frontiers | Frontiers in Imaging

TYPE Original Research
PUBLISHED 20 October 2023
DOI 10.3389/fimag.2023.1271885

OPEN ACCESS

EDITED BY
Hong Fu,
The Education University of Hong Kong, Hong Kong SAR, China

REVIEWED BY
Jiyong Wang,
Shanghai Pulse Medical Technology, Inc., China
Wanli Xie,
Qufu Normal University, China

*CORRESPONDENCE
Xiaomei Zhao
✉ zhaoxiaomei20@sdjzu.edu.cn

RECEIVED 04 August 2023
ACCEPTED 02 October 2023
PUBLISHED 20 October 2023

CITATION
Zhao X, Li H, Zhao Z and Li S (2023) Height reverse perspective transformation for crowd counting. *Front. Imaging.* 2:1271885.
doi: 10.3389/fimag.2023.1271885

Height reverse perspective transformation for crowd counting

Xiaomei Zhao*, Honggang Li, Zhan Zhao and Shuo Li

Shandong Key Laboratory of Intelligent Buildings Technology, School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan, China

Introduction: Crowd counting plays a critical role in the intelligent video surveillance of public areas. A significant challenge to this task is the perspective effect on human heads, which causes serious scale variations. Height reverse perspective transformation (HRPT) alleviates this problem by narrowing the height gap among human heads.

Methods: It employs depth maps to calculate the rescaling factors of image rows, and then it performs image transformation accordingly. HRPT enlarges small human heads in far areas to make them more noticeable and shrinks large human heads in closer areas to reduce redundant information. Then, convolutional neural networks can be used for crowd counting. Previous crowd-counting methods mainly solve the scale variation problem by designing specific networks, such as multi-scale or perspective-aware networks. These networks cannot be conveniently employed by other methods. In contrast, HRPT solves the scale variation problem through image transformation. It can be used as a preprocessing step and easily employed by other crowd-counting methods without changing their original structures.

Results and discussion: Experimental results show that HRPT successfully narrows the height gap among human heads and achieves state-of-the-art performance on a large crowd-counting RGB-D dataset.

KEYWORDS

crowd counting, scale variation problem, perspective effect, height reverse perspective transformation, RGB-D image

1. Introduction

Crowd-counting technology estimates the number of people in images, which is instrumental in intelligent video surveillance. It plays a vital role in safeguarding public safety by recognizing abnormal crowd gatherings automatically.

Over the past decades, countless researchers have focused on improving the accuracy of crowd-counting techniques (Sindagi and Patel, 2018; Gao et al., 2020; Fan et al., 2022). One standard method is to count the number of human heads captured by surveillance cameras, since occlusions on human heads are less severe than the other parts of human bodies. However, the perspective effect can cause significant scale variations in human head size, posing a critical challenge to accurate counting. This study proposes a novel height reverse perspective transformation (HRPT) method to alleviate the scale variation problem. This technique creatively narrows the height gap among human heads, particularly enlarging small human heads in far areas and shrinking large human heads in closer areas. Figure 1 displays a group of crowd RGB images to demonstrate the effect of HRPT.

Existing crowd-counting methods are categorized into detection-based, regression-based, and density map-based methods. Detection-based methods generally count people by detecting people or their heads (Idrees et al., 2015; Stewart et al., 2016; Liu Y. et al., 2019), while regression-based methods extract image features from the whole crowd image and

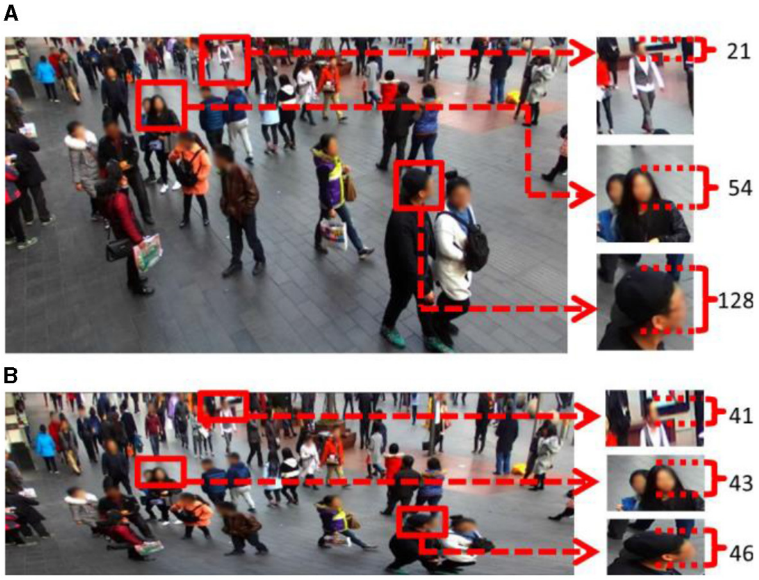

FIGURE 1
A group of crowd RGB images to show the effect of HRPT: **(A)** original crowd RGB image; **(B)** crowd RGB image transformed by HRPT. The heads of three pedestrians are marked in **(A, B)**. Their head heights, measured in pixels, are shown on the right of the images.

regress the number of people according to these features (Liu and Vasconcelos, 2015; Wang et al., 2015; Shang et al., 2016). Density map-based methods count people by estimating the density map and summing the density over the whole image (Ma et al., 2022; Wang et al., 2022, 2023; Yan et al., 2022). Detection-based methods usually perform poorly when dealing with dense crowds far from the camera (Liu et al., 2018; Xu et al., 2019; Fan et al., 2022), and regression-based methods ignore spatial information (Gao et al., 2020). Therefore, density map-based methods are more popular than the other two types of methods.

Many density map-based methods solve the scale variation problem using multi-scale networks (Liu W. et al., 2019; Ma et al., 2022; Wang et al., 2022, 2023). However, these networks only consider a finite number of discrete scales, limiting their ability to handle scale variations (Yan et al., 2022). Therefore, many researchers have focused on solving the scale variation problem by utilizing perspective-aware approaches (Yan et al., 2022; Zhang and Li, 2022). Perspective-aware approaches generally extract perspective information from RGB images (Yan et al., 2022; Zhang and Li, 2022). In contrast, a depth map can be directly used as perspective information, and it is more accurate than the perspective information extracted from an RGB image. Using a depth map to its maximum capacity can alleviate the scale variation problem caused by the perspective effect.

RGB-D cameras capture both RGB images and depth maps, which contain complementary information. Applying multiple complementary information is popular in many areas, such as automatic malfunction detection (Jing et al., 2017) and automatic driving (Wang et al., 2019; Zhuang et al., 2021). This is because using multiple types of complementary information can improve the robustness and accuracy of automatic systems. At present, although the RGB-D camera is less popular than the RGB camera owing to its higher cost, as the economy develops, the RGB-D camera is expected to be used more widely.

This study proposes the HRPT method, developed based on RGB-D images, to solve the scale variation problem in the crowd-counting task. By narrowing the height gap among human heads according to the perspective information in the depth map, HRPT alleviates the scale variation problem, reduces redundant information in near areas, and makes small human heads in faraway areas more visible. As shown in Figure 1, HRPT successfully narrows the height gap among human heads. It enlarges small human heads in far areas, making them more visible, and it shrinks large human heads in closer areas to reduce redundant information, making the counting network focus more attention on remote areas where the human heads are denser and harder to detect.

Previously developed crowd-counting methods often use specialized networks to tackle scale variation, employing multi-scale or perspective-aware strategies. Other methods must change their original network structures to employ these strategies. In contrast, the proposed HRPT can be used as a preprocessing step and easily integrated into existing methods without modifying their original network structures. This study combines HRPT with four well-performing crowd-counting methods: CSRNet (Li et al., 2018), DM-Count (Wang B. et al., 2020), GL (Wan et al., 2021), and CLTR (Liang et al., 2022) to produce promising results.

In addition to narrowing height gap among human heads by HRPT, we also try to narrow the width gap among human heads. However, narrowing the width gap does not achieve promising results. Section 5.2.1 details the discussion. Furthermore, if we ignore the human height, the perspective effect on the 2D ground can be eliminated using a homograph (Hartley and Zisserman, 2003). However, experimental results show that even though the homograph successfully eliminates the perspective effect on the 2D ground, it cannot achieve satisfactory results in alleviating the scale variations of human heads. Section 5.2.2 presents a more detailed discussion.

In summary, our main contributions are given below.

1) HRPT is a novel technique to alleviate the scale variation problem in the crowd-counting task. It uses the perspective information in depth maps and creatively narrows the height gap among human heads via image transformation. After HRPT, small human heads in outlying areas are enlarged and become more apparent, and large human heads in closer areas are shrunken to reduce redundant information.
2) HRPT is integrated with the following well-performing crowd-counting methods: CSRNet (Li et al., 2018), DM-Count (Wang B. et al., 2020), GL (Wan et al., 2021), and CLTR (Liang et al., 2022), and experimental results demonstrate that HRPT successfully improves crowd-counting performance. Our method (GL+HRPT) achieves state-of-the-art results and outperforms other methods that employ depth maps.
3) Another two image transformation methods (i.e. a shape-changing method that narrows the width gap among human heads and a geometric method that attempts to eliminate the perspective effect) are also experimented to solve the scale variation problem. Experimental results show that these two methods have poorer performance than the proposed HRPT.

