

Cross-View Cross-Scene Multi-View Crowd Counting

Qi Zhang¹, Wei Lin², Antoni B. Chan¹

¹ Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China

{qzhang364-c@my., abchan@}cityu.edu.hk

²School of Computer Science and School of Artificial Intelligence,
Northwestern Polytechnical University, Xi'an, Shaanxi, China.

elonlin24@gmail.com

Abstract

Multi-view crowd counting has been previously proposed to utilize multi-cameras to extend the field-of-view of a single camera, capturing more people in the scene, and improve counting performance for occluded people or those in low resolution. However, the current multi-view paradigm trains and tests on the same single scene and camera-views, which limits its practical application. In this paper, we propose a cross-view cross-scene (CVCS) multi-view crowd counting paradigm, where the training and testing occur on different scenes with arbitrary camera layouts. To dynamically handle the challenge of optimal view fusion under scene and camera layout change and non-correspondence noise due to camera calibration errors or erroneous features, we propose a CVCS model that attentively selects and fuses multiple views together using camera layout geometry, and a noise view regularization method to train the model to handle non-correspondence errors. We also generate a large synthetic multi-camera crowd counting dataset with a large number of scenes and camera views to capture many possible variations, which avoids the difficulty of collecting and annotating such a large real dataset. We then test our trained CVCS model on real multi-view counting datasets, by using unsupervised domain transfer. The proposed CVCS model trained on synthetic data outperforms the same model trained only on real data, and achieves promising performance compared to fully supervised methods that train and test on the same single scene.

1. Introduction

Deep neural network-based multi-view (MV) crowd counting [56, 57] was recently proposed to count people in wide-area scenes that cannot be covered by a single camera. In these works, feature maps from multiple camera views are fused together and decoded to predict a scene-level crowd density map. However, one major disadvantage of the current MV paradigm is the models are trained and

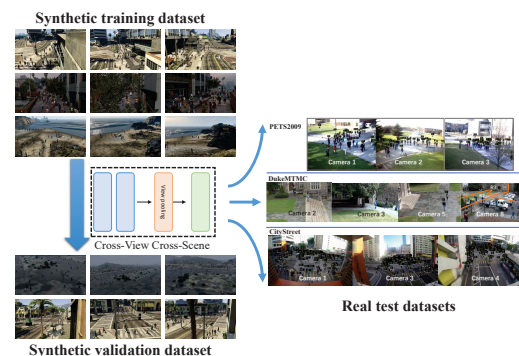


Figure 1: Cross-view cross-scene (CVCS) multi-view crowd counting. The CVCS model is trained and validated on synthetic multi-view crowd scenes, where the scenes and camera-views are different between the training and validation sets. To test on a real scene, unsupervised domain adaptation is applied to the trained CVCS model, where only the real images are used to fine-tune the model.

tested on the same single scene and a fixed camera layout, and thus the trained models do not generalize well to other scenes or other camera layouts.

In this work, we propose a new paradigm of *cross-view cross-scene* multi-view counting (CVCS), where MV counting models are trained and tested on different scenes and arbitrary camera layouts. This paradigm is challenging because both the scene and camera layout (including number of cameras) change at test time. In particular, in single-scene MV counting, the optimal selection of features from each camera and the handling of non-correspondence errors (caused by camera calibration errors or erroneous features) can be directly learned by the MV model (in its network parameters), since it is trained and tested on the same scene/cameras. In contrast, for CVCS MV counting, because the camera positions, camera orders, and scenes are all varying, the MV counting model must learn to dynamically handle different camera layouts and non-correspondence noise. To address these two issues, we propose a CVCS counting model, which attentively selects and fuses features from multiple cameras using the camera lay-

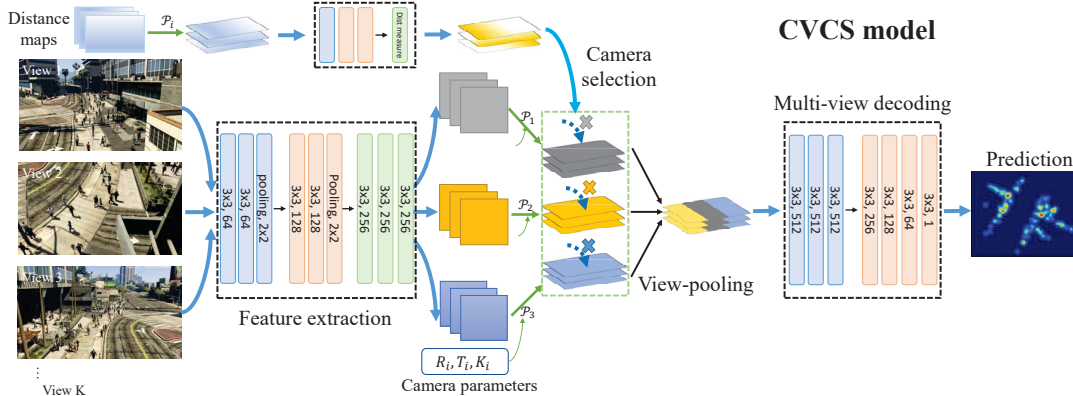


Figure 2: The pipeline of the cross-view cross-scene multi-view counting model (CVCS): 1) **Single-view feature extraction**: The first 7 layers of VGG-Net extracts the single-view features; 2) **Feature projection**: The extracted single-view features are projected to the average height-plane by camera projection; 3) **Multi-camera selection and fusion**: An adaptive CNN subnet selects among the multi-view feature maps, through an attention mechanism guided by the object-to-camera distance, and a max-pooling layer fuses projected camera-view features; 4) **Multi-view decoding**: The fused projected features are decoded to predict the scene-level density maps.

out (object-to-camera distances), and a noise-injection regularization scheme, which simulates non-correspondence errors, improving model generalization.

