Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

Annotation-Efficient Crowd Counting: A Survey with a Focus on Count-Level Weak Supervision

Yushan Xu *, Minmin Li, Yanan Zhang, Qiankun Ji, Qiaoqiao Feng, Fucheng Zhang

Anhui Xinhua University, Hefei, Anhui, 230000, China

*** Corresponding author:** Yushan Xu

Abstract: The current crowd counting models are trained from point level head annotation but only the total crowd count is assessed. It is expensive to close this gap: It takes about 2, 000 human-hours to annotate UCF-QNRF. This survey covers the topic of annotation-free/annotation-efficient crowd counting methods particularly with count-level weak supervision, which train from only image-level counts. We categorize the literature by the level of annotation efficiency, and argue that there are a 4-tiered taxonomy and two annotated efficiency pathways—one to enhance the use of annotation, and the other to diminish the need for it. The first pathway includes annotation-efficient losses under point supervision and semi-supervised methods which utilize unlabeled images using uncertainty-aware teacher student training, ranking and surrogate tasks. The second focuses on weakly supervised methods at the count level – from early running ranking networks, MAPS, and MATT, to most recent transformer, MLP and hybrid architectures. On four benchmarks, it is compared with a semi-supervised method and the results demonstrate that count-level weak supervision is competitive to semi-supervision without any semantic location annotations, but with a performance gap to full supervision on dense datasets. With that said, the possibilities of merging distinct approaches to annotation, annotation noise, and foundation models are open challenges.

Keywords: Crowd Counting; Weakly Supervised Learning; Count-Level Supervision; Annotation Efficiency; Semi-Supervised Learning; Density Map Estimation.

1. Introduction

Crowd counting aims to estimate the number of people in an image or video, and it underpins a wide range of applications in public safety, transportation management, and large-scale event surveillance [1], [2]. In dense scenarios such as stations, stadiums, and religious gatherings, a single image may contain thousands of individuals. Training a modern counting model on such images requires annotating every person with a point at the head center, and this annotation process is both labor-intensive and time-consuming. For example, annotating the UCF-QNRF dataset, which contains 1,535 images and approximately 1.25 million annotated heads [3], requires roughly 2,000 human-hours [4]; the NWPU-Crowd dataset contains 5,109 images with over 2.13 million annotations and a maximum of 20,033 heads in a single image [5]. The cost of acquiring such location-level annotations grows with the number of heads to be marked rather than with the number of images (see Section 6.1), which makes fully supervised counting impractical in many real-world deployments.

One of the recent observations that inspired this survey is the total number of people counted, which is the only issue in crowd counting evaluation. The standard metrics, mean absolute error (MAE) and mean squared error (MSE), do not account for spatial accuracy of individual predictions [8]. In fact, mainstream approaches still need point-level location annotation to produce the density maps to serve as intermediate supervision targets during training [9], [10]. This creates a mismatch between what the task is evaluated on and what the training procedure demands [8], [11]. At the core of the issues discussed in this survey is not to get better at counting but to make an impact on reducing this mismatch.

Most of the existing surveys have taken a traditional approach to arranging crowd counting methods based on the supervision type. The last and most relevant one from Wang et al. [1] deals with fully supervised, unsupervised and semi-supervised approaches with special focus on the latter two types of approaches. It has a limited discussion regarding count-level weak supervision, with the MATT model being mentioned only once, and classified as semi-supervised learning, and the others not discussed at all [1]. This survey has a different organizing-aspect. Instead of organizing methods according to the type of supervision, we use annotation efficiency as the main axis to group available methods that eased the annotation effort or took full advantage of the annotation effort. The case of count level weak supervision, namely learning from the total number of images without location description, is emphasized here since it is the most annotation efficient one existing to date.

This survey has a very limited set of questions. We aim to discuss only as background some traditional approaches for crowd counting, like MCNN and CSRNet, and do not attempt a general review of network architectures [10],[12]. Multimodal fusion and video-based counting is also omitted since they are different and require special consideration. The rest of this paper proceeds in the following way. Section 2 quantifies the counting problem and proposes that there are four levels of supervision, along with two pathways towards efficiency, which organize the remainder of the survey. Section 3 discusses point supervision-based annotation-efficient losses. Section 4 deals with semi-supervised approaches which make use of unlabeled images. The main part of this survey, Section 5, is devoted to weakly supervised approaches at the count level. Section 6 provides summary of datasets, annotation statistics and evaluation protocols. Open challenges and future directions are presented in Section 7 and the conclusion is presented in Section 8.

2. Problem Formulation and Supervision Taxonomy

2.1. Problem Formulation

We consider the problem of estimating the number of people in a single image. Let I denote an input image, and let N denote the ground-truth number of people in I. In the density-map paradigm, the annotation is a set of point coordinates $P=\{p_i=(x_i, y_i)\}$, where each p_i marks the head center of one person. A ground-truth density map D is obtained by convolving the point set with a Gaussian kernel [9], and the total count is recovered by integrating the density map over the image plane:

$$N=\sum_{(x,y)} D(x, y)$$

A counting model F_θ parameterized by θ takes I as input and produces either a predicted density map $\widehat{D}$ or a scalar count $\widehat{N}$, depending on the training paradigm.

The supervision signal varies across paradigms. Under point-level full supervision, every person in every training image is annotated with a point, and the model is trained to minimize the discrepancy between $\widehat{D}$ and D [10]. Under semi-supervision, only a small subset of training images carries point annotations, while the remaining images are unlabeled [7]. Under count-level weak supervision, each training image is annotated only with its total count N, and no location information is available [6]. These three levels of supervision differ not only in the volume of annotation but also in the type of information they convey, and this distinction is the basis of the taxonomy introduced below.

