Nvidia 3090 GPU. CLIP with ViT-B/16 backbone is used. In the fine-tuning phase, the parameters of the text encoder are frozen, and the RAdam optimizer [24] with a learning rate of 1e-4 is used to optimize the image encoder. The number of training epochs is set to 100. The M and N are set to 6. For the ranking text prompts, we set the basic reference count R_0 as 20 and the counting interval K as 35, i.e., the text prompts are "There are [20/55/90/125/160/195] persons in the crowd." Note that the ranking text prompts we used are kept the same in all datasets. During the testing phase, we set the P as 4 for UCF-QNRF and UCF_CC_50 datasets and set P as 3 for the rest datasets. For the large-scale datasets (i.e., UCF-QNRF, JHU-Crowd++), we make the longer size less than 2048, keeping the original aspect ratio.

Table 1. Comparison of the counting performance on the UCF-QNRF, JHU-Crowd++, ShanghaiTech Part A, Part B, and UCF_CC_50 datasets. Random* denotes that we randomly select a value from the pre-defined rank number R for each cropped patch.

| Method	Year	Label	QNRF		JHU		Part A		Part B		UCF_CC_50	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Zhang *et al.* [60]	CVPR 15	Point	-	-	-	-	181.8	277.7	32.0	49.8	467.0	498.5
MCNN [62]	CVPR 16	Point	277.0	426.0	188.9	483.4	110.2	173.2	26.4	41.3	377.6	509.1
Switch CNN [4]	CVPR 17	Point	228.0	445.0	-	-	90.4	135.0	21.6	33.4	318.1	439.2
LSC-CNN [36]	TPAMI 21	Point	120.5	218.2	112.7	454.4	66.4	117.0	8.1	12.7	225.6	302.7
CLTR [20]	ECCV 22	Point	85.8	141.3	59.5	240.6	56.9	95.2	6.5	10.2	-	-
CSS-CCNN-Random [3]	ECCV 22	None	718.7	1036.3	320.3	793.5	431.1	559.0	-	-	1279.3	1567.9
Random*	-	None	633.6	978.9	297.5	801.6	411.5	511.1	158.7	287.4	1251.6	1497.8
CSS-CCNN [3]	ECCV 22	None	437.0	722.3	217.6	651.3	197.3	295.9	-	-	564.9	959.4
CrowdCLIP (**ours**)	-	None	**283.3**	**488.7**	**213.7**	**576.1**	**146.1**	**236.3**	**69.3**	**85.8**	**438.3**	**604.7**
Improvement	-	-	*35.2%*	*32.3%*	*1.8%*	*11.5%*	*26.0%*	*20.1%*	*56.3%*	*70.1%*	*22.4%*	*37.0%*
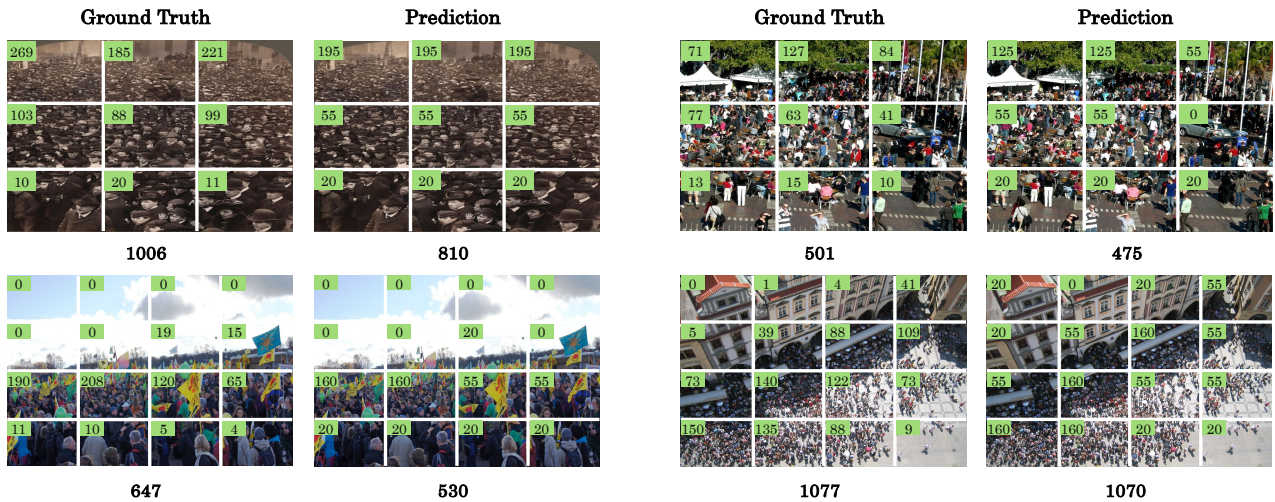
Figure 3. The first and second rows are selected from the ShanghaiTech Part A and UCF-QNRF datasets, respectively.

