MSMVD: Exploiting Multi-scale Image Features via Multi-scale BEV Features for Multi-view Pedestrian Detection

Taiga Yamane
taiga.yamane@ntt.com
Satoshi Suzuki
satoshixv.suzuki@ntt.com
Ryo Masumura
ryo.masumura@ntt.com
Shota Orihashi
shota.orihashi@ntt.com
Tomohiro Tanaka
tomohiro.tanaka@ntt.com
Mana Ihori
mana.ihori@ntt.com
Naoki Makishima
naoki.makishima@ntt.com
Naotaka Kawata
naotaka.kawata@ntt.com

NTT Human Informatics Laboratories
NTT Corporation
Yokosuka, Japan

Abstract

Multi-View Pedestrian Detection (MVPD) aims to detect pedestrians in the form of a bird's eye view (BEV) from multi-view images. In MVPD, end-to-end trainable deep learning methods have progressed greatly. However, they often struggle to detect pedestrians with consistently small or large scales in views or with vastly different scales between views. This is because they do not exploit multi-scale image features to generate the BEV feature and detect pedestrians. To overcome this problem, we propose a novel MVPD method, called Multi-Scale Multi-View Detection (MSMVD). MSMVD generates multi-scale BEV features by projecting multi-scale image features extracted from individual views into the BEV space, scale-by-scale. Each of these BEV features inherits the properties of its corresponding scale image features from multiple views. Therefore, these BEV features help the precise detection of pedestrians with consistently small or large scales in views. Then, MSMVD combines information at different scales of multiple views by processing the multi-scale BEV features using a feature pyramid network. This improves the detection of pedestrians with vastly different scales between views. Extensive experiments demonstrate that exploiting multi-scale image features via multi-scale BEV features greatly improves the detection performance, and MSMVD outperforms the previous highest MODA by 4.5 points on the GMVD dataset.

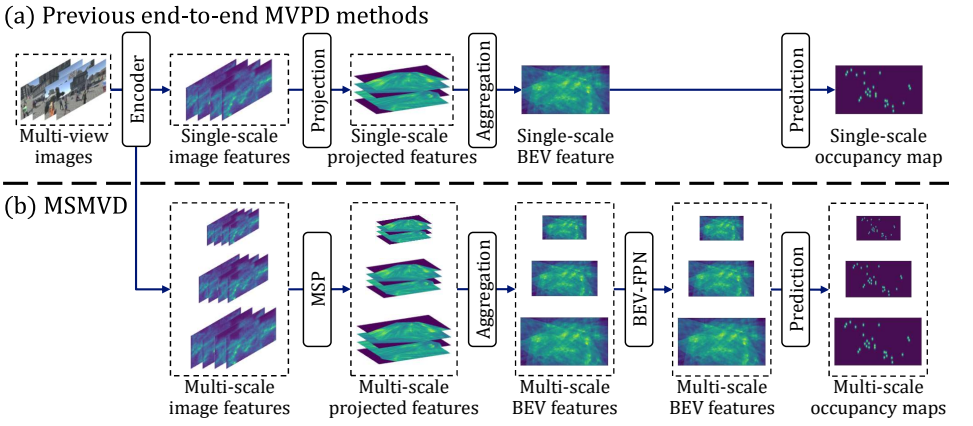


Figure 1: (a) Previous end-to-end MVPD methods generate a single-scale BEV feature from single-scale image features and predict one occupancy map. (b) MSMVD generates multi-scale BEV features from multi-scale image features and predicts multiple occupancy maps. In MSMVD, the image encoder consists of the backbone and image-FPN.

1 Introduction

Pedestrian detection from images is a crucial component in various applications such as surveillance systems [11], sports analyses [7], and human-computer interactions [37]. In pedestrian detection, occlusion is a serious problem because it hinders detecting pedestrians hidden behind obstacles or other pedestrians [5, 13]. As a potential solution, many studies have focused on *Multi-View Pedestrian Detection* (MVPD) [5, 17, 36]. This task aims to detect pedestrians in the form of a bird's eye view (BEV) from multiple camera views. In contrast to traditional monocular pedestrian detection [10, 40], MVPD utilizes multiple calibrated cameras with partially overlapping fields of view. MVPD is therefore expected to be more robust to occlusion in the overlapping areas than monocular pedestrian detection by aggregating complementary information across multiple views.