2.2. Supervision Taxonomy

We organize crowd counting methods into four levels of supervision, as summarized in Table 1. Level 1 is point-level full supervision, in which every training image is annotated with per-head point coordinates and the model is trained to regress a pixel-wise density map. This is the dominant paradigm in the literature, and representative methods include MCNN [10] and CSRNet [12]. Level 2 keeps the same point-level annotation but redesigns the loss function to tolerate annotation noise and reduce reliance on heuristic density-map construction. Representative methods include Bayesian Loss [13], DM-Count [14], and P2PNet [8]. Level 3 is semi-supervision, in which a small number of images carry point annotations and a large number of images are unlabeled; the unlabeled images are exploited through self-training, consistency regularization, or surrogate tasks. Representative methods include IRAST [7] and SUA [15]. Level 4 is count-level weak supervision, in which each training image is annotated only with its total count and no location information is used during training. Representative methods include MATT [6], TransCrowd [16], CrowdMLP [17], and JCTNet [18].

The four levels form a spectrum of decreasing annotation cost and, correspondingly, decreasing supervision strength. Level 1 requires the most annotation effort but provides the strongest spatial signal. Level 4 requires the least annotation effort, since the total count of an image can often be obtained through non-visual means such as turnstiles, sensors, or manual estimation, but it provides no spatial signal at all. Levels 2 and 3 lie in between. We note that MATT is not a purely count-level method: it combines a small number of point-annotated images with a large number of count-annotated images. We list it under Level 4 because its training is predominantly driven by count-level supervision, but this hybrid nature should be kept in mind when interpreting its results.

Table 1. Four-level supervision taxonomy for crowd counting.

Supervision level	Training signal	Annotati-on form	Representa-tive methods	Annotat-ion cost
Level 1: Point-level full supervision	Per-head point coordinates	All head locations annotated in every image	MCNN, CSRNet	Highest
Level 2: Point supervision + efficient loss	Per-head point coordinates	Same as Level 1, but loss tolerates noise	Bayesian Loss, DM-Count, P2PNet	High
Level 3: Semi-supervision	Few point-annotated + many unlabeled	Partially annotated images plus unlabeled ones	IRAST, SUA	Medium
Level 4: Count-level weak supervision	Total count per image	Count only, no location information	MATT, TransCrowd, CrowdMLP, JCTNet	Lowest

Note: MATT uses a hybrid setting with a small number of point-annotated images and a large number of count-annotated images. It is listed under Level 4 because its training is predominantly driven by count-level supervision, but it is not a purely count-level method.

2.3. Two Efficiency Pathways

The taxonomy outlined above points to two main directions for improving annotation efficiency. First, we can keep the annotation volume as it is, but try to get more out of it—this is what Level 2 methods do. While they still depend on point-level annotations, they redesign the loss function so that the same labels result in a more robust model. Bayesian Loss [13], for example, does not rely on Gaussian-smoothed density maps; instead, it uses a probabilistic formulation to supervise point annotations directly, thus avoiding errors introduced by the heuristic kernel bandwidth. DM-Count [14] takes this a step further by using optimal transport to align distributions, completely skipping the density-map construction. P2PNet [8], on the other hand, drops the density map entirely and models counting and localization as a single point-prediction

problem.

The second direction is to reduce the amount of annotation needed, which aligns with Level 3 and 4 methods. Semi-supervised learning makes use of abundant unlabeled images through self-training, consistency regularization, or surrogate tasks, meaning only a small portion of the dataset needs labels [7], [15], [19]. Count-level weakly supervised methods push this to the limit by requiring only the total count per image, eliminating spatial annotations completely [6], [16], [17], [18]. In short, the first route enhances annotation utilization, while the second directly reduces the annotation burden. The two are complementary, and a practical system may combine both.

This survey focuses on the second pathway, and in particular on count-level weak supervision at Level 4. The reason is twofold. First, the first pathway has been extensively studied and is already well covered by existing surveys, including the one by Wang et al. [1]. Second, count-level weak supervision represents the most annotation-efficient regime currently available and has seen rapid progress in recent years, yet it remains under-discussed in existing reviews: as noted in Section 1, most of the methods reviewed in Section 5 are not covered by [1]. Our focus on Level 4 is therefore intended to complement, rather than duplicate, existing survey efforts.

3. Annotation-Efficient Losses under Point Supervision

3.1. Limitations of Gaussian-Kernel Density Maps

The novel approach of Bayesian Loss [13] to the training problem recasts the problem as a distribution over the image plane, with its mean at each of the annotated points, and they supervise the network with the likelihood that any given pixel belongs to any given head. This eliminates the Gaussian smoothed density map as an explicit target, thus with it the kernel bandwidth as a free parameter. The annotation is used as-is – each point defines the distribution and the network is used to learn to match the distribution. The consequence is that, for the same point annotation, a more informative training signal is provided especially in areas where the heuristic kernel would have given a confusing supervision signal.

DM-Count [14] approaches the problem in a different way and poses it as distribution matching. The predicted density map is not compared to a ground truth density map that is obtained by Gaussian smoothing the distribution of annotated points, but is instead matched to the empirical distribution of such points with optimal transport, while taking into account a total-count consistency constraint as well. This dispenses with constructing the density map altogether and offers a generalization-bound with a relation between the counting error and the transport cost. The advantage is similar to that of Bayesian Loss: The point annotation is utilized directly without the Gaussian smoothing and the training signal is less dependent on the choice of the kernel bandwidth.