Effectively training cross-scene counting models requires a large dataset of scenes in order to capture the many possible variations of camera poses and scenes. For example, cross-scene models for *single-view* counting [3, 58, 31, 37, 1] are trained on large single-view datasets, such as ShanghaiTech [58], UCF-QNRF [12], JHU-CROWD++ [44] or NWPU-CROWD [51], which contain thousands of images, each of a different scene. However, current MV counting datasets, such as PETS2009 [8], DukeMTMC [33], and CityStreet [56] only contain 2 to 4 views of a single scene, and combining these three datasets only yields 3 scenes and 3 camera layouts, which is not enough to train a CVCS model. Collecting and annotating a large-scale MV dataset, comprising a large number of scenes taken with many synchronized cameras, is a time-consuming and laborious task, and is further complicated due to personal privacy issues and social-distancing in the current pandemic situation. To avoid such limitations, we generate a synthetic CVCS multi-view dataset of 31 scenes containing around 100 camera views with 100 frames in each view. The large number of camera views for each scene is sufficient for generalizing across camera layouts.

We use the synthetic dataset to train our CVCS model, and directly applying the trained CVCS model to real-world multi-view counting datasets, yielding promising results. The results are further improved by using unsupervised domain adaptation to fine-tune the trained model on only the test images (and not crowd labels).

In summary, the contributions of this paper are 3-fold:

1. We propose a cross-view cross-scene multi-view counting DNN model (CVCS), which adaptively selects and fuses multi-cameras, and a noise view regularization method to improve generalization. To our

knowledge, this is the first study on the cross-view cross-scene multi-view problem in crowd counting.

2. We propose a large synthetic multi-view crowd counting dataset, which contains a large number of camera views, scene variations and frames. This is the first large synthetic dataset for multi-view counting, which enables research on cross-scene cross-view problems.
3. The proposed CVCS model outperforms existing state-of-the-art MV models in the cross-view cross-scene paradigm. Furthermore, the CVCS model, trained on synthetic scenes and adapted to a real-world test scene with unsupervised domain adaptation, achieves promising performance compared with MV models trained on single-scenes.

2. Related Work

Multi-view crowd counting. Traditional multi-view (MV) counting methods [20, 28, 34, 47, 9] rely on foreground extraction techniques and hand-crafted features, and frequently train on PETS2009, which only contains 2 to 4 camera views and hundreds of frames. More recently, [56] proposed a multi-view multi-scale (MVMS) counting model, which fuses multiple views and multiple scales of feature maps into a scene-level feature map, and then decodes it to predict a scene-level density map. [56] collected a new MV counting dataset CityStreet for large-crowd single-scene training. Follow-up work [57] proposed to use 3D ground-truth for MV counting to improve counting performance. However, both these works are trained and tested on single scenes, and do not generalize to cross-scene tasks. Furthermore, the existing MV counting datasets have too few views for cross-view training. In contrast, in our paper we address learning multi-view models that generalize well to new scenes and camera-view layouts, and generate a new large-scale synthetic dataset for training.

Cross-scene single-view crowd counting. Single-

view counting algorithms [49] achieve cross-scene ability mainly by training DNNs on single-view counting datasets, which contain a large number of images capturing different scenes from different view angles [55, 58, 12, 44, 51, 11]. Unsupervised/semi-supervised [36, 43, 26] or weakly supervised methods [25, 48, 54, 59] have also been proposed to improve the cross-scene counting performance on the existing single-view datasets. Various works propose multi-scale DNN models to handle perspective/scale variations across scenes [42, 17, 15, 24, 53, 16, 27]. [18] proposed an adaptive convolution neural network (ACNN), which utilizes the context (camera height and angle) as an auxiliary input to make the model adaptive to different perspectives. [40] integrated the perspective information to provide additional knowledge of a person’s scale change in an image. [50] utilized a large synthetic single-view crowd counting dataset to train counting models, and used a CycleGAN transfer model to apply these models to real-world datasets, improving performance.

Synthetic datasets. Synthetic data is an increasingly popular tool for training deep learning models for various computer vision tasks [29], such as single-view crowd counting [50, 6, 5], automatic driving [21, 35, 10], image segmentation [35], and indoor navigation [39, 38, 46, 52]. Synthetic data is beneficial to real-world computer vision tasks when the real data is insufficient or difficult to acquire, or hard to annotate. To our knowledge, our generated dataset is the first large-scale synthetic dataset for the multi-view crowd counting problem.

Cross-view cross-scene in other multi-camera tasks. Cross-view cross-scene is also an important issue for other multi-camera or cross-camera related vision tasks, such as multi-view tracking/detection [2, 4], 3D reconstruction [45], 3D human pose estimation [13], or person ReID [60]. To obtain cross-view or cross-scene generalization, these methods rely on large training datasets [13], image style transfer [60] or adaptive view or scene modeling [7, 45]. [13] presented two solutions for multi-view 3D human pose estimation based on learnable triangulation methods, by combining 3D information from multiple 2D views, and showed that the model trained on Human3.6m generalized to other 3D human pose datasets. [60] proposed to use CycleGAN [61] to smooth the camera style disparities. [45] proposed Scene Representation Networks, which map world coordinates to a feature representation of local scene properties, and generalize to other scenes by assuming the same class has common shape and appearance properties that are characterized by latent variables. [7] introduced the Generative Query Network to predict unobserved viewpoints, in which the viewpoint parameters are network inputs and output is the view-adaptive feature representation.

Similar to these approaches, we also need large-scale data for training the cross-view cross-scene MV counting

models, and thus we generate a large synthetic dataset for the CVCS multi-view counting task. Furthermore, we propose an adaptive camera selection module to fuse multi-cameras, guided by the object-to-distance information. Unlike [60], we directly adapt the trained CVCS model to real-world datasets with unsupervised domain adaptation.

3. CVCS Multi-View Counting

In this section, we describe the new cross-view cross-scene (CVCS) multi-view counting task. We follow the camera settings of [56, 57], in which the input multi-camera views are synchronized and calibrated. However different from [56, 57], which assume a fixed number of fixed camera locations, CVCS assumes that the camera locations change, and the numbers and order of the cameras vary. Most importantly, the model is trained and tested on distinct scenes and distinct camera layouts, so as to understand the cross-view and cross-scene generalization performance.