2. Related work

The proposed method aims to alleviate the scale variation problem in crowd-counting tasks by making use of depth maps and image transformation. This problem has traditionally been tackled by employing multi-scale or perspective-aware crowd-counting methods. The following section first introduces these two types of previous methods, and then presents other related crowd-counting methods that also employ depth maps and image transformation.

2.1. Multi-scale crowd-counting methods

Multi-scale crowd-counting methods solve the scale variation problem by employing multi-scale network structures. Wang et al. (2022) and Wang et al. (2023) built multi-scale networks by employing multiple branches with different convolutional dilation rates. Liu W. et al. (2019) built a multi-scale network by using multiple branches with different pooling sizes. Jiang et al. (2019) and Ma et al. (2022) built multi-scale networks by combining image features of multiple layers. Jiang et al. (2020) and Du et al. (2023) built multi-scale networks by combining the estimated crowd density maps of multiple scales. However, multi-scale methods only consider a finite number of discrete scales, limiting their potential to solve the scale variation problem (Yan et al., 2022).

2.2. Perspective-aware crowd-counting methods

Many impressive perspective-aware methods have been proposed. For example, Zhang et al. (2015) estimated the number of pixels representing one meter and used this perspective information to normalize the density map. Yan et al. (2022) estimated the same perspective information as Zhang et al. (2015) and used it to select different dilation kernels. Zhang and Li (2022) embedded perspective information into a point-supervised network to better handle the scaling problem. Wan et al. (2021) used a perspective-guided cost function with a larger penalty to density far from the camera. Zhao et al. (2019) used the depth map predicted from RGB image as perspective information and embedded it into their density map prediction network. The abovementioned methods extract perspective information from RGB images, which is complicated and inaccurate. In contrast, depth maps can be directly used as accurate perspective information. In the following subsection, we introduce methods that employ depth maps.

2.3. Depth maps and crowd counting

Depth maps are the source of information, more accurate than the perspective information extracted from RGB images. Thanks to the current development of RGB-D cameras, several excellent crowd-counting methods have emerged to take advantage of depth maps. As density map-based crowd-counting methods have better performance in dealing with far-view areas, and as detection-based methods have better performance in dealing with near-view areas, Xu et al. (2019) used depth maps to segment RGB images into the far-view and near-view areas, and used density map-based and detection-based methods to deal with these two areas, respectively. However, the density map-based and detection-based methods employed in their framework do not use depth maps during counting. Liu et al. (2021, 2023), Zhang et al. (2021), and Li et al. (2023) proposed cross-model frameworks to estimate density maps, fusing image features extracted from RGB images and depth maps to make use of the complementary information in these two kinds of images. They only used depth maps as input and did not explicitly utilize the perspective information contained in depth maps. Lian et al. (2019) used depth-adaptive Gaussian kernels and depth-aware anchors to improve crowd-counting and localization results, using the perspective information in depth maps to improve the quality of ground-truth density maps and human head anchors. Additionally, Lian et al. (2022) used depth-guided dynamic dilated convolution to further improve the method proposed by Lian et al. (2019). Compared with the above methods, the proposed HRPT utilizes the perspective information in depth maps more intuitively, narrowing the height gap among human heads through image transformation to alleviate the scale variation problem in crowd counting.

2.4. Image transformation and crowd counting

Yang et al. (2020) proposed a reverse perspective network to evaluate and correct the perspective distortion in crowd images. Both Yang et al.'s (2020) method and our HRPT attempt to narrow the scale gap among human heads by image transformation. However, the two methods use different perspective information. Theirs uses the perspective information extracted from RGB images; our HRPT uses the perspective information in depth maps. The perspective information in depth maps is more accurate than that extracted from RGB images. Moreover, Yang et al. (2020) designed a specific network structure to estimate and correct perspective distortion. Other methods must change their original network structures to employ this approach. In contrast, they can easily employ our HRPT approach without changing any part of their original network structures.

3. Methods

Figure 2 depicts the flowchart of our crowd-counting framework. As shown in this figure, the original RGB-D image is first sent into the proposed HRPT module. HRPT employs the perspective information in the depth map to narrow the height gap among human heads in the RGB image by image transformation. Afterward, the transformed RGB image is sent into a density map-based crowd-counting network to estimate the crowd density map, and the counting result is calculated by summing the density over the whole image. The proposed HRPT comprises two main steps: rescaling factor calculation and height transformation. In the following, we first introduce each step of HRPT in detail and then briefly introduce the crowd-counting networks employed in our framework.

3.1. Rescaling factor calculation

The proposed HRPT is designed to narrow the height gap among the human heads in RGB images. To accomplish this goal, the rescaling factor should be inversely proportional to the height of the head in the original RGB image:

$$s = \frac{a_1}{h}, \tag{1}$$

where s denotes the rescaling factor; h denotes the head height in the original RGB image; and a_1 is a hyper-parameter that equals the rescaled head height.

According to Lian et al. (2019), the head height h is inversely proportional to the depth d. We formulate their relationship as follows:

$$h = \frac{a_2}{d}, \tag{2}$$

where a_2 is a parameter related to camera intrinsic parameters, such as focal length. After combining Equations (1, 2), we deduce the relationship between the rescaling factor s and depth d as follows:

$$s = \frac{a_1}{a_2} \cdot d. \tag{3}$$

As each pixel has a different depth value and the depth map always misses the depth value of some areas, such as the top-left area in the original depth map shown in Figure 2, transforming the crowd image according to the rescaling factors calculated by Equation (3) is hard. Fortunately, the depth usually complies with the following rule: pixels with smaller y-coordinates generally have higher depth values. Considering Equation (3) and the above rule, we conclude that it is possible to find the general relationship between the rescaling factor and y-coordinate. According to Rodriguez et al. (2011), under the assumption that people stand on the ground plane and the camera has no horizontal or in-plane rotation, the relationship between head height and y-coordinate is formulated as follows:

$$h = a_3 \cdot \left(y_o - \bar{y}_o\right), \tag{4}$$

where y_o denotes the y-coordinate in the original RGB image; $\bar{y}_o$ denotes the y-coordinate of the horizon in the original RGB image; and a_3 is a parameter related to camera extrinsic parameters, such as the camera height above the ground. We obtain the relationship between y-coordinate y_o and depth d by substituting Equation (4) into Equation (2):

$$\frac{1}{d} = \frac{a_3}{a_2} \cdot y_o - \frac{a_3}{a_2} \cdot \bar{y}_o. \tag{5}$$

We set $\frac{a_3}{a_2} = a_4$ and $-\frac{a_3}{a_2} \cdot \bar{y}_o = b_4$ to simplify Equation (5). Then, Equation (5) is rewritten as follows:

$$\frac{1}{d} = a_4 \cdot y_o + b_4. \tag{6}$$

According to Equation (6), $\frac{1}{d}$ is positively correlated with y_o. Next, we obtain the relationship between y-coordinate y_o and rescaling factor s by substituting Equation (6) into Equation (3):

$$s = \frac{a_1}{a_2 \cdot a_4} \cdot \frac{1}{y_o + \frac{b_4}{a_4}}. \tag{7}$$

We compute rescaling factors using Equation (7), where a_1 is a parameter whose value is determined by experience; a_2 is related to intrinsic camera parameters; and a_4 and b_4 are related to both extrinsic and intrinsic camera parameters. In general, camera intrinsic parameters can remain unchanged. The value of a_2 is the same for images captured by cameras with identical intrinsic parameters, and it can be determined by fitting Equation (2). In contrast, it is difficult to keep the camera's extrinsic parameters constant. As a result, the values of a_4 and b_4 should be recalculated for each image by fitting Equation (6) based on its associated depth map.