Another issue that has been tackled is the noise coming from the act of annotating itself, as done by Wan and Chan [20]. They note that annotators often lose or misannotate points in dense areas and make the same observation of noise in the annotation explicit during training so that the same nominal annotation provides a more accurate trained model.

3.2. Purely Point-Based Framework

A one step further redesign can be seen in the P2PNet [8] where they get rid of the density map completely. It reframes crowd counting and localization as a single point-prediction problem, which is feasible, as the network directly generates a set of point proposals along with the corresponding confidences and Hungarian matching is used to obtain the one-to-one mapping between the proposals and the ground-truth points. Thresholding the confidence values results in a total count, the number of proposals which survived the thresholding. It has two implications for the efficiency of annotation. It is directly used as the training target, e.g., point annotation, without intermediate density map. Second, the localization information in the annotation is no longer ignored when calculating the accuracy at the time of evaluation: The authors propose a new measure, density Normalized Average Precision (nAP) that considers both counting and localization. So the annotation is applied for the purpose it has been collected.

Further research has further developed the point-based approach in a couple of different ways. To boost both the counting and localization accuracy of the purely point-based formulation, PET [21] proposes the reject and re-allocate query decision at grid level. APGCC [22] addresses a known instability in the proposal-target matching process by introducing auxiliary positive and negative points that guide the matching, and by developing an implicit feature interpolation scheme that enables feature extraction at arbitrary positions. Both methods remain within the point-based paradigm and confirm that the reformulation introduced by P2PNet is not a one-off design but a general framework for annotation-efficient counting.

3.3. Summary

These methods improve the utilization of point annotations by removing the heuristic Gaussian smoothing or by explicitly accounting for annotation noise, but they do not reduce the volume of annotation: point-level annotation is still required for every training image. Reducing the volume of annotation itself is the subject of the next two sections.

4. Semi-Supervised Learning and Exploitation of Unlabeled Data

4.1. Ranking-Based Self-Supervision

The first pathway to reducing annotation volume is to exploit unlabeled images. A natural starting point is to extract supervision signals from the structural relationships that hold among images, without requiring any manual annotation. Learning-to-Rank [19] takes this approach by exploiting the fact that a sub-image of a crowd scene contains no more people than the full image from which it is cropped. Given an unlabeled image, the method generates a set of progressively smaller sub-images and trains the network with a pairwise ranking loss that enforces this ordering. The network learns to assign larger counts to larger regions, and this ordering signal is sufficient to train a feature representation that transfers to the counting task.

The annotation efficiency of this approach is clear: no location annotation is required for the unlabeled images, only the ordering relationship that is derived automatically from the cropping procedure. However, the method has an inherent limitation. The ordering relationship provides only a relative

constraint, not an absolute one. Essentially, this method provides the network with relative comparisons between regions (knowing where crowds are denser) but lacks the information of exact absolute counts. Therefore, the resulting feature representation is valuable mainly for pretraining and cannot independently support accurate counting. In practical implementations, Learning-to-Rank relies on a small amount of point-annotated data for absolute count calibration. This dependency classifies it as a semi-supervised technique rather than a fully unsupervised method.

4.2. Surrogate Tasks and Self-Training

A second approach to exploiting unlabeled images is to construct surrogate tasks whose supervision can be generated automatically, and to train the feature extractor on these tasks before or alongside the counting task. IRAST [7] exemplifies this approach. It derives a set of binary segmentation tasks from the density-map regression task: for each of several predefined thresholds, the network is trained to predict whether the density value at a given pixel exceeds that threshold. On labeled images, the segmentation targets are derived directly from the ground-truth density map. On unlabeled images, the network generates pseudo-labels through a self-training procedure, and a key contribution of the method is an inter-relationship rule that filters out pseudo-labels that are mutually inconsistent across thresholds.

The annotation efficiency of IRAST comes from the fact that the surrogate tasks operate at the pixel level and can therefore be trained on unlabeled images, while the density-map regressor is trained only on the labeled subset. Because the feature extractor is shared between the surrogate tasks and the counting task, improvements in the former translate into improvements in the latter. By annotating merely 10% of the training data on ShanghaiTech and 20% on UCF-QNRF, this method still achieves impressive accuracy. It highlights the fact that we can extract meaningful representations from unlabeled images without any manual labeling. The primary bottleneck, however, stems from the pseudo-labels assigned to these unlabeled images. In highly crowded scenes, ambiguous binary segmentation boundaries make such pseudo-labels highly unreliable. As training progresses, this label noise gradually accumulates and risks impairing the final model.

4.3. Spatial Uncertainty-Aware Teacher-Student Framework

The third is a direct solution to the noise problem: Make an estimate of the reliability of the pseudo-labels and use (relatively) reliable pseudo-labels to weight their contribution to training. SUA [15] proposes a framework of teacher-student for generating prediction on unlabeled images as well as a spatial uncertainty map to show the reliable region in the prediction. Uncertainty is taken to be the Shannon entropy of the binary segmentation output from several different forward passes using dropout. This way the student can only be trained to match the teacher's predictions in areas of low uncertainty thus effectively rejecting noisy pseudo-labels.

The efficiency of SUA in the annotation aspect is similar to that of IRAST in that unlabeled images are utilized by a surrogate task and the counting task is trained with a small set of labeled data. The uncertainty map introduces the new idea that one can trust pseudo-labels in a principled manner based on the uncertainty map, and not from a fixed threshold. The method is tested on 4 data sets and 50% of the data is labeled; for some of the data sets results are similar to fully-supervised methods. It is limited because this estimate of the uncertainty may also be unreliable in the boundary areas between crowd and background, where the segmentation output is also ambiguous for the teacher model.