3.1. CVCS multi-view counting model

Our proposed CVCS multi-view counting model consists of 4 stages (see Fig. 2), as follows.

(1) *Single-view feature extraction*: The first 7 layers of VGG-Net [41, 22] are used to extract the single-view features. To handle a variable number of views and input camera order, the feature extraction subnet is shared across all input camera views, which requires the feature extraction part to be general enough for different scenes and camera views. Thus, we choose VGG-Net as the feature extractor.

(2) *Single-view feature projection*: The extracted single-view features are projected to a common scene plane by a projection layer with variable camera parameters. As in other multi-view methods [56, 57, 13, 19], the projection is implemented with a spatial transformer net (STN) [14]. In contrast to [56, 57], which uses a fixed set of camera parameters in the projection, our model uses different camera parameters based on the current camera-views.

(3) *Multi-camera selection and fusion*: An adaptive CNN selects among the feature maps of the camera-views, based on an attention mechanism guided by the object-to-camera distance (more details in Sec. 3.2). The motivation of using a selection mechanism is to allow the network to learn, at each location in the scene-level plane, which camera should be more important during fusion. A view-wise max-pooling layer fuses the projected multi-view features. Since the output size of the max-pooling is fixed, the fusion stage is invariant to the number and order of the cameras.

(4) *Multi-view decoding*: the fused projected features are decoded to predict the scene-level density map.

The detailed layer settings of each module can be found in the supplemental. The CVCS model is trained using the MSE loss between the predicted and ground-truth scene-level density map. To make the CVCS model robust to

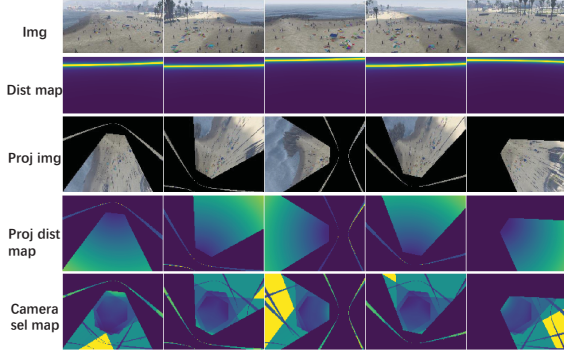


Figure 3: The distance maps and camera selection maps. Warmer colors indicate larger values. In the bottom row, yellow corresponds to 1.

non-correspondence errors arising from small camera calibration errors or erroneous features, we propose a noise-injection method that creates an extra camera view with noise (see Sec. 3.3).

A key difference of our CVCS model and the previous MV counting models [56] is that our model is specifically designed to handle different scenes and camera views, through: 1) layers that are invariant to camera order and camera view, and 2) a multi-view fusion model based on camera layout geometry, which adapts to each camera layout. Another key difference is the proposed noise-injection method to regularize the our model during training.

3.2. Camera selection module

Intuitively, when fusing multi-view features together to form the scene-plane feature map, cameras that are closer to a particular location in the scene-plane should be favored over other cameras, since closer cameras have a clearer view of the location and likely yield more reliable features. Thus, we propose a camera selection module that selects and fuses cameras based on the object-to-camera distance. The raw object-to-camera distance map is input into a subnet CNN, which maps the raw distance to a camera score. Next, using the camera score, a distance measure layer calculates the weights for each camera.

Formally, let F and P denote the feature extraction and projection layer. The K input camera views are $\{V_k\}_{k=1}^K$, their corresponding distance maps are $\{D_k\}_{k=1}^K$ (see Fig. 3) and extracted single-view features are $\{F_k\}_{k=1}^K$. The distance map of each camera view is computed (as in the side information used in [56] for scale selection):

$$D(x, y) = \log(\|RP(x, y, h_{avg}) + T\|_2), \quad (1)$$

where \mathcal{P} , R and T are the camera-view to world projection function, the rotation matrix, and translation, respectively, and h_{avg} is the average person height.

To perform camera selection, first the distance maps are passed through a shared CNN M , and then projected to the scene-plane, $M_k = P(M(D_k))$. Next, the distance mea-

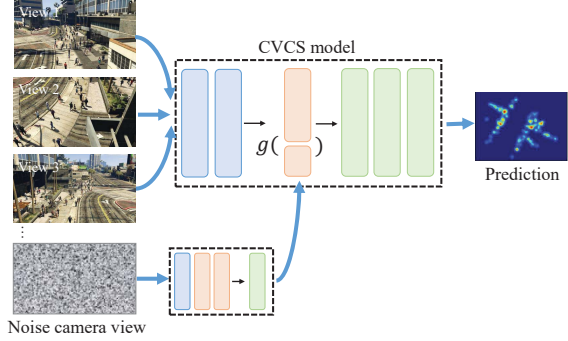


Figure 4: CVCS model with noise view regularization during training. The noise camera view can be added at different layers (see Tab. 1). During testing, the noise view is removed.

sure layer transforms M_k into weight for each camera:

$$\tilde{W}_k = 1 / \exp((M_k - \hat{M})^2), \quad (2)$$

where \hat{M} is the distance to the nearest camera for each pixel (i, j) on the scene-plane,

$$\hat{M}(i, j) = \min(M_1(i, j), \dots, M_K(i, j)). \quad (3)$$

In (2), the nearest camera is assigned weight 1 and other cameras are assigned weights proportionally to the distance ratio, yielding the weight map. Note that the pixels out of the camera views are masked out, and not involved in the camera selection map calculation. The weight maps are normalized across views, $W_k = \tilde{W}_k / \sum_{n=1}^K \tilde{W}_n$, and for each view k , the weight map is element-wise multiplied with the corresponding projected camera-view feature map $P(F_k)$, yielding the attended projected feature map $W_k \otimes P(F_k)$.