3.2. Height transformation

Height transformation is implemented according to the rescaling factors calculated by Equation (7), showing that each image row has a specific value of rescaling factor. Thus, height transformation can be performed by adjusting the height of each image row according to its rescaling factor. Figure 3 displays the height transformation principle using a toy example, where I_o denotes the original RGB image; I_t denotes the RGB image after height transformation. The heights and y-coordinates of image rows are displayed in I_o and I_t. These heights are measured in pixels. As shown in Figure 3, before height transformation, the height of each row is 1. After height transformation, the height of each row is multiplied by its corresponding rescaling factor. For example, in I_o, the height of the image row with y-coordinate y_o^* is 1. After

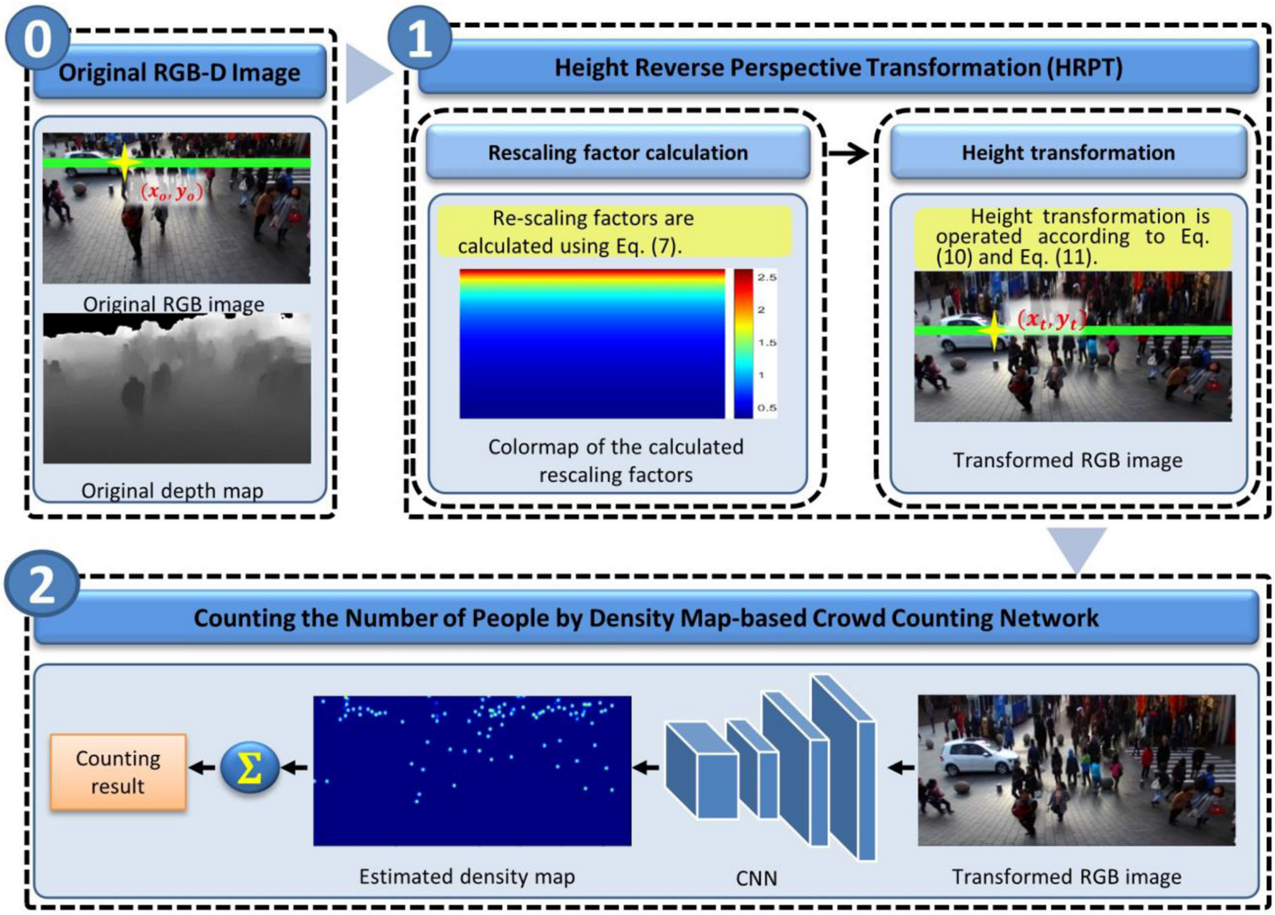
0
Original RGB-D Image
(x_o, y_o)
Original RGB image
Original depth map
1
Height Reverse Perspective Transformation (HRPT)
Rescaling factor calculation
Re-scaling factors are calculated using Eq. (7).
2.5
2
1.5
1
0.5
Colormap of the calculated rescaling factors
Height transformation
Height transformation is operated according to Eq. (10) and Eq. (11).
(x_t, y_t)
Transformed RGB image
2
Counting the Number of People by Density Map-based Crowd Counting Network
Counting result
Σ
Estimated density map
CNN
Transformed RGB image

FIGURE 2
Flowchart of our crowd-counting framework.

height transformation, the height of its corresponding row in I_t is changed to 1 $s\left(y_o^*\right) = s\left(y_o^*\right)$, and the y-coordinate is changed to $y_t^* = \sum_{y_o=1}^{y_o^*} s\left(y_o\right)$. $s\left(y_o\right)$ denotes the rescaling factor of the image row with y-coordinate y_o. It is calculated by Equation (7). $s\left(y_o^*\right) = s\left(y_o = y_o^*\right)$.

The toy example shown in Figure 3 depicts an ideal height transformation process. In the ideal process, the calculated heights and y-coordinates of image rows in I_t have a high probability of being decimals. However, in practice, they should be integers. Therefore, this ideal process cannot be implemented in practice. To solve this problem, we approximate this ideal height transformation process using some equations that can be easily implemented in practice.

As shown in Figure 3, the relationship between y_t^* and y_o^* is $y_t^* = \sum_{y_o=1}^{y_o^*} s\left(y_o\right)$. We use integration to replace summing:

$$y_t^* = \int_1^{y_o^*} s(y_o)dy_o. \tag{8}$$

Then, we substitute Equation (7) into Equation (8) and obtain the following expression:

$$y_t^* = \frac{a_1}{a_2 \cdot a_4} \cdot ln\left(\frac{a_4 \cdot y_o^* + b_4}{a_4 + b_4}\right). \tag{9}$$

In Equation (9), a_4, b_4, and y_o^* are above 0. Equation (9) can be reformulated as:

$$y_o^* = \frac{1}{a_4}\left(e^{\frac{a_2 \cdot a_4}{a_1} \cdot y_t^*}\left(a_4 + b_4\right) - b_4\right). \tag{10}$$

The y-coordinates of image rows in the RGB image after height transformation are integers, $y_t^* = 1, 2, 3, \ldots, H_t$, where H_t denotes the height of the RGB image after height transformation. For a particular y_t^*, we can use Equation (10) to calculate its corresponding y_o^*. Let us use $I_o(y_o^*)$ to denote the image row with y-coordinate y_o^* in I_o, and use $I_t(y_t^*)$ to denote the image row with y-coordinate y_t^* in I_t. In our height transformation process, the pixels in $I_o(y_o^*)$ are assigned to the pixels in $I_t(y_t^*)$. We use linear interpolation to calculate $I_o(y_o^*)$ since the calculated y_o^* has a high probability of being a decimal:

$$\begin{aligned} I_t\left(y_t^*\right) &= I_o\left(y_o^*\right) = \left(y_o^* - \lfloor y_o^* \rfloor\right) I_o\left(\lceil y_o^* \rceil\right) \\ &\quad + \left(\lceil y_o^* \rceil - y_o^*\right) I_o\left(\lfloor y_o^* \rfloor\right). \end{aligned} \tag{11}$$

where $\lfloor y_o^* \rfloor$ denotes the nearest integer lower than y_o^*; $\lceil y_o^* \rceil$ denotes the nearest integer higher than y_o^*, $\lceil y_o^* \rceil - \lfloor y_o^* \rfloor = 1$.

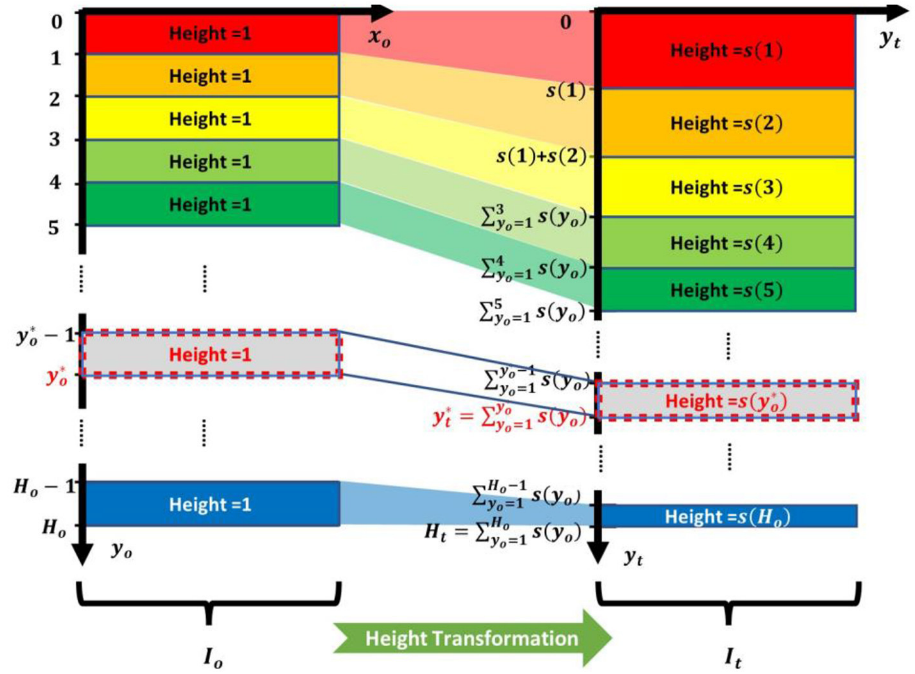

FIGURE 3
Height transformation process of a toy example. I_o denotes the original RGB image; H_o denotes the height of I_o; I_t denotes the RGB image after height transformation; $s(y_o)$ denotes the rescaling factor calculated using Equation (7), and $s(1) = s(y_o = 1), \ldots, s(y_o^*) = s(y_o = y_o^*), \ldots, s(H_o) = s(y_o = H_o)$. The height of I_t is $H_t = \sum_{y_o=1}^{H_o} s(y_o)$.