As with various computer vision tasks, end-to-end trainable deep learning methods are the dominant approach in MVPD [1, 2, 12, 16, 17, 18, 30, 32, 33, 36, 39]. They adopt a unified framework, as shown in Fig. 1(a). It first extracts image features for individual views from a specific layer of the image encoder (*e.g.*, the last convolutional layer in ResNet [15]). Then, it generates a BEV feature by projecting the image features to the BEV space based on calibrated camera parameters and predicts a BEV occupancy map from that BEV feature. This framework aggregates complementary information from multiple views through the BEV feature, enabling the model to train in an end-to-end manner.

End-to-end MVPD methods have outperformed non-end-to-end methods [3, 4, 5, 13, 31, 38]. However, as shown in the result of MVFP [1] in Fig. 2, they often struggle to detect two types of pedestrians: those with consistently small or large scales (i.e., apparent sizes in images) in views (yellow, pink, and purple rectangles), and those with vastly different scales between views (red and blue rectangles). This is because single-scale image features from a specific layer of the image encoder used in existing end-to-end methods to generate the BEV feature have limited ability to represent pedestrians with diverse scales. In contrast to end-to-end MVPD methods, many monocular methods [6, 14, 20, 21, 22] utilize multi-scale image

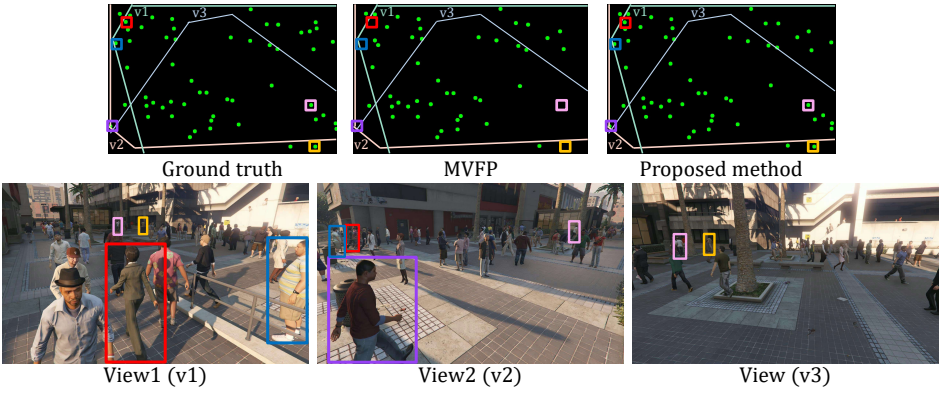


Figure 2: (Top) Comparison of the predicted BEV maps using MVFP [1] and our proposed method. MVFP corresponds to Fig. 1(a). Green circles represent detected pedestrian locations, and colored rectangles illustrate the differences between MVFP and our proposed method. (Bottom) View images. Pedestrians with rectangles correspond to those with the same color rectangles on the BEV maps.

features to effectively handle pedestrians with various scales, which are multiple features at different resolutions extracted from different layers in the image encoder. Small pedestrians are better represented in high-resolution features and vice versa, so conducting prediction for each of the multi-scale image features helps the detection of pedestrians across a wide range of scales. Despite their benefits, multi-scale image features have not been used in end-to-end MVPD methods. Furthermore, in MVPD, since pedestrians often have vastly different scales between views, the important scale information of each view for detecting such pedestrians varies from one view to another. Therefore, to fully utilize multi-scale image features in MVPD, it is important to combine information at different scales of multiple views.

In this paper, we propose a novel end-to-end MVPD method, called **M**ulti-**S**cale **M**ulti-**V**iew **D**etection (MSMVD), as shown in Fig. 1(b). The main idea of MSMVD is to effectively generate and leverage multi-scale BEV features that inherit the properties of multi-scale image features from multiple views. To this end, MSMVD newly introduces two key components: a multi-scale projection (MSP) and a BEV feature pyramid network (BEV-FPN). Unlike existing methods that project only single-scale image features to the BEV space, MSP projects the multi-scale image features extracted by the image encoder to that space, scale-by-scale, as shown in Fig. 3(a). This design allows MSMVD to generate multiple BEV features at different resolutions (i.e., multi-scale BEV features), each of which inherits the properties of the image features at its corresponding scale from multiple views. Therefore, utilizing these multi-scale BEV features leads to precisely detecting pedestrians with consistently small or large scales in views. BEV-FPN combines information at different scales of multiple views by processing the multi-scale BEV features utilizing an FPN-like network. This allows MSMVD to effectively detect pedestrians with vastly different scales across views by combining important scale information of such pedestrians across multiple views. Extensive experiments demonstrate that exploiting multi-scale image features of multiple views via multi-scale BEV features leads to better detection performances. Particularly in terms of MODA metric [5], MSMVD outperforms previous state-of-the-art methods by 4.5 points on GMVD [36] and 0.5 points on Wildtrack [5] and MultiviewX [17].