In summary, the camera selection module uses side information to dynamically select and fuse the camera views based on the geometry of the camera layout. For comparison, a scale selection module is used in [56], where the object-to-distance map selects the appropriate feature scale in the camera-view to obtain scale consistency across views and within images. In contrast, our camera selection module picks the appropriate camera when performing feature fusion at the scene-level, in order to keep the highest fidelity features. Adaptive camera selection is required for CVCS, since the camera layout changes. In contrast, for single-scene MV models, the camera selection is learned implicitly in the network parameters, which does not generalize to new camera layouts. We further compare these two selection modules in the ablation study.

3.3. Noise camera view regularization

Assuming a correct camera projection and noiseless feature extraction process, the projected features should all align on the scene-level plane. However, in a real system, camera calibration errors, imperfect projection operators, and spurious noise in the feature maps cause *non-correspondence errors* in the projected feature maps. In

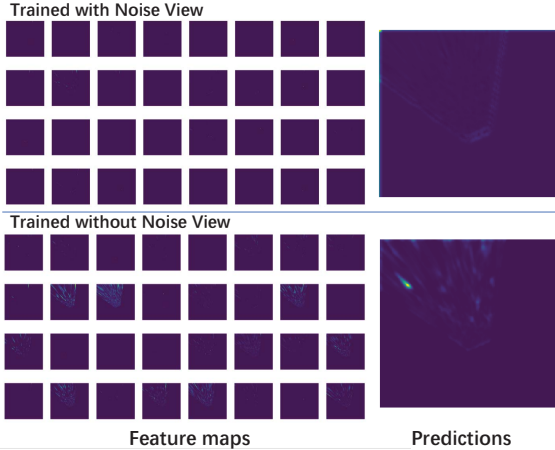


Figure 5: Robustness to noise: the feature maps and prediction for models trained with/without noise view regularization. The input is a Gaussian noise image.

single-scene MV counting, a *scene- and layout-dependent* CNN fusion module learns how to handle these non-correspondence errors. On the other hand, CVCS counting cannot use layout-dependent fusion modules, since the scene and camera layout change.

To make the CVCS model robust to non-correspondence error, we propose a noise-based regularization scheme, where an extra camera view consisting of Gaussian noise $\epsilon = N(0, 1)$ is input in the model at training stage together with the real camera views (see Fig. 4). The camera geometry of the virtual noise camera is the same as one of the input cameras. Note that the noise camera view and associated layers are removed in the testing stage.

Intuitively, the noise camera view simulates non-correspondence errors by randomly activating features in the map. Training with the noise camera view guides the model to reduce the influence of this type of noise on the final prediction, preventing overfitting. An example is seen in Fig. 5, where the model learns to ignore the noise when trained with noise-view regularization. The regularization effect can also be explained from other aspects:

1) **Data augmentation.** Denote the whole CVCS model as \mathcal{M} . The model with noise camera view is changed from $\mathcal{M}(x_1, \dots, x_K)$ to $\mathcal{M}(x_1, \dots, x_K, \epsilon)$, where x_k is a real input camera view and ϵ is the random noise view. During training, the input camera views can be the same but noise view ϵ is random, which is a type of data augmentation.

2) **Noise injection.** The proposed noise camera view regularization approach can also be explained as a new noise injection function [30] for improving model generalization. The training of a noise-injected neural networks is equivalent to optimizing the lower bound of the marginal likelihood over noise ϵ [30]. The difference between additive Gaussian noise, dropout and the proposed noise camera view is the form of the noise injection function $g(x, \epsilon)$. Different noise injection functions g can arise by changing

Table 1: Noise injection functions g that vary on where the noise is added. P is the projection layer, F is the image feature extractor, and H is a separate feature extractor for the noise.

$g(x, \epsilon)$	Noise view added ...
$\max(P(F(x)), \epsilon)$	after projection
$\max(P(F(x)), P(\epsilon))$	before projection
$\max(P(F(x)), P(F(\epsilon)))$	at input layer (same feature extractor)
$\max(P(F(x)), P(H(\epsilon)))$	at input layer (different feature extractor)
$\text{sum}(P(F(x)), P(F(\epsilon)))$	at input layer, sum (same feature extractor)
$\text{sum}(P(F(x)), P(H(\epsilon)))$	at input layer, sum (different feature extractor)

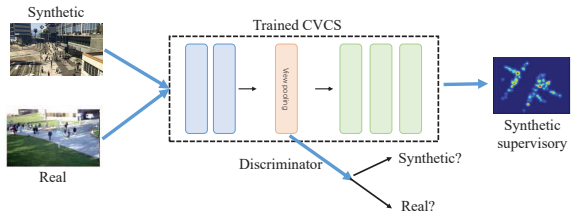


Figure 6: The unsupervised domain adaptation to real datasets.

the network layer to which the noise is injected (see Tab. 1). We compare these different noise injection functions in the ablation study (Sec. 5.3.2).

3.4. Unsupervised domain adaptation to real data

Our CVCS model is trained on a synthetic multi-view dataset (see Sec. 4). Directly applying the trained CVCS model to real multi-view datasets, such as PETS2009 and CityStreet, will be limited by the domain gap between synthetic and real scenes. To reduce the domain gap, we apply unsupervised domain adaptation to fine-tune the trained CVCS model on each test scene, using only the test images (without the crowd labels). In particular, we add a feature discriminator to the trained CVCS model to reduce the feature gap (see Fig. 6). The discriminator is inserted after the view-pooling layer in the CVCS model. During the fine-tuning stage, both the real and synthetic images are input to the model, then the real/synthetic features are fed to the discriminator to be classified. Only the synthetic features are sent to the multi-view decoder, since crowd labels for the real data are not available for training. The fine-tuning loss function combines the synthetic counting loss and the discriminator loss. Our procedure of training on synthetic data and then testing on real data after unsupervised domain adaptation is more useful for practical applications, compared to previous state-of-the-arts [56, 57], which require training and testing on the same real scene with fixed cameras.