3.3. Crowd-counting networks

The proposed HRPT can be used as an image preprocessing step in crowd-counting networks. Then, this study selects four well-performing crowd-counting methods: CSRNet (Li et al., 2018), DM-Count (Wang B. et al., 2020), GL (Wan et al., 2021), and CLTR (Liang et al., 2022) with their brief introductions presented below:

CSRNet (Li et al., 2018) is a representative crowd-counting method based on density map estimation. It generates ground-truth density maps by blurring the dot annotations of human heads with Gaussian kernels. CSRNet uses VGG-16 (Simonyan and Zisserman, 2015) as its backbone and uses the L2 loss between predicted density map and ground-truth density map as its loss function. A popular tactic used in density map-based methodologies is to generate ground-truth density maps using Gaussian kernels. Their counting performance is strongly associated with the quality of generated ground-truth density maps (Ma et al., 2019). However, a recent study indicates that using Gaussian kernels is detrimental to the generalization performance (Wang B. et al., 2020). DM-Count (Wang B. et al., 2020) is proposed to solve the above problem by abandoning Gaussian kernels. It does not generate any ground-truth density maps in advance, and it uses optimally balanced transport to calculate the training loss between the predicted density map and dot annotations of human heads. GL (Wan et al., 2021) offers a similar technique to DM-Count (Wang B. et al., 2020). Differently from DM-Count (Wang B. et al., 2020), GL (Wan et al., 2021) uses unbalanced optimal transport, which preserves the predicted and annotated counts and generates pixel and point-wise loss.

The above three methods are based on Convolutional Neural Networks (CNNs). In recent years, transformer has been successfully used in many computer vision tasks and has achieved higher performance than CNN (Han et al., 2023). Therefore, we also employ a transformer-based crowd counting method, CLTR (Liang et al., 2022). CLTR takes image features extracted by CNN and trainable embeddings as input of a transformer-decoder. It directly predicts the localizations of human heads.

3.4. The processing steps of our method

The processing steps of our method are shown in the following Algorithm 1, where n denotes the n^{th} crowd image in the dataset; N denotes the total number of images in the dataset. The meanings of a_1, a_2, a_4, and b_4 have been introduced in Section 3.1. The meanings of y_o^*, y_t^*, $I_o(y_o^*)$, $I_t(y_t^*)$, H_t, and I_t have been introduced in Section 3.2.

As shown in the above algorithm, a_2 is set by fitting Equation (2), and a_4 and b_4 are set by fitting Equation (6). We need to use the depths in depth maps to fit these two equations. Thus, the depth information is used in the above step (2) and step (4).

4. Experiment

The proposed HRPT requires the perspective information in depth maps to accomplish image transformation. However, most public crowd-counting datasets only contain RGB images

```
Input: Crowd RGBD dataset, where images are captured by cameras with same intrinsic parameters.
(1) Set $a_1$ by experience.
(2) Set $a_2$ by fitting Equation (2) on the training subset. The value of $a_2$ is the same for all images in this dataset. We need to use the depths in depth maps to fit Equation (2).
(3) For $n$=1 to $N$ do
(4)   Set $a_4$ and $b_4$ by fitting Equation (6). The values of $a_4$ and $b_4$ are recalculated for each image.
      We need to use the depths in depth maps to fit Equation (6).
(5)   For $y_t^*$=1 to $H_t$ do
(6)     Calculate the corresponding $y_o^*$ of each $y_t^*$ according to Equation (10).
(7)     Calculate $I_o(y_o^*)$ using linear interpolation according to Equation (11).
(8)     Assign $I_o(y_o^*)$ to $I_t\left(y_t^*\right)$.
(9)   Count the number of people by sending $I_t$ to the crowd counting networks.
Output: The crowd-counting results.
```

Algorithm 1. Our proposed crowd-counting method.

(Wang Q. et al., 2020). Fortunately, Lian et al. (2019) released a large RGB-D crowd-counting dataset called ShanghaiTechRGBD in 2019, comprising 1193 training images and 1000 testing images. Most of our experiments are done on this RGBD dataset. Besides, our method can be extended to the RGB dataset by predicting the depth maps of the RGB images. We choose ShanghaiTech PartB dataset (Zhang et al., 2016) to evaluate the performance of our method on the RGB dataset. Our experiments are implemented with Pytorch framework. We use one Nvidia RTX 2080ti GPU and one Intel Core i7 9700k CPU.

The image transformation and crowd-counting performances of our method are discussed in the subsequent subsections.

4.1. Performance of image transformation

We use our experience to set a_1 to 40. The counting performances with different values of a_1 are shown in Section 4.2.4. Then, a_2 is set to 350 by fitting Equation (2), and its value is the same for all images. Using the least-square algorithm, a_4 and b_4 are set by fitting Equation (6). Their values differ based on different images. The distributions of a_4 and b_4 are shown in Section 4.2.4. HRPT uses image processing to narrow the height gap among human heads. Figure 4 depicts three groups of crowd RGB images to demonstrate the effectiveness of HRPT. The heads of three pedestrians are marked in each RGB image. Their pixel-measured head heights are indicated on the left side of the images. HRPT stretches the top areas of RGB images and shrinks the bottom areas of RGB images, as shown in Figure 4. By comparing the head heights shown in Figure 4, we observe that HRPT successfully narrows the height gap among human heads. After HRPT, the head heights are approximated to the value of a_1.

In Figure 5, we compare the head heights to demonstrate the effectiveness of HRPT. As shown in this figure, HRPT successfully enlarges the heights of small heads in far areas, reduces the heights of large heads in near areas, and narrows the height gap among human heads.

4.2. Performance of crowd counting

4.2.1. Training details and metrics

To verify the effectiveness of HRPT on crowd counting, we combine four well-performing crowd-counting neural networks, CSRNet (Li et al., 2018), DM-Count (Wang B. et al., 2020), GL (Wan et al., 2021), and CLTR (Liang et al., 2022), with HRPT. During training of these four neural networks, Adam (Kingma and Ba, 2014) is used as the optimizer, and the learning rate is set to 1×10^{-5}. The performance of different methods is evaluated using the mean absolute error (MAE) and mean square error (MSE), as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i^p - C_i^{gt}\right|, \tag{12}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i^p - C_i^{gt}\right)^2}, \tag{13}$$

where N is the number of testing images; C_i^p and C_i^{gt} are the predicted and ground-truth number of people in i^{th} testing image, respectively.

4.2.2. Comparisons to the baselines

This study combines HRPT with four well-performing crowd-counting methods: CSRNet (Li et al., 2018), DM-Count (Wang B. et al., 2020), GL (Wan et al., 2021), and CLTR (Liang et al., 2022). Therefore, we use CSRNet, DM-Count, GL, and CLTR as our baselines. Comparisons to these four baselines are shown in Table 1.

As shown in Table 1, CLTR is built based on transformer (Liang et al., 2022), while CSRNet, DM-Count, and GL are built based on CNN. Many studies have proved that transformer has higher performance than other types of networks, such as CNN (Han et al., 2023). However, the transformer-based crowd-counting method CLTR has poorer performance than the other three CNN-based methods in our experiments. This is because, transformer models are more sensitive to the hyper-parameters for training, such as batch size, and are huger and more computationally expensive (Han et al., 2023). However, we only have one Nvidia RTX 2080ti GPU. To train CLTR on our device, we set a much smaller batch size than the authors of CLTR. As a result, the maximum performance of CLTR is not achieved. Even so, our experimental results still demonstrate that HRPT can improve the crowd-counting performance of CLTR.

Table 1 also demonstrates that HRPT offers a more significant improvement to CSRNet than to DM-Count and GL. This is because CSRNet uses Gaussian kernels to generate ground-truth density maps. Its performance is highly dependent on the quality of the generated ground-truth density maps, which is reduced due