A closely related area of research is based on the following approach: The uncertainty estimate is calibrated via direct supervision. Li et al. [4] introduce an uncertainty estimator which was trained by a supervised surrogate function according to the spatial matching distance between predicted and ground-truth point sets and then applied to a patch-wise structure. The main concept is that the labeled subset can be used to scale the uncertainty, instead of just take it from the model consistency on unlabeled images. This yields more dependable uncertainty estimates, which in turn, lead to better pseudo-labels. The method gets good results with 10% labelled data while SUA gets good results with 50%.

4.4. Summary

These studies confirm that unlabeled data can considerably ease the annotation burden, yet a minimal amount of point-annotated data remains necessary to calibrate the final counts. Moving forward, we shift our focus to methods that remove location annotations entirely and learn solely from image-level counts.

5. Count-Level Weak Supervision: Learning from Totals

5.1. Ranking Networks

Yang et al. [11] introduced the first counting network trained exclusively on image-level counts. The core idea stems from the fact that while spatial locations are missing, the relative ranking of crowd sizes among images is often accessible. Consequently, the network is equipped with a shared backbone and two parallel heads: one for count regression and another for sorting images based on crowd density. This sorting constraint forces the backbone to extract density-aware features, ultimately boosting regression performance.

Because it only requires a single scalar per image, the annotation cost is drastically reduced—these counts can be collected via turnstiles, sensors, or rough manual estimates without pixel-level labeling. Experimental results on WorldExpo'10 and UCSD demonstrate its competitiveness against fully supervised methods, proving that count-level supervision is viable. However, the approach is not without flaws. On one hand, the sorting head relies on manually designed soft labels for adjacent ranks, introducing unnecessary heuristics. On the other hand, the standard CNN backbone lacks the global receptive field needed to handle highly congested scenes.

5.2. MATT

MATT [6] tackles a core bottleneck of count-level weak supervision: a count loss alone is incapable of properly constraining the density map. Training a model solely to ensure that its density map sums up to the ground-truth count leaves the spatial distribution largely unconstrained; many irrelevant maps can easily satisfy this. MATT circumvents this issue by generating multiple density maps via parallel auxiliary branches. The training is guided by two rules—each branch must yield the correct integral count, and all branches must align with one another following a predefined

smoothing operation.The consistency constraint forces the shared feature extractor to encode spatial information that supports multiple density-map realizations, which in turn improves the primary branch.

The annotation efficiency of MATT is more nuanced than that of the sorting network. MATT uses a hybrid setting: a small number of images carry point-level annotations, while a large number of images carry only count-level annotations. The point-annotated images are used to train the primary branch with the standard density-map loss, and the count-annotated images are used to train the auxiliary branches with the count and consistency losses. This hybrid design places MATT between semi-supervision and count-level weak supervision, and we therefore list it under Level 4 (see the note to Table 1). The method demonstrates that a small amount of point annotation can be combined with a large amount of count annotation to achieve results that are close to fully supervised methods, and it establishes the auxiliary-task strategy as a general recipe for count-level weak supervision.

5.3. Transformer and MLP Paradigms

Since both the sorting network and MATT rely on CNN backbones, their limited receptive fields pose a fundamental constraint. Without spatial supervision, a model must depend heavily on global context to deduce density from the total count. Naturally, this motivates shifting towards architectures with a global receptive field.

TransCrowd [16] pioneered this shift by adopting a pure Transformer for count-level weakly supervised counting. It treats the task as a sequence-to-count problem: the image is divided into patches and embedded as tokens, which are processed by a Transformer encoder to regress the final count. The self-attention mechanism inherently provides a global receptive field, capturing the long-range dependencies crucial for dense scenes. Evaluated across five benchmarks, TransCrowd significantly outperforms earlier CNN-based models, highlighting the backbone choice as a critical factor.

Taking a different architectural route, CrowdMLP [17] substitutes self-attention with a multi-granularity MLP. The rationale is that the quadratic complexity and heavy parameter count of self-attention might be excessive for counting. CrowdMLP tokenizes images at various granularities, processes them via a shared MLP, and fuses the embeddings prior to regression. By construction, the MLP offers a global receptive field with fewer inductive biases—an advantage when spatial labels are missing. Additionally, the authors propose a self-supervised proxy task, Split-Counting, which enforces the sum of counts from two randomly split sub-regions to match the whole-image count. This cleverly injects spatial cues during training without needing point annotations.

Ultimately, TransCrowd and CrowdMLP reflect a shared insight: without spatial supervision, backbone design becomes the key lever for performance. A global receptive field is essential for inferring density from scalar counts, while the specific realization (self-attention or MLP) is largely secondary.

5.4. Cross-Dataset Generalization

Generalization behavior is often overlooked in count-level weakly supervised methods. According to CrowdFormer [23], a count-level weakly supervised model trained on one dataset frequently outperforms fully supervised methods in cross-dataset tests. This occurs because count-level supervision avoids forcing the model to fit the specific density-map patterns of the source dataset; instead, it learns a more abstract, transferable representation of crowd density.

This has two major implications. It indicates that count-level weak supervision is not just a cheaper alternative, but potentially a better strategy for deployment in new environments. It also reinforces this survey's central argument: the absence of spatial supervision, though detrimental to in-domain accuracy, can significantly boost generalization. This finding directly motivates the hybrid strategies discussed in Section 7, where a small amount of point annotation is paired with massive count annotation to achieve both optimal accuracy and robust generalization.

