Adapting Lightweight Image-based Counting Models for Video Crowd Counting

Weibo Shu Antoni B. Chan
Dept. of Computer Science, City University of Hong Kong
weiboshu2-c@my.cityu.edu.hk abchan@cityu.edu.hk

Abstract

Video crowd counting aims to predict the people count in each frame of a video. It requires effectively leveraging spatio-temporal (ST) information in videos while satisfying real-time constraints. However, most existing methods use ST information from neighboring frames through auxiliary extraction and fusion modules—resulting in large computational cost and the need to buffer multiple frames during inference. Such designs limit their practicality in real-world applications with limited computational resources or stringent real-time requirements. To address these issues, we revisit video crowd counting from the perspective of lightweight image-based counting models that enable real-time deployment under limited resources. We analytically define ST information in a model-independent and statistically interpretable manner, and incorporate it into training via a statistical regularizer that effectively enhances model performance without adding modules or inference overhead. Most framework hyperparameters are further formulated as statistical inference problems, allowing automatic estimation from data and consequently efficient adaptation to new scenarios. Our framework unifies video crowd counting and image-based counting models under a compact, principled formulation that is lightweight, portable, and efficient. We also establish theoretical foundations for adapting image-based counting models to video crowd counting and achieve state-of-the-art accuracy and efficiency across six benchmarks, including challenging DRONECROWD and VSCROWD. Our codes will be available at: wbshu/SR.

1. Introduction

Crowd analysis has wide applications in various fields such as surveillance [3, 50, 51], aerial photography [44, 45, 53], and thermal imaging [25]. Crowd counting is a fundamental task in crowd analysis, and its research mainly focuses on image crowd counting (ICC) due to the relative ease of collecting and annotating image data across a wide variety of scenes, as compared to video data. As a result, video crowd counting (VCC) remains relatively underexplored. Compared with ICC datasets, VCC datasets contain more images but come from fewer scenes, with consecutive frames exhibiting high temporal continuity and spatial homogeneity that provide rich spatio-temporal (ST) cues.

Due to the video modality, the crucial point in VCC methodology is to utilize the ST information in consecutive frames for improving prediction performance [8, 27, 28, 48, 49]. However, the ST information used in most existing works is implicitly and indirectly related to the counting task (e.g., low-level optical flow). The ST information is typically extracted and fused to the features for the count (or density map) prediction. As the extraction, fusion, and final prediction are all based on deep neural networks, how the ST information helps to improve the prediction is not clearly interpretable or understood. In addition, the extraction and fusion of auxiliary ST features increases the model complexity, resulting in higher storage requirements and lower computational efficiency which may not meet the real-time requirement of VCC. The strategy of predicting current frame's count by using extra neighbor frames also requires buffering more frames and features during predictions. Such restrictions impose challenges on deploying these models in real-world applications with limited computational resources and high real-time demands, which greatly limit their practicality.

One idea to tackle the above challenges is to exploit the lightweight subset of ICC models, which naturally inherit the comparatively low time and space complexity. In contrast to recent transformer-based or multi-branch models with heavy computation, these lightweight architectures (e.g., [16, 26, 31]) maintain compact designs and fast inference suitable for real-world deployment. Anchored in these practical challenges, this paper investigates how to enhance the VCC ability of such lightweight ICC models. To ground our framework, we establish a rigorous statistical formulation that quantifies the theoretical gap between ICC and VCC models, revealing when and why ICC models are capable of VCC. Methodologically, while prior work [37] introduced the characteristic function (ChF) as a static per-image representation for ICC, we examine ChF along a

different and previously unexplored dimension: its temporal evolution across frames. Our framework therefore starts with a precise analytical definition of ST information, explicitly tied to counting tasks and derived from temporal dynamics of the ChF, which makes its extraction independent of models and its usage principled. Then we use the ST information to devise a statistical regularizer to enhance the training of the lightweight ICC models on VCC tasks. To further improve adaptability and efficiency, we resort to statistical analysis. We reformulate the selections of key hyperparameters as statistical inference problems, enabling them to be estimated directly from the dataset and avoiding manual tuning. Moreover, the entire pipeline—from extracting to leveraging ST information—is fully formalized, allowing the framework to be distilled into a compact, portable training scheme with a single regularizer that generalizes to diverse ICC models. Our framework for VCC is unique from related works in three aspects:

1. The ST information used in previous works is indirectly related to counting tasks, which makes its usage require auxiliary modules for extraction and fusion. In contrast, we analytically and precisely define the ST information that is directly aligned with the counting task. The definition enables its extraction to rely solely on the training dataset, making it independent of specific models.
2. Instead of using ST information as auxiliary features for model fitting, we incorporate it as a statistical regularizer that constrains model complexity. This regularization enforces temporal consistency during training without adding modules or inference overhead, and the model is guided to search for solutions of controlled complexity, which leads to improved VCC performance.
3. In contrast to prior methods that rely on multiple consecutive frames for both training and inference, our framework only needs frame pairs for training and a single frame for inference. Moreover, its main hyperparameters are automatically determined via statistical inference from the dataset, avoiding manual tuning.

These three properties make our framework conceptually novel, statistically grounded, and practically efficient. In summary, the contributions of our paper are three-fold:

- **Establishing the theoretical and empirical foundations of ICC models for VCC.** To our best knowledge, this work is the first to systematically investigate adapting lightweight ICC models for VCC. On the theoretical side, we derive a statistical formulation that quantifies the gap between image-based and video-based counting models. It reveals how closely single-frame inference can approximate the optimal multi-frame estimation, formalizes the theoretical sufficiency of ICC models for VCC, and also provides a principled way to assess trade-offs between accuracy and efficiency under limited computational or data resources. On the practical side, experiments on six benchmark datasets demonstrate superior results of our framework compared with state-of-the-art (SOTA) VCC methods—the accuracy of lightweight ICC models is effectively improved while retaining their advantages of low time complexity, highlighting the value of this direction for real-time applications.
- **Analytical formalization of ST information.** We propose the first analytical definition of ST information that is explicitly tied to VCC. By characterizing the temporal dynamics of the ChF [37], we instantiate this ST information via a tight inequality and prove that it captures the statistical information of average local count changes across neighboring frames. This formalization allows ST information to be extracted directly from the dataset, independent of model structures, and used in a principled and interpretable manner. Moreover, the extraction is performed only once, rather than at every training step as in prior works, which greatly improves training efficiency.
- **A compact and portable framework.** Our framework is distilled into a concise structure: a training scheme with a statistical regularizer encapsulating ST information, as shown in Fig. 1. The framework hyperparameters are reformulated as statistical inference problems, enabling them to be automatically estimated from the training dataset in a single pass. This design makes the framework highly portable and easily adaptable to different lightweight ICC models and datasets, while preserving efficiency and reducing the need for costly validation experiments and manual hyper-parameter tuning.