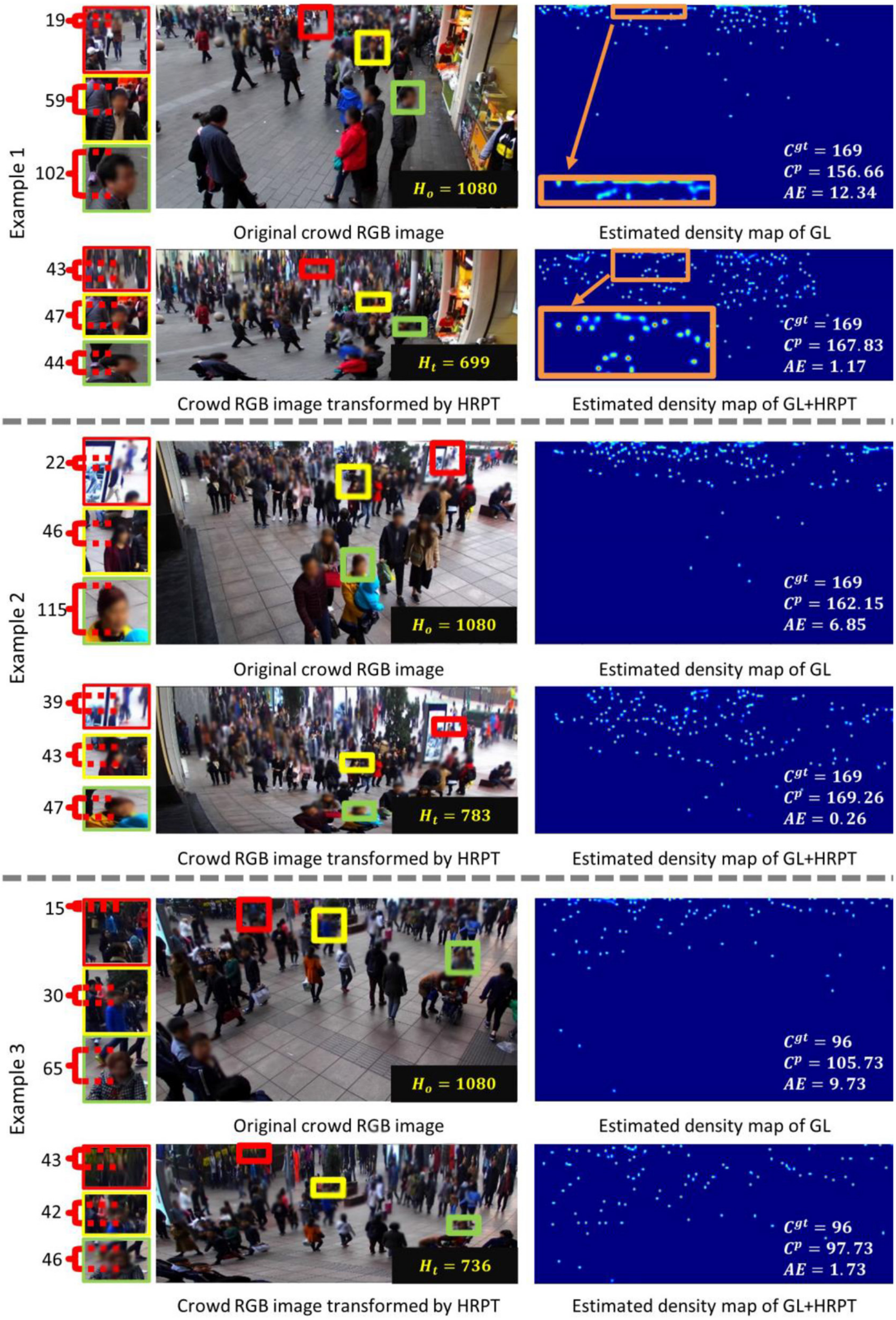

FIGURE 4
Three examples show the effectiveness of HRPT. The heads of three pedestrians are marked in each RGB image. Their head heights, measured in pixels, are shown on the left of the images. In this figure, H_o denotes the height of the original crowd RGB image; H_t denotes the height of the crowd RGB image transformed by HRPT; C^{gt} denotes the ground-truth number of people; C^p denotes the predicted number of people; and AE denotes the absolute error.

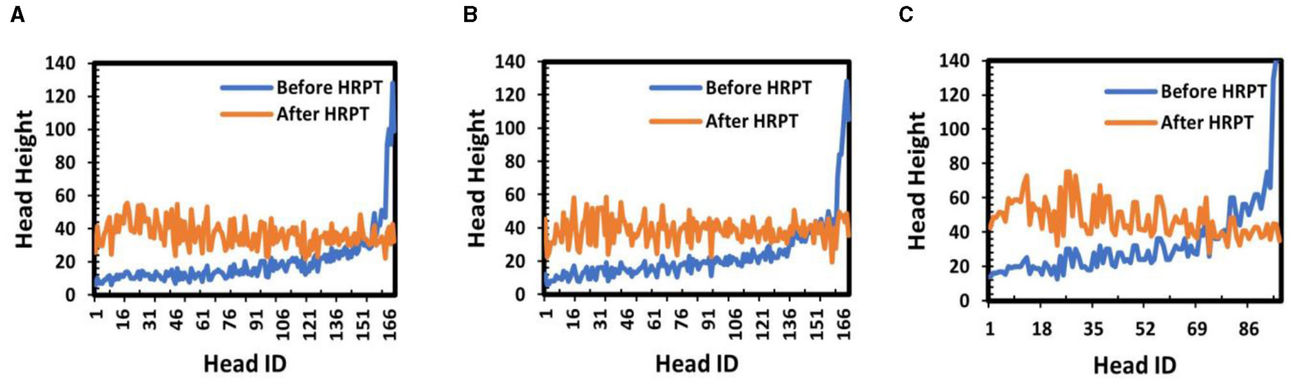

FIGURE 5
Comparisons of head heights to show the effectiveness of HRPT: **(A)** head heights in Example 1 of Figure 4; **(B)** head heights in Example 2 of Figure 4; and **(C)** head heights in Example 3 of Figure 4.

TABLE 1 **Comparisons to the baselines on the ShanghaiTechRGBD dataset.**

Methods		Architecture	Speed (ms)		MAE	MSE
			Preprocessing (CPU)	Neural network (GPU)		
Baselines	CLTR (Liang et al., 2022)	Transformer	0	249	5.29	7.70
	CSRNet (Li et al., 2018)	CNN	0	164	5.11	9.99
	DM-Count (Wang B. et al., 2020)	CNN	0	149	4.00	5.95
	GL (Wan et al., 2021)	CNN	0	152	3.96	5.97
Ours	**CLTR+HRPT**	Transformer	268	173	4.61	6.71
	CSRNet + HRPT	CNN	268	124	3.76	5.65
	DM-Count + HRPT	CNN	268	108	3.78	5.59
	GL + HRPT	CNN	268	111	**3.70**	**5.42**

Scores marked in bold indicate the best results on the corresponding metric.

to the scale variations of human heads (Ma et al., 2019). The above quality reduction is narrowed since HRPT can narrow the height gap among human heads. In contrast, DM-Count and GL do not use Gaussian kernels to generate ground-truth density maps. Their sensitivities to the scale variations of human heads are smaller than those of CSRNet. Therefore, HRPT offers a more considerable improvement to CSRNet than to DM-Count and GL. In addition, because GL uses unbalanced optimal transport, which preserves the predicted and annotated counts, it has better performance than CSRNet and DM-Count. Therefore, GL+HRPT achieves the best crowd counting performance.

Besides the evaluation scores of MAE and MSE, the processing speeds are also shown in Table 1. The baselines do not employ the preprocessing step. Therefore, they spend 0 ms in the preprocessing step. Our methods use HRPT as the preprocessing step. HRPT spends 268 ms for each image, which is a considerable amount of time. This is because our current version of HRPT is operated in CPU. In the future, we will put the for-loop in step (5) of Algorithm 1 in GPU to increase the processing speed. In addition, as shown in Table 1, after HRPT, the processing speeds of neural networks are faster than the baselines. This is because HRPT effectively reduces much redundancy information in the near areas. Image examples in Figure 4 show that, in the original crowd RGB images, the near areas contain much fewer people but take up much more image spaces. In contrast, in the crowd RGB images transformed by HRPT, the near areas are shrunken by a large margin. The average image size is reduced by HRPT.

Figure 4 also depicts the density map estimation results of GL and GL+HRPT, displaying the effectiveness of HRPT on crowd counting qualitatively. By comparing these density map estimation results, we can see that HRPT distributes human heads more evenly. In the density map estimation results of Example 1, a couple of corresponding regions located at the top of images are marked. After comparing these two regions, it becomes clear that without HRPT, it is hard to distinguish between human heads that are far from the camera. However, with HRPT, it is much easier to identify these heads based on the estimated density map. Although these two regions have different heights, they correspond to the same area in the actual scene.

As GL+HRPT achieves the best performance, in the following, we focus on the experiments of GL+HRPT, and use **ours** to denote GL+HRPT.

4.2.3. Comparisons to other methods

In this subsection, we compare our method with other crowd-counting methods that also use depth maps. Evaluation results of

TABLE 2 Comparisons to other crowd-counting methods that also use depth maps on the ShanghaiTechRGBD dataset.

Methods	Year	Backbones	MAE	MSE
RDNet (Lian et al., 2019)	2019	ResNet-101 + VGG-16	4.96	7.22
CSRNet + IADM (Liu et al., 2021)	2021	VGG-16 × 3	4.38	7.06
DPDNet (Lian et al., 2022)	2022	ResNet-101 + VGG-16	4.23	6.75
Cross-model (Zhang et al., 2021)	2021	VGG-16 × 2	3.76	5.46
Li et al. (2023)	2023	VGG-16 × 2	4.03	5.81
CCANet (Liu et al., 2023)	2023	VGG-16 + Designed	3.78	5.56
Ours	2023	VGG-19	**3.70**	**5.42**

Scores marked in bold indicate the best results on the corresponding metric.

different methods on the ShanghaiTechRGBD dataset are shown in the following Table 2.

