# Active View Selection for Scene-level Multi-view Crowd Counting and Localization with Limited Labeling Budget

Qi Zhang, *Member, IEEE,* Bin Li, Antoni B. Chan, *Senior Member, IEEE,* and Hui Huang*, *Senior Member, IEEE*

***Abstract*—Multi-view crowd counting and localization fuse the input multi-camera views for estimating the crowd number or locations on the ground. Existing methods mainly focus on accurately predicting on the crowd shown in the input views, which neglects the problem of choosing the 'best' camera views to perceive all crowds well in the scene. Besides, existing view selection methods require massive labeled views and images, and lack the ability for cross-scene settings, reducing their application scenarios. Thus, in this paper, we study the view selection issue for better scene-level multi-view crowd counting and localization results with cross-scene ability and limited label demand, instead of input-view-level results. We first propose a baseline view selection method (IVS) that considers view and scene geometries in the view selection strategy and conducts the view selection, labeling, and downstream tasks independently. Based on the IVS baseline, we put forward an active view selection method (AVS) that jointly conducts the view selection, labeling, and downstream tasks. In AVS, we actively select the labeled views and consider both the view/scene geometries and the predictions of the downstream task models in the view selection process under a limited labeling budget. Experiments on multi-view counting and localization tasks demonstrate the cross-scene and the limited label demand advantages of the proposed active view selection method (AVS), outperforming existing methods and with wider application scenarios.**

***Index Terms*—Multi-view, crowd counting, crowd localization, view selection, limited label.**

## I. Introduction

Multi-view crowd counting and localization leverage multiple camera views to predict the crowd counts or locations on the ground, alleviating the issue of severe occlusions in large and wide scenes. However, existing multi-view crowd counting and localization methods mainly focus on designing models for accurate estimation of the crowd covered by a randomly selected set of input views [1], which may not contain or perceive all the crowds well in the scene, resulting in an incorrect prediction of the crowd in the scene. As in Fig. 1 top, these frameworks are trained and tested using the ground truth (GT) constructed from the randomly selected views, *i.e.*, not tested on the whole scene.

Thus, for a complete multi-view vision system, we not only need to design better downstream task models (*e.g.* counting, localization) but also select the best views for better scene-level downstream task performance. A simple two-stage solution is to conduct the view selection first, label the selected views, and then train the downstream multi-view models based on the labeled views, called independent view selection. As shown in Fig. 1 bottom, the view-selection-based multi-view framework is trained with GT constructed with selected views, and tested with the scene-level GT, where *the GT constructed with selected views should be close to the scene-level GT*. Recent path planning methods [2]–[4] adopted the scene/view geometries in view selection for better 3D reconstruction.

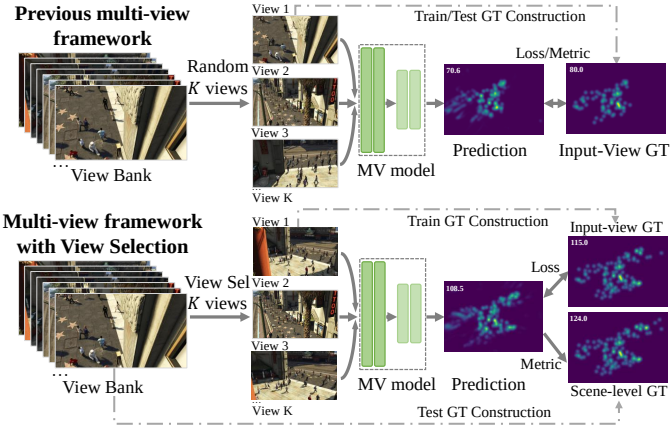



Fig. 1: The comparison of the existing multi-view counting/localization frameworks and the scene-level multi-view framework with view selection.

Unfortunately, few research works have studied the view selection issue in multi-view counting and localization for whole scene performance. In addition, the two-stage solution divides the view selection and downstream task model training into 2 separate stages, where the priors used for view selection may not be optimal settings for the downstream tasks. Furthermore, to reduce the annotation effort, sometimes only limited views are labeled, causing extra difficulties for downstream model learning. A recent method MVSelect [5] proposed a reinforcement learning (RL) framework for view selection and multi-view tasks. However, MVSelect requires GT annotations of *all views*, making it impractical for selecting among large numbers of views to save annotation budget, and has a weak generalization ability due to the limitations of the proposed framework, *making it not applicable to novel scenes*.

To address the mentioned issues, in this paper, we first propose an independent view selection **baseline** (IVS), based on which we further put forward an active view selection framework (AVS), requiring only limited labeled views and with cross-scene abilities. IVS proposes a view selection score

Qi Zhang, Bin Li, and Hui Huang are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Antoni B. Chan is with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China. E-mail: qi.zhang.opt@gmail.com, azswerfr@gmail.com, abchan@cityu.edu.hk, hhzhiyan@gmail.com

*Corresponding author: Hui Huang

Manuscript received xx, 2026; revised xx, 20xx

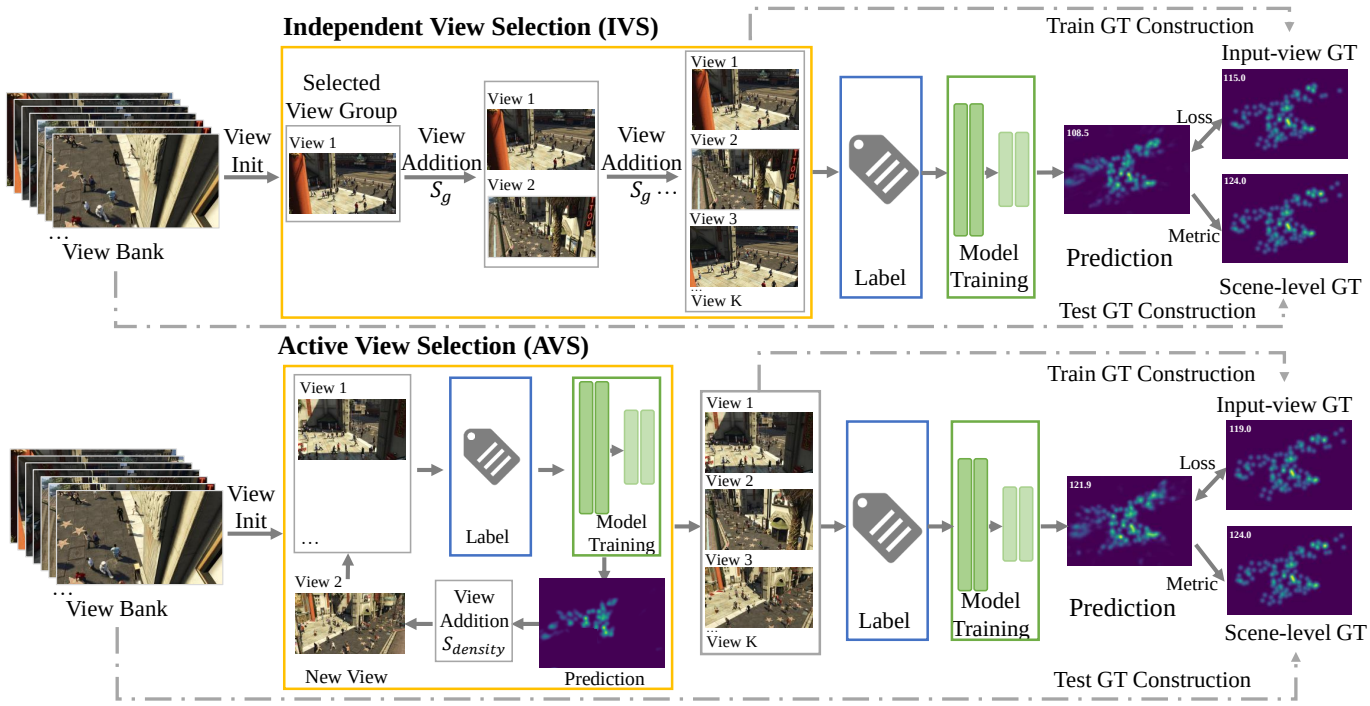



Fig. 2: The pipeline of the proposed independent view selection baseline (IVS) and active view selection framework (AVS). IVS (top) separates the view selection (selecting views one by one with view selection score $S_g$ based on scene/view geometries) and downstream model training, while AVS (bottom) jointly conducts view selection and downstream model training: In the view selection process, the downstream model's prediction together and scene/view geometries are both used in view selection score $S_{density}$ to select new views, then the downstream model is trained with the updated view group, repeating the process until finishing selecting $K$ views, finally the downstream model is trained again with the selected and labeled views.

equation $S_g$ based on view and scene geometries, including 3 terms: *the scene coverage, the average distance, and the view diversity*, working together to cover the crowd in the scene mostly and clearly. The 3 terms are combined to select views that cover most scene areas to contain as many crowds as possible, have the lower average person-to-camera distance to "see" the crowd clearly, and have the higher view diversity instead of placing all cameras at the same locations. As illustrated in the top part of Fig. 2, we first perform view selection to expand the selected view set from a single view to $K$ views based on the proposed view selection score equation $S_g$. The selected views are then annotated and used to train the multi-view task model. During training, GT is constructed from the selected input views, whereas during testing, the GT is derived from all available views.

Furthermore, in contrast to optimizing the view selection and the downstream model independently as in IVS, AVS jointly optimizes the view selection and the downstream tasks by introducing the downstream task predictions in the view selection process. As shown in Fig. 2 bottom, the view selection score $S_{denisty}$ considers both the view/scene geometries (similar in IVS) and the prediction results of the downstream task models when expanding the selected view group from 1 to $K$ views, after which the downstream model is trained with the labeled selected views. Thus, the view selection, the data labeling, and the downstream model training are jointly conducted. Once $K$ views have been selected, the downstream model is retrained on the labeled views. Additionally, to reduce annotation costs, we further introduce a novel pseudo-labeling strategy that is applied during both the view selection and downstream training stages.

The contributions of this paper are summarized as follows.

- Few research works have studied the view selection problem for scene-level multi-view crowd counting and localization. We propose a novel independent view selection baseline IVS for scene-level downstream tasks, utilizing view and scene geometries in the view selection.
- Based on IVS, we propose an active framework AVS, where the view selection step and the downstream task models are jointly optimized, with better performance than IVS. And we only require limited view labels from the selected views, via adopting pseudo labels on candidate views in the training.
- Experiments demonstrate that our method can apply to novel new scenes, with wider application scenarios, and outperform comparison methods on both multi-view counting and localization tasks.

## II. Related Work

**Multi-view crowd counting.** Compared to single-image counting [6]–[13], multi-view crowd counting is proposed to handle scenes with large areas, severe occlusions, or irregular shapes by fusing multiple synchronized and calibrated camera views. Traditional methods [14], [15] employed foreground extraction techniques and hand-crafted features, with limited

TABLE I: Summary of main notations.

| Symbol | Meaning | Symbol | Meaning |
|---|---|---|---|
| $F$ | The number of selected frames | $H_i$ | Visible scene region by view $i$ |
| $K$ | The number of selected views | $h$ | Scene height |
| $v_{max}$ | The view with the largest FOV | $w$ | Scene width |
| $V_{select}$ | Selected view group | $D_p$ | Sum of inverse distance |
| $S$ | View selection score equation | $S_{sc}$ | $S$ with scene coverage |
| $S_g$ | $S$ with view/scene geometries | $S_{ad}$ | $S$ with average distance |
| $S_{mask}$ | Mask-indicated $S$ | $S_{vd}$ | $S$ with view diversity |
| $S_{density}$ | Density-indicated $S$ | $B_k$ | Crowd density map mask |
| $N$ | Downstream task model | $M_k$ | Crowd density map |
| $\emptyset$ | No $N$ | $D_p^{den}$ | $D_p$ with $M_k$ |
| $H_s$ | Scene region | $V_{select}^k$ | $k$ selected views |
| $H_v^k$ | Visible scene region by $V_{select}$ | $V^g$ | View set of scene's $g$ |
| $G$ | Scene set | $E$ | The number of training epoch |
| $V_{select}^g$ | Selected views of scene $g$ | | |

performance and generalization abilities. Recently, deep learning methods [16]–[19] have been introduced, trained, and tested with ground-plane density maps based on the crowds appearing in the input camera views. CVCS [1] proposed a camera selection model for cross-view cross-scene multi-view counting with a large synthetic cross-scene multi-view dataset. [20] put forward a transformer model with attention-mechanism-based 2D-3D feature lifting. Overall, existing multi-view counting methods focus on the accurate crowd number estimation of the people contained in a randomly selected set of input camera views. *Current SOTAs have not yet explored the problem of selecting the best views for scene-level multi-view counting*. Moreover, existing pretrained single-image models can serve as backbones for multi-view models, thereby improving training speed and effectiveness. And it can directly help us obtain crowd information to enhance perception, such as the initialization process of the proposed methods.