2 Related Work

Monocular pedestrian detection. One of the main difficulties in monocular pedestrian detection is effectively handling pedestrians at various scales. Traditional methods tackled this difficulty by generating hand-crafted features such as HOG features [8] or SIFT features [28] for multiple images with different resolutions. Modern deep learning methods [20, 21, 22, 24, 25, 26, 34, 41, 43] usually exploit multi-scale image features extracted from different layers in the backbone (*e.g.*, ResNet [15]), where small pedestrians are better represented and detected in high-resolution features and vice versa. As a pioneering work, SSD [25] leverages multi-scale image features and demonstrates their effectiveness. FPN [22] enhances multi-scale image features using a top-down path that gradually fuses low-resolution image features with higher-resolution image features. PAFPN [6, 14, 24, 29] further improved FPN by adding an extra bottom-up path that gradually fuses output multi-scale features of FPN from high to low resolution. EfficientDet [34] improves the detection performance by repeating a simplified version of PAFPN several times. Deformable DETR [43] and RT-DETR [41] utilize Transformer [35] to generate effective multi-scale image features. CSP [26], F2DNet [20], and LSFM [21] generate a powerful single-scale image feature by merging multi-scale image features of FPN into one feature. Due to its simplicity yet effectiveness, PAFPN is utilized by our method for the image encoder and BEV-FPN.

Multi-view pedestrian detection. In MVPD, aggregating information on multiple views is essential to reduce the effect of occlusion because each view has an overlapping but different field of view [5, 17, 36]. Early methods [3, 4, 5, 13, 31, 38] detected pedestrians for each view using monocular pedestrian detection and aggregated the detection results of multiple views. While they were important early attempts in MVPD, they still relied on monocular detection, so the occlusion problem remained. To overcome this problem, end-to-end trainable deep learning methods that do not rely on monocular detection have become the dominant approach [1, 2, 12, 16, 17, 18, 30, 32, 33, 36, 39]. They generate a representation of the BEV space, called the BEV feature, by projecting image features of multiple views to the BEV space and predict a BEV occupancy map from the BEV feature. The information on multiple views is aggregated through this BEV feature. MVDet [17] is a pioneering end-to-end method that generates the BEV feature using inverse perspective mapping. To further improve detection performance, recent studies have proposed more effective projections [1, 16, 18, 30, 32], data augmentations [12, 16, 30, 33], network architectures [1, 2, 12], and training strategies [36, 39]. While they have improved the detection performances, exploring the effective way to exploit multi-scale image features of multiple views remains an open problem.

3 Proposed Method

3.1 Overall Architecture

Figure 1(b) shows an overview of MSMVD. This takes N view images $\{I^n\}_{n=1}^N$, where $I^n \in \mathbb{R}^{3\times H\times W}$ is the image from the n-th view. Here, H and W are the height and width of the image, respectively. The image encoder, consisting of the backbone and image feature pyramid network (image-FPN), extracts multi-scale image features for individual views. Specifically, when using the ResNet [15] backbone, it first extracts image features $\{F_3^n, F_4^n, F_5^n\}_{n=1}^N$,

where $F_l^n \in \mathbb{R}^{C_l \times \frac{H}{2^l} \times \frac{W}{2^l}}$ is the n-th view image feature from the l-th stage in ResNet, and C_l is the channel size in the l-th stage. Then, as in many monocular methods [6, 14, 29], image-FPN enhances these multi-scale image features, resulting in $\{\tilde{F}_3^n, \tilde{F}_4^n, \tilde{F}_5^n\}_{n=1}^N$, where $\tilde{F}_l^n \in \mathbb{R}^{C \times \frac{H}{2^l} \times \frac{W}{2^l}}$ is the enhanced n-th view image feature, and C is the channel size of the image-FPN. For these multi-scale image features, small pedestrians are better represented in $\tilde{F}_3^n$, and large pedestrians are better represented in $\tilde{F}_5^n$.