4. Synthetic CVCS Dataset

The proposed large synthetic multi-view crowd counting dataset is generated using GCC-CL [50], which works as a plug-in for the game “Grand Theft Auto V”. The generating process consists of two parts: scene simulation and multi-view recording. First, crowd scenes are simulated, through the selection of the background selected, region of interest (ROI), weather condition, human models and postures, etc. Next, cameras are placed at various locations to record the

Table 2: Comparison of multi-view crowd datasets.

Dataset	Imgs. (train / test)	Scenes	Counts	Views	Image Res.
PETS2009	1105 / 794	1	20-40	3	768×576
DukeMTMC	700 / 289	1	10-30	4	1920×1080
CityStreet	300 / 200	1	70-150	3	2704×1520
Synthetic (ours)	200,000 / 80,000	31	90-180	60-120	1920×1080

crowd scene from various perspectives. Birds-eye views are also collected for visualization. Each person has a specific ID for mapping their coordinates in the world coordinate system and their locations in each camera-view image. The camera parameters, such as coordinates, deflection angles and fields-of-view, are also recorded.

In total, the whole synthetic MV counting dataset contains 31 scenes. For each scene, around 100 camera views are set for multi-view recording. The multi-view recording is performed 100 times with different crowd distributions in the scene, i.e., each scene contains 100 multi-view frames, with each frame comprising 60 to 120 camera-views. The image resolution is 1920×1080. Compared with other MV counting datasets, like PETS2009 [8], DukeMTMC [33] and CityStreet [56], our proposed synthetic dataset contains more scenes, more camera views variations, and more total images (see Tab. 2), which makes it more amenable for training and validating CVCS multi-view counting. Example images from the proposed synthetic multi-view counting dataset are shown in Fig. 1 and the Supplemental.

5. Experiments

In this section, we conduct experiments on CVCS multi-view counting using our proposed model.

5.1. Test datasets

The real test datasets are PETS2009 [8], DukeMTMC [33] and CityStreet [56]. We use the same dataset settings as previous MV counting [56]. The dataset information is shown in Tab. 2. The input images are downsampled to 384×288, 640×360, and 676×380, for PETS2009, DukeMTMC, and CityStreet, respectively. The ground-plane scene-level density maps have resolutions of 152×177, 160×120, and 160×192 for the three datasets.

5.2. Experiment settings

The synthetic dataset contains 31 scenes in total, of which 23 scenes are used for training and the remaining 8 scenes are used for testing. During training, we randomly select $K = 5$ views for $P = 5$ times in each iteration per frame of each scene. For evaluation, we randomly select $K = 5$ views for $P = 21$ times ($V/K + 1$, $V = 100$ camera views) per frame of each scene, in order to test on more camera layouts. The input image resolution is 640×360 (resized to 1/3 of the original resolution). Ground-plane patch-based training is used instead of complete image training, where 5 patches are extracted from the view-pooled features, corresponding to the 5 patches ex-

Table 3: Experiment results of our CVCS model with different modules.

Model	MAE	NAE
Backbone	14.13	0.115
Backbone+MVMS	9.30	0.080
Backbone+CamSel	8.63	0.074
Backbone+NoiseV	7.94	0.069
CVCS (Backbone+CamSel+NoiseV)	7.22	0.062

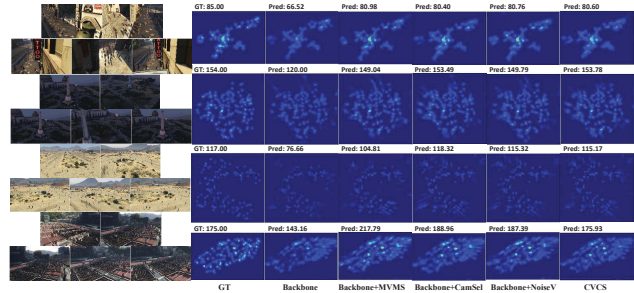


Figure 7: The results of CVCS variations on synthetic datasets. Using camera selection and/or noise-view regularization (CVCS, Backbone+CamSel, Backbone+NoiseV) are more accurate than the backbone or backbone with MVMS (Backbone+MVMS). See Supplemental for full-size figure.

tracted from the ground-truth scene-level density maps. The patch size is 160×180, and 1 pixel is equal to 0.5m in the real world. The ground-truth density maps are generated by convolving a Gaussian kernel with the ground-truth person annotation dot map. The learning rate is 1e-3, with learning decay 1e-4, and weight decay is 1e-4. The single-view feature extractor is pretrained via a single-view counting task on the synthetic dataset. Two evaluation metrics are used, the mean absolute error (MAE) and mean normalized absolute error (NAE) of the predicted counts on the test set.

5.3. Experiment results

We first report results on CVCS multi-view counting on the synthetic dataset, followed by the ablation study. Finally, we present results on the real datasets.

5.3.1 CVCS performance

We test 5 variations of our CVCS model on the synthetic dataset. The first method is the backbone model of our CVCS model without the camera selection and noise camera view (denoted as “Backbone”). The second method is the backbone model with the multi-view multi-scale selection architecture from [56] (denoted as “Backbone+MVMS”), where a 3-scale pyramid is used in the feature extraction part and camera view distance maps are used to fuse the multi-scales before projection. The next 2 methods add either the camera selection module (Backbone+CamSel) or the noise camera view (Backbone+NoiseV) to the backbone model. Finally, adding both modules to the backbone yields our full model (CVCS).

The results on the synthetic dataset are shown in Table 3 and Fig. 7. Using the proposed camera selection module

Table 4: Ablation study on the camera selection module.

Model	MAE	NAE
Backbone	14.13	0.115
+CamSel (no conv)	10.77	0.089
+CamSel (1 × 1 conv)	8.63	0.074
+CamSel (3 conv)	8.15	0.069

Table 5: Ablation study of the noise view regularization.