6. Results and Analysis

6.1. Comparison with the State-of-the-Arts

As reported in Tab. 1, the proposed CrowdCLIP goes beyond the state-of-the-art method [3] by a large margin in all evaluated datasets. Typically, for the UCF-QNRF, an extremely dense dataset, CrowdCLIP outperforms CSS-CCNN by 35.2% improvement for MAE and 32.3% improvement for MSE. For the degraded images (JHU-Crowd++ dataset), our method achieves 213.7 MAE and 576.1 MSE. Our method also reports remarkable performance on two sparse datasets, ShanghaiTech Part A and ShanghaiTech Part B. These impressive results verify that our method is robust in various complex conditions. We further provide some qualitative visualizations to analyze the effectiveness of our method, as shown in Fig. 3. Crowd-CLIP performs well in different scale scenes, although it is a pure unsupervised method.

Additionally, we can find that our unsupervised method still presents highly competitive performance compared with some popular fully-supervised methods [60, 62]. Specifically, on the UCF-QNRF dataset, our method is very close to the MCNN [62] in terms of MAE (283.3 vs. 277.0). An interesting phenomenon is that our method achieves better performance than the method in [60] on the ShanghaiTech Part A and UCF_CC_50 datasets.

6.2. Cross-Dataset Validation

Generally, scene variation could easily cause significant performance drops, while a crowd counting method with strong generalization ability is usually expected. So we adopt cross-dataset evaluation to demonstrate the generalization ability of CrowdCLIP. In this setting, the model is trained on one dataset while testing on the other.

Compared with fully-supervised methods. We first make comparisons between our CrowdCLIP and two fully-supervised methods [40, 62], as depicted in Tab. 2. Although our method does not require any annotated label,

Table 2. Experimental results on the transferability of our method and popular fully-supervised methods under cross-dataset evaluation.

Method	Year	Label	Part B→Part A		Part A→Part B	
			MAE	MSE	MAE	MSE
MCNN [62]	CVPR 16	Point	221.4	357.8	85.2	142.3
D-ConvNet [40]	ECCV 18	Point	**140.4**	**226.1**	**49.1**	99.2
CrowdCLIP (ours)	-	None	217.0	322.7	69.6	**80.7**

Table 3. Experimental results on the transferability of unsupervised methods under cross-dataset evaluation.

Method	Year	Label	Part A→QNRF		QNRF→Part A		Part A→JHU		JHU→Part A	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSS-CCNN [3]	ECCV 22	None	472.4	-	235.7	-	251.3	-	266.3	-
CrowdCLIP (ours)	-	None	**294.9**	**498.7**	**148.2**	**227.3**	**212.8**	**508.5**	**253.2**	**393.2**
Improvement	*-*	*-*	*37.6%*	*-*	*37.1%*	*-*	*15.3%*	*-*	*4.9%*	*-*

it still achieves competitive transfer abilities. Our method even outperforms MCNN [62] by 18.3% in MAE and outperforms D-ConvNet [40] by 18.6% in MSE when adapting ShanghaiTech Part A to Part B.

Compared with unsupervised methods, our method significantly outperforms the state-of-the-art [3] on various transfer settings, as shown in Tab. 3. Specifically, Crowd-CLIP significantly improves the CSS-CCNN by 37.6% MAE on Part A crossing to UCF-QNRF and 37.1% MAE on UCF-QNRF crossing Part A. For the rest crossing setting, CrowdCLIP also achieves the best results.

6.3. Ablation Study

The following experiments are conducted on the UCF-QNRF [12] dataset, a large and dense dataset that can effectively avoid overfitting.

The effectiveness of the progressive filtering strategy and fine-tuning. We first study the effectiveness of the proposed progressive filtering strategy and fine-tuning, and the results are listed in Tab. 4. We can make the following observations: 1) As expected, directly using the original CLIP (*i.e.*, zero-shot CLIP without the first two text encoders) reports unsatisfactory performance, only 528.7 MAE, significantly worse than the current SOTA [3] (437.0 MAE). This result also verifies that the original CLIP can not directly work well in the crowd counting task, *i.e.*, it is non-trivial to apply CLIP for our task; 2) Using the proposed inference strategy, we improve the performance of the original CLIP from 528.7 MAE to 414.7 MAE, which is slightly better than CSS-CCNN. 3) When we adopt the fine-tuned image encoder $\mathcal{E}_f$ for the last stage to map the crowd patches into count intervals, we get a significant performance gain. This highlights the effectiveness of ranking-based contrastive fine-tuning. The following ablation studies are organized using the setting of the last line of Tab. 4.