2. Related Works

Image crowd counting (ICC). Early research in ICC was based on "detect then count" methods [10, 15, 52], which use extracted features to detect people in images and then accumulate the detection number as the final people count, or "image to count" methods [3, 6, 23], which regress the total people count directly from the input image. More recently, ICC research is mainly divided into the density-map-based methods [11, 20, 34, 36] and the point-based methods [4, 17, 24, 39]. The density map methods rely on the heat map representation of the ground truth, i.e., the density map is a 2D heat map obtained by convolving the labeled dot map (each dot corresponds a people head in that position) with a Gaussian kernel[1], which represents the distribution of people in the spatial domain. The counting model is trained to predict density maps using image-density map pairs for supervision. The point-based methods resort to predicting the position and confidence value of people proposals, where the proposals are preset at the center of each local pattern reserved for the people in that local region. Generally, ICC works can also be roughly sorted into loss function

[1]The convolution kernels could also be learned [40, 41].

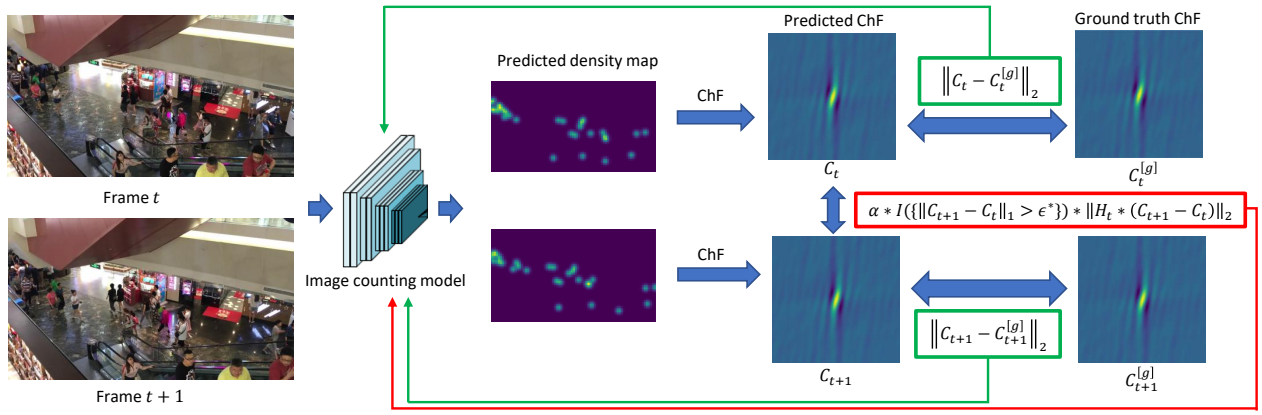

Figure 1. Two adjacent frames are fed into a lightweight image-based counting model to generate separate density maps. The training loss has two parts: (1) supervision from the ground-truth characteristic functions [37] (ChFs; green boxes), and (2) our statistical regularizer applied between the ChFs of two adjacent predictions (red box). α is the balance factor. The regularization only applies during training. In the inference stage, a single frame is used to predict the density map of the current frame, i.e., the same mode as image counting. Distinct from prior works and from the static use of ChF in [37, 38], the regularizer leverages analytically defined spatial–temporal information derived from the ChF's temporal dynamics, expressed through a tight inequality inside the indicator function $I(.)$, with its bound ϵ^* estimated adaptively from the training dataset via statistical inference. The weighting function H_t, also derived analytically and estimated automatically from the training data, refines the constraint by tolerating expected people motion. All components are compactly unified into a single portable regularizer, distilling all the theoretical insights into a simple training scheme. As it requires no extra modules, minimal manual hyperparameter tuning, and only single-frame input at inference, the framework is efficient while retaining accuracy.

design [31, 38, 42, 43] and model design [12, 33, 35, 46].

Video crowd counting (VCC). Currently, most VCC methods focus on perfecting density map generation by leveraging ST information from neighboring frames [1, 7–9, 30, 47–49, 53]. These methods differ in how to extract and exploit the information. [49] extends a language model (a temporal model) to videos. [8, 9, 47] directly merge the features extracted from the neighboring frames into the density map prediction pipeline of the current frame, e.g., [1, 53] uses an optical flow network to extract optical flow features from consecutive frames. [14, 48] uses a transformer structure to capture the difference and relevance between adjacent frames to improve the model's prediction ability. [27, 28] uses the features of adjacent frames to predict people flow between them. These related works focus on using the ST information to provide additional features or guide model architectures. In contrast, we employ ST information in a fundamentally different and more interpretable way. Instead of embedding it as auxiliary features or designing complex architectures around it, we analytically define ST information to be directly aligned with the counting objective, and extract it once from the dataset without relying on additional networks. This precise formalization enables us to use ST information not as extra inputs, but as the foundation of a statistical regularizer that constrains model training. In this way, our framework avoids additional modules, and guides the learning process toward solutions of controlled complexity— thus improving accuracy while preserving efficiency.

3. Methodology

We first elucidate the definition and extraction of our ST information, propose our statistical regularizer and training framework, and finally establish the theoretical foundation.

A VCC dataset comprises several groups of consecutive frames. Each group of frames is extracted from a video clip with a fixed framerate. We use the following notation:

- $I_i^{(k)}$: ith frame from the kth group (video clip).
- $D_i^{(k)}$: density map of the ith frame from the kth group.
- $C_i^{(k)}$: characteristic function of the density map of the ith frame from the kth group.
- M: total number of groups of frames (clips).
- N_k: number of frames in the kth group (kth clip).