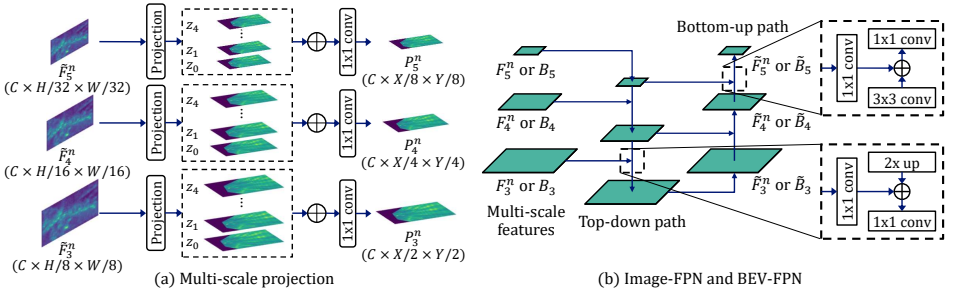
After obtaining multi-scale image features, our multi-scale projection (MSP) and subsequent max pooling generate multi-scale BEV features from those image features, and then, BEV-FPN combines information at different scales of multiple views. MSP projects multi-scale image features $\{\tilde{F}_3^n, \tilde{F}_4^n, \tilde{F}_5^n\}_{n=1}^N$ to the BEV space based on camera parameters, scale-by-scale. We denote these projected features as $\{P_3^n, P_4^n, P_5^n\}_{n=1}^N$, where $P_l^n \in \mathbb{R}^{C \times \frac{X}{2^{l-2}} \times \frac{Y}{2^{l-2}}}$ is obtained from $\tilde{F}_l^n$. Here, X and Y are the width and height of the ground truth BEV map. This MSP allows the model to generate multi-scale BEV features that inherit the properties of multi-scale image features of multiple views. We formulate MSP in Sec. 3.2. To aggregate the information at each scale of multiple views, we perform max pooling along the view direction, scale-by-scale, and generate multi-scale BEV features $\{B_3, B_4, B_5\}$. Here, $B_l \in \mathbb{R}^{C \times \frac{X}{2^{l-2}} \times \frac{Y}{2^{l-2}}}$ is generated from $\{P_l^n\}_{n=1}^N$. Namely, the information on small pedestrians of multiple views is aggregated to B_3, and that on large pedestrians is aggregated to B_5. BEV-FPN combines the information at different scales of multiple views by processing multi-scale BEV features, resulting in refined multi-scale BEV features $\{\tilde{B}_3, \tilde{B}_4, \tilde{B}_5\}$, where $\tilde{B}_l \in \mathbb{R}^{C \times \frac{X}{2^{l-2}} \times \frac{Y}{2^{l-2}}}$. We describe the detailed structure of the BEV-FPN in Sec. 3.3.

Then, prediction networks followed by the sigmoid function predict multi-scale BEV occupancy maps $\{M_3, M_4, M_5\}$ from the multi-scale BEV features, where $M_l \in \mathbb{R}^{1 \times \frac{X}{2^{l-2}} \times \frac{Y}{2^{l-2}}}$ is predicted from $\tilde{B}_l$. Because the predicted occupancy maps have lower resolutions than the ground truth, we also predict offset maps $\{O_3, O_4, O_5\}$ to recover the discretization errors of pedestrian locations, following MVDeTr [16]. Here, $O_l \in \mathbb{R}^{2 \times \frac{X}{2^{l-2}} \times \frac{Y}{2^{l-2}}}$ is predicted from $\tilde{B}_l$. MSMVD is optimized by calculating the loss for all occupancy maps and offset maps during training and predicts the final occupancy map by merging multi-scale occupancy maps during inference. We describe these training and inference procedures in Sec. 3.4.

3.2 Multi-scale Projection

Figure 3(a) shows our multi-scale projection (MSP). To inherit the properties of multi-scale image features of multiple views to the BEV feature, MSP projects multi-scale image features $\{\tilde{F}_3^n, \tilde{F}_4^n, \tilde{F}_5^n\}_{n=1}^N$ to the BEV space and output projected features $\{P_3^n, P_4^n, P_5^n\}_{n=1}^N$. Let $K^n \in \mathbb{R}^{3\times3}$ and $[R^n|T^n] \in \mathbb{R}^{3\times4}$ be the intrinsic and extrinsic camera parameter matrices of the n-th view, respectively. Here, R^n represents the rotation, and T^n represents the translation. We project n-th view multi-scale image features to the BEV space using the correspondence [16, 17, 30, 32] between the n-th view 2D image pixel coordinates (u^n, v^n) and the 3D world position (x, y, z), as

$$\gamma^n \begin{pmatrix} u^n \\ v^n \\ 1 \end{pmatrix} = K^n [R^n|T^n] \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}, \tag{1}$$



(a) Multi-scale projection (b) Image-FPN and BEV-FPN

Figure 3: (a) Multi-scale projection (MSP). (b) Image-FPN and BEV-FPN consist of top-down and bottom-up paths. $\oplus$ indicates a concatenation. Image-FPN processes multi-scale image features, and BEV-FPN processes multi-scale BEV features.