**Multi-view crowd localization.** Multi-view crowd localization estimates the crowd locations on the ground in the scene. Early methods' performance is limited [21], [22] due to no view feature alignment. Recent methods [23]–[31] put forward end-to-end frameworks with better performance. MVDet [32] used feature perspective transformations to fuse multi-views. [29] extends multi-view crowd localization to pedestrian occupancy prediction, achieving SOTA performance on the proposed benchmark. CaMuViD [33] facilitates flexible transformation and improves feature fusing across views, removing the need for BEV representation and achieving better detection accuracy. Similarly, *most SOTA multi-view crowd localization methods also focus on estimating crowds 'seen' in the input camera views, not targeting all crowds in the scene*. MVSelect [5] is the most related to our paper, and it proposed a reinforcement learning (RL) framework for view selection and downstream tasks. However, MVSelect requires annotations of all views to train the model, and has a weak generalization ability to apply to novel scenes. *In contrast, our method only needs to label the selected views, and with the aid of the proposed pseudo labels, it can be well applied to novel news scenes in the test stage* (see experiments).

**View selection for other multi-view tasks.** View selection is also vital in many other multi-view vision tasks [34]–[42], such as path planning, or multi-view object classification. [43] measured the reconstructability in a learning way and designed an interactive path planning framework for view selection. MVTN [44] directly regresses optimal viewpoints for 3D shape recognition with an MLP. [42] proposed a reinforcement learning-based framework for multi-view active fine-grained visual recognition. [45] proposed a unified framework for view selection methods and devised a thorough benchmark to assess its impact on neural rendering. By considering the reconstruction errors' spatial distribution of the 3D predictions, a reconstruction error-guided view selection method [46] has been proposed and achieves better reconstruction performance. [2] designs a comprehensive texture quality assessment system to guide view planning, and proposes an innovative aerial path planning framework designed to co-capture images for reconstructing both structured geometry and high-fidelity textures. Furthermore, [3] introduces a novel changeability heuristic to evaluate the likelihood of scene changes, driving the planning of two flight paths to select better viewpoints for 3D reconstruction. *It is a trend in other multi-view tasks to jointly conduct the view selection and the downstream task. However, there is little research on view selection for scene-level multi-view crowd counting and localization tasks*, which is a relatively unexplored area.

## III. Active View Selection Framework

We first propose a novel independent view selection **baseline** (IVS) adopting a two-stage process for scene-level downstream tasks. Next, based on IVS, we propose the active view selection framework (AVS) for jointly optimizing view selection and downstream task model training. Pseudo labels are adopted to enhance the model's cross-scene generalization abilities. For both IVS and AVS, we assume an annotation budget of $F$ frames per view and $K$ views of each scene, or a total $FK$ images per scene. We provide a summary table of the main notations shown in Table I for symbol querying.

### A. Independent View Selection Baseline (IVS)

In IVS, we first initialize the selected frames and the first view, then start to add new views one by one with the proposed view selection score equation based on scene/view geometries, expanding the selected view number from 1 to $K$.

**Algorithm 1** Independent View Selection Framework

**Input**: each scene's total views $V^g \in \{v_1^g, ..., v_n^g\}$, all the scenes $G = \{g\}$, max selected view number $K$, view selection score equation $S = S_g$, task model $N$.
initialize frames and the selected view group $\{V_{select}^g\}$.
**for** $g \in G$ **do**
  **for** $k \in \{2, \ldots, K\}$ **do**
    view_addition($V^g, V_{select}^g$, $S$, N);
  **end for**
**end for**
label all the selected views $\{V_{select}^g\}$;
model_training($N, \{V_{select}^g\}$).

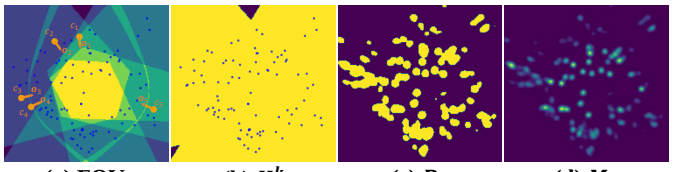
**Algorithm 2** View Addition

**Input**: all views $V^g \in \{v_1^g, ..., v_n^g\}$ of scene $g$, selected view group $V_{select}^g$ of scene $g$, view selection score equation $S \in \{S_g, S_{mask}, S_{density}\}$, and task model $N$ ($= \emptyset$ if $S = S_g$).
$s\_v_{select} = -\inf$;
**for** $v \in V^g \setminus V_{select}^g$ **do**
  $s\_v = S(\{V_{select}^g, v\}, N)$;
  **if** $s\_v > s\_v_{select}$ **then**
    $v_{select} = v$;
    $s\_v_{select} = s\_v$;
  **end if**
**end for**
$V_{select}^g = \{V_{select}^g, v_{select}\}$.

*1) Initialization:* Initialization has two stages: First, selecting the $F$ frames to be processed and selecting the first view. For frame selection, we first find the view $v_{max}$ with the largest field-of-view (FOV) area on the ground. Then, we select the first frame as the one with the largest predicted crowd count in view $v_{max}$ using DM-Count [47], a pre-trained single-image counting model that can help effectively perceive crowd information in a label-efficient manner. Next, given the diversity in the limited labels, we select the rest frames with the lowest cosine similarity between the selected frames and candidate frames of view $v_{max}$. This process is repeated until $F$ frames are selected. For view initialization, the view with the largest crowd count sum across all selected $F$ frames is selected as the first view.

*2) View Addition: $S_g$:* With the selected frames generated by frame initialization and the selected view group $V_{select}$ including the first view from view initialization above, a view selection score equation $S_g$ is proposed for view addition. For each iteration, the $S_g$ score of each candidate view together with the current $V_{select}$ is calculated, and then we select the new view with the largest score. The process is repeated until the specified $K$ views are selected, as shown in Algorithm 2. $S_g$ consists of 3 terms: Scene Coverage, Average Distance, and View Diversity, which are as follows.

**Scene coverage** score term $S_{sc}$ indicates whether the selected views can cover all crowds in the scene as much as possible. We first use the scene ground plane map as the scene region $H_s$, whose area size is $Area(H_s) = hw$, and $h$ and $w$ are the height and width of the map. Then, we calculate the visible scene areas covered by the selected views as $H_v^k = \{H_1 \cup H_2 ... H_k\}$, which is the combined FOV region of $k$ views (see Fig. 3 (a) and (b)), and $H_i$ is view $i$'s FOV covering region. Thus, the area ratio of the visible region $H_v^k$ and the scene region $H_s$ is defined as $S_{sc}$:



(a) FOVs (b) $H_v^k$ (c) $B_k$ (d) $M_k$

Fig. 3: (a) The combined FOVs of selected views; (b) The FOV mask; (c) The crowd mask; (d) The crowd density. The dots in (a) and (b) are ground-truth crowd locations.

$$S_{sc} = \sum H_v^k / Area(H_s) = \sum \{H_1 \cup H_2 ... H_k\} / (hw), \quad (1)$$

where a larger $S_{sc}$ indicates a higher probability of covering all crowds by the selected views.

**Average distance** score term $S_{ad}$ considers the average person-to-camera distance in the view selection, indicating whether the selected views can 'see' the crowd clearly. Specifically, for each location $p$ in the visible region $H_v^k$, the person's inverse distance to the currently selected $k$ cameras is calculated as $D_p = \sum_{i=1}^{k} 1/\|p - c_i\|$, where $c_i$ is the $i$-th camera's location on the ground (see Fig. 3 (a)) and $p$ is in $H_i$, where higher values indicates shorter distance to the selected cameras. Combining all crowds, $S_{ad}$ is:

$$S_{ad} = \frac{\sum_{p \in H_v^k} D_p}{\sum H_v^k}. \quad (2)$$

As the crowd locations are not known, all locations in the approximate visible region $H_v^k$ are used in the calculation. Similarly, a higher $S_{ad}$ forces the selected cameras to be close to the crowds and captures the crowds more clearly.

**View diversity** score term $S_{vd}$ avoids the setting that all cameras are located at the same place and point out. Because we require multi-cameras to be placed at different corners facing each other to make full use of their occlusion handling potential [16]. Thus, we adopt a similarity measure [48] to calculate $S_{vd}$, which is defined as follows:

$$S_{vd} = \exp\left(-\lambda \sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{\mathbf{o}_i \cdot \mathbf{o}_j}{\|c_i - c_j\| + \epsilon}\right), \quad (3)$$

where $\mathbf{o}_i$ and $\mathbf{o}_j$ are the camera optical axis directions (see Fig. 3 (a)), $c_i$ and $c_j$ are camera locations, $\lambda$ is a hypeparameter, and $\epsilon$ is a small value to avoid zero denominator. When the selected cameras have larger view direction and location differences, i.e., view diversity, $S_{vd}$ is higher.

By combining the 3 terms, we obtain the independent view selection score equation $S_g$ by only considering the view and scene geometries.

$$S_g = S_{sc} * S_{ad} * S_{vd} \quad (4)$$

$$= \frac{\sum_{p \in H_v^k} D_p}{Area(H_s)} \exp\left(-\lambda \sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{\mathbf{o}_i \cdot \mathbf{o}_j}{\|c_i - c_j\| + \epsilon}\right). \quad (5)$$

Once the required frames and views are selected, they are annotated, and then the downstream task model is trained on the labeled data (see Fig. 2 top right). The main **weakness** of

**Algorithm 3** Active View Selection Framework

**Input**: each scene's total views $V^g \in \{v_1^g, ..., v_n^g\}$, all the scenes $G = \{g\}$, max selected view number $K$, training epochs $E$, threshold $\tau$ to add view, view selection score equation $S \in \{S_{mask}, S_{density}\}$, and task model $N$.
initialize frames and the selected view group $\{V_{select}^g\}$.
label all the selected views $\{V_{select}^g\}$.
**for** $e \in \{1, \ldots, E\}$ **do**
  $metric$ = model_training($N$,$\{V_{select}^g\}$);
  **if** $metric > \tau$ **and** $len(V_{select}^g) < K$ **then**
    **for** $g \in G$ **do**
      view_addition($V^g, V_{select}^g, S, N$);
    **end for**
    label all the added new views;
  **end if**
**end for**

the independent view selection baseline (IVS) is that the view selection and downstream model training are separated, which does not ensure an optimal result for scene-level tasks. For example, the selected views are not necessarily suitable for downstream model training due to multi-view counting and localization models being sensitive to view angles, heights, or other properties.