3.1. Formalization of ST information

We define ST information as an ST statistical consistency constraint. After progressive theoretical analysis, the formal definition arises as a tight inequality.

ST statistical consistency. Because a video captures the temporal evolution of a scene, the crowd distribution between consecutive frames should not vary arbitrarily—instead, it follows a bounded statistical variation. In the context of counting tasks, such variation should reflect the change in local people counts between adjacent frames.

Formally, let $f(t)$ denote the *local* people count in a region at time t. We require that the local count evolution

satisfies:

$$|f(t+1) - f(t)| \leq \epsilon, \tag{1}$$

where ϵ is a data-driven upper bound extracted from the training set.

Challenges. Defining the function f properly is nontrivial. First, although people's motion is continuous in the real world, the pixel-level representation of density maps exhibits discontinuous variations, even for slow movements, especially under the dot-map representation. Hence, defining consistency directly in the spatial domain based on pixel values is unreliable. Second, the diversity of scenes, crowd densities, and camera perspectives makes it difficult to choose a proper region size and shape for local counting.

Frequency-domain ST consistency. Given these difficulties, We therefore bypass the spatial density map by using the characteristic function (ChF) [37], a frequency-domain information carrier for people's distribution.

Whereas prior works treated the ChF purely as a static frequency-domain descriptor [37, 38], we take a different perspective: we analyze its temporal evolution, revealing that the ChF encodes highly structured and mathematically tractable ST information. In particular, we establish two new theorems that characterize the temporal dynamics of ChF values between consecutive frames, providing a principled foundation for defining ST consistency directly in the frequency domain. Proofs are in the supplemental.

Theorem 1 *Let Q denote the number of people in the scene at time t, and let their spatial positions be the vector $\mathbf{X}_t = (\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(Q)})^T$. Denote by $C_t(\mathbf{w})$ the characteristic function of the corresponding density map, where $\mathbf{w}$ is the 2D frequency coordinate. Then, for any $\mathbf{w}$,*

$$C_{t+\Delta t}(\mathbf{w}) - C_t(\mathbf{w}) \xrightarrow{\Delta t \to 0} A_{t,\mathbf{w}}(\mathbf{X}_{t+\Delta t} - \mathbf{X}_t),$$

where $A_{t,\mathbf{w}}$ is a row vector with $\|A_{t,\mathbf{w}}\|_2 \leq Q\|\mathbf{w}\|_2$.

Theorem 1 unveils a previously unexplored property of the ChF: although pixel-level density maps change discontinuously even under smooth motion, *their ChFs evolve in a temporal-locally linear manner with respect to people's movement*. This finding extends prior work that used ChF only as a static representation for image counting, by revealing its dynamic behavior across time. Leveraging this property, temporal variation can be directly measured through the magnitude of ChF changes between consecutive frames.

Theorem 2 *Let D_{t+1} and D_t be the density maps of two consecutive frames at time $t+1$ and t, with their corresponding characteristic functions C_{t+1} and C_t. If $\|C_{t+1} - C_t\|_1 \leq \epsilon$, then the average local people count change over any region R,*

$$\Delta_R(t, t+1) = \frac{\left|\int_R (D_{t+1}(\mathbf{x}) - D_t(\mathbf{x}))\, d\mathbf{x}\right|}{m(R)},$$

satisfies $\Delta_R(t, t+1) \leq (2\pi)^{-2}\epsilon$, where $m(R)$ denotes the area of R.

Theorem 2 holds for any local region regardless of its shape and size, thereby addressing the second challenge regarding the selection of local counting region. Together, Theorems 1 and 2 establish a unified theoretical foundation that links spatial representation and temporal evolution, enabling ST consistency to be rigorously defined in the frequency domain.

Building upon these results, we formalize the temporal variation bound of the local count in (1) to all regions simultaneously as:

$$\|C_{t+1} - C_t\|_1 \leq \epsilon, \tag{2}$$

where $\|C_{t+1} - C_t\|_1 = \int |C_{t+1}(\mathbf{w}) - C_t(\mathbf{w})| d\mathbf{w}$ accumulates temporal ChF changes over all frequency coordinates. By Theorem 1, it directly reflects people's motion magnitude, while Theorem 2 ensures that bounding such variation also bounds all local count changes. Note that $f(t)$ in (1) is a scalar, while now in (2), C_t is a function defined over the frequency domain, providing a functional-level characterization of ST consistency.

ST information. The complete definition of ST information requires specifying the bound ϵ in (2). Given a training video dataset $\{C_i^{(k)}\}_{i,k}$ (notations defined earlier), we define the empirical temporal variation bound as:

$$\epsilon^* = \max_{k \in [M]} \max_{i \in [N_k - 1]} \|C_{i+1}^{(k)} - C_i^{(k)}\|_1, \tag{3}$$

where $[N] = 1, \cdots, N$. If the training data are regarded as samples from the underlying population distribution, ϵ^* serves as a *data-driven statistic* that estimates the intrinsic upper bound of temporal variation in the ground truth data distribution. In practice, a slightly modified version of (3) is used for robustness to outlier frames (see the supplemental).

We thus define the ST information as an ST consistency constraint that incorporates this statistical prior:

$$\|C_{t+1} - C_t\|_1 \leq \epsilon^* \tag{4}$$

where ϵ^* is defined in (3) and computed purely from ground truth density maps without any network processing. Thus the extracted ST information represents an intrinsic property of the dataset itself, which is independent of specific models. This formulation captures the *expected magnitude of statistically admissible temporal variation* in a theoretically grounded and model-agnostic manner, forming the foundation for our subsequent regularization design.