where γ^n is a scaling factor that accounts for a down-sampling of image features and determines the spatial resolution of the BEV feature. For each of n-th view multi-scale image features, we set γ^n to the value that makes the resolution ratio of projected features $\{P_3^n, P_4^n, P_5^n\}$ and that of the image features $\{\tilde{F}_3^n, \tilde{F}_4^n, \tilde{F}_5^n\}$ the same. Specifically, for $\tilde{F}_l^n$, γ^n is set so that the spatial resolution of P_l^n is $\frac{X}{2^{l-2}} \times \frac{Y}{2^{l-2}}$. This allows the model to inherit the properties of multi-scale image features to multi-scale BEV features. We project each image feature $\tilde{F}_l^n$ with multiple heights $\{z_0, z_1, z_2, z_3, z_4\}$, like 3DROM [30], to utilize information about the entire pedestrian body, where z_i corresponds to the height of $30 \times i$ cm. Then, projected features of multiple heights are concatenated along the channel direction and processed by a 1×1 convolutional layer, resulting in a final C-channel projected feature P_l^n.

3.3 BEV-FPN

BEV-FPN combines information at different scales of multiple views to precisely detect pedestrians whose scales are vastly different between views because important scale features of such pedestrians differ from one view to another. To this end, we utilize PAFPN [24] for BEV-FPN, which consists of the top-down and bottom-up paths, as shown in Fig. 3(b). The top-down path takes multiple features at different resolutions as input and gradually fuses low-resolution features with higher-resolution features. This process combines important information in low-resolution features with that in higher-resolution features. The bottom-up path takes the multiple features output from the top-down path as input and gradually fuses high-resolution features with lower-resolution features. This process combines important information in high-resolution features with that in lower-resolution features. Since the information of $\{P_l^n\}_{n=1}^N$ is aggregated to B_l by MSP and max pooling, BEV-FPN combines information at different scales of multiple views by processing multi-scale BEV features $\{B_3, B_4, B_5\}$. As a result, BEV-FPN outputs enhanced multi-scale BEV features $\{\tilde{B}_3, \tilde{B}_4, \tilde{B}_5\}$, and they improve the detection of pedestrians with vastly different scales between views.

3.4 Training and Inference

During training, we calculate the loss for multi-scale BEV occupancy maps $\{M_3, M_4, M_5\}$ and multi-scale offset maps $\{O_3, O_4, O_5\}$. Following MVDeTr [16], we used the Focal loss [23] and L1 loss variant [26, 42] for the occupancy maps and offset maps, respectively.

Let $\mathcal{L}_{\text{det}}(\cdot,\cdot)$ be the loss function for the occupancy maps, and $\mathcal{L}_{\text{off}}(\cdot,\cdot)$ be the loss function for the offset maps. Our overall loss $\mathcal{L}_{\text{all}}$ is

$$\mathcal{L}_{\text{all}} = \sum_{l\in\{3,4,5\}} \{\mathcal{L}_{\text{det}}(M_l,\hat{M}_l) + \mathcal{L}_{\text{off}}(O_l,\hat{O}_l)\}. \tag{2}$$

Here, $\hat{M}_l$ is the ground truth occupancy map down-sampled to match with M_l and smoothed using a Gaussian kernel. $\hat{O}_l$ is the ground truth offset map corresponding to the discretization error of $\hat{M}_l$. While we do not use O_4 and O_5 for inference, we calculate the losses for them as auxiliary losses. We optimize MSMVD using $\mathcal{L}_{\text{all}}$ as the overall training objective.

During inference, we merge predicted multiple occupancy maps $\{M_3,M_4,M_5\}$ into one occupancy map $M \in \mathbb{R}^{1\times\frac{X}{2}\times\frac{Y}{2}}$ to combine pedestrians with various scales, as

$$M = \frac{1}{3}\{M_3 + \text{Upsample}(M_4) + \text{Upsample}(M_5)\}, \tag{3}$$

where the Upsample function indicates up-sampling of the occupancy map to the same spatial resolution as M_3 using a bilinear interpolation. Since down-sampling high-resolution occupancy maps could lose detailed location information of pedestrians, we up-sample low-resolution occupancy maps to the highest resolution. After the merging, we extract pedestrians over the threshold and convert their locations to those in the original $X \times Y$ BEV map using the offset map O_3 in the same way as MVDeTr [16].