### B. Active View Selection (AVS)

To address the weakness of the IVS baseline–optimizing the view selection and downstream model separately, we propose the AVS framework that jointly optimizes the view selection and downstream task models as in Fig. 2 bottom. In the view selection process, the intermediate model's prediction together with the scene/view geometries is adopted in the view selection score $S_{density}$, and selected new views are labeled and used to train the downstream model, which is repeated until the desired view number is reached. Finally, the downstream model is trained again with the selected and labeled views. The complete algorithm procedure is presented in Algorithm 3. The initialization process is the same as IVS. We update the view selection score equation in IVS by introducing the downstream task predictions, denoted as $S_{mask}$ and $S_{density}$, with details as follows.

**Mask-indicated view selection $S_{mask}$.** The visible region in (5) is defined by the combination of the FOVs of the selected cameras, which neglects the actual crowd regions in the scene. Therefore, we rely on the prediction density maps $M_k$ from the downstream model $N$ of the selected $k$ views $\{v_1, v_2, ..., v_k\}$ to more accurately indicate the appearing crowds' regions in the scenes: $M_k = N(v_1, v_2, ..., v_k)$. Specifically, we binarize $M_k$ with a threshold $\sigma$ to obtain a crowd density map mask $B_k$ (see Fig. 3 (c)), which is used to replace $H_v^k$ in (1), (2) and (5). Thus, we obtain

$$S_{mask} = \frac{\sum_{p \in B_k} D_p}{Area(H_s)} \exp\left(-\lambda \sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{\mathbf{o}_i \cdot \mathbf{o}_j}{\|c_i - c_j\| + \epsilon}\right), \quad (6)$$

where $B_k$ indicates *the crowd location information, which is utilized in the view selection process*. Note that the downstream counting or localization model $N$ is involved in the view selection score term $S_{mask}$, while the newly selected views during the view selection process could be fed into and train the downstream model. Therefore, the view selection and downstream task model are interacting in these two steps and thus influence each other.

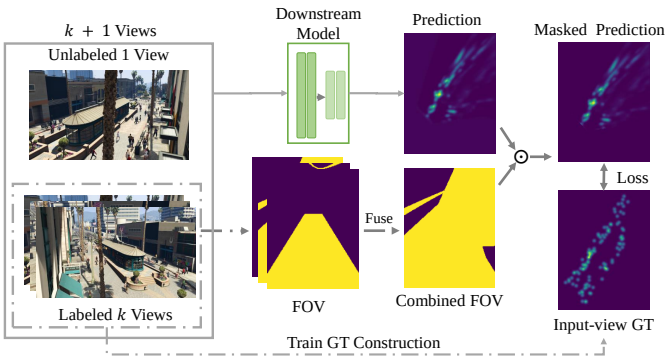



Fig. 4: The pseudo label during view selection. $\odot$ denotes the element-wise multiplication of matrices.

**Density-indicated view selection $S_{density}$.** The mask-indicated view selection uses the binarized density map $B_k$ to indicate the crowd-visible regions, which neglects the crowd density's influence on the view selection score term. In other words, the crowded areas with higher densities should have higher weights in the view selection process. Therefore, we propose the density-indicated view selection score $S_{density}$ by introducing the density prediction $M_k$ (see Fig. 3 (d)) of the downstream task model into $D_p$ in (2), which is rewritten as $D_p^{den} = \sum_i^k \frac{M_k(p)}{\|p - c_i\|}$, where $M_k(p)$ indicates the density value of point $p$. Thus, *by updating $S_{ad}$ with $D_p^{den}$, and replacing $H_v^k$ with $B_k$ in $S_{sc}$ and $S_{ad}$*, the view selection score term in (6), we obtain:

$$S_{density} = \frac{\sum_{p \in B_k} D_p^{den}}{Area(H_s)} \exp\left(-\lambda \sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{\mathbf{o}_i \cdot \mathbf{o}_j}{\|c_i - c_j\| + \epsilon}\right). \quad (7)$$

The view selection score term $S_{density}$ considers both *the view/scene geometries, and the crowds' density level and location information*, where the view selection and downstream model training are conducted jointly.

### C. Pseudo Labels and Training

To enhance the model's generalization ability, we utilize novel pseudo labels to train the downstream model better. During view selection, the currently selected views $V_{select}^k$ and a random unselected view are combined as pseudo inputs to train the model, whose GT is ground-plane density maps of crowds covered by $V_{select}^k$. Specifically, as in Fig. 4, for the pseudo inputs in the view selection, labeled $k$ views and 1 random view from the remaining unlabeled views will be together as the model's inputs (total $k+1$ views) to obtain the predicted ground-plane density map, and the predicted density map is further masked by the combined FOV mask of the labeled $k$ views. The corresponding GT ground-plane density map is constructed from the labeled $k$ views, which is accurate for supervising the prediction masked by the combined FOV of the selected $k$ views. Thus, additional unlabeled views

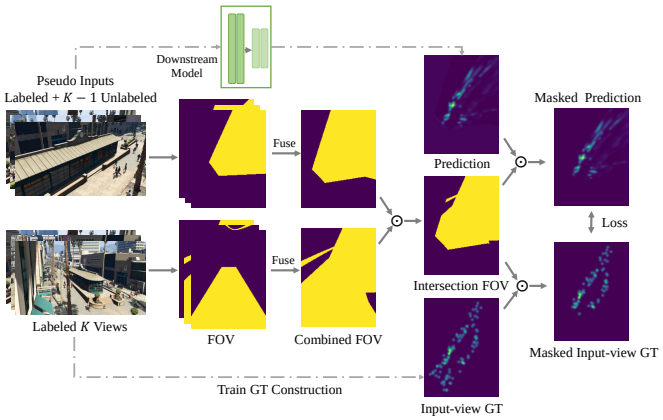



Fig. 5: The pseudo label in the final downstream model training after view selection.

TABLE II: The label cost of selected view number $K$ and selected $F$ on different datasets.

| Dataset | View | Scene | Original Training Frame | $K$ | $F$ |
|---|---|---|---|---|---|
| CityStreet | 3 | 1 | 300 | 2 | 60 |
| CVCS | 60-120 | 31 | 100 | 5 | 20 |
| Wildtrack | 7 | 1 | 360 | 3 | 360 |
| MultiviewX | 6 | 1 | 360 | 3 | 360 |

can be incorporated into downstream model training, thereby enhancing its generalization to new views.

Besides, after the view selection, the selected views $V^K_{select}$ (note that $K$ is the final number of the selected views) of the $F$ selected frames are used for downstream model training. In addition to that, we also add pseudo inputs in training, which is a mix of 1 selected view and $K-1$ unselected random views, whose pseudo-GT is the $K$ selected views' GT ground-plane density maps masked by the intersection of $H^K_v$ and the pseudo input views' combined FOV mask, and the prediction is also masked by the intersection FOV. Specifically, as shown in Fig. 5, for the pseudo inputs after view selection, 1 selected view (random chosen from $V^K_{select}$) and $K-1$ random views from the remaining unlabeled views are combined and regarded as the model's inputs to predict the ground-plane density map. The GT ground-plane density map is constructed from the labeled $K$ views. Since the pseudo inputs and GT are from different views, we can only supervise the common regions covered by the pseudo input views and the $K$-labeled views constructing the GT. Thus, we add a FOV intersection mask on both the prediction and GT density map in the loss calculation, where the intersection mask is the common region of the combined FOVs of the pseudo-input views and the $K$-labeled views. Note that both pseudo labels, which are used during the view selection stage or applied after view selection, are generated based on the selected $F$ multi-frames and the labeled views.

By using pseudo labels, a large number of unlabeled views are incorporated into model training, significantly enhancing the model's generalization. *Both IVS and AVS adopted pseudo labels in the model training.* For IVS, because it does not involve joint training with downstream tasks during view selection, the pseudo-labels are used only in the final training of the downstream model after view selection. Additionally, for both IVS and AVS, the training set samples are doubled during the final model training stage after view selection. Specifically, compared with the original training method, which doesn't use the pseudo inputs, the pseudo inputs are used as additional samples, *i.e.* in the ratio of 1:1 for the $K$-labeled view inputs and the pseudo-label view inputs.

TABLE III: Comparison of the multi-view counting results on the CVCS dataset.

| Method | MAE ↓ | MSE ↓ | NAE ↓ | CoverRate ↑ |
|---|---|---|---|---|
| MVMS [16] (Random) | 37.22 | 44.74 | 0.276 | 0.872 |
| CVCS [1] (Random) | 30.97 | 36.47 | 0.228 | 0.901 |
| Uniform | 21.76 | 25.75 | 0.163 | 0.945 |
| Random | 36.59 | 42.06 | 0.271 | 0.885 |
| Random (Pseudo) | 28.22 | 33.73 | 0.208 | 0.885 |
| MVSelect [5] | 13.58 | 17.94 | 0.099 | 0.899 |
| IVS_$S_g$ (Ours) | 14.81 | 18.47 | 0.110 | 0.954 |
| AVS_$S_{mask}$ (Ours) | 12.53 | 15.33 | 0.093 | 0.955 |
| AVS_$S_{density}$ (Ours) | **10.99** | **13.57** | **0.083** | **0.960** |

TABLE IV: Comparison of the multi-view counting results on the CityStreet dataset.

| Method | MAE↓ | MSE↓ | NAE↓ | CoverRate↓ |
|---|---|---|---|---|
| MVMS [16] (Random) | 16.82 | 20.68 | 0.194 | 0.961 |
| CVCS [1] (Random) | 15.14 | 18.25 | 0.180 | 0.948 |
| Uniform | 12.91 | 15.20 | 0.172 | **0.983** |
| Random | 13.48 | 16.47 | 0.166 | 0.965 |
| Random (Pseudo) | 11.01 | 13.66 | 0.128 | 0.965 |
| MVSelect [5] | 10.39 | 13.04 | **0.116** | 0.957 |
| IVS_$S_g$ (Ours) | 11.28 | 14.36 | 0.128 | 0.946 |
| AVS_$S_{mask}$ (Ours) | 10.26 | 12.52 | 0.121 | 0.946 |
| AVS_$S_{density}$ (Ours) | **9.80** | **11.93** | 0.118 | 0.946 |

## IV. Experiments and Results

### A. Experiment Settings

**Experiment design.** In the training, IVS conducts the view selection, labeling, and downstream model training independently, while AVS conducts them jointly until the view number reaches $K$ and then trains the downstream task model on the labeled $K$ views. In the testing, no model training is needed for IVS or AVS, where the same view selection process is conducted with all testing frames, and the downstream model prediction is directly used in the view selection score.

**Dataset.** We use publicly available datasets and comply with the original licenses and terms of use. The multi-view crowd counting task is conducted on CVCS [1] and CityStreet [16] datasets. The multi-view crowd localization task is conducted on Wildtrack [49] and MultiviewX [32] datasets.

- **CVCS** is a multi-scene dataset which includes 23 scenes for training and 8 scenes for testing, with 60-120 camera views in each scene with calibrations, which is challenging and suitable to *validate the cross-scene generalization of the proposed frameworks*.
- **CityStreet** is a single-scene real-world dataset and contains 3 camera views and 300 frames for training and 200 for testing.
- **Wildtrack** and **MultiviewX** are two single-scene datasets, where the same settings are used as in MVSelect [5]: 360 frames are all used for model training and 40 frames are for testing, without frame selection.

The label costs of each dataset are shown in Table II.