3.2. Statistical Regularizer and Framework

Basic regularizer and framework. We use (4) to design a regularizer to improve the VCC ability of lightweight ICC

models. In our training framework, each input consists of a pair of adjacent frames, which are passed independently to the ICC model to produce two predicted density maps, D_t and D_{t+1}, whose ChFs are C_t and C_{t+1} respectively. Using the ground-truth ChFs, $C_{t+1}^{[g]}$ and $C_t^{[g]}$, we define the following loss function for training with the $(t, t+1)$ pair,

$$\mathcal{L} = \underbrace{\|C_t - C_t^{[g]}\|_2 + \|C_{t+1} - C_{t+1}^{[g]}\|_2}_{\mathcal{L}_g} + \underbrace{\alpha\, \mathbb{1}(\|C_{t+1} - C_t\|_1 > \epsilon^*)\|C_{t+1} - C_t\|_2}_{\mathcal{L}_c}, \tag{5}$$

where ϵ^* is defined in (3), α is the balance factor, $\|\cdot\|_2 = \sqrt{\int |h(\mathbf{w})|^2 d\mathbf{w}}$ is the L_2-norm of a function, and $\mathbb{1}(x)$ is the indicator function taking value 1 if condition x is true and 0 otherwise. Here we use the L_2-norm for the loss calculation due to its better training stability (see details in [38], Sec. IV.B). The training framework is shown in Fig. 1.

The first term $\mathcal{L}_g$ in (5) is the regular supervision using the ground truth. The second term $\mathcal{L}_c$ serves as our basic regularizer that penalizes deviations violating the statistical ST constraint. When $\|C_{t+1} - C_t\|_1$ exceeds the data-driven bound ϵ^*, which means that *the temporal variation of predicted local counts may exceed the statistical boundary from the data distribution*, the regularizer imposes a penalty; otherwise, it remains inactive as the variation is consistent with the statistical bound. This design encourages predictions that remain consistent with the intrinsic temporal statistics of the data, thereby constraining unnecessary model complexity without introducing any extra modules or inference overhead.

Our regularizer is applied during training only. In the inference stage, the density map is predicted from a single-frame input, similar to a standard image counting framework.

Motion-tolerant regularizer. The change in local people count between two adjacent predictions originates from two potential sources:

1. the inconsistency between predictions;
2. the natural movement of people between frames.

An effective regularizer should penalize the former while tolerating the latter, i.e., it should be motion-tolerant.

To achieve this, we reweight the regularization term using a frequency-dependent weight function:

$$\mathcal{L}_m = \mathbb{1}(\|C_{t+1} - C_t\|_1 > \epsilon^*)\|H_t * (C_{t+1} - C_t)\|_2, \tag{6}$$

where $H_t(\mathbf{w})$ adjusts the penalty at each frequency $\mathbf{w}$ according to how sensitive that frequency is to normal motion, and $*$ is element-wise multiplication. The choice of $H_t(\mathbf{w})$ is theoretically grounded by the following theorem.

Theorem 3 *Assume there are two adjacent frames I_t and I_{t+1}, their corresponding ground-truth density maps D_t and D_{t+1} are obtained by convolving their ground-truth dot maps with a Gaussian kernel $\mathcal{N}(0, \mathbf{\Sigma})$, and their corresponding characteristic functions are C_t and C_{t+1}. Suppose that the total people count is Q, and that each individual's motion between two frames is a 2D vector sample drawn from the asymptotic overall motion distribution, whose covariance matrix is denoted by $\mathbf{\Lambda}$. Then for each frequency coordinates $\mathbf{w}$, the asymptotic distribution of $C_{t+1}(\mathbf{w}) - C_t(\mathbf{w})$ has standard deviation $\sqrt{Q(1 - \exp(-\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}))\exp(-\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})}$.*

The proof is in the supplemental. Theorem 3 characterizes the normal-motion-induced oscillation of ChF values, as quantified by the derived standard deviation. We therefore design $H_t(\mathbf{w})$ in (7) based on the reciprocal of the standard deviation—giving higher weight to frequencies whose ChF values should remain stable even when people move, and lower weight to frequencies where normal motion naturally induces changes. This weighting makes the regularizer focus on "unjustified" inconsistencies rather than expected dynamics.

$$H_t(\mathbf{w}) = \frac{1}{\sqrt{\min\{C_t(\mathbf{0}), C_{t+1}(\mathbf{0})\} * h_{\mathbf{\Lambda}}(\mathbf{w}) + 1}}, \tag{7}$$

$$h_{\mathbf{\Lambda}}(\mathbf{w}) = (1 - \exp(-\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}))\exp(-\gamma^2\mathbf{w}^T\mathbf{w}), \tag{8}$$

where $C_t(\mathbf{0})$ and $C_{t+1}(\mathbf{0})$ are the total people count for D_t and D_{t+1} (by definition in [37]), γ is the bandwidth hyperparameter of the Gaussian kernel for the density map where $\mathbf{\Sigma} = \gamma^2\boldsymbol{I}$, and $*$ is element-wise multiplication. Here in (7), we use the smaller of the two total counts ($\min\{C_t(\mathbf{0}), C_{t+1}(\mathbf{0})\}$) for conservative normalization, and add $+1$ in the denominator for numerical stability.

Estimation of $\mathbf{\Lambda}$. The key parameter $\mathbf{\Lambda}$, governing the motion-induced ChF variation (Theorem 3), encodes dataset-specific motion patterns via its covariance structure. Therefore, we treat the choice of $\mathbf{\Lambda}$ as a data-driven statistical inference problem rather than as hyperparameter tuning. Since explicit motion data (e.g., point correspondences between frames) are annotation-expensive and unavailable in most VCC datasets, we estimate $\mathbf{\Lambda}$ indirectly through frequency-domain statistics rather than from explicit motion annotations.