4 Experiments

4.1 Datasets

Wildtrack [5]. This is a real-world dataset consisting of images captured with 7 cameras and comprises 400 frames. This dataset covers a 12 m $\times$ 36 m region quantized into a 480×1440 grid using square grid cells of 2.5 cm^2. Each frame includes 20 pedestrians on average. Wildtrack splits them into 360 frames for training and the remaining 40 frames for testing. We use the randomly sampled 40 frames in the training split as the validation data.

MultiviewX [17]. This is a synthetic dataset using the Unity engine and closely follows the style of Wildtrack. This dataset comprises 400 frames captured by 6 cameras and covers a 16 m $\times$ 25 m region. The region is quantized into a 640×1000 grid. Each frame includes 40 pedestrians on average. MultiviewX also splits them into 360 frames for training and 40 frames for testing. We use 40 frames in the training split as the validation data.

GMVD [36]. This is a large-scale synthetic dataset and includes 7 scenes. Each scene varies the number of cameras and camera layouts, which makes the dataset more challenging than a constrained camera setup like Wildtrack or MultiviewX. Except for the camera setup and the size of the covered region, all parameters follow those of MultiviewX. Each frame includes 20-40 pedestrians on average. GMVD splits them into 6 scenes with 4983 frames for training and 1 scene with 1012 frames for testing. Following Vora *et al.* [36], we use MultiviewX as the validation data during training on the GMVD train split and evaluate trained models on the GMVD test split. Since GMVD is the largest dataset, we set it as the main dataset for our experiments.

Method	MODA	MODP	Precision	Recall
Vora+ [36]	68.2	76.3	91.5	75.5
SHOT† [32]	71.5	77.9	93.7	76.7
Suzuki+ [33]	72.3	77.1	91.3	77.1
3DROM† [30]	73.7	77.3	92.2	80.5
MVAug† [12]	73.8	77.9	91.9	80.8
OmniOcc† [2]	75.1	76.9	92.3	82.0
MVFP† [1]	75.7	78.2	94.3	80.5
Ours	**80.2**	**81.3**	**95.7**	**83.9**

Table 1: Comparison on GMVD. † indicates our re-implementation results.

Method	Wildtrack				MultiviewX			
	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
Vora+ [36]	86.7	76.2	95.1	91.4	88.2	79.9	96.8	91.2
MVDet [17]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
SHOT [32]	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
MVDeTr [16]	91.5	82.1	**97.4**	94.0	93.7	**91.3**	99.5	94.2
M-MVOT [39]	92.1	81.3	94.5	97.8	96.7	86.1	98.8	**97.9**
MVAug [12]	93.2	79.8	96.3	97.0	95.3	89.7	99.4	95.9
3DROM [30]	93.5	75.9	97.2	96.2	95.0	84.9	99.0	96.1
OmniOcc [2]	93.5	81.5	94.9	97.8	94.4	88.7	98.7	95.6
MVFP [1]	94.1	78.8	96.4	97.7	95.7	85.1	98.4	97.2
Ours	**94.6**	**83.3**	95.9	**98.8**	**97.2**	**91.3**	**99.8**	97.4

Table 2: Comparison on Wildtrack and MultiviewX. OmniOcc is re-implemented by us.

4.2 Implementation Details and Evaluation Metrics

Unless otherwise mentioned, we employed ResNet18 [15] and PAFPN [24] for the backbone and image-FPN in the image encoder, respectively. ResNet18 was pretrained on ImageNet [9]. We use four convolutional layers for each prediction network. We set the channel size C to 256 and the detection threshold to 0.4.

Input images were resized to 720×1280. Following MVDeTr [16], we applied random resizing and cropping to images as the data augmentation during training. We optimized the model using the Adam optimizer [27]. We set the batch size to 1 and accumulated the gradient over 16 batches. The learning rate was initialized to 1.0×10^{-3} and decayed to 1.0×10^{-6} following a cosine schedule. We set the training epoch to 10 for GMVD and 50 for Wildtrack and MultiviewX. More training details are provided in Supp. A.

Following previous studies [17, 36], we used four standard metrics provided by Chavdarova *et al*. [5] and Kasturi *et al*. [19]: multiple object detection accuracy (MODA), multiple object detection precision (MODP), recall, and precision. A detected pedestrian was classified as a true positive if its distance from the ground truth was within 0.5 meters. We used MODA as the primary performance indicator following previous studies [17, 36].