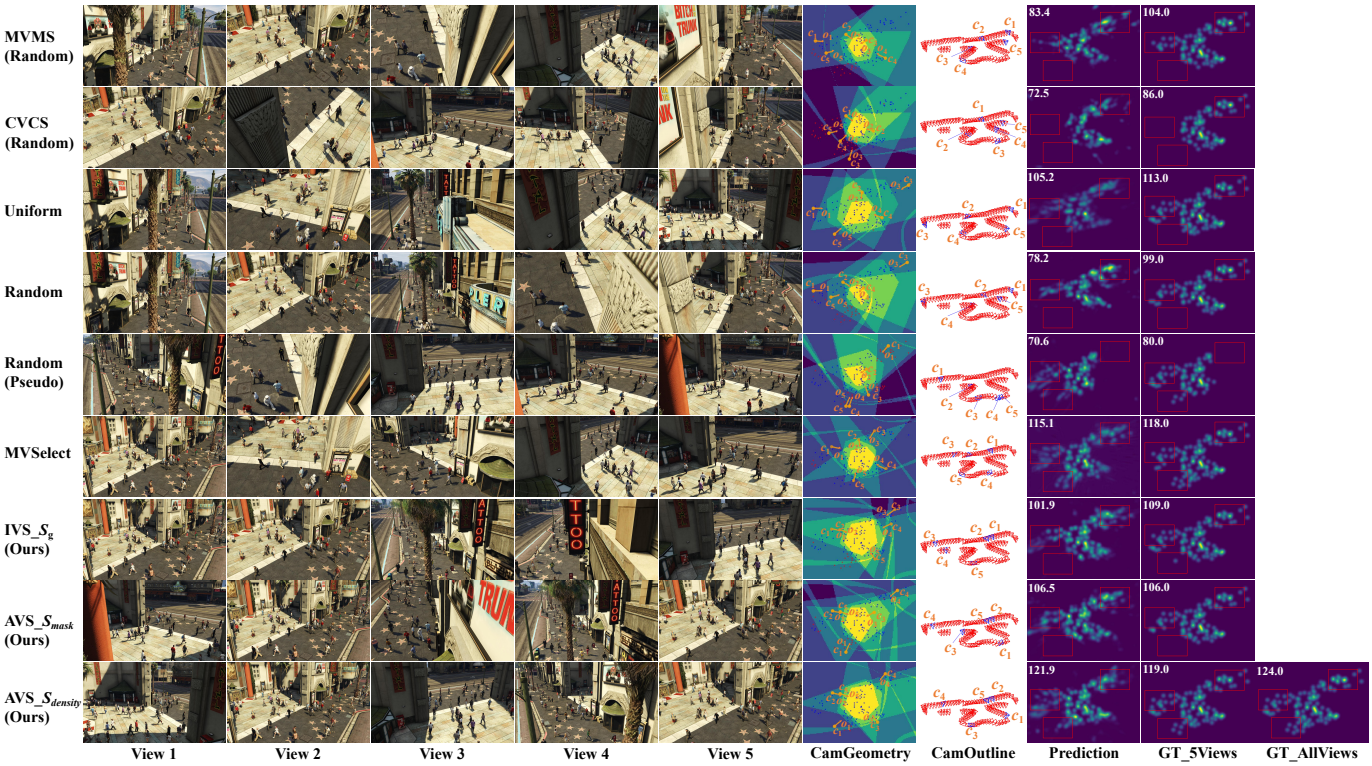



Fig. 6: The view selection and multi-view counting results on CVCS: our methods select better views covering the whole scene and predict better density maps close to the scene-level crowd GT (GT_AllViews). Blue camera view indicates the selected view in column 'CamOutline'.

TABLE V: The ablation study on the terms of the AVS score equation $S_{density}$ on CVCS dataset.

| Term | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| $S_{sc}$ | 16.44 | 21.01 | 0.125 |
| $S_{sc} * S_{ad}$ | 18.56 | 23.39 | 0.138 |
| $S_{sc} * S_{vd}$ | 14.77 | 18.31 | 0.108 |
| All ($S_{density}$) | **10.99** | **13.57** | **0.083** |

**Comparison methods.** For the *multi-view counting task*, we compare the proposed AVS with the IVS baseline, 'Random', 'Uniform', 'Random (Pseudo)', and MVSelect.

- 'Random' randomly selects $K$ views at once, trains on the selected views, and shares the same multi-view counting model architecture as ours.
- 'Uniform' uses the same multi-view counting model as ours, shares the same multi-view counting model architecture as ours, but replaces the view selection method with the uniform view sampling from all views.
- 'Random (Pseudo)' selects views in the 'Random' way but adopts pseudo-label training.
- MVSelect [5] is a view selection method based on the RL framework, which needs all the labels for model training.

We also compare with previous SOTAs CVCS [1] and MVMS [16] with the same labeling budget using the random selection way, denoted as 'CVCS (Random)' and 'MVMS (Random)'. Since MVSelect cannot be used in the cross-view cross-scene dataset (CVCS), we simply use the number of the prediction action in DQN greater than the maximum action of each scene, and then use a view mask to guarantee a valid action prediction for each scene.

For the *multi-view localization task*, we still compare the proposed AVS with the IVS baseline, 'Random', 'Uniform', 'Random (Pseudo)', and MVSelect. Moreover, we also compare with recent SOTAs trained with full labels, such as MVDet [32], SHOT [23], MVDeTr [24], 3DROM [25], and MMVOT [28]. These methods represent strong fully-supervised baselines with diverse modeling paradigms, providing a comprehensive benchmark for evaluation. Note that MVSelect uses *all* labels for joint view selection and crowd localization model training, while we only need to label the selected $K$ views. This setting significantly reduces annotation cost and better reflects practical deployment scenarios where full multi-view annotations are often unavailable or prohibitively expensive, while still maintaining competitive performance.

**Implementation details.** For the multi-view counting task, the input image resolution is 640x360 on CVCS and CityStreet, and a random 160x180 cropping strategy on the scene map is adopted in the training on CVCS. *For the view numbers and the values of $K$ and $F$ in each dataset, see Table II for details*. For AVS, during each view expansion iteration, an MAE threshold $\tau$ of 20 is adopted to stop the multi-view counting model training for conducting the next view addition. For the multi-view counting model, we use the backbone model in CVCS method with a feature pyramid fusion net (FPN) as the downstream multi-view counting model. The model is trained using the SGD optimizer with a learning rate of 1e-3. $\epsilon$ is 1e-10 and $\lambda$ is 0.1 in (3), and threshold $\sigma$ is the mean of density map $M_k$.

For the multi-view localization task, during the view selec-

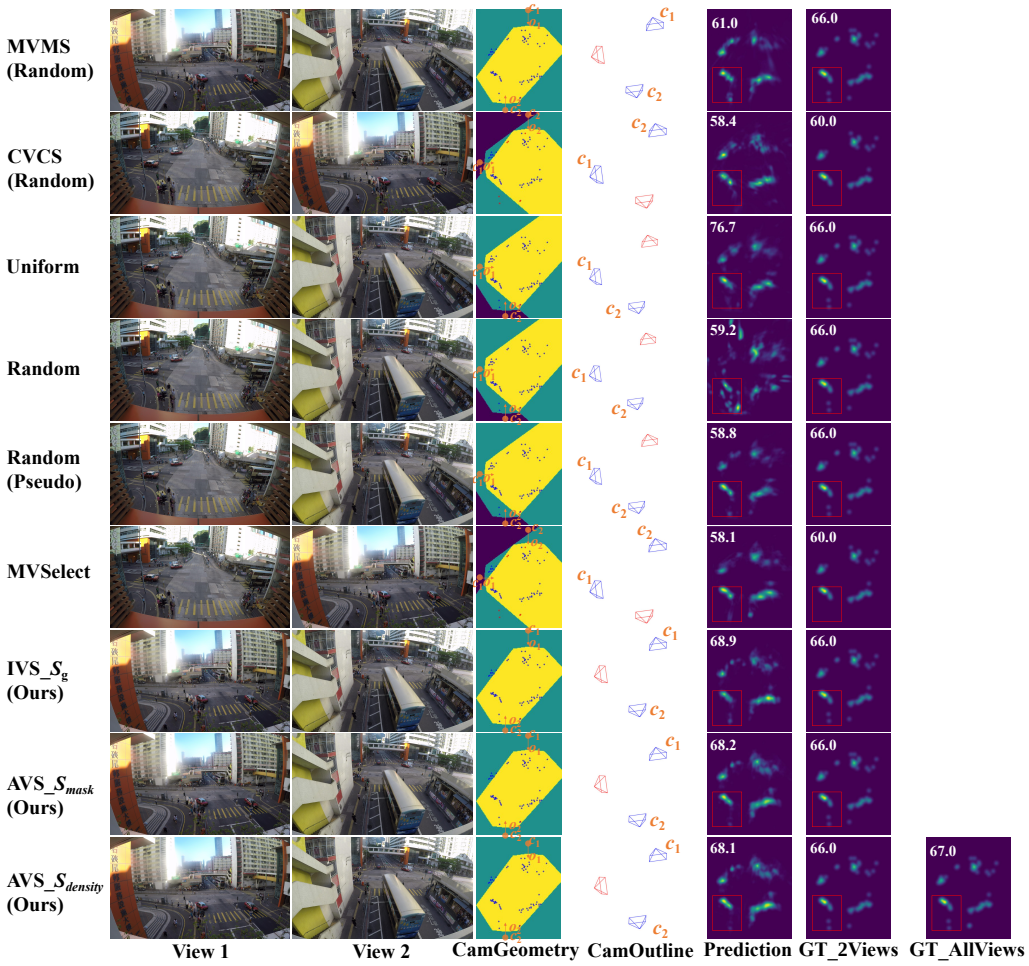



Fig. 7: The view selection and multi-view counting results on CityStreet. Since the CityStreet scene is small, the results of different view selection methods are similar. Our methods still achieve better scene-level prediction results.

tion process, the multi-view crowd localization model training threshold $\tau$ is set to MODA $= 40$: the model training stops when MODA reaches 40, and then we add the next view. This progressive strategy ensures that each newly introduced view contributes complementary information while avoiding unnecessary training overhead and redundancy among views. Unlike MVSelect, which uses annotations from all views for training, we label only the selected views and apply pseudo-labels to incorporate unlabeled views, thereby reducing annotation cost while maintaining competitive performance. We use the same MVDet implemented in MVSelect as the downstream multi-view localization model, trained with data augmentation and focal loss in MVDet. The model is trained using the SGD optimizer, with learning rates of 1e-2 and 5e-2 for Wildtrack and MultiviewX, respectively. $\sigma$ is 0.6 in (6), and $\lambda$ and $\epsilon$ are the same as in multi-view counting settings to ensure consistency across tasks.