Analysis on ranking prompts design. We then analyze the influence of different ranking text prompt designs, as shown in Tab. 5. We observe that when the ranking prompts are set to ['20', '55', '90', '125', '160', '195'], *i.e.*, $R_0 = 20$ and $K = 35$, the CrowdCLIP achieves the best performance. An interesting phenomenon is that when we set an abstract ranking prompt $R_0 = A + 20$ and $K = 35$, where 'A' is the original Latin alphabet, the CrowdCLIP also achieves superior performance. It indicates that CrowdCLIP can study the potential ranking representation from the ordinal language space. Using different ranking prompts to fine-tune the image encoder, the performance is always better than the zero-shot CLIP [34] and CSS-CCNN [3], which demonstrates the effectiveness of our CrowdCLIP. The following ablation studies are organized using $R_0 = 20$ and $K = 35$.

Analysis on fixing different encoders. We further study the effect of fixing different encoders of CLIP [34], listed in Tab. 6. If we directly use the original CLIP with the proposed inference strategy (*i.e.*, the parameters of the image encoder and text encoder are not tuned), the MAE is only 414.7. When we only fine-tune the image encoder, we observe a significant improvement, where MAE is from 414.7 to 283.3. However, if we try to fine-tune the text encoder, we observe a significant performance drop. We argue the main reasons are as follows: 1) The natural language knowledge learned from large-scale pre-training encodes rich language priors, and fine-tuning will break the language priors. 2) As the text encoder is fixed, mapping image features into ranking language space will have a strong reference.

The influence of patch number. We next study the influence of using different patch number P during the testing phase. As shown in Tab. 7, the MAE and MSE achieve the best when the P is set to 4 and 5, respectively. Note that whether it is set to 3, 4 or 5, the performance of CrowdCLIP always significantly outperforms the zero-shot CLIP [34].

The influence of data size. Finally, we explore the influence of using different data sizes. Fig. 4 shows that per-

Table 4. The effectiveness of our progressive filtering strategy and fine-tuning. $\mathcal{E}_o$ and $\mathcal{E}_f$ refer to the original image encoder and fine-tuned image encoder, respectively. $\mathcal{T}_o$, $\mathcal{T}_1$, and $\mathcal{T}_2$ denote the text encoder of the first, second and third text encoder. Random* denotes that we randomly select a value from the pre-defined set of rank numbers R for each cropped patch.

Methods	First stage	Second stage	Third stage	MAE	MSE
Random*	-	-	-	633.6	978.9
CSS-CCNN [3]	-	-	-	437.0	722.3
Zero-Shot CLIP [34]	-	-	$\mathcal{E}_o, \mathcal{T}_2$	528.7	690.7
CrowdCLIP	-	-	$\mathcal{E}_f, \mathcal{T}_2$	318.8	491.9
Zero-Shot CLIP [34]	$\mathcal{E}_o, \mathcal{T}_0$	-	$\mathcal{E}_o, \mathcal{T}_2$	437.7	623.1
CrowdCLIP	$\mathcal{E}_o, \mathcal{T}_0$	-	$\mathcal{E}_f, \mathcal{T}_2$	286.8	490.4
Zero-Shot CLIP [34]	$\mathcal{E}_o, \mathcal{T}_0$	$\mathcal{E}_o, \mathcal{T}_1$	$\mathcal{E}_o, \mathcal{T}_2$	414.7	612.4
CrowdCLIP (**ours**)	$\mathcal{E}_o, \mathcal{T}_0$	$\mathcal{E}_o, \mathcal{T}_1$	$\mathcal{E}_f, \mathcal{T}_2$	**283.3**	**488.7**

Table 5. Ablation study on ranking prompts design (including the basic reference count R_0 and counting interval K). The ranking prompts are defined as "There are [class] persons in the crowd."

R_0	K	Prompts	MAE	MSE
20	30	[20, 50, ..., 140, 170]	324.1	569.8
20	35	[20, 55, ..., 160, 195]	**283.3**	**488.7**
20	40	[20, 60, ..., 180, 220]	358.4	602.9
A + 20	35	[A + 20, ..., A + 195]	316.2	515.3
10	35	[10, 45, ..., 150, 185]	374.5	602.5
30	35	[30, 65, ..., 170, 205]	373.4	633.7

Table 6. The influence of fixing different encoders.