4.3 Comparison with Previous State-of-the-Art Methods

To verify the effectiveness of MSMVD, we compared it with previous state-of-the-art methods on GMVD, as shown in Table 1. Previous methods generate a single-scale BEV feature from single-scale image features, as shown in Fig. 1(a). Since several previous methods had been evaluated only on Wildtrack and MultiviewX, we re-implemented them for GMVD. Results of other methods are directly copied from the original papers. MSMVD significantly outperformed the previous methods on all metrics. In particular, it achieved MODA

	MODA	MODP	Precision	Recall
Baseline	71.8	80.0	93.6	76.9
+ MSP	74.9	80.5	94.3	80.3
+ MSP and BEV-FPN	**80.2**	**81.3**	**95.7**	**83.9**

(a)

	MODA	MODP	Precision	Recall
M_3&O_3	79.1	81.2	94.9	83.6
M_4&O_4	78.1	80.3	95.0	82.4
M_5&O_5	77.6	79.5	95.0	81.9
Ours	**80.2**	**81.3**	**95.7**	**83.9**

(b)

Table 3: (a) Effect of each component in MSMVD. (b) Comparison of our multi-scale inference and single-scale inference.

Method	Backbone	MODA	MODP	Precision	Recall	Params (M)
Baseline	ResNet18	71.8	80.0	93.6	76.9	11.7
	ResNet34	72.1	78.2	95.5	75.6	21.8
	ResNet50	74.3	79.2	94.1	79.2	24.2
	ResNet101	75.8	78.9	96.8	78.3	43.2
Ours	ResNet18	80.2	81.3	95.7	83.9	22.9
	ResNet34	81.0	81.5	94.8	**85.7**	33.0
	ResNet50	82.0	81.3	94.9	85.6	35.9
	ResNet101	**82.2**	**82.6**	**97.2**	85.6	54.9

Table 4: Effect of backbone size on the baseline model and our proposed method.

and MODP 4.5 and 3.1 points higher than the previous highest scores, respectively. These results demonstrate that exploiting and combining multi-scale image features of multiple views leads to a better detection performance and show that MSMVD is superior to the other methods. Figure 2 visually compares the results of MSMVD and MVFP [1]. While MVFP failed to detect pedestrians with consistently small (yellow and pink rectangles) or large (a purple rectangle) scales in views and pedestrians with vastly different scales between views (red and blue rectangles), our MSMVD accurately detected these pedestrians. This indicates the advantages of MSMVD over MVFP. Specifically, the detection of the former pedestrians indicates the advantage of exploiting multi-scale image features and conducting prediction for each of the multi-scale BEV features, and that of the latter pedestrians indicates the advantage of combining the information at different scales of multiple views by BEV-FPN.

To validate the versatility of MSMVD, which is not limited to GMVD, we also compared it with previous methods on Wildtrack and MultiviewX, as shown in Table 2. The result of OmniOcc [2] is achieved by our re-implementation, and the results of the other methods are directly copied from the original papers. MSMVD outperformed the previous methods on almost all metrics. In particular, it exceeded the previous highest MODA by 0.5 points on both datasets. These results demonstrate that our method is effective for various datasets.

4.4 Ablation Study

In this subsection, we investigate the effect of MSP, BEV-FPN, our inference procedure, and the backbone size. We set the model that removes MSP and BEV-FPN from MSMVD and generates a single-scale BEV feature from single-scale image features extracted by the last layer of the image encoder as the baseline. All experiments were conducted on GMVD. In Supp. B, we also investigate the effect of the training with O_4 and O_5, the offset prediction, the pooling method, the architecture of image-FPN and BEV-FPN, and the scaling factor γ^n in MSP.

Effect of each component. To investigate the contribution of MSP and BEV-FPN, we gradually added them to the baseline, as shown in Table 3(a). Adding MSP significantly improved MODA and recall and outperformed all methods in Table 1 on MODA, other than

OmniOcc and MVFP. This indicates that adding MSP reduces missed detection and demonstrates the effectiveness of utilizing multi-scale image features via multi-scale BEV features. Adding BEV-FPN further improved the detection performance and outperformed all methods in Table 1. This shows that adding BEV-FPN enhances multi-scale BEV features and demonstrates the importance of combining information at different scales of multiple views.

Effect of merging multi-scale occupancy maps. To investigate the effect of merging predictions at multiple scales during inference, we compared our inference procedure with using only a single-scale occupancy map and the corresponding offset map, as shown in Table 3(b). When employing our inference procedure, we achieved a better detection result than when using only a single-scale occupancy map and its offset map. This demonstrates the advantage of combining predictions at multiple scales. In addition, even when using only a single-scale occupancy map and offset map, MSMVD outperformed all methods in Table 1. This also indicates that MSMVD's BEV feature is superior to that of other methods.