**Evaluation metrics.** For the multi-view counting task, We use mean absolute error (MAE), root mean squared error (MSE), and normalized absolute error (NAE) of the predicted crowd count and the *scene-level* ground-truth count (all crowds in the scene) as metrics:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{gt} - \hat{y}_i\right|, \tag{8}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|y_i^{gt} - \hat{y}_i\right|^2}, \tag{9}$$

$$NAE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i^{gt} - \hat{y}_i\right|}{y_i^{gt}}, \tag{10}$$

where $N$ is the number of the samples, and $\hat{y}_i$ and $y_i^{gt}$ are the predicted count and the corresponding ground truth (GT) count of the $i$-th sample, respectively. In evaluation, the GT count refers to the crowd number in all views, not the crowd count of the selected views, to indicate the scene-level counting performance. Thus, the metrics not only assess the performance of the counting model but also reflect whether the selected views can adequately cover all crowds. The percentage of the crowds covered by the selected views among all crowds in the scene is used to evaluate different view

TABLE VI: The ablation study on $\lambda$ of $S_{density}$ on the CVCS dataset.

| $\lambda$ | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| 0.05 | 12.88 | 16.24 | 0.096 |
| 0.1 | **10.99** | **13.57** | **0.083** |
| 0.5 | 15.07 | 18.18 | 0.112 |
| 1 | 12.74 | 15.87 | 0.093 |

TABLE VII: The ablation study on the threshold $\tau$ to conduct view addition on the CVCS dataset..

| $\tau$ | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| 15 | 11.47 | 13.99 | 0.086 |
| 20 | **10.99** | **13.57** | **0.083** |
| 30 | 12.79 | 15.66 | 0.096 |

TABLE VIII: The ablation study on the terms of the independent view selection score equation $S_g$ on the CVCS dataset.

| Term | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| $S_{sc}$ | 21.65 | 27.46 | 0.159 |
| $S_{sc} * S_{ad}$ | 19.62 | 25.09 | 0.144 |
| $S_{sc} * S_{vd}$ | 19.45 | 24.68 | 0.143 |
| All ($S_g$) | **14.81** | **18.47** | **0.110** |

TABLE IX: The ablation study on when to add pseudo label training for AVS_$S_{density}$ (Ours) on CVCS dataset.

| Pseudo | MAE ↓ | MSE ↓ | NAE ↓ |
|---|---|---|---|
| None | 20.17 | 24.77 | 0.156 |
| ViewSel | 19.89 | 24.62 | 0.154 |
| ModelTrain | 11.32 | 14.69 | **0.083** |
| Both (Ours) | **10.99** | **13.57** | **0.083** |

selection methods, and is denoted as 'CoverRate':

$$CoverRate = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i^{gt5}}{y_i^{gt}}, \tag{11}$$

where $y_i^{gt5}$ and $y_i^{gt}$ denote the crowd number in the selected views and all views, respectively. 'CoverRate' could indicate the crowd coverage performance of the selected views.

For the multi-view localization task, we use Multiple Object Detection Accuracy (MODA, MA.), Multiple Object Detection Precision (MODP, MP.), Precision (P), Recall (R), and F1_score (F1) as metrics. Specifically, $MODA = 1-(FP+FN)/(TP+FN)$ measures the overall performance. $MODP = (\sum(1-d[d<t]/t))/TP$ measures the localization precision, where $d$ is the distance from a detected person point to its ground truth and $t$ is the distance threshold set to 0.5m as in [32]. $P = TP/(FP+TP)$, $R = TP/(TP+FN)$, and $F1 = 2P*R/(P+R)$. $TP$, $FP$, and $FN$ are the number of true positives, false positives, and false negatives, respectively.

### *B. Multi-view Crowd Counting Results*

We compare the proposed AVS with the IVS baseline and other comparison methods in Table III on the CVCS dataset, and AVS achieves the best performance. 'MVMS (Random)' and 'CVCS (Random)' achieve much worse results because they input random views without good view selection for scene-level counting. Since the efficiency of the pseudo label training, the results of 'Random (Pseudo)' are better than 'Random'. Compared to 'Random' view selection, 'Uniform' achieves better performance. But our methods consider view/scene geometries and interaction with downstream models, achieving the best results. Compared to MVSelect, which uses all labels for training, our methods still achieve better results, further demonstrating the effectiveness of the proposed methods. **Compared to IVS**, AVS is much better, either with $S_{mask}$ or $S_{density}$. Even though IVS_$S_g$ has a close CoverRate to AVS_$S_{density}$, its scene-level counting performance is much worse than AVS_$S_{density}$. This shows the superiority of AVS, which optimizes the view selection and the multi-view counting model training jointly for better scene-level results. $S_{density}$ is better than $S_{mask}$ because $S_{density}$ considers the crowd density levels as well as the location information in view selection, while $S_{mask}$ only utilizes the location information. Note that the CVCS dataset is a multi-scene dataset, and our methods could perform *cross-scene training and testing*, demonstrating the flexibility and generalization ability. We also compare the proposed view selection methods with the comparison methods on CityStreet in Table IV. The table shows that results with view selection methods are better than those without. However, the proposed methods overall achieve the best results, further indicating that equipping a well-designed view selection method is essential for scene-level tasks with limited labels.

The **visualizations** on CVCS are shown in Fig. 6, where the inputs are variable for different methods. 'GT_$K$Views' and 'GT_AllViews' are the GT constructed from the $K$ selected views and all views. The former is used for training, and the

TABLE X: The ablation on the fixed labeled view number in the pseudo inputs after view selection on the CVCS dataset.

| # Fixed view | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| 0 | 11.07 | 14.27 | **0.083** |
| 1 | **10.99** | **13.57** | **0.083** |
| 2 | 11.09 | 13.83 | 0.084 |
| 3 | 12.15 | 15.82 | 0.091 |
| 4 | 12.10 | 14.91 | 0.090 |

TABLE XI: The ablation on the ratio between labeled inputs and pseudo inputs after view selection on the CVCS dataset.

| Ratio | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| Only Labeled | 12.65 | 15.35 | 0.096 |
| 1:1 | 10.99 | 13.57 | 0.083 |
| 1:2 | **10.85** | **13.56** | **0.080** |

TABLE XII: The ablation study on the selected view/frame number $K$ (keep $F=20$)/$F$ (keep $K=5$) on CVCS dataset.

| $K$ | MAE | MSE | NAE | $F$ | MAE | MSE | NAE |
|---|---|---|---|---|---|---|---|
| 3 | 15.95 | 20.13 | 0.118 | 5 | 19.01 | 22.91 | 0.144 |
| 5 | 10.99 | 13.57 | 0.083 | 10 | 11.69 | 14.56 | 0.088 |
| 7 | 10.57 | 13.04 | 0.079 | 20 | 10.99 | 13.57 | 0.083 |
| 9 | **9.82** | **12.24** | **0.072** | 40 | **10.31** | **12.87** | **0.077** |

TABLE XIII: The ablation study on the test of different selected view numbers $K$ using the AVS_$S_{density}$ model trained with $F = 20$ and $K = 5$ on the CVCS dataset.

| $K$ | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| 3 | 15.06 | 18.88 | 0.112 |
| 5 | 10.99 | 13.57 | 0.083 |
| 7 | **10.53** | **13.09** | **0.080** |

TABLE XIV: The ablation study on the multi-view counting models using AVS_$S_{density}$ on the CVCS dataset.

| Model | MAE ↓ | MSE ↓ | NAE ↓ |
|---|---|---|---|
| Backbone | 11.20 | 14.70 | 0.083 |
| Backbone+CVCS | 15.25 | 18.63 | 0.113 |
| Backbone+MVMS | **9.74** | **13.31** | **0.074** |
| Backbone+FPN | 10.99 | 13.57 | 0.083 |

TABLE XV: The ablation study on the frame number used for view selection when testing AVS_$S_{density}$ on the CVCS dataset. Note that the counting model still tests on all frames (100) for the result report.

| Frames | MAE↓ | MSE↓ | NAE↓ | Time (h)↓ |
|---|---|---|---|---|
| 5 | 11.43 | 14.20 | 0.085 | **2.0** |
| 20 | 10.97 | 13.60 | **0.082** | 7.1 |
| 50 | **10.91** | **13.54** | **0.082** | 15.6 |
| 100 | 10.99 | 13.57 | 0.083 | 30.7 |

latter is for evaluation. It's observed that the 'GT_$K$Views' of our method 'AVS_$S_{density}$' contains the most crowds, and our method can also cover more crowds in the scene (red dots in 'CamGeometry' indicate crowds not covered by the selected views), indicating the efficacy of our view selection method. The predictions further highlight the advantages of our method, whereas the comparison methods fail to capture the regions highlighted by the red boxes. Fig. 7 shows the visualizations on CityStreet. Because the CityStreet scene is small, the GT between selected views and the scene is similar. The figure shows that our methods can select more adaptive views, enabling the counting model to achieve better scene-level performance, thereby demonstrating the advantages of the proposed methods.

### C. Ablation Study on Multi-view Crowd Counting

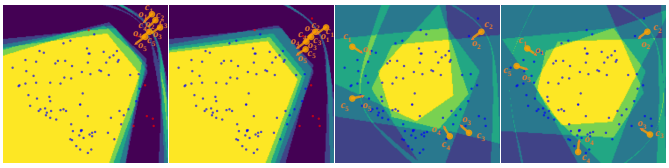
To better validate the efficiency of our proposed methods, the experiments are mainly conducted on CVCS dataset. The results are as follows.

**Ablation on the terms of $S_{density}$.** We perform ablation studies on the usage of the 3 terms in $S_{density}$ in Table V: using $S_{sc}$, using $S_{sc} * S_{ad}$ or $S_{sc} * S_{vd}$, or using all 3 terms ($S_{density}$). Compared to only using $S_{sc}$, adding $S_{vd}$ improves the results, while adding $S_{ad}$ without the view diversity term $S_{vd}$ achieves worse results, due to selected views being placed at the similar locations and directions, reducing the multi-view fusion performance (as shown in Fig. 8 (b)). Using all 3 terms is the best because it can select views covering most of the crowd with a larger overlapping area for better multi-view fusion (see Fig. 8 (d)), which indicates each term's contribution to the final view selection performance.

**Ablation study on $\lambda$ in $S_{density}$.** We conduct experiments on the hyper-parameter $\lambda$ in $S_{density}$ to evaluate the sensitivity of the term $S_{vd}$. As shown in Table VI, when $\lambda$ is too small, the view diversity constraint becomes weak, which may lead to suboptimal view sets (see Fig. 8 (a) and (b)). In contrast, excessively large $\lambda$ values $(0.5, 1)$ overemphasize view diversity, diminishing the contributions of $S_{sc}$ and $S_{ad}$ and resulting in inferior performance. Therefore, compared with other settings, $\lambda = 0.1$, which provides a balanced sensitivity, is more suitable for the proposed scoring function.

TABLE XVI: The ablation studies on the random frame selection method on the CVCS dataset ($F = 20$).

| Frame Select | MAE↓ | MSE↓ | NAE↓ |
|---|---|---|---|
| Random/Uniform | 12.28 | 15.46 | 0.092 |
| AVS (Ours) | **10.99** | **13.57** | **0.083** |



**(a) $S_{sc}$** **(b) $S_{sc}$* $S_{ad}$** **(c) $S_{sc}$* $S_{vd}$** **(d) All (Ours)**

Fig. 8: The selected view positions ($c_j$) and directions ($o_j$) for different terms. Red dots are uncovered crowds.

**Ablation study on the threshold $\tau$.** We experiment with the downstream model performance threshold $\tau$ for view addition. As shown in Table VII, $\tau = 20$ is more suitable for our AVS framework. When $\tau$ is too large, the counting model is under training, which may result in bad final performance; When $\tau$ is too small, the counting model may overfit at each step of adding views, leading to poor final performance. $\tau = 20$ achieves a balance between the current and final counting model performance.

**Ablation study on the terms of $S_g$.** In addition to the term ablation study for $S_{density}$ of AVS framework in the manuscript, we conduct additional ablation experiments on the 3 terms in $S_g$ of the IVS framework in Table VIII: using $S_{sc}$, $S_{sc} * S_{ad}$, $S_{sc} * S_{vd}$, or using all 3 terms (namely $S_g$). The results are similar to the experiments on $S_{density}$ and demonstrate that each term in $S_g$ contributes to the final performance by leveraging the scene and view geometries.

**Ablation study on pseudo labels.** The ablation studies on the pseudo labels for AVS_$S_{density}$ are shown in Table IX. We compare the method without using the pseudo labels in the model training (None), using pseudo labels only at the view selection stage (ViewSel), using pseudo labels only at the final model training stage when view selection is finished (ModelTrain), or using pseudo labels at both stages (Ours). The results show that pseudo labels can indeed improve the performance when added at any stage, and adding pseudo labels at both stages can obtain the best performance.