5.5. Hybrid Architectures and Recent Advances

The most recent methods in this category combine the strengths of different architectural families. JCTNet [18] addresses a specific limitation of the pure Transformer approach: when the input image is partitioned into non-overlapping patches, individual heads may be split across patch boundaries, which leads to over- or under-estimation. JCTNet addresses this by using a CNN front-end to extract features before Transformer processing, so that the patch partitioning operates on feature maps rather than on the raw image. The hybrid architecture combines the strengths of both worlds. A CNN front-end is utilized to introduce the inductive bias that pure Transformers lack, which proves essential under data-scarce conditions. Following this, the Transformer back-end extracts global context. Ultimately, this combination delivers competitive accuracy on five benchmarks, all while keeping the parameter count significantly lower than that of TransCrowd and its contemporaries.

Along similar lines, CrowdNeXt [24] introduces dual-path feature aggregation together with a robust loss function to further strengthen count-level weakly supervised counting, and SR2 [25] enhances MLP-based count-level counting through scale reasoning and ranking. Together with JCTNet, these methods indicate that the count-level paradigm continues to attract active methodological development.

5.6. Summary and Transition

The methods reviewed in this section represent the second efficiency pathway identified in Section 2.3: they eliminate location annotations entirely and use only the total count per image as supervision. The evolution from the sorting network [11] to MATT [6] to TransCrowd [16] and CrowdMLP [17] to JCTNet [18] reflects a progressive refinement of the count-level weak supervision paradigm: the sorting network establishes feasibility, MATT addresses the under-constrained density map problem through auxiliary tasks, TransCrowd and CrowdMLP show that the backbone architecture is a critical lever, and JCTNet combines the strengths of multiple families. The performance of the most recent methods is within a few MAE points of fully supervised methods on several benchmarks, and the cross-dataset results reported by CrowdFormer [23] suggest that count-level weak supervision may have an advantage in generalization. However, a gap remains between count-level weak supervision and full supervision, particularly in dense scenes, and bridging this gap is the subject of Section 7.

6. Datasets, Annotation Statistics, and Evaluation Protocols

6.1. Mainstream Datasets and Annotation Scales

The datasets in crowd counting exhibit substantial differences in both scene coverage and annotation effort. For this survey, we focus on two dimensions: image quantity and the total count of annotated heads, which jointly dictate the annotation cost. The latter metric, notably, scales non-linearly with density.

ShanghaiTech [10] comprises 1,198 images (330,165 heads), partitioned into a dense Internet-sourced Part A and a sparser street-scene Part B. UCF_CC_50 [26] is highly compact but extremely dense, containing merely 50 images yet averaging 1,280 individuals per frame. JHU-Crowd++ [27], meanwhile, offers broader diversity with 4,372 images spanning diverse weather and lighting conditions.

UCF-QNRF [3] pushes the difficulty further with 1,535 high-resolution images and about 1.25 million annotated heads. It covers a wide array of scenes—streets, festivals, religious gatherings, and sports events—with densities ranging from dozens to over 12,000 people. The annotation process is estimated to have consumed roughly 2,000 human-hours [4], rendering it one of the most labor-intensive datasets available.

NWPU-Crowd [5] is the largest public benchmark in terms of both image count and annotation volume. It contains 5,109 high-resolution images with over 2.13 million annotated heads, and the maximum count in a single image reaches 20,033. The dataset covers a wide range of scenes and is split into training, validation, and test sets with 3,109, 500, and 1,500 images respectively.

The annotation statistics summarized above have a direct implication for count-level weak supervision. The total annotation cost of a dataset is proportional to the number of annotated heads, not to the number of images. For UCF-QNRF, the average number of heads per image is over 800, which means that the cost of annotating a single image is very high. For datasets such as ShanghaiTech Part B, the average is closer to 120 heads per image. This variation matters because count-level weak supervision replaces per-head annotation with a single scalar per image, so its relative benefit over full supervision is larger on datasets with higher crowd density.

6.2. Evaluation Metrics

The standard metrics for crowd counting are mean absolute error (MAE) and mean squared error (MSE). Following the convention in the crowd counting literature, MSE is defined here as the square root of the mean squared error, i.e., the root mean squared error (RMSE); this differs from the standard statistical definition of MSE. Both metrics are computed at the image level by comparing the predicted count to the ground-truth count for each test image, with lower values indicating superior performance.

These two metrics capture different error characteristics. MAE treats all deviations equally, whereas MSE is more sensitive to outliers due to the squaring operation. Consequently, a model with a low MAE but high MSE likely suffers from a few catastrophic errors, usually on images with extreme crowd densities. Reporting both is thus a standard practice, which we adopt in this survey.

To further evaluate spatial localization, P2PNet [8] introduced density Normalized Average Precision (nAP). This metric matches predicted points to ground-truth points using a density-aware distance threshold and computes the area under the precision-recall curve. While MAE and MSE only assess global counts, nAP verifies whether the predicted points are spatially accurate. Notably, nAP is largely inapplicable to count-level weak supervision since most methods in this category do not generate point predictions. The lack of a dedicated localization metric for count-level approaches remains an open challenge, which we address in Section 7.

6.3. Weakly-Supervised vs. Fully-Supervised: A Performance Comparison

Table 2 summarizes the counting accuracy of representative methods across four benchmarks: ShanghaiTech A, ShanghaiTech B, UCF-QNRF, and JHU-Crowd++. The table is divided into three groups by supervision type: fully supervised (FS), semi-supervised (SS), and count-level weakly supervised (CL).