Fixed text encoder	Fixed image encoder	MAE	MSE
✓	✓	414.7	612.4
✓	-	**283.3**	**488.7**
-	✓	523.4	870.4
-	-	416.6	723.5

Table 7. Ablation study on the patch number design.

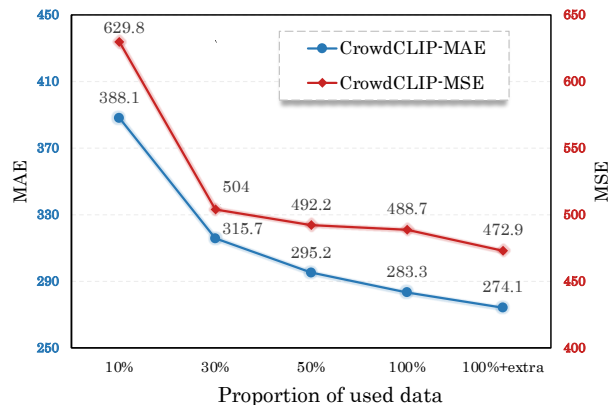| Setting	Patch Number	MAE	MSE
I	3	325.0	534.6
II	4	**283.3**	488.7
III	5	305.2	**477.8**

Figure 4. Comparison in MAE and MSE of using different data size for training on UCF-QNRF dataset. We use ShanghaiTech Part A as the extra data.

level information is also useful to help analyze the crowd. In the future, we would like to explore localization in an unsupervised manner for crowd counting.

7. Conclusion

In this work, we propose a new framework, Crowd-CLIP, to transfer the knowledge from the vision-language pre-trained model (CLIP) to the unsupervised crowd counting task. By fine-tuning the image encoder through multi-modal ranking loss and using a progressive filtering strategy, the performance of CrowdCLIP is largely improved. We conduct extensive experiments on challenging datasets to demonstrate the state-of-the-art performance of our CrowdCLIP. We hope our method can provide a new perspective for the crowd counting task.

formance improves steadily as the data number increases. Additionally, since our method is in an unsupervised setting, we can further utilize extra data to improve the performance. For example, when adopting ShanghaiTech Part A as the extra data, the proposed CrowdCLIP achieves considerable performance gain. These impressive results emphasize the practical utility of our CrowdCLIP.

6.4. Limitation

The main limitation is that our method only provides the count-level information for a given image. However, point-

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[2] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 2021. 2

[3] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A Sindagi, R Venkatesh Babu, and Vishal M Patel. Completely self-supervised crowd counting via distribution matching. In *European Conference on Computer Vision*, pages 186–204. Springer, 2022. 2, 3, 6, 7, 8

[4] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5744–5752, 2017. 2, 6

[5] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2020. 1

[6] Binghui Chen, Zhaoyi Yan, Ke Li, Pengyu Li, Biao Wang, Wangmeng Zuo, and Lei Zhang. Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 16065–16075, 2021. 2

[7] Yajie Chen, Dingkang Liang, Xiang Bai, Yongchao Xu, and Xin Yang. Cell localization and counting using direction field map. *IEEE Journal of Biomedical and Health Informatics*, 26(1):359–368, 2021. 2

[8] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 3

[9] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 976–980. IEEE, 2022. 5

[10] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions ON Graphics*, 2022. 2, 3

[11] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2013. 5

[12] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proc. of European Conference on Computer Vision*, pages 532–546, 2018. 5, 7

[13] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4706–4715, 2020. 2

[14] Di Kang, Zheng Ma, and Antoni 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*, 29(5):1408–1422, 2018. 1

[15] Xiyu Kong, Muming Zhao, Hao Zhou, and Chongyang Zhang. Weakly supervised crowd-wise attention for robust crowd counting. In *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 2722–2726. IEEE, 2020. 3

[16] Yinjie Lei, Yan Liu, Pingping Zhang, and Lingqiao Liu. Towards using count-level weak supervision for crowd counting. *Pattern Recognition*, 109:107616, 2021. 1, 3

[17] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In *Proc. of Advances in Neural Information Processing Systems*, 2022. 3

[18] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 1, 2

[19] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6):1–14, 2022. 1, 3

[20] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *Proc. of European Conference on Computer Vision*, 2022. 2, 6

[21] Dingkang Liang, Wei Xu, Yingying Zhu, and Yu Zhou. Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia*, 2022. 2

[22] Hui Lin, Zhiheng Ma, Xiaopeng Hong, Yaowei Wang, and Zhou Su. Semi-supervised crowd counting via density agency. In *Proc. of ACM Multimedia*, pages 1416–1426, 2022. 1