**Ablation study on the number of labeled views for pseudo inputs.** We conduct experiments on the number of labeled views in pseudo inputs after view addition. Specifically, we vary the proportion of labeled views used to construct pseudo inputs to analyze their impact on model performance. With more labeled views in pseudo inputs, the randomness of the pseudo inputs is reduced, leading to less diverse training signals and consequently worse generalization ability, as shown in Table X. This observation suggests that excessive reliance on labeled views may cause the model to overfit to limited view configurations. Hence, we only retain one labeled view in pseudo inputs, which preserves sufficient randomness and diversity, achieving better performance and improved robustness across different view combinations.

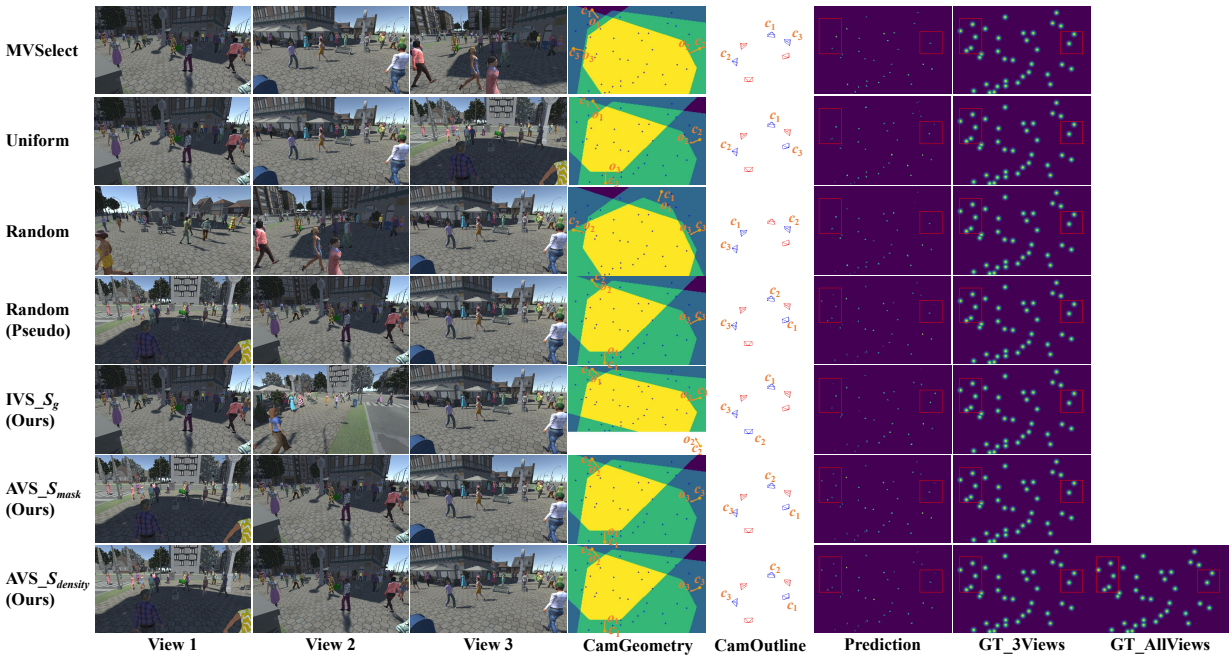



Fig. 9: The view selection and multi-view localization results on MultiviewX. Our methods select views that achieve better prediction results. Blue camera view indicates the selected view in column 'CamOutline'.

TABLE XVII: The ablation on the comparison method with uniform view selection and pseudo-labels on CVCS dataset.

| Method | MAE ↓ | MSE ↓ | NAE ↓ | CoverRate ↑ |
|---|---|---|---|---|
| MVMS (Random) | 37.22 | 44.74 | 0.276 | 0.872 |
| MVMS (Random, Pseu.) | 31.11 | 38.29 | 0.229 | 0.872 |
| MVMS (Uniform) | 30.38 | 36.42 | 0.225 | 0.945 |
| MVMS (Uniform, Pseu.) | 23.37 | 28.49 | 0.173 | 0.945 |
| CVCS (Random) | 30.97 | 36.47 | 0.228 | 0.901 |
| CVCS (Random, Pseu.) | 24.89 | 30.54 | 0.182 | 0.901 |
| CVCS (Uniform) | 28.06 | 32.91 | 0.209 | 0.945 |
| CVCS (Uniform, Pseu.) | 22.28 | 27.97 | 0.164 | 0.945 |
| Random | 36.59 | 42.06 | 0.271 | 0.885 |
| Random (Pseu.) | 28.22 | 33.73 | 0.208 | 0.885 |
| Uniform | 21.76 | 25.75 | 0.163 | 0.945 |
| Uniform (Pseu.) | 15.69 | 19.92 | 0.115 | 0.945 |
| IVS_$S_g$ (Baseline, Ours) | 14.81 | 18.47 | 0.110 | 0.954 |
| AVS_$S_{mask}$ (Ours) | 12.53 | 15.33 | 0.093 | 0.955 |
| AVS_$S_{density}$ (Ours) | **10.99** | **13.57** | **0.083** | **0.960** |

**Ablation study on the ratio pseudo input.** We conduct experiments on the ratio between pseudo inputs and labeled inputs. Generally speaking, using more pseudo inputs can improve the model's robustness. However, because the labels are imperfect, we cannot rely solely on the pseudo inputs for model training. As shown in Table XI, the ratio of 1:2 between labeled inputs and pseudo inputs achieves the best result, but we use the 1:1 ratio between labeled inputs and pseudo inputs for time tradeoff.

**Different view and frame number.** We conduct ablation studies on the selected view number $K$ and frame number $F$ for AVS_$S_{density}$ in Table XII, with other settings kept the same (except $\tau = 30$ for 5 frames for its poor performance). As $K$ increases, more views are provided to cover the whole scene, generally achieving better scene-level counting performance. With more frames, the multi-view counting model is trained on more labeled data, yielding better results.

**Ablation study on the testing view number.** We conduct extra experiments on the testing with different view numbers $K$ (3, 5, and 7) using the AVS_$S_{density}$ model trained with $K = 5$ views. As shown in Table XIII, with the view number increasing, the model's performance improves relatively, as more regions are covered, demonstrating the good generalization ability of the proposed AVS framework to variable numbers of input views.

**Ablation on the different multi-view counting models.** The ablation studies on the multi-view counting models in the active view selection framework AVS_$S_{density}$ are shown in Table XIV. The 'Backbone' model is the same backbone model of CVCS [1]. '+MVMS','+CVCS', and '+FPN' mean adding the multi-view multi-scale selection [16], camera view selection [1], and the feature pyramid fusion module [50] to the 'Backbone' model, respectively. From the table, using 'Backbone+MVMS' achieves the best multi-view counting results. 'Backbone+CVCS' achieves worse results than 'Backbone', perhaps due to the requirement of more labeled data to learn a good camera selection module, whereas our task setting provides limited labeled data. We adopt the 'Backbone+FPN' model as the multi-view counting model in our experiments to balance training efficiency and performance.

**Ablation study on the frame number for view selection and computation time in testing on the CVCS dataset.** We conduct experiments on using different frame numbers in view selection when testing AVS_$S_{density}$ trained with labeled $F = 20$ frames and $K = 5$ views. Table XV indicates that only using 20 frames at testing to conduct view selection can achieve almost similar results compared with using 50 frames or 100 frames, with much less computation time, though. This demonstrates that the adopted frame initialization approach can effectively select representative frames for view selection

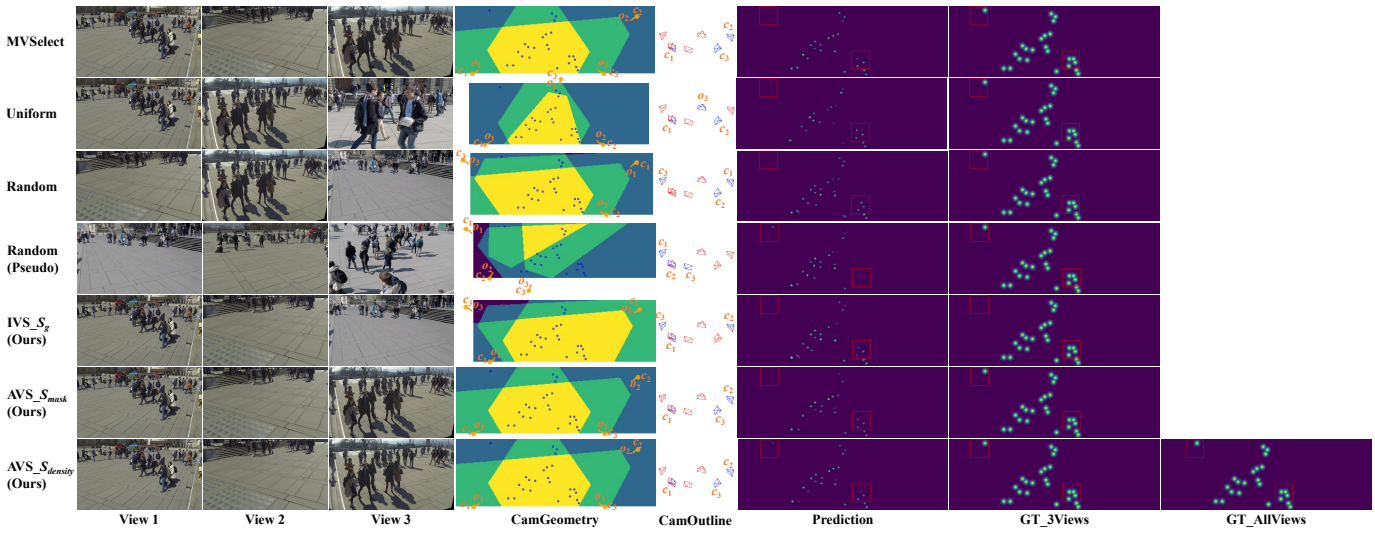



Fig. 10: The view selection and multi-view localization results on Wildtrack. Our methods still select views that achieve better prediction results, further demonstrating the efficiency of the proposed method.

TABLE XVIII: Comparison of the multi-view localization results on Wildtrack and MultiviewX. AVS achieves the best performance among all partial-labeled methods (3 views) and outperforms MVSelect trained with the ground truth of full labels (all views). Bold indicates the best result among all partial-labeled methods.

| | Dataset | MultiviewX | | | | | Wildtrack | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Label | Method | MODA↑ | MODP↑ | P↑ | R↑ | F1↑ | MODA↑ | MODP↑ | P↑ | R↑ | F1↑ |
| Full | MVDet [32] | 83.9 | 79.6 | 96.8 | 86.7 | 91.5 | 88.2 | 75.7 | 94.7 | 93.6 | 94.1 |
| | SHOT [23] | 88.3 | 82.0 | 96.6 | 91.5 | 94.0 | 90.2 | 76.5 | 96.1 | 94.0 | 95.0 |
| | MVDeTr [24] | 93.7 | 91.3 | 99.5 | 94.2 | 97.8 | 91.5 | 82.1 | 97.4 | 94.0 | 95.7 |
| | 3DROM [25] | 95.0 | 84.9 | 99.0 | 96.1 | 97.5 | 93.5 | 75.9 | 97.2 | 96.2 | 96.7 |
| | M-MVOT [28] | 96.7 | 86.1 | 98.8 | 97.9 | 98.3 | 92.1 | 81.3 | 94.5 | 97.8 | 96.1 |
| | MVSelect [5] | 88.1 | 89.8 | 98.2 | 89.7 | 93.8 | 88.6 | 79.9 | 93.3 | 94.2 | 93.7 |
| Partial | Random | 85.3 | 80.8 | 97.3 | 87.7 | 92.2 | 80.6 | 75.8 | 93.0 | 87.1 | 89.8 |
| | Random (Pseudo) | 85.5 | 81.1 | 97.5 | 87.7 | 92.4 | 82.8 | 75.4 | 93.8 | 88.5 | 91.0 |
| | Uniform | 82.6 | **87.3** | 96.4 | 85.8 | 90.8 | 84.6 | **79.7** | 95.1 | 89.2 | 92.0 |
| | IVS_$S_g$ (Ours) | 87.1 | 83.3 | 97.5 | 89.4 | 93.3 | 87.4 | 75.7 | 93.2 | **94.2** | 93.7 |
| | AVS_$S_{mask}$ (Ours) | 87.9 | 80.5 | 97.3 | 90.4 | 93.7 | 87.7 | 77.0 | 95.5 | 92.0 | 93.7 |
| | AVS_$S_{density}$ (Ours) | **89.2** | 82.1 | **98.0** | **91.0** | **94.4** | **89.6** | 76.7 | **96.1** | 93.4 | **94.7** |

and reduce testing computation time. Note that the counting model still tests on all frames (100) for the performance report.