The FS group establishes the upper bound of performance achievable with full point annotations. Among the four methods listed, APGCC [22] achieves the best MAE on ShanghaiTech A (48.8), ShanghaiTech B (5.6), and JHU-Crowd++ (54.3), while PET [21] achieves the best MAE on UCF-QNRF (79.53, versus 80.1 for APGCC). These methods represent the state of the art in fully supervised counting and are used as the reference point for the other two groups.

The SS (Semi-Supervised) group illustrates the impact of reducing the number of annotated images while maintaining the annotation type. However, because the three methods rely on different fractions of labeled data—IRAST [7] uses 10% (ShanghaiTech) and 20% (UCF-QNRF), SUA [15] uses 50%, and Li et al. [4] use 10%—their results are not strictly comparable (as indicated in the table). Generally, while semi-supervised methods rival fully supervised ones on the sparser ShanghaiTech B, their performance gap widens on the denser benchmarks like UCF-QNRF and JHU-Crowd++.

In contrast, the CL (Count-Level) group—the core focus of this survey—demonstrates the effect of removing location annotations entirely. Among the six methods evaluated, CrowdMLP [17] achieves the lowest MAE on ShanghaiTech A (57.83) and JHU-Crowd++ (67.57), whereas JCTNet [18] leads on ShanghaiTech B (7.2) and UCF-QNRF (90.0). The progression from the earliest method in this group, Yang et al. [11], to the most recent, JCTNet [18], reflects a substantial improvement: the MAE on ShanghaiTech A drops from 104.6 to 62.8, and the MAE on UCF-QNRF drops from approximately 97 to 90. The most recent methods are within a small margin of the semi-supervised methods on ShanghaiTech, and CrowdMLP even outperforms every semi-supervised method on JHU-Crowd++ (67.57, versus 74.87 for Li et al.).

Two observations follow from Table 2. First, count-level weak supervision has progressed to the point where it is competitive with semi-supervision on some benchmarks, while requiring no location annotation at all. Second, a gap remains between count-level weak supervision and full supervision, and this gap is largest on the densest benchmark, UCF-QNRF, where the best count-level method (JCTNet, MAE 90.0) is still more than 10 MAE points behind the best fully supervised method (PET, MAE 79.53). This gap is the subject of the next section.

Table 2. Performance comparison: weakly-supervised vs. fully-supervised methods.

Method	Supervisi-on	SHTec-h A (MAE)	SHTec-h B (MAE)	UCF-QNRF (MAE)	JHU++ (MAE)
Fully Supervised (FS)					
CSRNet	FS	68.2	10.6	—	—
P2PNet	FS (point)	52.74	6.25	85.32	—
PET	FS (point)	49.34	6.19	79.53	58.5
APGCC	FS (point)	48.8	5.6	80.1	54.3
Semi-Supervised (SS)					
IRAST (10% SHTech / 20% QNRF)	SS	86.9	14.7	135.6	—
SUA (50%)	SS	68.5	14.1	130.3	80.7
Li et al. (10%)	SS	70.76	9.71	104.04	74.87
Count-Level Weakly Supervised (CL) — focus of this survey					
Yang et al.	CL	104.6	12.3	—	—
MATT	CL	80.1	11.7	—	—
TransCrowd	CL	66.1	9.3	97.2	74.9
CrowdMLP	CL	57.828	7.667	94.130	67.568
CrowdFormer	CL	62.1	8.5	93.3	—
JCTNet	CL	62.8	7.2	90.0	—

Notes. FS = fully supervised; SS = semi-supervised; CL = count-level weakly supervised. MAE: lower is better. All values are transcribed from the original papers. For semi-supervised methods, the percentage of labeled data is indicated in the Method column; note that these percentages differ across methods, so results are not directly comparable across the SS group without accounting for this.

7. Challenges and Future Directions

7.1. Inherent Limitation of Count-Level Supervision

The most fundamental limitation of count-level weak supervision is the absence of spatial supervision. A count-level loss constrains only the integral of the predicted density map or the scalar output of a regression head; it does not constrain how the predicted mass is distributed across the image. As a result, two very different crowd distributions, one uniform and one highly clustered, can produce the same total count and therefore the same training loss. Without additional spatial cues, the network has no way to distinguish between them.

This ambiguity is not uniformly distributed across scenes. In sparse scenes, the crowd is spread relatively uniformly and spatial distributions are not necessarily very meaningful other than in terms of the number of persons in the crowd. On the other hand, in dense scenes the crowd is very non-uniform, with some region having hundreds of people congested and others having almost no people in the region. The lack of understanding of these regions is actually the reason for the gap in accuracy between the best count-level method and the best fully-supervised method that is reported in Section 6.3, which is over 10 MAE points on UCF-QNRF. One of the core open issues of count-level weak supervision is tackling this inherent limitation.

There are three directions which look promising. The first is based on the notion of introducing auxiliary self-supervised tasks which induce spatial structure, yet do not need annotation. An early example is CrowdMLP's Split-Counting [17] which splits the image into two disjoint regions with random sampling randomly and requires that the number of counts produced by the two regions equals the number of counts produced by the entire image. The similar ideas might be extended to multi-scale splitting, region-based consistency or contrastive objectives between different crops related to the same scene. The other line is using foundation models with spatial priors obtained through broad pretraining. See Section 7.4 for a further discussion of this. The third one is a mixed-up approach that includes a small quantity of point annotation and quantity of the count annotation, refer the latter in section 7.2.