Specifically, we model the empirical frequency-wise variance of ChF differences as a proxy signal for the underlying motion covariance. For each video k and frame pair $(i, i+1)$, we compute (notations defined earlier)

$$\delta_i(\mathbf{w}) = C_{i+1}^{(k)}(\mathbf{w}) - C_i^{(k)}(\mathbf{w}), \tag{9}$$

$$E_\delta(\mathbf{w}) = \frac{\sum_{k=1}^{M}\sum_{i=1}^{N_k-1}\delta_i(\mathbf{w})}{\sum_{k=1}^{M}(N_k-1)}, \tag{10}$$

$$S(\mathbf{w}) = \frac{1}{\sum_{k=1}^{M}(N_k-1)}\sum_{k=1}^{M}\sum_{i=1}^{N_k-1}\frac{|\delta_i(\mathbf{w}) - E_\delta(\mathbf{w})|^2}{\min\{C_{i+1}^{(k)}(\mathbf{0}), C_i^{(k)}(\mathbf{0})\}+1}. \tag{11}$$

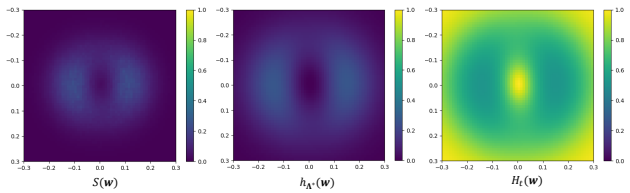

Figure 2. Visualization of the empirical normalized variance function $S(\mathbf{w})$, the best-fit $h_{\mathbf{\Lambda}*}(\mathbf{w})$ defined in (8) with $\mathbf{\Lambda}^*$ defined in (12), and the weight function $H_t(\mathbf{w})$ defined in (7) with $\min\{C_t(\mathbf{0}), C_{t+1}(\mathbf{0})\} = 10$ and $\mathbf{\Lambda}$ set as (12). The dataset is DRONECROWD [45] and γ is set as 8 pixels.

In (11), $S(\mathbf{w})$ is the empirical normalized variance of frequency-wise value changes due to normal people motion, where the denominator removes scale effects caused by varying scene densities and '$+1$' is the stabilizer in case of 0 denominator.

Then, we estimate $\mathbf{\Lambda}$ by fitting the theoretical normalized variance function $h_{\mathbf{\Lambda}}(\mathbf{w})$ in (8) to the empirical normalized variance $S(\mathbf{w})$:

$$\mathbf{\Lambda}^* = \text{argmin}_{\mathbf{\Lambda} \succeq 0} \int |h_{\mathbf{\Lambda}}(\mathbf{w}) - S(\mathbf{w})|^2 \, d\mathbf{w}. \tag{12}$$

This fitting procedure allows $\mathbf{\Lambda}$ to be inferred in a fully data-driven manner—bypassing the need for explicit motion tracking while preserving the statistical semantics of motion covariance in the frequency domain. Fig. 2 visualizes the best-fit $h_{\mathbf{\Lambda}}(\mathbf{w})$ and the empirical normalized variance function $S(\mathbf{w})$, as well as an instantiation of the weight function $H_t(\mathbf{w})$.

Finally, for training, the basic regularizer is replaced with the motion-tolerant regularizer:

$$\mathcal{L}' = \mathcal{L}_g + \alpha \mathcal{L}_m, \tag{13}$$

where $\mathcal{L}_g$ is in (5) and $\mathcal{L}_m$ is in (6) with the estimated $\mathbf{\Lambda}^*$ (12) used in $H_t(\mathbf{w})$.

3.3. Theoretical analysis

This subsection provides the theoretical foundation for adapting ICC models to VCC. In general, ICC models predict crowd counts from a single frame, whereas VCC models leverage multiple consecutive frames. We aim to quantify the statistical discrepancy between these two paradigms, and to characterize the conditions under which single-frame inference is theoretically sufficient.

To formalize this, we compare the optimal mean-squared estimators under different information sets.

Theorem 4 *Let a video be denoted by* $\mathcal{V} = (\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N)$ *where each frame* $\mathcal{I}_t$ *is a random tensor representing the* t*-th frame. Let* $\mathcal{C}_t$ *be the random variable representing the total number of people at frame* t*. We define two classes of estimators:*

$$\mathcal{F}_{\text{img}} = \{f : \mathbb{R}^d \to \mathbb{R} \mid f \text{ measurable}\},$$

$$\mathcal{F}_{\text{vid}}^{(l,r)} = \{f : \mathbb{R}^{(r+l+1)d} \to \mathbb{R} \mid f \text{ measurable}\}$$

corresponding respectively to models that rely only on a single (current) frame $\mathcal{I}_t$ *or on a temporal window* $(\mathcal{I}_{t-l}, \ldots, \mathcal{I}_{t+r})$*, where* d *is the flattened dimension of the random frame tensor,* $l, r \in \mathbb{N}$ *are the number of previous and future frames models used for inference, which depends on the concrete model and strategy.*

We define the ***theoretical discrepancy*** *between the two classes of estimators as*

$$\Delta_{l,r} = \mathbb{E}\Bigg[\Big(\arg\min_{f \in \mathcal{F}_{\text{img}}} \mathbb{E}\big[|f(\mathcal{I}_t) - \mathcal{C}_t|^2\big] - \arg\min_{f \in \mathcal{F}_{\text{vid}}^{(l,r)}} \mathbb{E}\big[|f(\mathcal{I}_{t-l}, \ldots, \mathcal{I}_{t+r}) - \mathcal{C}_t|^2\big]\Big)^2\Bigg].$$

Then, the theoretical discrepancy admits an equivalent formulation in terms of conditional expectations

$$\Delta_{l,r} = \mathbb{E}\Big[\big|\mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_{t-l}, \ldots, \mathcal{I}_{t+r}] - \mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_t]\big|^2\Big].$$

In particular, the theoretical discrepancy vanishes under any of the following conditions:

1. ***Temporal redundancy****:*

$$\mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_{t-l}, \ldots, \mathcal{I}_{t+r}] = \mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_t];$$

2. ***Full observability from a single frame****:*

$$\mathbb{E}[\mathcal{C}_t \mid \mathcal{V}] = \mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_t];$$

3. ***Deterministic target****:*

$$\mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_t] = \mathcal{C}_t.$$

The proof is in the supplemental. Intuitively, $\Delta_{l,r}$ measures the residual uncertainty reduction that a VCC model could theoretically achieve beyond an ICC model, assuming both are trained to their respective optimal predictors with the smallest generalization error. When $\Delta_{l,r}$ is small, temporal context provides negligible new information about the current-frame count. When $\Delta_{l,r} = 0$, both paradigms are theoretically equivalent in attainable accuracy.

The three sufficient conditions correspond respectively to:

1. the current frame encapsulates all relevant visual information available in adjacent frames for $\mathcal{C}_t$;
2. the full video contains no additional statistical information beyond the current frame for estimating $\mathcal{C}_t$;
3. the current frame already determines $\mathcal{C}_t$ exactly.