Model	Type	MAE	NAE
Backbone		14.13	0.115
+Dropout		13.16	0.111
+NoiseV (A)	$\max(x, \epsilon)$	19.91	0.163
+NoiseV (B)	$\max(P(F(x)), \epsilon)$	9.64	0.084
+NoiseV (C)	$\max(P(F(x)), P(\epsilon))$	9.42	0.079
+NoiseV (D)	$\max(P(F(x)), P(F(\epsilon)))$	7.94	0.069
+NoiseV (E)	$\text{sum}(P(F(x)), P(F(\epsilon)))$	8.57	0.076
+NoiseV (F)	$\max(P(F(x)), P(H(\epsilon)))$	8.48	0.074
+NoiseV (G)	$\text{sum}(P(F(x)), P(H(\epsilon)))$	8.54	0.076

or noise view regularization yields large reduction of errors over the Backbone, while their combination further reduces the errors, showing that the two modules can work together to improve CVCS performance in terms of different aspects. The MVMS architecture also reduces the error when used with the Backbone, as the multi-scale selection can promote scale consistency of the features across camera views. However, the error reduction using MVMS is not as large as using CamSel and NoiseV. The main reason is that the camera selection and noise view regularization modules are designed for better cross-view cross-scene performance.

5.3.2 Ablation study

In the ablation study we consider various permutations of our CVCS model.

Camera selection module. We consider 3 settings of the CNN mapping in the camera selection module: no convolution layers, i.e., passing the distance map directly to the distance measure layer (denoted as “no conv”), a 1×1 conv layer, and a 3-layer CNN. The results are shown Table 4. Compared to the Backbone model, all the camera selection settings reduce the counting error; CNN achieves the best error reduction due to its flexibility in mapping the distance information to a suitable weight.

Noise view regularization. We next experiment with different types of the noise view regularization added to the Backbone, as shown Table 5. Generally, adding noise-based regularization can improve the performance, except when directly corrupting the input image (NoiseV, Type A). The best regularization occurs when the noise is passed through a feature extractor and projection (Types D-G). These noise injection functions better simulate the noisy feature extraction process and how the noise is projected into the scene-level feature map. Dropout regularization is also tested with the Backbone, where 2 dropout layers are added after the 2nd and 4th layers of the feature extractor. Dropout can also improve the performance slightly, but not as much as the noise camera view.

Table 6: Ablation study combining the camera selection module and noise view regularization. Noise types are defined in Table 5.

Model	MAE	NAE
Backbone	14.13	0.115
+CamSel (1 × 1 conv)+NoiseV (Type D)	7.22	0.062
+CamSel (1 × 1 conv)+NoiseV (Type E)	9.98	0.087
+CamSel (1 × 1 conv)+NoiseV (Type F)	8.34	0.074
+CamSel (1 × 1 conv)+NoiseV (Type G)	8.32	0.074
+CamSel (3 conv)+NoiseV (Type D)	7.96	0.069
+CamSel (3 conv)+NoiseV (Type E)	8.22	0.072
+CamSel (3 conv)+NoiseV (Type F)	7.56	0.066
+CamSel (3 conv)+NoiseV (Type G)	7.44	0.065

Table 7: Ablation study on different numbers of input camera views.

No. Views	Backbone		CVCS	
	MAE	NAE	MAE	NAE
3	14.28	0.130	7.24	0.071
5	14.13	0.115	7.22	0.062
7	14.35	0.113	7.07	0.058
9	14.56	0.112	7.04	0.056
11	15.15	0.115	7.00	0.055

Combining camera selection and noise view regularization. From the previous experiments, we find that the camera selection and noise view regularization can both improve the performance of Backbone model. Since the two modules address different aspects, we combine the two modules together to further improve the performance. Specifically, the camera selection modules are combined with different noise view regularization methods and the results are shown in Table 6. The best combination uses 1×1 conv in the camera selection module and noise injection in the input view (Type D), $\max(P(F(x)), P(F(\epsilon)))$. Generally, the various combinations perform better than the Backbone and Backbone+MVMS models.

Variable camera number. Our CVCS model is specifically designed to handle any number of camera views at test time. To show the influence of number of camera views, we test CVCS with different number of input cameras. Note that the models are trained with 5 input camera views, and tested on different number of views. The results are presented in in Table 7. The performance with different numbers of cameras is stable for both the Backbone and the full model CVCS. Increasing the camera views from 5 to 11, the full CVCS achieves slightly lower error because the extra cameras provide more information in some regions that were poorly covered with only 5-cameras.

5.3.3 Cross-view cross-scene on real data

We apply the trained CVCS model to real multi-view counting datasets. Unlike previous state-of-the-arts [56, 57] that train and test on the same single real scene with fixed camera views, we train on synthetic scenes and test on real scenes (cross-scene setting) with different camera views (cross-view). We allow unsupervised domain adaptation, which uses the images of the test scene to fine-tune the feature extractor, and does not use crowd labels. We believe

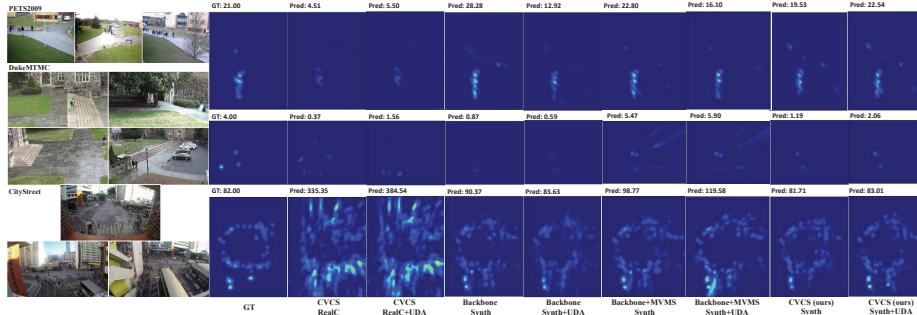


Figure 8: The cross-view cross-scene results on real datasets. Our CVCS model trained on the synthetic data shows better performance than CVCS trained on real data. Applying unsupervised domain adaptation (UDA) to our CVCS improves the performance. See Supp. for full-sized figure.