7.2. Bridging the Accuracy Gap

The accuracy gap between count-level weak supervision and full supervision is unlikely to be closed by count-level supervision alone. A more realistic target is to reduce the gap through hybrid annotation strategies that allocate annotation effort where it is most valuable.

Two hybrid strategies appear particularly promising. The first is image-level mixing, in which a small fraction of training images carries full point annotations and the remaining images carry only counts. This is the setting adopted by MATT [6], and its results show that a small number of point-annotated images, on the order of a few percent of the training set, is sufficient to substantially improve the accuracy of a count-level model. The point-annotated images calibrate the spatial reasoning of the network, while the count-annotated images provide the broad coverage that improves generalization. The relative benefit of

this strategy is largest on datasets with high crowd density, where the cost of annotating a single image is high.

The second strategy is region-level mixing, in which each training image is annotated with point-level annotations in a randomly selected sub-region and count-level annotations elsewhere. This approach, explored in the partial annotation setting [28], reduces the per-image annotation cost while still providing spatial supervision within the annotated region. The combination of region-level mixing with count-level supervision has not been systematically studied and represents a promising direction for future work.

7.3. Annotation Noise and Annotator Variability

One issue that has been widely faced but not much talked about in crowd counting is Annotation noise. Immediately obvious is that annotators often miss heads or put markings slightly to the wrong location, and that such errors are correlated in space – they tend to occur in the densest areas – the areas where they are most difficult to see. Not much is known about the size and scope of this noise, nor is there a standardized way of quantifying annotation quality. However, there is a question of annotator variability, that is, counting the people in images can be subject to error due to different annotators giving different results for the same image, especially when the image is densely populated as the decision on whether a partial head should be counted is subjective. Inter-annotator agreement generally has not been reported for most of the datasets and little is understood about how this affects the training models.

This problem is especially pertinent to the case of weak supervision at the "count-level", as the "count-level" represents the "weakly-supervised signal" that is obtained from annotation. When the noise in the annotation is noise, it occurs on the count, which in turn leads to noise on the training objective. Another approach; such as Wan and Chan [20] which models the annotation noise explicitly, has only been explored in the context of point-level supervision and is not mentioned in the literature of count-level supervision. In the count-level view the noise is different—perhaps a few persons are under- or over-counted due to missed head counts or simply because of noisy data in scenarios where there are many persons present. It might be a helpful direction to take to have a strong loss function that is not very sensitive to the small number of errors.

7.4. Foundation Models and Self-Supervised Pre-Training

Recent breakthroughs in computer vision largely stem from foundation models trained on massive datasets. Models like CLIP [29] and the Segment Anything Model (SAM) [30] offer highly transferable visual representations, allowing them to be adapted to downstream tasks with minimal task-specific labels. Although applying them to crowd counting remains nascent, they present an intriguing avenue for count-level weak supervision.

The relevance of foundation models to count-level weak supervision is twofold. First, they provide spatial priors learned from large-scale data: a model pretrained on diverse visual scenes has internalized some notion of what constitutes a crowd and how people are distributed within a scene. These priors can compensate for the spatial supervision that is absent in count-level training. Second, they enable zero-shot or few-shot adaptation: a foundation model can be applied to a new scene with minimal or no fine-tuning, which reduces the annotation requirement even further.

A notable early work is CrowdCLIP [31], which applies a pretrained vision-language model to perform unsupervised crowd counting via image-text alignment. While it operates in an unsupervised rather than a count-level weakly supervised paradigm, its use of foundation models to anchor spatial information is highly instructive. A natural progression would be to adapt this to count-level supervision—for example, using foundation models to draft pseudo-labels that are subsequently refined by count-level loss. The primary hurdle, however, is computational: the inference cost of large foundation models clashes with the real-time requirements typical of crowd counting. Distillation and efficient fine-tuning are therefore being actively explored to bridge this gap.

A parallel direction lies in self-supervised pre-training on unlabeled crowd images. Instead of relying on annotations, one could pretrain a counting-specific representation using the abundant unlabeled data available in the wild. Methods like Learning-to-Rank [19] and the surrogate tasks of IRAST [7] serve as early prototypes of this strategy. Upgrading them to a foundation-model scale—with modern architectures and massive pre-training—remains a highly promising yet insufficiently explored frontier.

8. Conclusion

This survey has given an overview of the aspect of annotation-efficient crowd counting focusing on count-level weak supervision. We started with a discrepancy: evaluation of crowd counting is based on just the count of the crowd while standard training procedure involves training with point-level annotation. This mismatch instigated a four-level supervision taxonomy (point-level, count-level, weak supervision and full supervision) as well as a distinction between two efficiency pathways: using the annotation more efficiently and requiring less annotation.We reviewed methods along both pathways, with a detailed treatment of the count-level weakly supervised methods that constitute the second pathway.

The evidence reviewed in this survey supports three conclusions. First, count-level weak supervision has progressed substantially in recent years. The most recent methods achieve results that are competitive with semi-supervised methods on several benchmarks, while requiring no location annotation at all. Second, a gap remains between count-level weak supervision and full supervision, and this gap is largest on the densest benchmarks. The absence of spatial supervision is a fundamental limitation that is unlikely to be resolved by count-level supervision alone. Third, combining count-level weak supervision with a small amount of point annotation, with auxiliary self-supervised tasks, or with foundation models is a promising direction for closing this gap.