Importantly, the last condition does not require a perfect scene without heavy occlusions or blur. It only fails when the occlusion or blur is information-theoretically complete — that is, all observable evidence of a person is entirely absent (no visible body parts, silhouettes, or contextual cues).

Datasets	Resolution	View	Frames	Clips	Avg
UCSD [3]	238×158	Surveillance	800/1200	4/6	24.9
MALL [5]	640×480	Surveillance	800/1200	1/1	31.2
FDST [8]	1920×1080	Surveillance	9000/6000	60/40	26.7
VENICE [27]	1280×720	Camera	80/87	1/3	215.0
DRONECROWD [45]	1920×1080	Aerial	24600/9000	82/30	144.8
VSCROWD [14]	1920×1080	Surveillance	49036/13902	497/137	37.2

Table 1. Datasets. 'Frames' is the number of training/test frames, 'Clips' is the number of training/test clips, 'Avg' is the average people count per frame.

α	0		0.6		0.8		1.0		3.0	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
	18.1	26.5	15.1	23.0	**14.1**	**19.9**	15.2	24.2	16.5	23.4

Table 2. The effect of balance factor α using the loss with our statistical regularizer (Eq. (13)) on DRONECROWD.

In realistic scenes, even under severe occlusion, residual visual cues (e.g., partial body contours or shadows) usually suffice for both humans and models to infer a person's presence. Thus, $\mathbb{E}[\mathcal{C}_t \mid \mathcal{I}_t]$ remains a statistically sufficient representation for most observable frames.

From a practical standpoint, this theoretical lens suggests that improving the informativeness of single frames via higher image quality, better viewpoints, or multi-camera setups—directly reduces $\Delta_{l,r}$ and narrows the potential advantage of multi-frame models. Hence, under resource constraints, enhancing input quality may provide a better accuracy–efficiency trade-off than increasing temporal model complexity.

In summary, our analysis does not assert that single-frame inference always suffices for VCC in practice; rather, it delineates the information-theoretic boundary where temporal context ceases to add value, providing a theoretically quantifiable criterion ($\Delta_{l,r}$) for assessing the sufficiency of ICC models for VCC.

4. Experiments

In this section we present experiments using our framework with statistical regularizer (denoted as SR) for VCC.

4.1. Setup

Our experiments are conducted on six benchmark datasets, with details shown in Table 1. Our evaluation adheres to the standard experimental protocol of VCC and covers all representative benchmark datasets commonly adopted in the literature [7, 14, 21, 22, 27, 30, 32, 45, 48, 49]. These benchmarks collectively span diverse viewpoints (bird-view, camera-view, and surveillance-view) and motion dynamics such as smooth directed motion, abrupt local changes, and irregular variations, covering the typical motion patterns investigated in VCC studies. Note that, among these datasets, DRONECROWD is the most challenging, as people appear only as small-scale, low-resolution instances due to the aerial viewpoint, making visual cues for identification extremely subtle and sparse. Furthermore, the scenes in the test set of DRONECROWD are the most different from the scenes in its training set.

technique used	MAE	MSE
baseline ($\alpha = 0$)	18.1	26.5
+ basic regularizer (5)	15.3	22.7
+ motion-tolerant regularizer (13)	**14.1**	**19.9**

Table 3. Ablation study on different type of regularizers. The test is executed on the DRONECROWD dataset.

	MCNN [51]		CSRNet [16]		CAN [26]		VGG19 [37]		MAN [18]	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
w/o SR	34.7	42.5	19.8	25.6	22.1	33.4	18.1	26.5	18.7	23.4
w/ SR	**30.6**	**38.5**	**17.1**	**24.5**	**16.9**	**22.3**	**14.1**	**19.9**	**14.9**	**21.7**

Table 4. Using our statistical regularizer (SR) with other lightweight image-based counting models on DRONECROWD.

The ICC model in our framework is the same VGG19-based model used in [31, 37]. We train the model with the motion-tolerant regularizer, i.e., $\mathcal{L}'$ in (13). The balance factor in (13) is set as 0.8.

For generating ground truth (GT) density maps, the bandwidth γ of Gaussian kernel is set as the conventional 8 pixels. For the data augmentation, we follow the convention and use the random crop (with probability 1) and random horizontal flip (with probability 0.5). The optimizer is Adam with a $1e-5$ learning rate and a $1e-4$ weight decay. We use a batch size of 8 (4 frame pairs in a batch). For approximating the integral calculation in the loss, we use the integral range $[-0.3, 0.3]^2$ and the Riemann sum granularity 0.01, which follows [37].

The evaluation metrics are the standard MAE (mean absolute error) and MSE (root mean square error).

4.2. Ablation Study

The ablation focuses on three aspects: (1) the influence of the balance factor α, (2) the effect of different regularizer forms, and (3) the generality of the proposed regularizer across network architectures. All studies are conducted on the more challenging DRONECROWD dataset.

As for ϵ^* and $\mathbf{\Lambda}^*$, they are statistically inferred from the training data rather than manually tuned. ϵ^* is derived in closed form, and $\mathbf{\Lambda}^*$ is obtained as the optimum of a well-posed data-driven problem with stable solutions. This formulation yields a stable and statistically consistent solution not involving tuning choices, making both parameters data-determined and inherently robust across datasets.

Balance factor α. We study the effect of α using the statistical regularizer in (13). The results for different α are presented in Table 2, and $\alpha = 0.8$ is the best balance factor.

Form of regularizer. We compare different versions of our regularizer in Table 3. Compared to the baseline ($\alpha = 0$), adding the basic regularizer using the loss in (5) decreases the MAE from 18.1 to 15.2. Modifying the basic regularizer to include the tolerance for expected motions further decreases the MAE to 14.1, which demonstrates the efficacy of our SR.