Table 8: Results testing on real datasets. Different training schemes are used: “RealS” means training and testing on the same real scene (single-scene MV); “RealC” means cross-scene training on 2 real scenes and testing on the other; “Synth” means cross-scene training on synthetic dataset; “+UDA” adds unsupervised domain adaptation.

Model	Training	Test Scene					
		PETS2009		DukeMTMC		CityStreet	
		MAE	NAE	MAE	NAE	MAE	NAE
Dmap_wtd [34]	RealS	7.51	0.261	2.12	0.255	11.10	0.121
Dect+ReID [56]	RealS	9.41	0.289	2.20	0.342	27.60	0.385
LateFusion [56]	RealS	3.92	0.138	1.27	0.198	8.12	0.097
EarlyFusion [56]	RealS	5.43	0.199	1.25	0.220	8.10	0.096
MVMS [56]	RealS	3.49	0.124	1.03	0.170	8.01	0.096
3D [57]	RealS	3.15	0.113	1.37	0.244	7.54	0.091
CVCS	RealC	23.34	0.729	5.28	0.623	215.23	2.700
CVCS	RealC+UDA	20.11	0.636	5.34	0.628	249.25	3.110
Backbone	Synth	8.05	0.257	4.19	0.913	11.57	0.156
Backbone+MVMS	Synth	6.03	0.191	3.07	0.553	14.02	0.194
CVCS (ours)	Synth	5.33	0.174	2.85	0.546	11.09	0.124
Backbone	Synth+UDA	5.91	0.200	3.11	0.551	10.09	0.117
Backbone+MVMS	Synth+UDA	5.28	0.175	3.00	0.585	12.05	0.157
CVCS (ours)	Synth+UDA	5.17	0.165	2.83	0.525	9.58	0.117

this testing paradigm is more practical, since annotations of the target scenes are not required.

We test three groups of methods, with different training setups. The first group is trained and tested on *single* real scenes (denoted as “RealS”), and include 2 traditional multi-view counting methods, Dmap_wtd [34, 56] and Dect+ReID [32, 23, 56], 3 DNN-based fusion methods (EarlyFusion, LateFusion, and MVMS) [56], and 3D counting [57]. The second group uses cross-scene training on the real datasets (denoted as “RealC”), where our CVCS model (Backbone+CamSel+NoiseV) is trained on 2 real scenes, and tested on the remaining scene. The third group uses cross-scene training on our synthetic dataset and directly tests on the real data (denoted as “Synth”). We test 3 models: Backbone, Backbone+MVMS, and our CVCS model. We also add unsupervised domain adaptation using images from the test scene (“+UDA”). Note that no crowd labels are used from the test scene for UDA.

The test performances of these methods on the 3 real datasets are shown in Table 8. Directly testing our synthetically trained CVCS on real scenes achieves promising performance, which is better or competitive to the 2

traditional multi-view counting methods, Dmap_wtd and Dect+ReID. Furthermore, adding unsupervised domain adaptation (“Synthetic+UDA”) effectively reduces the domain gap between synthetic and real data, yielding lower errors compared to without UDA. Our CVCS also outperforms Backbone and Backbone+MVMS, for both Synth training and Synth+UDA training, which shows the advantage of our cross-scene cross-view specific modules, camera selection and noise-view regularization.

Training using only the real scenes in a cross-scene manner (RealC) yields very large errors, compared to training with the synthetic data (Synth), showing that the model overfits when there are too few training scenes, and that there is significant benefit in training on more scenes/views even if they are synthetic. Finally, our CVCS model using UDA is slightly worse than the MVMS model [56] trained and tested on the same-scene (MAE increases by 1-2), which shows the promise of our approach and the CVCS paradigm. Example visualizations of the results are presented in Fig. 8.

6. Conclusion

In this paper, we propose the task of cross-view cross-scene (CVCS) multi-view counting, where models are trained to generalize to different scenes and camera layouts. We propose a CVCS multi-view DNN with a camera selection and fusion module and noise-view regularization, to adapt the network to different camera layouts and to learn to ignore non-correspondence errors. We collect a large-scale synthetic dataset with large numbers of camera views and scenes for training and evaluating the CVCS multi-view counting. Furthermore, we show that the synthetically-trained CVCS model can be applied to real scenes via unsupervised domain adaptation, which only uses images from the test scene. Overall, our work advances research on multi-view crowd counting from the single-scene fixed-camera setting to cross-view cross-scene setting, which is more practical for deployment.

Acknowledgments. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11212518).