In addition to the evaluation results of different methods, Table 2 also displays their type of backbones. In Table 2, "Designed" denotes that the authors designed the network backbone; "VGG-16 × 3" depicts that the network contains three branches whose backbones are VGG-16 (Simonyan and Zisserman, 2015), as do "VGG-16 × 2"; "ResNet-101 + VGG-16" means that the network contains two branches whose backbones are ResNet-101 (He et al., 2016) and VGG-16. Similarly, "VGG-16 + Designed" means that the network contains two branches whose backbones are VGG-16 and Designed.

As shown in Table 2, our method outperforms other methods that also employ depth maps. This demonstrate that, although employing depth maps improves the crowd-counting results, different methods have different performances. Compared with other methods, the proposed HRPT more efficiently uses depth maps and improves the crowd-counting performance. Moreover, other depth map methods use complex network backbones because they employ additional network branches to deal with depth maps. In contrast, our method uses depth maps in the HRPT module, serving as a preprocessing step of an excellent crowd-counting network. Therefore, our method can use depth maps without changing any part of the original network structure. Thus, the backbone used in our method is much simpler than those used in other methods, which also use depth maps.

4.2.4. Study on image transformation parameters

a_1 is a hyper-parameter of HRPT. As shown in Equation (1), it represents the head height after HRPT. To study its effectiveness on crowd-counting performance, we change its value to 10, 20, 30, and 40, and then evaluate the corresponding crowd-counting results. Evaluation results with different values of a_1 are shown in Table 3.

Table 3 shows that with the increase of a_1, the crowd-counting performance improves. a_1 represents the head heights after HRPT. The smaller the value of a_1, the more image details are lost by HRPT, which is detrimental to crowd counting. When we set a_1 to 50, some images in our experimental dataset will become very large and cannot be processed by the crowd-counting network on our device. Therefore, we finally set a_1 to 40.

In addition, the processing speeds in Table 3 show that, with the increase of a_1, the processing speed becomes slower and slower. This is because, the larger the value of a_1, the larger the images outputted by HRPT. Then, the preprocessing step and neural network need to spend more time to deal with these images.

TABLE 3 Evaluation results with different values of a_1 on the ShanghaiTechRGBD dataset.

a_1	Speed (ms)		MAE	MSE
	Preprocessing (CPU)	Neural network (GPU)		
10	147	29	4.43	6.63
20	166	55	3.74	5.51
30	188	81	3.70	5.60
40	268	111	**3.70**	**5.42**

Scores marked in bold indicate the best results on the corresponding metric.

a_4 and b_4 are set by fitting Equation (6). Their values differ based on different images. Their distributions on the ShanghaiTechRGBD dataset are shown in Figure 6. As shown in this figure, in both the training and testing datasets, the maximum number of a_4 falls into $(3, 4) \times 10^{-4}$, and the maximum number of b_4 falls into $[2.5, 4.0] \times 10^{-2}$. In addition, the distribution of a_4 on the training subset is similar to the distribution of a_4 on the testing subset; the distribution of b_4 on the training subset is similar to the distribution of b_4 on the testing subset.

4.2.5. Evaluation on RGB dataset

Our method can be extended to the RGB dataset by predicting the depth maps of the RGB images (Lian et al., 2022). HRPT estimates the relationship between the y-coordinate and rescaling factor. It is built under the assumption that in each image, people stand on the same ground plane and the camera has no horizontal rotation. Therefore, HRPT suits images captured from flat areas with surveillance views, and it requires the horizontal lines of captured images to be parallel to the image rows. In addition, HRPT narrows the head height gap by stretching the far areas. If some rows in the image are close to or above the horizontal lines, the depths of these rows are infinite. According to Equation (3), the rescaling factors of these rows are also infinite. Thus, HRPT does not suit images that contain image rows close to or above the horizontal lines. Images in the ShanghaiTech PartB dataset (Zhang et al., 2016) satisfy the above requirements. Therefore, we choose

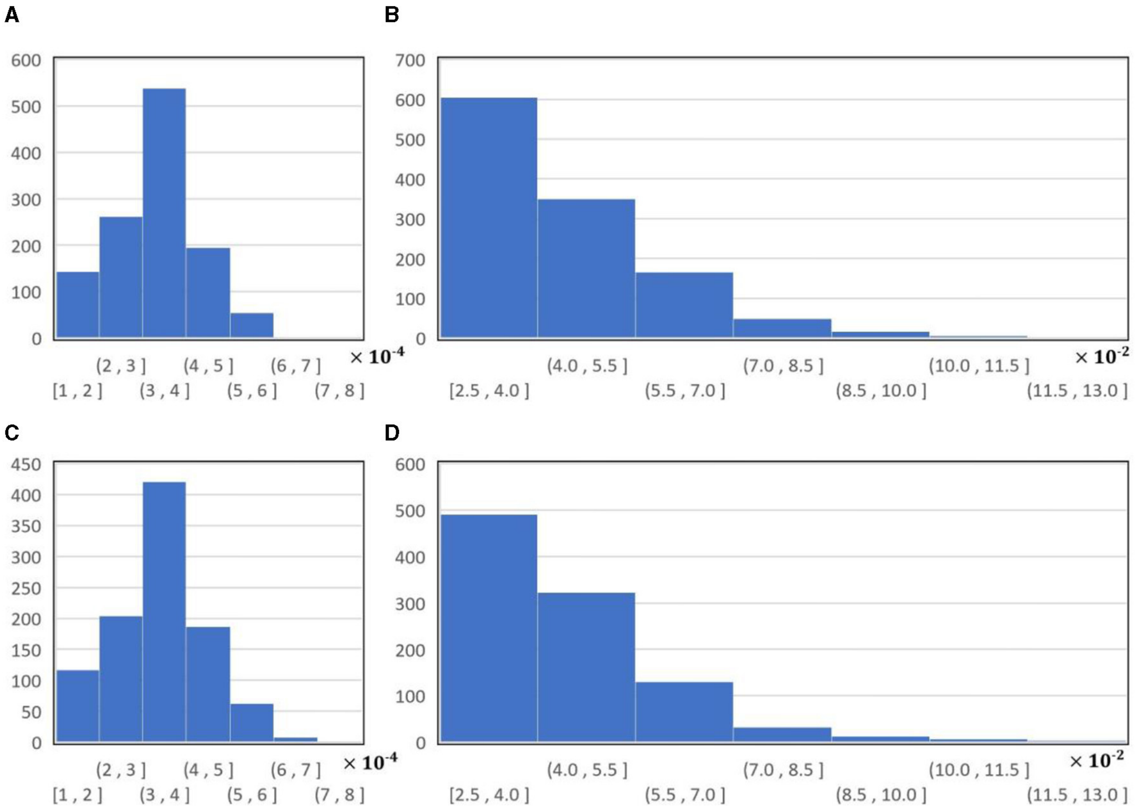

FIGURE 6
The histograms of a_4 and b_4 on ShanghaiTechRGBD dataset: **(A)** the histogram of a_4 on the training subset; **(B)** the histogram of b_4 on the training subset; **(C)** the histogram of a_4 on the testing subset; **(D)** the histogram of b_4 on the testing subset.

ShanghaiTech PartB to evaluate the performance of our method on the RGB dataset. Evaluation results of our method and many other well-performing methods are shown in Table 4. As shown in this table, our method achieves high performance.

TABLE 4 **Evaluation results of different methods on the ShanghaiTech PartB dataset.**

Methods	MAE	MSE
MCNN (Zhang et al., 2016)	26.4	41.3
DecideNet (Liu et al., 2018)	20.8	29.4
CSRNet (Li et al., 2018)	10.6	16.0
RDNet (Lian et al., 2019)	8.8	12.9
DM-Count (Wang B. et al., 2020)	7.4	11.8
GL (Wan et al., 2021)	7.3	11.7
Cross-model (Zhang et al., 2021)	8.3	12.9
CCANet (Liu et al., 2023)	8.1	13.5
DPDNet (Lian et al., 2022)	7.9	12.4
AutoScale (Xu et al., 2022)	**6.8**	11.3
Ours	**6.8**	**11.2**

Scores marked in bold indicate the best results on the corresponding metric.

5. Discussion

In this section, we first analyze why the proposed HRPT can improve the crowd-counting performance from another point of view. Afterward, we discuss another two image transformation methods, the drawbacks of our methods and our future work.

5.1. Analysis

The proposed HRPT is designed to improve the crowd-counting performance by alleviating the scale variation problem. In the following, we analyze why the proposed HRPT can improve the crowd-counting performance from another point of view.

First, we analyze the effect of HRPT on the bottom areas of crowd RGB images. Human heads in the bottom areas are large and sparse before HRPT, as shown in Figure 4. Larger human heads contain much more detailed information. However, current crowd-counting methods do not require so much detailed information to pick out human heads. Therefore, the bottom areas of crowd RGB images contain much redundant information. Moreover, large and sparse human heads occupy too much image space, making counting networks spend too much energy on those "easy samples." After HRPT, the bottom areas are shrunken to shorten the heights of human heads. Shrinking the bottom areas reduces redundant information and compels counting networks to pay more attention to the top areas with more "hard samples."