	Tr	Inf	UCSD		VENICE		MALL		FDST		DRONECROWD		VSCROWD	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [16] (2018)	I	I	1.16	1.47	35.8	50.0	2.46	4.7	3.69	4.82	19.8	25.6	13.8	21.1
BL [31] (2019)	I	I	-	-	-	-	1.96	2.49	2.85	3.74	-	-	8.7	11.8
P2PNet [39] (2021)	I	I	2.07	2.70	12.2	16.8	1.92	2.49	2.29	3.04	18.6	26.9	7.0	12.1
MAN [18] (2022)	I	I	1.60	1.90	17.7	22.2	1.70	2.17	1.90	2.36	18.7	23.4	8.3	10.4
ChfL [37] (2022)	I	I	0.92	1.20	13.2	16.9	1.63	2.07	1.63	2.15	18.1	26.5	7.4	13.5
PET [24] (2023)	I	I	1.99	2.49	14.7	20.4	1.77	2.29	1.74	2.37	20.0	28.4	7.3	11.7
Gramformer [19] (2024)	I	I	-	-	21.3	28.6	1.69	2.14	5.15	6.32	-	-	8.1	15.7
ConvLSTM [49] (2017)	V	V	1.30	1.79	-	-	2.24	8.50	4.48	5.82	-	-	28.0	46.1
TAN [47] (2020)	V	V	1.08	1.41	-	-	2.03	2.60	-	-	-	-	-	-
MLSTN [9] (2020)	V	V	1.02	1.32	-	-	1.80	2.42	2.35	3.02	-	-	22.9	41.1
MOPN [13] (2020)	V	V	0.97	1.22	-	-	1.78	2.25	1.76	2.25	-	-	-	-
Monet [2] (2021)	V	V	1.17	1.45	-	-	1.54	2.02	-	-	-	-	-	-
EPF [28] (2021)	V	V	0.81	1.07	14.2	18.4	4.04	5.03	2.10	2.46	28.1	36.5	10.4	14.6
PHNet [32] (2021)	V	V	0.82	1.05	18.1	25.1	-	-	1.65	2.16	-	-	-	-
STDNet [30] (2021)	V	V	0.76	1.01	-	-	1.47	1.88	-	-	-	-	-	-
STNNet [45] (2021)	V	V	-	-	-	-	-	-	-	-	15.8	**18.7**	12.0	18.6
GNANet [14] (2022)	V	V	-	-	-	-	-	-	2.1	2.9	-	-	8.2	10.2
STGN [48] (2022)	V	V	0.82	1.04	14.1	20.1	1.53	1.97	1.38	1.82	22.5	29.0	9.6	12.5
CLRNet [7] (2022)	V	V	**0.72**	**0.94**	-	-	1.45	1.84	1.45	1.97	17.3	23.4	-	-
MFA [21] (2023)	V	V	0.94	1.15	22.4	29.7	1.51	1.93	1.90*	2.60*	25.8	32.4	8.4	19.1
DACM [22] (2025)	V	V	0.94	1.22	11.1	14.3	1.44	1.85	1.31	1.75	18.2	27.3	7.1	14.7
SUCC [29] (2025)	V	V	1.30	1.70	-	-	1.80	2.30	3.10	4.30	-	-	-	-
SR (ours)	V	I	0.75	0.97	**8.2**	**10.5**	**1.43**	**1.82**	**1.27**	**1.61**	**14.1**	19.9	**5.4**	**9.5**

Table 5. Comparison with state-of-the-art video counting methods. We compare with both (top) image-based and (bottom) video-based counting methods. The best result is in bolded, and the second best is underlined. 'Tr.' and 'Inf.' indicate the type of input data used during training and inference: 'V' means the method uses consecutive video frames as input, and 'I' means using a single image as input. ($*$: The reported MAE&MSE are different in [21] and [22]. We ran with the official codes and adopted the results from [22].)

Algorithm	training time per epoch	inference time per image	fps
EPF [28]	49 min	0.043 s	23.5
STGN [48]	476.1 s	0.017 s	58.5
MFA [21]	242.0 s	0.033 s	30.1
DACM [22]	435.3 s	0.016 s	61.3
SR (ours)	85.5 s	0.010 s	99.5

Table 6. Runtime comparison on FDST. All times are obtained using PyTorch on a Linux system with an RTX3090 TI. The reported time is the average over 50 epochs.

Generality of regularizer. We select five typical lightweight counting networks to demonstrate that SR is generally effective regardless of network structures. The results in Table 4 show that our SR framework generally enhances the VCC ability of ICC models.

4.3. Comparison with SOTAs

Next we compare our method (SR) with SOTA VCC methods. We conduct comparisons from two aspects, counting performance and efficiency.

Counting performance. Table 5 compares the counting performance on six benchmark datasets. SR obtains the best MAE&MSE on MALL, VENICE, FDST, and VSCROWD, as well as the 2nd best MAE&MSE on UCSD dataset. On the challenging DRONECROWD, SR obtains the best MAE and 2nd best MSE. On the two larger datasets, DRONECROWD and VSCROWD, our method's MAE greatly outperforms the SOTA methods. On VENICE, which has relatively less training data, our strategy has a significant advantage on the counting performance.

Efficiency. We also compare the efficiency of our method with four VCC methods with available codes. We test on FDST and the image size is resized to 640×360 according to the convention. Table 6 shows the result. The efficiency advantage of our method during training becomes more evident as the dataset becomes larger. Our inference time advantage will be more evident as the size of the input frame increases. In addition, because our framework uses single-frame based inference, the frame rate for inference is substantially higher than video-based methods, which better satisfies the real-time demand of VCC.

5. Conclusion

We approached VCC from a lightweight and statistically grounded perspective. By analytically defining the ST information relevant to counting, we established a principled way to incorporate temporal cues into ICC models through a statistical regularizer, without introducing extra modules or inference overhead. Our theoretical analysis further quantified the attainable gap between image-based and video-based predictors, providing a clear criterion for when single-frame inference achieves video-level optimality. Comprehensive experiments on six diverse benchmarks verified that our framework preserves the efficiency of lightweight ICC models while achieving SOTA accuracy, demonstrating its scalability and robustness across architectures and scenes.