**Different frame selection methods.** We conduct an additional random frame selection method, and our proposed frame selection method outperforms the random or uniform frame selection shown in Table XVI. The reason is that our method selects frames across various lighting conditions using cosine similarity, thereby enhancing the sample diversity and the network's robustness. For uniform frame sampling, because of unordered frames on CVCS dataset, random frame sampling is equivalent to uniform frame sampling.

**Ablation study on comparison method with uniform view selection and pseudo label.** As shown in Table XVII, uniform view selection consistently achieves better scene-level performance than random selection, highlighting the importance of a reasonable view selection method. However, our methods further outperform the uniform strategy, indicating that explicitly modeling view/scene geometries as well as crowd density and localization information enables the selection of views that are more suitable for accurate prediction. In addition, we incorporate pseudo-labels (Pseu.) into the comparison methods. The results show that introducing pseudo-labels consistently improves their performance, demonstrating the effectiveness of this strategy. Nevertheless, our methods still achieve superior results, underscoring that a well-designed, geometry- and density-aware view selection plays a more critical role in overall performance improvement.

### *D. Multi-view Crowd Localization Results*

As shown in Table XVIII, we compare our methods with other view-selection-based methods (MVSelect, 'Random', and 'Random (Pseudo)', 'Uniform') and full-label supervised methods. The results show that AVS_$S_{density}$ outperforms the random view selection methods 'Random', 'Random (Pseudo)', 'Uniform', and MVSelect, demonstrating the advantages of using joint optimization of the view selection and downstream model training. Compared to IVS, AVS achieves better performance, either with $S_{mask}$ or $S_{density}$. $S_{density}$ is better than $S_{mask}$, which also proves AVS's effectiveness due to considering both crowd density-level information and location information, and the view/scene geometries in the view selection. Compared to MVDet, SHOT, MVDeTr, 3DROM, and M-MVOT, which are trained on all input views and labels, the proposed active view selection framework

TABLE XIX: The ablation study on the pseudo labels.

| Dataset | MultiviewX | | | | | Wildtrack | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pseudo | MA. | MP. | P | R | F1 | MA. | MP. | P | R | F1 |
| None | 86.2 | 73.0 | **98.4** | 87.6 | 92.7 | 79.6 | **77.6** | 94.2 | 84.9 | 89.3 |
| ViewSel | 86.9 | 79.8 | 97.6 | 89.1 | 93.1 | 83.6 | 75.9 | 94.7 | 88.6 | 91.5 |
| ModelTrain | 87.8 | 81.6 | 98.0 | 89.7 | 93.6 | 79.9 | 77.5 | 95.8 | 83.6 | 89.3 |
| Both (Ours) | **89.2** | **82.1** | 98.0 | **91.0** | **94.4** | **89.6** | 76.7 | **96.1** | **93.4** | **94.7** |

TABLE XX: The ablation study on the threshold $\tau$ to conduct view addition on the MultiviewX and Wildtrack.

| Dataset | MultiviewX | | | | | Wildtrack | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | MA. | MP. | P | R | F1 | MA. | MP. | P | R | F1 |
| 30 | 87.5 | 81.3 | 97.3 | 90.0 | 93.5 | 85.3 | 76.3 | 96.0 | 89.0 | 92.4 |
| 40 | **89.2** | **82.1** | 98.0 | **91.0** | **94.4** | **89.6** | **76.7** | 96.1 | **93.4** | **94.7** |
| 50 | 87.4 | 78.3 | **98.1** | 89.1 | 93.4 | 87.4 | 76.1 | **96.4** | 90.8 | 93.5 |
| 60 | 87.0 | **82.1** | 97.9 | 88.9 | 93.2 | 86.9 | 73.6 | 96.0 | 90.7 | 93.2 |

($S_{density}$) outperforms MVDet on both MultiviewX and Wildtrack with limited label costs, also proving the advantages of our methods. Note that MVSelect also relies on all camera view labels (annotations and calibrations) in the model training and cannot perform on novel new scenes with different views and scene settings, while *our methods only rely on limited view labels with wider application scenarios (as on CVCS)*.

We visualize the results from the comparison methods and the proposed methods on MultiviewX and Wildtrack in Fig. 9 and 10. Because both datasets are smaller than CVCS, a few views can easily cover the crowd on the ground. Our method achieves better scene-level localization performance than the comparisons, as shown in the red box regions, where our methods have fewer missing points. Furthermore, our AVS framework jointly optimizes view selection and multi-view localization model, and employs novel pseudo-labels during model training to improve localization performance.

### *E. Ablation Study on Multi-view Crowd Localization*

**Ablation study on pseudo labels.** The ablation studies on the pseudo labels for the active view selection framework ($S_{density}$) are shown in Table XIX: no pseudo labels (None), adding at view selection (ViewSel) or final model training stage (ModelTrain) after view selection, or both (Ours). Similarly, we add the pseudo-label training at different stages and compare their influence on the performance. The results show that regardless of which stage, pseudo labels can improve the performance. On Wildtrack, adding pseudo labels is more effective at the view selection stage. The possible reason is that the view difference is larger in Wildtrack, and thus the pseudo-label training is more useful for the model to generalize to new views. Anyway, adding pseudo-label training at both stages can achieve the best performance.

**Ablation study on threshold $\tau$.** As a hyperparameter, the threshold $\tau$ controls when to select the next view. For the multi-view crowd localization task on Wildtrack and MultiviewX, AVS framework selects the next view when the MODA during training exceeds the threshold $\tau$. AVS framework jointly optimizes view selection and downstream task models. Hence, a well-trained model can produce a more accurate density map, thereby reducing prediction errors during view selection. Consequently, we need an appropriate model to select views by controlling the threshold $\tau$. As shown in Table XX, threshold $\tau = 40$ is more suitable on Wildtrack and MultiviewX, achieving the better scene-level performance.

TABLE XXI: The ablation study on the training frame number $F$ on the MultiviewX dataset.

| Method | AVS_$S_{density}$ | | | | | MVSelect | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Frame | MA. | MP. | P | R | F1 | MA. | MP. | P | R | F1 |
| 36 | 73.6 | 76.4 | 95.8 | 77.0 | 85.4 | 61.5 | 73.4 | 96.9 | 63.5 | 76.7 |
| 72 | 81.2 | 71.6 | 95.9 | 84.8 | 90.0 | 69.8 | 51.2 | 96.6 | 72.4 | 82.8 |
| 180 | 86.0 | 80.8 | 96.7 | 89.1 | 92.7 | 77.6 | 60.0 | 97.1 | 80.0 | 87.7 |
| 360 | **89.2** | **82.1** | **98.0** | **91.0** | **94.4** | **88.1** | **89.8** | **98.2** | **89.7** | **93.8** |

TABLE XXII: The ablation study on the selected view number $K$ on the MultiviewX dataset, training and testing using the same $K$ views.

| Method | AVS_$S_{density}$ | | | | | MVSelect | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| View | MA. | MP. | P | R | F1 | MA. | MP. | P | R | F1 |
| 2 | 81.7 | 80.4 | 97.1 | 84.3 | 90.2 | 77.5 | 82.9 | 97.4 | 79.6 | 87.6 |
| 3 | 89.2 | **82.1** | 98.0 | 91.0 | 94.4 | 88.1 | **89.8** | **98.2** | 89.7 | 93.8 |
| 4 | 92.0 | 78.7 | 98.2 | 93.7 | 95.9 | 89.6 | 87.9 | 97.5 | 92.0 | 94.7 |
| 5 | **93.4** | 79.2 | **98.3** | **95.1** | **96.6** | **92.3** | 89.1 | 97.8 | **94.5** | **96.1** |

**Comparison of training with different frame numbers.** We compare the proposed AVS_$S_{density}$ and MVSelect by training with different numbers of frames (36, 72, 180, and 360) on MultiviewX in Table XXI. The testing set is the same (40 frames). As the number of training frames increases, performance improves, and our method outperforms MVSelect across a range of frame counts, demonstrating its efficiency relative to MVSelect.

**Comparison of training and testing with different view numbers.** We conduct experiments on the proposed method and MVSelect, training, or testing with different view numbers. For the results shown in Table XXII, the view number used during both the training and testing processes is the same. As the number of views increases, the performance gradually becomes better, and it is nearly converged when $K \geq 3$. Furthermore, regardless of the number of views used, our method still outperforms MVSelect, indicating the advantage of the proposed method. The results shown in Table XXIII use the model trained on 3 views to test with different view numbers, $K = 2, 3, 4, 5$. Our method generally outperforms MVSelect on all view numbers, showing the proposed method's generalization ability to different input view numbers. Note that MVSelect utilizes labels from all views for training and is challenging to apply to new scenes due to the reinforcement learning framework.

**Comparison of frame selection methods.** Compared with the uniform and random frame selection methods in Table XXIV, our frame selection approach is still better. As mentioned above, our method selects frames with various lighting conditions using cosine similarity, which can enhance the diversity of the sample and the robustness of the network.

**Comparison of the model costs.** As shown in Table XXV, due to using pseudo labels and model predictions for view selection, our training time is higher than MVSelect and Random, but comparable to others. Yet, our test speed is similar to baselines and faster than MVSelect since no extra reinforcement learning network is needed.

TABLE XXIII: The ablation study on the selected view number in testing on the MultiviewX dataset, training with 3 views and testing with $K$ (2, 3, 4, and 5) views.

| Method | AVS_$S_{density}$ | | | | | MVSelect | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| View | MA. | MP. | P | R | F1 | MA. | MP. | P | R | F1 |
| 2 | 82.8 | 80.4 | 97.0 | 85.5 | 90.9 | 77.1 | 83.4 | 96.5 | 80.0 | 87.5 |
| 3 | 89.2 | 82.1 | 98.0 | 91.0 | 94.4 | 88.1 | **89.8** | 98.2 | 89.7 | 93.8 |
| 4 | 92.7 | **82.6** | 98.5 | 94.2 | 96.3 | 91.2 | 87.8 | 98.3 | 92.9 | 95.5 |
| 5 | **93.4** | 82.4 | **98.6** | **94.7** | **96.6** | **93.1** | 88.9 | **98.6** | **94.5** | **96.5** |

TABLE XXIV: The ablation study on the same training frame number $F = 72$ with different frame selection methods. IDdiff consists of the selected frame ID mean and standard deviation.

| Dataset | Wildtrack | | | MultiviewX | | |
|---|---|---|---|---|---|---|
| Frame Select | MA. | F1 | IDdiff | MA. | F1 | IDdiff |
| Random | 77.5 | 87.8 | 24.7± 20.7 | 80.7 | 89.6 | 4.9±4.1 |
| Uniform | 77.0 | 87.5 | 25.0± 0.0 | 80.8 | 89.6 | 5.0±0.0 |
| AVS(Ours) | **79.2** | **88.9** | 18.0 ±52.3 | **81.2** | **90.0** | 5.0±6.0 |

## V. Conclusion

In this paper, we focus on the view selection issue for scene-level multi-view crowd counting and localization tasks. We first propose the independent view selection baseline (IVS), which considers view and scene geometries. Then, based on IVS, we propose the active view selection method (AVS), which incorporates downstream model predictions into view selection and jointly optimizes both view selection and downstream tasks. Extensive experiments on the two tasks demonstrate the advantages of the proposed AVS method over all comparison methods. The proposed method can be applied to novel scenes with limited labels, demonstrating its superior generalization and broader applicability. In the future, the method could also be extended to other BEV-based or 3D reconstruction tasks to reduce labeling costs.