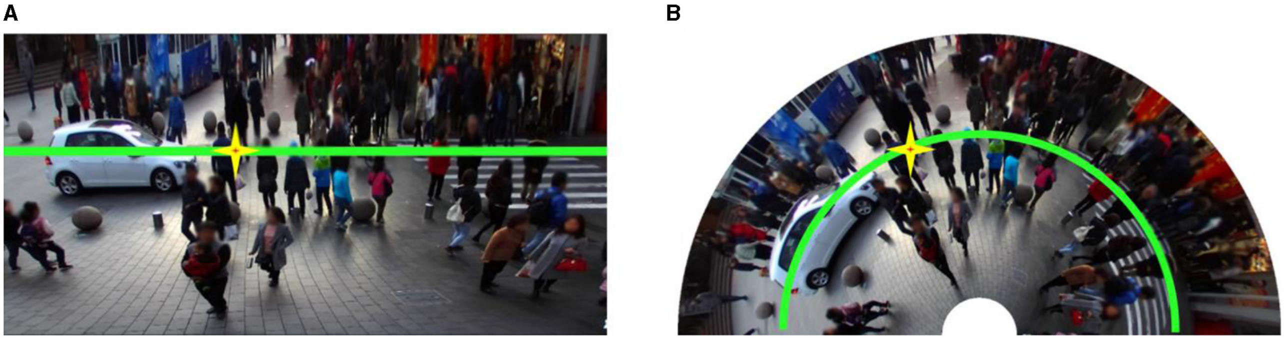

FIGURE 7
A group of images to show the effect of narrowing the width gap among human heads by changing a rectangle-shaped image into a fan-shaped image: **(A)** before shape changing; **(B)** after shape changing. In this figure, the height gap among human heads has been narrowed by HRPT.

Next, we analyze the effect of HRPT on the top areas of crowd RGB images. As shown in Figure 4, before HRPT, human heads in the top areas are tiny and dense. Information about these small human heads may be lost when extracting high-level image features through crowd-counting networks. After HRPT, the top areas are stretched to enlarge the heights of human heads. Even though human heads in these top areas become thinner, they can still be identified as human heads. The possibility of losing their information while extracting high-level image features becomes much smaller. Therefore, the proposed HRPT can improve crowd-counting performance.

5.2. Other two image transformation methods

5.2.1. Shape-changing method

Section 4 reports the experimental results demonstrating that narrowing the height gap among human heads helps improve crowd-counting performance. What about narrowing the width gap among human heads? If we attempt to narrow the width gap by changing each row's width according to the rescaling factor calculated from Equation (7), the shape of the RGB image will be changed from rectangle to trapezoid. Moreover, human heads in the top-right and top-left regions will be seriously deformed and lose their essential characteristics. To avoid this severe deformation, we narrow the width gap among human heads by changing rectangle-shaped images into fan-shaped images. This shape-changing method naturally enlarges the widths of human heads in the top regions and shortens the widths in the bottom regions. Figure 7 displays a group of crowd RGB images to show the effect of narrowing the width gap among human heads through shape changing.

As shown in Figure 7, this shape-changing method is operated by changing the shape of each image row from straight to semi-circle, and the width gap among human heads is successfully narrowed after shape changing. However, this shape-changing method has a shortcoming: it adds additional rotation. We combine this shape-changing method with GL and GL + HRPT to evaluate its effect on crowd-counting performance. The evaluation results are shown in Table 5.

TABLE 5 **Evaluation results of shaping changing method on ShanghaiTechRGBD dataset.**

Methods	MAE	MSE
GL	3.96	5.97
GL + Shape-Changing	4.36	6.43
GL + HRPT	**3.70**	**5.42**
GL + HRPT + Shape-Changing	4.27	6.30

Scores marked in bold indicate the best results on the corresponding metric.

As shown in Table 5, after we combine this shape-changing method with GL, MAE and MSE rise to 4.36 and 6.43, respectively; after we combine this shape-changing method with GL + HRPT, MAE and MSE rise to 4.27 and 6.30, respectively. These experimental results demonstrate that this shape changing negatively affects crowd counting, implying that the crowd-counting network is not robust enough to handle the additional rotation.

5.2.2. Geometric method

The perspective effect mainly causes the scale variation problem in crowd images. Accurately eliminating the perspective effect on the 3D world is very hard in surveillance scenes. In contrast, homographs can quickly eliminate the perspective effect on the 2D ground (Hartley and Zisserman, 2003). Here, we test whether the homograph can narrow the scale gap (both height and width gap) among human heads by eliminating the perspective effect on the 2D ground. Figure 8 displays a group of crowd RGB images to show the effect of the homograph.

We zoom in on two areas of Figure 8B to display the effect of the homograph. We can observe the shape of floor tiles in these two areas. As shown in Figure 8, the shape of the floor tiles is trapezoid before the homograph, and these tiles have different sizes. After homograph, the shape of the floor tiles returns to square, and these floor tiles have the same size. This result demonstrates that the homograph can eliminate the perspective effect on the 2D ground and narrow the scale gap (both height gap and width gap) among floor tiles. However, as shown in Figure 8B, homograph cannot narrow the scale gap among human heads. After the

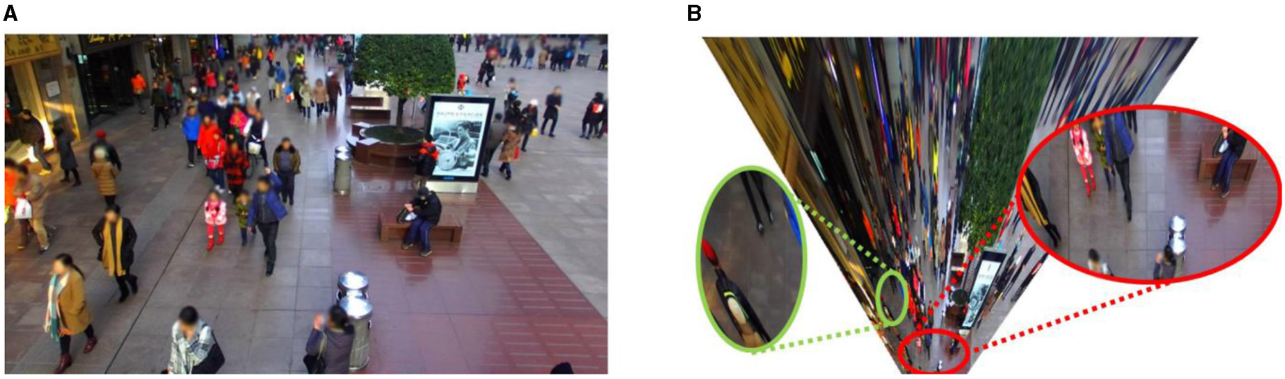

FIGURE 8
A group of crowd RGB images to show the effect of eliminating the perspective effect by homograph: **(A)** the original crowd RGB image; **(B)** the crowd RGB image transformed by homograph.

homograph, the human heads far from the camera are stretched too much, and those near the camera are shrunken too much. Additionally, human heads at the top-right and top-left regions are seriously deformed and lose their essential characteristics. Severe deformation is harmful to crowd counting. The homograph transformation result shown in Figure 8B is similar to a frame of crowd video rectified by estimated scene geometry in Rodriguez et al. (2011).

5.3. Drawbacks and future works

This study proposes HRPT to narrow the height gap among human heads. HRPT narrows the head height gap by stretching the far areas. If some rows in the image are close to or above the horizontal lines, the depths of these rows are infinite. According to Equation (3), the rescaling factors of these rows are also infinite. Thus, we cannot transform them by our HRPT. Furthermore, HRPT is built under the assumption that in each image, people stand on the same ground plane and the camera has no horizontal rotation. Therefore, it only suits images captured from flat areas with surveillance views, and it requires the horizontal lines of captured images to be parallel to the image rows. The above requirements limit the wide usage of HRPT. In the future, we plan to address the above first problem by a segmentation method to remove image rows above the horizontal lines. Moreover, we plan to address the above second problem by building a more advanced image transformation model that suits more scenarios and does not require the horizontal lines of captured images to be parallel to the image rows.

In addition, as mentioned above, we do not achieve satisfactory results when narrowing the width gap among human heads. In the future, we will improve crowd-counting performance by proposing a more efficient image transformation method to narrow the width gap among human heads and an efficient crowd-counting network to handle the additional rotation, as shown in Figure 7B.

As shown in Section 4.2.5, our method can be extended to the RGB dataset by predicting the depth maps of the RGB images. However, owing to the limitation of monocular vision, single image-based depth prediction methods only provide relative depth information (Lian et al., 2022), which limits their depth prediction accuracy. Previous research has demonstrated that embedding focal length can overcome the problem of single image-based depth prediction and acquire accurate depths (He et al., 2018). In the future, we will build a crowd-counting dataset with known focal lengths, then use these focal lengths to predict accurate depths, and finally associate these accurate depth prediction results with our method to further improve the crowd-counting accuracy on RGB images.

Besides, as shown in Table 1, HRPT spends 268 ms for each image, which is a considerable amount of time. Our current version of HRPT is operated in CPU. In the future, we will put the for-loop in step (5) of Algorithm 1 in GPU to increase the processing speed of HRPT.