References

[1] Marco Avvenuti, Marco Bongiovanni, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Nicola Messina. A spatio-temporal attentive network for video-based crowd counting. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2022. 3

[2] Haoyue Bai and S-H Gary Chan. Motion-guided non-local spatial-temporal network for video crowd counting. *arXiv preprint arXiv:2104.13946*, 2021. 8

[3] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008. 1, 2, 7

[4] I Chen, Wei-Ting Chen, Yu-Wei Liu, Ming-Hsuan Yang, Sy-Yen Kuo, et al. Improving point-based crowd counting and localization based on auxiliary point guidance. *arXiv preprint arXiv:2405.10589*, 2024. 2

[5] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, page 3, 2012. 7

[6] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013. 2

[7] Li Dong, Haijun Zhang, Jianghong Ma, Xiaofei Xu, Yimin Yang, and QM Jonathan Wu. Clrnet: A cross locality relation network for crowd counting in videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3, 7, 8

[8] Yanyan Fang, Biyun Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. Locality-constrained spatial transformer network for video crowd counting. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 814–819. IEEE, 2019. 1, 3, 7

[9] Yanyan Fang, Shenghua Gao, Jing Li, Weixin Luo, Linfang He, and Bo Hu. Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting. *Neurocomputing*, 392:98–107, 2020. 3, 8

[10] Weina Ge and Robert T Collins. Marked point processes for crowd counting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920. IEEE, 2009. 2

[11] Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. Regressor-segmenter mutual prompt learning for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28380–28389, 2024. 2

[12] Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21848–21859, 2023. 3

[13] Mohammad Asiful Hossain, Kevin Cannons, Daesik Jang, Fabio Cuzzolin, and Zhan Xu. Video-based crowd counting using a multi-scale optical flow pyramid network. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 8

[14] Haopeng Li, Lingbo Liu, Kunlin Yang, Shinan Liu, Junyu Gao, Bin Zhao, Rui Zhang, and Jun Hou. Video crowd localization with multifocus gaussian neighborhood attention and a large-scale benchmark. *IEEE Transactions on Image Processing*, 31:6032–6047, 2022. 3, 7, 8

[15] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th international conference on pattern recognition*, pages 1–4. IEEE, 2008. 2

[16] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 1, 7, 8

[17] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022. 2

[18] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19628–19637, 2022. 7, 8

[19] Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. Gramformer: Learning crowd counting via graph-modulated transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3395–3403, 2024. 8

[20] Wei Lin and Antoni B Chan. Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21663–21673, 2023. 2

[21] Miaogen Ling, Tianhang Pan, Yi Ren, Ke Wang, and Xin Geng. Motional foreground attention-based video crowd counting. *Pattern Recognition*, 144:109891, 2023. 7, 8

[22] Miaogen Ling, Jixuan Chen, Yongwen Liu, Wei Fang, and Xin Geng. Dual-branch adjacent connection and channel mixing network for video crowd counting. *Pattern Recognition*, page 111709, 2025. 7, 8

[23] Bo Liu and Nuno Vasconcelos. Bayesian model adaptation for crowd counts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4175–4183, 2015. 2

[24] Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-query quadtree for crowd counting, localization, and more. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1676–1685, 2023. 2, 8

[25] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4823–4833, 2021. 1

[26] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 5099–5108, 2019. 1, 7

[27] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Estimating people flows to better count them in crowded scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 723–740. Springer, 2020. 1, 3, 7

[28] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Counting people by estimating people flows. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8151–8166, 2021. 1, 3, 8

[29] Rui Ma, Yi Hou, Chenxuan Li, Huizhu Jia, and Xiaodong Xie. Scene-adaptive unsupervised crowd counting for video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 8

[30] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. *IEEE Transactions on Multimedia*, 24:261–273, 2021. 3, 7, 8

[31] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6142–6151, 2019. 1, 3, 7, 8

[32] Shiqiao Meng, Jiajie Li, Weiwei Guo, Lai Ye, and Jinfeng Jiang. Phnet: Parasite-host network for video crowd counting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1956–1963. IEEE, 2021. 7, 8

[33] Hong Mo, Xiong Zhang, Jianchao Tan, Cheng Yang, Qiong Gu, Bo Hang, and Wenqi Ren. Countformer: Multi-view crowd counting transformer. *arXiv preprint arXiv:2407.02047*, 2024. 3

[34] Zhuoxuan Peng and S-H Gary Chan. Single domain generalization for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28025–28034, 2024. 2

[35] Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Diffuse-denoise-count: Accurate crowd-counting with diffusion models. *arXiv preprint arXiv:2303.12790*, 2023. 3

[36] Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Crowddiff: Multi-hypothesis crowd density estimation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12809–12819, 2024. 2

[37] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19618–19627, 2022. 1, 2, 3, 4, 5, 7, 8

[38] Weibo Shu, Jia Wan, and Antoni B Chan. Generalized characteristic function loss for crowd analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 4, 5

[39] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021. 2, 8

[40] Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1130–1139, 2019. 2

[41] Jia Wan, Qingzhong Wang, and Antoni B Chan. Kernel-based density map generation for dense object counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[42] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021. 3

[43] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. *arXiv preprint arXiv:2009.13077*, 2020. 3

[44] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network. *arXiv preprint arXiv:1912.01811*, 2019. 1

[45] 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. 1, 6, 7, 8

[46] Shaokai Wu and Fengyu Yang. Boosting detection in crowd analysis via underutilized output features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15609–15618, 2023. 3

[47] Xingjiao Wu, Baohan Xu, Yingbin Zheng, Hao Ye, Jing Yang, and Liang He. Fast video crowd counting with a temporal aware network. *Neurocomputing*, 403:13–20, 2020. 3, 8

[48] Zhe Wu, Xinfeng Zhang, Geng Tian, Yaowei Wang, and Qingming Huang. Spatial-temporal graph network for video crowd counting. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):228–241, 2022. 1, 3, 7, 8

[49] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5151–5159, 2017. 1, 3, 7, 8

[50] Qi Zhang and Antoni B Chan. 3d crowd counting via multi-view fusion with 3d gaussian kernels. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12837–12844, 2020. 1

[51] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 1, 7

[52] Tao Zhao and Ramakant Nevatia. Bayesian human segmentation in crowded situations. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–459. IEEE, 2003. 2

[53] Zhiyuan Zhao, Tao Han, Junyu Gao, Qi Wang, and Xuelong Li. A flow base bi-path network for cross-scene video crowd understanding in aerial view. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 574–587. Springer, 2020. 1, 3