Several open problems remain. The inherent limitation of count-level supervision, the design of hybrid annotation strategies, the characterization of annotation noise, and the adaptation of foundation models to the counting task are all active areas of research. We hope that the taxonomy and the efficiency framework presented in this survey provide a useful organizing perspective for future work in this direction.

Acknowledgments

The authors gratefully acknowledge the financial support

from the Scientific Research Project of Anhui Xinhua University (2025zrqyb01), the College Students' Innovation and Entrepreneurship Training Program of Anhui Province (S202512216078), the Quality Engineering Project of Anhui Province (2025wkc070), the Scientific Research Project of Anhui Xinhua University (2026zrqyb07), the Scientific Research Project of Anhui Xinhua University (2025zrqyb02), the Scientific Research Project of Anhui Xinhua University (2025zx11), and the Scientific Research Project of Anhui Province (2024AH050607).

References

[1] Wang, J., Gao, M., Li, Q., Kim, H., & Jeon, G. (2024). A survey on supervised, unsupervised, and semi-supervised approaches in crowd counting. CMC-Computers, Materials & Continua, 81(3), 3561–3582.

[2] Gao, G., Gao, J., Liu, Q., Wang, Q., & Wang, Y. (2020). CNN-based density estimation and crowd counting: A survey. arXiv preprint arXiv:2003.12783.

[3] Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 532–546).

[4] Li, C., Hu, X., Abousamra, S., & Chen, C. (2023). Calibrating uncertainty for semi-supervised crowd counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 16731–16741).

[5] Wang, Q., Gao, J., Lin, W., & Li, X. (2021). NWPU-Crowd: A large-scale benchmark for crowd counting and localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), 2141–2149.

[6] Lei, Y., Liu, Y., Zhang, P., & Liu, L. (2021). Towards using count-level weak supervision for crowd counting. Pattern Recognition, 109, 107616.

[7] Liu, Y., Liu, L., Wang, P., Zhang, P., & Lei, Y. (2020). Semi-supervised crowd counting via self-training on surrogate tasks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 242–259).

[8] Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., & Wu, Y. (2021). Rethinking counting and localization in crowds: A purely point-based framework. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3365–3374).

[9] Lempitsky, V., & Zisserman, A. (2010). Learning to count objects in images. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 1324–1332).

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

[11] Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., & Sebe, N. (2020). Weakly-supervised crowd counting learns from sorting rather than locations. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 1–17).

[12] Li, Y., Zhang, X., & Chen, D. (2018). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1091–1100).

[13] Ma, Z., Wei, X., Hong, X., & Gong, Y. (2019). Bayesian loss for crowd count estimation with point supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 6142–6151).

[14] Wang, B., Liu, H., Samaras, D., & Hoai, M. (2020). Distribution matching for crowd counting. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33, pp. 1595–1607).

[15] Meng, Y., Zhang, H., Zhao, Y., Yang, X., Qian, X., Huang, X., & Zheng, Y. (2021). Spatial uncertainty-aware semi-supervised crowd counting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 15529–15539).

[16] Liang, D., Chen, X., Xu, W., Zhou, Y., & Bai, X. (2022). TransCrowd: Weakly-supervised crowd counting with transformers. Science China Information Sciences, 65(6), 160104.

[17] Wang, M., Zhou, J., Cai, H., & Gong, M. (2023). CrowdMLP: Weakly-supervised crowd counting via multi-granularity MLP. Pattern Recognition, 144, 109830.

[18] Wang, F., Liu, K., Wei, N., Sang, N., Xia, X., & Sang, J. (2026). Weakly supervised crowd counting with joint CNN and transformer network. Pattern Recognition, 171, 112077.

[19] Liu, X., van de Weijer, J., & Bagdanov, A. D. (2018). Leveraging unlabeled data for crowd counting by learning to rank. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7661–7669).

[20] Wan, J., & Chan, A. B. (2020). Modeling noisy annotations for crowd counting. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33, pp. 3386–3396).

[21] Liu, C., Lu, H., Cao, Z., & Liu, T. (2023). Point-query quadtree for crowd counting, localization, and more. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1676–1685).

[22] Chen, I.-H., Chen, W.-T., Liu, Y.-W., Yang, M.-H., & Kuo, S.-Y. (2024). Improving point-based crowd counting and localization based on auxiliary point guidance. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 428–444).

[23] Savner, S. S., & Kanhangad, V. (2023). CrowdFormer: Weakly-supervised crowd counting with improved generalizability. Journal of Visual Communication and Image Representation, 94, 103853.

[24] Savner, S. S., & Kanhangad, V. (2025). CrowdNeXt: Boosting weakly supervised crowd counting with dual-path feature aggregation and a robust loss function. IEEE Transactions on Instrumentation and Measurement, 74, 1–16.

[25] Gao, M., Deng, M., Zhao, H., Chen, Y., & Chen, Y. (2024). Improving MLP-based weakly supervised crowd-counting network via scale reasoning and ranking. Electronics, 13(3), 471.

[26] Idrees, H., Saleemi, I., Seibert, C., & Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2547–2554).

[27] Sindagi, V. A., Yasarla, R., & Patel, V. M. (2022). JHU-Crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5), 2594–2609.

[28] Xu, Y., Zhong, Z., Lian, D., Li, J., Li, Z., Xu, X., & Gao, S. (2021). Crowd counting with partial annotations in an image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 15570–15579).

[29] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.,

Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML) (pp. 8748–8763).

[30] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 4015–4026).

[31] Liang, D., Xie, J., Zou, Z., Ye, X., Xu, W., & Bai, X. (2023). CrowdCLIP: Unsupervised crowd counting via vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2893–2903).