Effect of the backbone size. MSMVD has more parameters than the baseline due to MSP and BEV-FPN. To show that the performance improvement by MSMVD is not simply due to this parameter increase, we applied backbones with different sizes and investigated their effect on the detection performance, as shown in Table 4. While larger backbones improved the detection performance of the baseline, MSMVD with ResNet18 achieved a better result than the baseline with the same or more parameters. This demonstrates that MSMVD is superior to a simple parameter increase by larger backbones. Larger backbones are also beneficial to our method, and MSMVD with the largest ResNet101 achieved 82.2 MODA and 82.6 MODP, which significantly outperformed existing methods in Table 1.

5 Conclusion

We proposed a novel end-to-end multi-view pedestrian detection (MVPD) method, called MSMVD. In contrast to previous methods that rely only on single-scale image features, MSMVD exploits multi-scale image features via multi-scale bird's eye view (BEV) features to effectively detect pedestrians at various scales. In addition, MSMVD combines information at different scales of multiple views by processing multi-scale BEV features using a feature pyramid network. This makes the model more robust to pedestrians with vastly different scales between views. Extensive experiments demonstrate that exploiting multi-scale image features via multi-scale BEV features improves the detection performance, and MSMVD achieved a new state-of-the-art result on three major MVPD datasets. We hope that our work opens new possibilities in exploring end-to-end MVPD.

References

[1] Sithu Aung, Haesol Park, Hyungjoo Jung, and Junghyun Cho. Enhancing multi-view pedestrian detection through generalized 3D feature pulling. In *WACV*, 2024.

[2] Sithu Aung, Min-cheol Sagong, and Junghyun Cho. Multi-view pedestrian occupancy prediction with a novel synthetic dataset. In *AAAI*, 2025.

[3] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *ICCV*, 2017.

[4] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *ICMLA*, 2017.

[5] 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 *CVPR*, 2018.

[6] Yuming Chen, Xinbin Yuan, Jiabao Wang, Ruiqi Wu, Xiang Li, Qibin Hou, and Ming-Ming Cheng. YOLO-MS: rethinking multi-scale representation learning for real-time object detection. *TPAMI*, 2025.

[7] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, 2023.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 2011.

[11] Mohamed Elhoseny. Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems. *CSSP*, 2020.

[12] Martin Engilberge, Haixin Shi, Zhiye Wang, and Pascal Fua. Two-level data augmentation for calibrated multi-view detection. In *WACV*, 2023.

[13] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *TPAMI*, 2007.

[14] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv:2107.08430*, 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *ACMMM*, 2021.

[17] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *ECCV*, 2020.

[18] Jinwoo Hwang, Philipp Benz, and Pete Kim. Booster-SHOT: Boosting stacked homography transformations for multiview pedestrian detection with attention. In *WACV*, 2024.

[19] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *TPAMI*, 2008.

[20] Abdul Hannan Khan, Mohsin Munir, Ludger van Elst, and Andreas Dengel. F2DNet: Fast focal detection network for pedestrian detection. In *ICPR*, 2022.

[21] Abdul Hannan Khan, Mohammed Shariq Nawaz, and Andreas Dengel. Localized semantic feature mixers for efficient pedestrian detection in autonomous driving. In *CVPR*, 2023.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[24] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[26] Wei Liu, Irtiza Hasan, and Shengcai Liao. Center and scale prediction: Anchor-free approach for pedestrian and face detection. *PR*, 2023.

[27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

[28] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[29] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. RTMDet: An empirical study of designing real-time object detectors. *arXiv:2212.07784*, 2022.

[30] Rui Qiu, Ming Xu, Yuyao Yan, Jeremy S Smith, and Xi Yang. 3D random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In *ECCV*, 2022.

[31] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object detection. In *ICCV*, 2011.

[32] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *ICCV*, 2021.

[33] Satoshi Suzuki, Shotaro Tora, and Ryo Masumura. Scene generalized multi-view pedestrian detection with rotation-based augmentation and regularization. In *ICIP*, 2024.

[34] Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet: Scalable and efficient object detection. In *CVPR*, 2020.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

[36] Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Bringing generalization to deep multi-view pedestrian detection. In *WACV*, 2023.

[37] Tim Wengefeld, Steffen Müller, Benjamin Lewandowski, and Horst-Michael Gross. A multi modal people tracker for real time human robot interaction. In *RO-MAN*, 2019.

[38] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *CVPR*, 2016.

[39] Qi Zhang, Kaiyi Zhang, Antoni B Chan, and Hui Huang. Mahalanobis distance-based multi-view optimal transport for multi-view crowd localization. In *ECCV*, 2024.

[40] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017.

[41] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *CVPR*, 2024.

[42] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv:1904.07850*, 2019.

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.