[23] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 19628–19637, 2022. 2

[24] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *Proc. of Intl. Conf. on Learning Representations*, 2020. 5

[25] Xinyan Liu, Guorong Li, Zhenjun Han, Weigang Zhang, Yifan Yang, Qingming Huang, and Nicu Sebe. Exploiting sample correlation for crowd counting with multi-expert network. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 3215–3224, 2021. 2

[26] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 5

[27] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1862–1878, 2019. 5

[28] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *Proc. of European Conference on Computer Vision*, pages 242–259. Springer, 2020. 1, 3

[29] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 6469–6478, 2019. 2

[30] Zhihao Liu, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, et al. Visdrone-cc2021: the vision meets drone crowd counting challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2830–2838, 2021. 2

[31] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 15549–15559, 2021. 3

[32] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. Shallow feature based dense attention network for crowd counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 2020. 2

[33] Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E Green, and Nassir Navab. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. *arXiv preprint arXiv:2107.12518*, 2021. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Intl. Conf. on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4, 7, 8

[35] Usman Sajid, Hasan Sajid, Hongcheng Wang, and Guanghui Wang. Zoomcount: A zooming mechanism for crowd counting in static images. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3499–3512, 2020. 2

[36] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and Venkatesh Babu Radhakrishnan. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2739–2751, 2020. 2, 6

[37] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. Almost unsupervised learning for dense crowd counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, pages 8868–8875, 2019. 3

[38] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022. 2, 3

[39] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Counting with focus for free. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 4200–4209, 2019. 2

[40] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2018. 2, 6, 7

[41] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Proc. of IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, 2017. 1, 2

[42] Vishwanath A Sindagi and Vishal M Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, 2018. 1

[43] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 5

[44] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 3365–3374, 2021. 2

[45] Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose or to fuse? scale selection for crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2576–2583, 2021. 2

[46] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021. 2

[47] Changan Wang, Qingyu Song, Boshen Zhang, Yabiao Wang, Ying Tai, Xuyi Hu, Chengjie Wang, Jilin Li, Jiayi Ma, and Yang Wu. Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 3234–3242, 2021. 2

[48] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpucrowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2141–2149, 2020. 1

[49] Yi Wang, Junhui Hou, Xinyu Hou, and Lap-Pui Chau. A self-training approach for point-supervised object detection and counting in crowds. *IEEE Transactions on Image Processing*, 30:2876–2887, 2021. 2

[50] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2021. 2

[51] Jiahao Xie, Wei Xu, Dingkang Liang, Zhanyu Ma, Kongming Liang, Weidong Liu, Rui Wang, and Ling Jin. Superresolution information enhancement for crowd counting. *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2023. 2

[52] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Porc. of IEEE Intl. Conf. on Computer Vision*, 2019. 2

[53] Haipeng Xiong and Angela Yao. Discrete-constrained regression for local counting models. In *Proc. of European*

Conference on Computer Vision, pages 621–636. Springer, 2022. 2

[54] Chenfeng Xu, Dingkang Liang, Yongchao Xu, Song Bai, Wei Zhan, Xiang Bai, and Masayoshi Tomizuka. Autoscale: Learning to scale for crowd counting. *International Journal of Computer Vision*, pages 1–30, 2022. 2

[55] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 8382–8390, 2019. 2

[56] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2, 3

[57] Wei Xu, Dingkang Liang, Yixiao Zheng, Jiahao Xie, and Zhanyu Ma. Dilated-scale-aware category-attention convnet for multi-class object counting. *IEEE Signal Processing Letters*, 28:1570–1574, 2021. 2

[58] Yanyu Xu, Ziming Zhong, Dongze Lian, Jing Li, Zhengxin Li, Xinxing Xu, and Shenghua Gao. Crowd counting with partial annotations in an image. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 15570–15579, 2021. 3

[59] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *Proc. of European Conference on Computer Vision*, pages 1–17, 2020. 1, 3

[60] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 833–841, 2015. 3, 6

[61] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 5

[62] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 589–597, 2016. 1, 2, 3, 5, 6, 7

[63] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3

[64] Zhikang Zou, Yu Cheng, Xiaoye Qu, Shouling Ji, Xiaoxiao Guo, and Pan Zhou. Attend to count: Crowd counting with adaptive capacity multi-scale cnns. *Neurocomputing*, 367:75–83, 2019. 2

[65] Zhikang Zou, Xiaoye Qu, Pan Zhou, Shuangjie Xu, Xiaoqing Ye, Wenhao Wu, and Jin Ye. Coarse to fine: Domain adaptive crowd counting via adversarial scoring network. In *Proc. of ACM Multimedia*, pages 2185–2194, 2021. 2