6. Conclusions

This study uses HRPT to alleviate the scale variation problem in crowd-counting tasks. HRPT creatively narrows the height gap among human heads by using the perspective information contained in the depth map. Moreover, it enlarges small human heads in far areas to make them more visible, and it shrinks large human heads in closer areas to reduce redundant information. Other excellent crowd-counting methods can easily employ HRPT as a preprocessing step. Experimental results show that our method achieves high crowd-counting performance.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/svip-lab/RGBD-counting.

Author contributions

XZ: Funding acquisition, Methodology, Project administration, Software, Writing—original draft. HL: Methodology, Software, Writing—original draft. ZZ: Formal analysis, Writing—review and editing. SL: Formal analysis, Writing—review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Natural Science Foundation of Shandong Province (grant ZR2021QF094) and the Youth Innovation Team Technology Project of Higher School in Shandong Province (2022KJ204).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Du, Z., Shi, M., Deng, J., and Zafeiriou, S. (2023). Redesigning multi-scale neural network for crowd counting. *IEEE T Image Proc.* 32, 3664–3678. doi: 10.1109/TIP.2023.3289290

Fan, Z., Zhang, H., Zhang, Z., Lu, G., Zhang, Y., Wang, Y., et al. (2022). A survey of crowd counting and density estimation based on convolutional neural network. *Neurocomputing* 472, 224–251. doi: 10.1016/j.neucom.2021.02.103

Gao, G., Gao, J., Liu, Q., Wang, Q., and Wang, Y. (2020). CNN-based density estimation and crowd counting: a survey. *arXiv preprint* arXiv:2003.12783. doi: 10.48550/arXiv.2003.12783

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2023). A survey on vision transformer. *IEEE T Pattern Anal.* 45, 87–110. doi: 10.1109/TPAMI.2022.3152247

Hartley, R., and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge, MA: Cambridge University Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

He, L., Wang, G., and Hu, Z. (2018). Learning depth from single images with deep neural network embedding focal length. *IEEE T Image Process.* 27, 4676–4689. doi: 10.1109/TIP.2018.2832296

Idrees, H., Soomro, K., and Shah, M. (2015). Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE T Pattern Anal.* 37, 1986–1998. doi: 10.1109/TPAMI.2015.2396051

Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., et al. (2019). "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 6126–6135.

Jiang, X., Zhang, L., Lv, P., Guo, Y., Zhu, R., Li, Y., et al. (2020). Learning multi-level density maps for crowd counting. *IEEE T Neur. Net. Lear.* 31, 2705–2715. doi: 10.1109/TNNLS.2019.2933920

Jing, L., Wang, T., Zhao, M., and Wang, P. (2017). An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. *Sensors* 17, 414. doi: 10.3390/s17020414

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980

Li, H., Zhang, S., and Kong, W. (2023). RGB-D crowd counting with cross-modal cycle-attention fusion and fine-coarse supervision. *IEEE T Ind. Inform.* 19, 306–316. doi: 10.1109/TII.2022.3171352

Li, Y., Zhang, X., and Chen, D. (2018). "CSRNet: dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 1091–1100.

Lian, D., Chen, X., Li, J., Luo, W., and Gao, S. (2022). Locating and counting heads in crowds with a depth prior. *IEEE T Pattern Anal.* 44, 9056–9072. doi: 10.1109/TPAMI.2021.3124956

Lian, D., Li, J., Zheng, J., Luo, W., and Gao, S. (2019). "Density map regression guided detection network for RGB-D crowd counting and localization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 1821–1830.

Liang, D., Xu, W., and Bai, X. (2022). "An end-to-end transformer model for crowd localization," in *Proceedings of the European Conference on Computer Vision* (Cham: Springer), 38–54.

Liu, B., and Vasconcelos, N. (2015). "Bayesian model adaptation for crowd counts," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 4175–4183.

Liu, J., Gao, C., Meng, D., and Hauptmann, A. G. (2018). "Decidenet: counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 5197–5206.

Liu, L., Chen, J., Wu, H., Li, G., Li, C., and Lin, L. (2021). "Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 4821–4831.

Liu, W., Salzmann, M., and Fua, P. (2019). "Context-aware crowd counting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 5094–5103.

Liu, Y., Cao, G., Shi, B., and Hu, Y. (2023). CCANet: a collaborative cross-modal attention network for RGB-D crowd counting. *IEEE T Multimedia* 25, 1–12. doi: 10.1109/TMM.2023.3262978

Liu, Y., Shi, M., Zhao, Q., and Wang, X. (2019). "Point in, box out: beyond counting persons in crowds," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 6462–6471.

Ma, Y., Sanchez, V., and Guha, T. (2022). "Fusioncount: efficient crowd counting via multiscale feature fusion," in *Proceedings of the International Conference on Image Processing* (Bordeaux: IEEE), 3256–3260.

Ma, Z., Wei, X., Hong, X., and Gong, Y. (2019). "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE International Conference on Computer Vision* (Seoul: IEEE), 6141–6150.

Rodriguez, M., Laptev, I., Sivic, J., and Audibert, J.-Y. (2011). "Density-aware person detection and tracking in crowds," in *Proceedings of the IEEE International Conference on Computer Vision* (Barcelona: IEEE), 2423–2430.

Shang, C., Ai, H., and Bai, B. (2016). "End-to-end crowd counting via joint learning local and global count," in *Proceedings of the International Conference on Image Processing* (Phoenix, AZ: IEEE), 1215–1219.

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556

Sindagi, V. A., and Patel, V. M. (2018). A survey of recent advances in CNN-based single image crowd counting and density estimation. *Pattern Recogn Lett.* 107, 3–16. doi: 10.1016/j.patrec.2017.07.007

Stewart, R., Andriluka, M., and Ng, A. Y. (2016). "End-to-end people detection in crowded scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2325–2333.

Wan, J., Liu, Z., and Chan, A. B. (2021). "A generalized loss function for crowd counting and localization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 1974–1983. doi: 10.1109/CVPR46437.2021.00201

Wang, B., Liu, H., Samaras, D., and Nguyen, M. H. (2020). "Distribution matching for crowd counting," in *Advances in Neural Information Processing Systems* (Beijing).

Wang, C., Zhang, H., Yang, L., Liu, S., and Cao, X. (2015). "Deep people counting in extremely dense crowds," in *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference,* 1299–1302.

Wang, M., Cai, H., Han, X., Zhou, J., and Gong, M. (2023). STNet: scale tree network with multi-level auxiliator for crowd counting. *IEEE T Multimedia*. 25, 2074–2084. doi: 10.1109/TMM.2022.3142398

Wang, Q., Gao, J., Lin, W., and Li, X. (2020). NWPU-crowd: a large-scale benchmark for crowd counting and localization. *IEEE T Patt. Anal.* 43, 2141–2149. doi: 10.1109/TPAMI.2020.3013269

Wang, W., Liu, Q., and Wang, W. (2022). Pyramid-dilated deep convolutional neural network for crowd counting. *Appl. Intell.* 52, 1825–1837. doi: 10.1007/s10489-021-02537-6

Wang, Z., Wu, Y., and Niu, Q. (2019). Multi-sensor fusion in automated driving: a survey. *IEEE Access*. 8, 2847–2868. doi: 10.1109/ACCESS.2019.2962554

Xu, C., Liang, D., Xu, Y., Bai, S., Zhan, W., Bai, X., et al. (2022). Autoscale: learning to scale for crowd counting. *Int. J. Comput. Vision*. 130, 405–434. doi: 10.1007/s11263-021-01542-z

Xu, M., Ge, Z., Jiang, X., Cui, G., Lv, P., Zhou, B., et al. (2019). Depth information guided crowd counting for complex crowd scenes. *Pattern Recogn. Lett*. 125, 563–569. doi: 10.1016/j.patrec.2019.02.026

Yan, Z., Zhang, R., Zhang, H., Zhang, Q., and Zuo, W. (2022). Crowd counting via perspective-guided fractional-dilation convolution. *IEEE T Multimedia*. 24, 2633–2647. doi: 10.1109/TMM.2021.3086709

Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., and Sebe, N. (2020). "Reverse perspective network for perspective-aware object counting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 4373–4382.

Zhang, C., Li, H., Wang, X., and Yang, X. (2015). "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 433–841.

Zhang, J., and Li, J. (2022). "Perspective-guided point supervision network for crowd counting," in *2022 International Conference on High Performance Big Data and Intelligent Systems, HDIS 2022* (Tianjin: IEEE), 212–217.

Zhang, S., Li, H., and Kong, W. (2021). A cross-modal fusion based approach with scale-aware deep representation for RGB-D crowd counting and density estimation. *Exp. Syst. Appl.* 180, 115071. doi: 10.1016/j.eswa.2021.115071

Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 589–597.

Zhao, M., Zhang, C., Zhang, J., Porikli, F., Ni, B., Zhang, W., et al. (2019). Scale-aware crowd counting via depth-embedded convolutional neural networks. *IEEE T Circ. Syst. Vid.* 30, 3651–3662. doi: 10.1109/TCSVT.2019.2943010

Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., and Tan, M. (2021). "Perception-aware multi-sensor fusion for 3D lidar semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision* (Montreal, QC: IEEE), 16260–16270.