References

- [1] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2020. 2
- [2] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 271–279, 2017. 3
- [3] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM Multimedia Conference*, pages 640–644. ACM, 2016. 2
- [4] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. 3
- [5] Ernest Cheung, Tsan Kwong Wong, Aniket Bera, Xiaogang Wang, and Dinesh Manocha. Lcrowdv: Generating labeled videos for simulation-based crowd behavior learning. In *14th European Conference on Computer Vision (ECCV)*, pages 709–727, 2016. 3
- [6] Nicolas Courty, Pierre Allain, Clement Creusot, and Thomas Corpetti. Using the agoraset dataset: Assessing for the quality of crowd video analysis methods. *Pattern Recognition Letters*, 44:161 – 170, 2014. 3
- [7] S M Ali Eslami, Rezende, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 3
- [8] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–6, 2009. 2, 6
- [9] Weina Ge and Robert T. Collins. Crowd detection with a multiview sampler. In *ECCV*, pages 324–337, 2010. 2
- [10] Xiaoxia Hu, Xuefeng Liu, Zhenming He, and Jiahua Zhang. Batch modeling of 3d city based on esri cityengine. In *IET International Conference on Smart and Sustainable City (IC-SSC)*, pages 69–73, 2013. 3
- [11] Y. Hu, X. Jiang, X. Liu, B. Zhang, J. Han, X. Cao, and D. Doermann. Nas-count: Counting-by-density with neural architecture search. In *ECCV*, 2020. 3
- [12] Haroon Idrees and et al. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, pages 532–546, 2018. 2, 3
- [13] Karim Isakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7718–7727, 2019. 3
- [14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3
- [15] Xiaolong Jiang and et al. Crowd counting and density estimation by trellis encoder-decoder networks. In *CVPR*, pages 6133–6142, 2019. 3
- [16] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3
- [17] Di Kang and Antoni B. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. In *BMVC*, 2018. 3
- [18] Di Kang, Debarun Dhar, and Antoni B. Chan. Incorporating side information by adaptive convolution. In *Advances in Neural Information Processing Systems*, pages 3867–3877, 2017. 3
- [19] Abhishek Kar, Christian Háne, and Jitendra Malik. Learning a multi-view stereo machine. In *NIPS*, pages 365–376, 2017. 3
- [20] Jingwen Li, Lei Huang, and Changping Liu. People counting across multiple cameras for intelligent video surveillance. In *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 178–183. IEEE, 2012. 2
- [21] Wei Li, Chengwei Pan, Rong Zhang, Jiaping Ren, Yuexin Ma, Jin Fang, Feilong Yan, Qichuan Geng, Xinyu Huang, Huajun Gong, Weiwei Xu, Guoping Wang, Dinesh Manocha, and Ruigang Yang. Aads: Augmented autonomous driving simulation using data-driven algorithms. *arXiv preprint arXiv:1901.07849*, 4(28), 2019. 3
- [22] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018. 3
- [23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 8
- [24] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *CVPR*, pages 5099–5108, 2019. 3
- [25] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1862–1878, 2019. 3
- [26] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. *arXiv preprint arXiv:2007.03207*, 2020. 3
- [27] Zhiheng Ma, X. Wei, Xiaopeng Hong, and Y. Gong. Bayesian loss for crowd count estimation with point supervision. pages 6141–6150, 2019. 3
- [28] L. Maddalena, A. Petrosino, and F. Russo. People counting by learning their appearance in a multi-view camera environment. *Pattern Recognition Letters*, 36:125–134, 2014. 2
- [29] Sergey I. Nikolenko. Synthetic data for deep learning. *ArXiv*, abs/1909.11512, 2019. 3

- [30] Hyeonwoo Noh, Tackgeun You, Jonghwan Mun, and Bohyung Han. Regularizing deep neural networks by noise: Its interpretation and optimization. In *Advances in Neural Information Processing Systems*, pages 5109–5118, 2017. 5
- [31] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, pages 615–629. Springer, 2016. 2
- [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 8
- [33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 2, 6
- [34] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Scene invariant multi camera crowd counting. *Pattern Recognition Letters*, 44(8):98–112, 2014. 2, 8
- [35] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M. Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 84–100, 2018. 3
- [36] Deepak Babu Sam, Neeraj Nagaraj Sajjan, Himanshu Maurya, and Venkatesh Babu Radhakrishnan. Almost unsupervised learning for dense crowd counting. *Thirty-Third AAAI Conference on Artificial Intelligence*, 33(1):8868–8875, 2019. 3
- [37] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, pages 4031–4039, 2017. 2
- [38] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas A. Funkhouser, and Vladlen Koltun. Minos: Multimodal indoor simulator for navigation in complex environments. *ArXiv*, abs/1712.03931, 2017. 3
- [39] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9338–9346, 2019. 3
- [40] Miaoqing Shi, Zhaohui Yang, and et al. Revisiting perspective information for efficient crowd counting. In *CVPR*, pages 7279–7288, 2019. 3
- [41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3
- [42] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, pages 1879–1888, 2017. 3
- [43] Vishwanath A Sindagi, Rajeev Yasarla, Deepak Sam Babu, R Venkatesh Babu, and Vishal M Patel. Learning to count in the crowd from limited labeled data. *arXiv preprint arXiv:2007.03195*, 2020. 3
- [44] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1221–1231, 2019. 2, 3
- [45] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1121–1132, 2019. 3
- [46] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017. 3
- [47] N. Tang, Y. Y. Lin, M. F. Weng, and H. Y. Liao. Cross-camera knowledge transfer for multiview people counting. *IEEE Transactions on Image Processing*, 24(1):80–93, 2014. 2
- [48] Matthias von Borstel, Melih Kandemir, Philip Schmidt, Madhavi K. Rao, Kumar T. Rajamani, and Fred A. Hamprecht. Gaussian process density counting from weak supervision. In *European Conference on Computer Vision*, pages 365–380, 2016. 3
- [49] Jia Wan and Antoni B. Chan. Adaptive density map generation for crowd counting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1130–1139, 2019. 3
- [50] Qi Wang, Junyu Gao, and et al. Learning from synthetic data for crowd counting in the wild. In *CVPR*, pages 8198–8207, 2019. 3, 5
- [51] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting. *arXiv preprint arXiv:2001.03360*, 2020. 2, 3
- [52] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *ArXiv*, abs/1801.02209, 2018. 3
- [53] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4374–4383, 2020. 3
- [54] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. *ECCV*, 2020. 3
- [55] Cong Zhang, Hongsheng Li, and et al. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841, 2015. 3
- [56] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Computer Vision and Pattern Recognition*, pages 8297–8306, 2019. 1, 2, 3, 4, 5, 6, 7, 8
- [57] Qi Zhang and Antoni B Chan. 3d crowd counting via multi-view fusion with 3d gaussian kernels. In *AAAI Conference on Artificial Intelligence*, pages 12837–12844, 2020. 1, 2, 3, 5, 7, 8
- [58] Yingying Zhang and et al. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016. 2, 3

- [59] Zhen Zhao, Miaoqing Shi, Xiaoxiao Zhao, and Li Li. Active crowd counting with limited supervision. *arXiv preprint arXiv:2007.06334*, 2020. 3
- [60] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Computer Vision and Pattern Recognition*, pages 5157–5166, 2018. 3
- [61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 3