## Acknowledgements

This work was supported in parts by NSFC (62202312), City University of Hong Kong Strategic Research (7005665), Shenzhen Science and Technology Program (KJZD20240903100022028), and Scientific Development Funds from Shenzhen University.

## References

[1] Q. Zhang, W. Lin, and A. B. Chan, "Cross-view cross-scene multi-view crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 557–567.

[2] W. Xiong, B. Zeng, Z. Hu, J. Guo, K. Xie, and H. Huang, "Aerial path planning for urban geometry and texture co-capture," *ACM Trans. Graph.*, vol. 44, no. 6, Dec. 2025. [Online]. Available: https://doi.org/10.1145/3763292

[3] M. Tang, N. Wang, Z. Xie, J. Hu, K. Xie, X. Guo, and H. Huang, "Aerial path online planning for urban scene updation," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, ser. SIGGRAPH Conference Papers '25. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3721238.3730639

[4] Z. Liang, L. Yang, Z. Liang, J. C. F. Chan, Z. Zhang, M. Wang, and J. C. Cheng, "Optimized uav view planning for high-quality 3d reconstruction of buildings using a modified sparrow search algorithm," *Adv. Eng. Inform.*, vol. 65, no. PD, May 2025. [Online]. Available: https://doi.org/10.1016/j.aei.2025.103344

TABLE XXV: Model cost comparison on MultiviewX.

| Method | Memory(GB) | FLOPs(G) | Train(s) | Test(s) |
|---|---|---|---|---|
| MVSelect | **16.879** | 532.200 | **1200** | 10 |
| Random | 17.594 | **530.703** | 1700 | **7** |
| Random (Pseudo) | 17.594 | **530.703** | 7600 | **7** |
| IVS_$S_g$ (Ours) | 17.594 | **530.703** | 8006 | **7** |
| AVS_$S_{mask}$ (Ours) | 17.594 | **530.703** | 9172 | **7** |
| AVS_$S_{density}$ (Ours) | 17.594 | **530.703** | 9194 | **7** |

[5] Y. Hou, S. Gould, and L. Zheng, "Learning to select views for efficient multi-view understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 20 135–20 144.

[6] D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 38–54.

[7] T. Han, L. Bai, L. Liu, and W. Ouyang, "Steerer: Resolving scale variations for counting and localization via selective inheritance learning," in *2023 IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 791–21 802.

[8] Z. Zhao, M. Shi, X. Zhao, and L. Li, "Active crowd counting with limited supervision," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 565–581.

[9] S. S. Savner and V. Kanhangad, "Crowd counting from limited labeled data using active learning," *IEEE Signal Processing Letters*, vol. 30, pp. 1662–1666, 2023.

[10] S. Zhang, W. Ke, S. Liu, X. Hong, and T. Zhang, "Boosting semi-supervised crowd counting with scale-based active learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 8681–8690.

[11] W. Lin, C. Zhao, and A. B. Chan, "Point-to-region loss for semi-supervised point-based crowd counting," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 363–29 373.

[12] J. Zhou, J. Zhang, and Y. Gui, "Crowd counting in domain generalization based on multi-scale attention and hierarchy level enhancement," *Scientific Reports*, vol. 15, no. 1, p. 155, 2025.

[13] M. Shang and X. Hong, "2d gaussians spatial transport for point-supervised density regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 11, 2026, pp. 8824–8832.

[14] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Scene invariant multi camera crowd counting," *Pattern Recognition Letters*, vol. 44, pp. 98–112, 2014.

[15] N. C. Tang, Y.-Y. Lin, M.-F. Weng, and H.-Y. M. Liao, "Cross-camera knowledge transfer for multiview people counting," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 80–93, 2015.

[16] Q. Zhang and A. B. Chan, "Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8297–8306.

[17] L. Zheng, Y. Li, and Y. Mu, "Learning factorized cross-view fusion for multi-view crowd counting," in *2021 IEEE International Conference on Multimedia and Expo*. IEEE, 2021, pp. 1–6.

[18] Q. Zhai, F. Yang, X. Li, G.-S. Xie, H. Cheng, and Z. Liu, "Co-communication graph convolutional network for multi-view crowd counting," *IEEE Transactions on Multimedia*, vol. 25, pp. 5813–5825, 2022.

[19] B. Li, D. Chen, and Q. Zhang, "Wscf-mvcc: Weakly-supervised calibration-free multi-view crowd counting," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2025, p. 83–98.

[20] H. Mo, X. Zhang, J. Tan, C. Yang, Q. Gu, B. Hang, and W. Ren, "Countformer: Multi-view crowd counting transformer," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LII*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 20–40.

[21] T. Chavdarova and F. Fleuret, "Deep multi-camera people detection," in *2017 16th IEEE International Conference on Machine Learning and Applications*. IEEE, 2017, pp. 848–853.

[22] P. Baqué, F. Fleuret, and P. Fua, "Deep occlusion reasoning for multi-camera multi-target detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 271–279.

[23] L. Song, J. Wu, M. Yang, Q. Zhang, Y. Li, and J. Yuan, "Stacked homography transformations for multi-view pedestrian detection," in

*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6029–6037.

[24] Y. Hou and L. Zheng, "Multiview detection with shadow transformer (and view-coherent data augmentation)," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1673–1682.

[25] R. Qiu, M. Xu, Y. Yan, J. S. Smith, and X. Yang, "3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 695–710.

[26] M. Liu, C. Zhu, S. Ren, and X.-C. Yin, "Unsupervised multi-view pedestrian detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1034–1042.

[27] Q. Zhang, Y. Gong, D. Chen, A. B. Chan, and H. Huang, "Multi-view people detection in large scenes via supervised view-wise contribution weighting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7242–7250.

[28] Q. Zhang, K. Zhang, A. B. Chan, and H. Huang, "Mahalanobis distance-based multi-view optimal transport for multi-view crowd localization," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IV*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 19–36.

[29] S. Aung, M.-C. Sagong, and J. Cho, "Multi-view pedestrian occupancy prediction with a novel synthetic dataset," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1782–1790.

[30] J. Ma, T. Wang, M. Liu, D. Ahmedt-Aristizabal, and C. Nguyen, "Dchm: Depth-consistent human modeling for multiview detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 7731–7740.

[31] R. Alturki, A. Hilton, and J.-Y. Guillemaut, "Enhanced multi-view pedestrian detection using probabilistic occupancy volume," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3377–3386.

[32] Y. Hou, L. Zheng, and S. Gould, "Multiview detection with feature perspective transformation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 1–18.

[33] A. E. Daryani, M. U. M. Bhutta, B. Hernandez, and H. Medeiros, "Camuvid: Calibration-free multi-view detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, June 2025, pp. 1220–1229.

[34] Z. Luo, Q. Yu, and Q. Zhang, "Robust single-view 3d object reconstruction with stable diffusion generation and farthest view selection," in *Image and Graphics - 13th International Conference, ICIG 2025, Xuzhou, China, October 31 - November 2, 2025, Proceedings, Part II*, ser. Lecture Notes in Computer Science, Z. Lin, L. Wang, Y. Jiang, X. Wang, S. Liao, S. Shan, R. Liu, J. Dong, and X. Yu, Eds. Springer, 2025, pp. 529–540. [Online]. Available: https://doi.org/10.1007/978-981-95-3393-0_43

[35] S. Majumder, T. Nagarajan, Z. Al-Halah, R. Pradhan, and K. Grauman, "Which viewpoint shows it best? language for weakly supervising view selection in multi-view instructional videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 016–29 028.

[36] L. Di Giammarino, B. Sun, G. Grisetti, M. Pollefeys, H. Blum, and D. Barath, "Learning where to look: Self-supervised viewpoint selection for active localization using geometrical information," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVI*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 188–205.

[37] S. Li, A. Guedon, C. Boittiaux, S. Chen, and V. Lepetit, "Nextbestpath: Efficient 3d mapping of unseen environments," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=7WaRh4gCXp

[38] S. Kiciroglu, H. Rhodin, S. N. Sinha, M. Salzmann, and P. Fua, "Activemocap: Optimized viewpoint selection for active human motion capture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020, pp. 100–109.

[39] S. Zheng, B. Zhou, R. Shao, B. Liu, S. Zhang, L. Nie, and Y. Liu, "Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 19 680–19 690.

[40] Y. Sun, Q. Huang, D.-Y. Hsiao, L. Guan, and G. Hua, "Learning view selection for 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 14 459–14 468.

[41] S. Ruan, Y. Dong, H. Su, J. Peng, N. Chen, and X. Wei, "Towards viewpoint-invariant visual recognition via adversarial training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 4686–4696.

[42] R. Du, W. Yu, H. Wang, T.-E. Lin, D. Chang, and Z. Ma, "Multi-view active fine-grained visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 1568–1578.

[43] Y. Liu, L. Lin, Y. Hu, K. Xie, C.-W. Fu, H. Zhang, and H. Huang, "Learning reconstructability for drone aerial path planning," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 197:1–197:17, 2022.

[44] A. Hamdi, S. Giancola, and B. Ghanem, "Mvtn: Multi-view transformation network for 3d shape recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1–11.

[45] W. Xiao, R. S. Cruz, D. Ahmedt-Aristizabal, O. Salvado, C. Fookes, and L. Lebrat, "Nerf director: Revisiting view selection in neural volume rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 20 742–20 751.

[46] Q. Zhang, Z. Luo, T. Yu, and H. Huang, "View transformation robustness for multi-view 3d object reconstruction with reconstruction error-guided view selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 076–10 084.

[47] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," *Advances in neural information processing systems*, vol. 33, pp. 1595–1607, 2020.

[48] X. Zhou, K. Xie, K. Huang, Y. Liu, Y. Zhou, M. Gong, and H. Huang, "Offsite aerial path planning for efficient urban scene reconstruction," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 192:1–192:16, 2020.

[49] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5030–5039.

[50] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp. 936–944.

**Qi Zhang** received the Ph.D. degree in Computer Science from the City University of Hong Kong, Hong Kong SAR, China, in 2021. He is currently a Tenured Associate Professor at the Visual Computing Research Center, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include crowd analysis and simulation, 3D reconstruction and generation, autonomous driving, and urban scene point cloud analysis.

**Bin Li** received the B.E. degree in Software Engineering from Jiangxi University of Science and Technology in 2022, China. He is currently pursuing the M.S. degree at the Visual Computing Research Center, College of Computer Science and Software Engineering, Shenzhen University. His research interests include 3D vision and crowd analysis.

**Antoni B. Chan** is currently a Full Professor in the Department of Computer Science, City University of Hong Kong, and is also the Associate Dean of the College of Computing, City University of Hong Kong, Hong Kong SAR, China. His research interests include computer vision, machine learning, pattern recognition, and music analysis.

**Hui Huang** is a Distinguished TFA Professor of Shenzhen University, where she directs the Visual Computing Research Center and is also the Dean of the College of Computer Science and Software Engineering, Shenzhen University, China. Her research interests span computer graphics, 3D vision, and visualization.