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

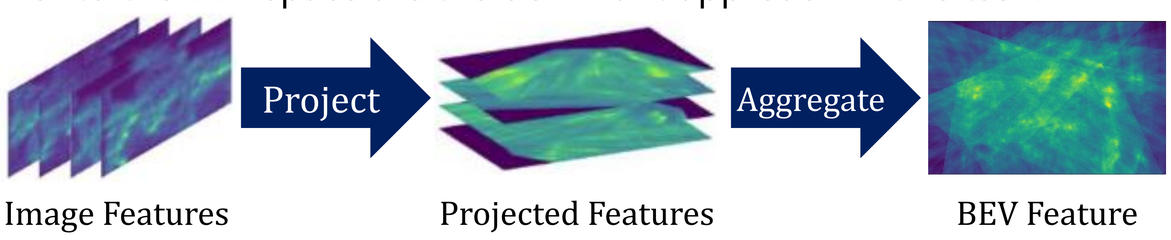
Taiga Yamane, Satoshi Suzuki, Ryo Masumura, Shota Orihashi, Tomohiro Tanaka, Mana Ihori, Naoki Makishima, Naotaka Kawata
(Human Informatics Laboratories, NTT, Inc., Japan)

BMVC 2025

Background

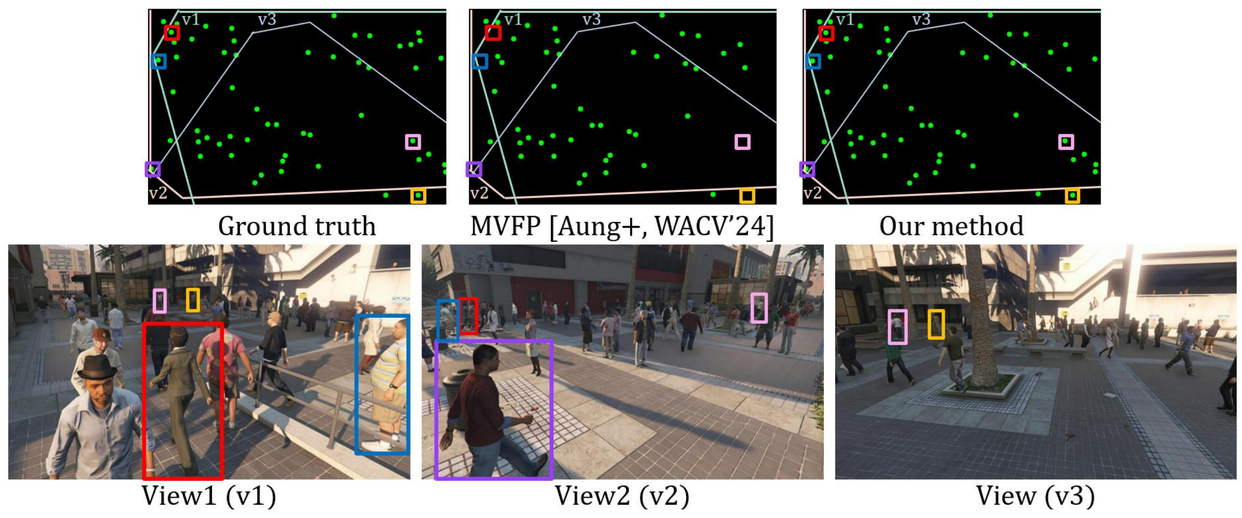
- **Multi-view Pedestrian Detection** aims to detect pedestrians in the form of a bird's eye view (BEV) from multi-view images.

- **End-to-end deep learning methods** that project image features onto the BEV space are the dominant approach in this task.



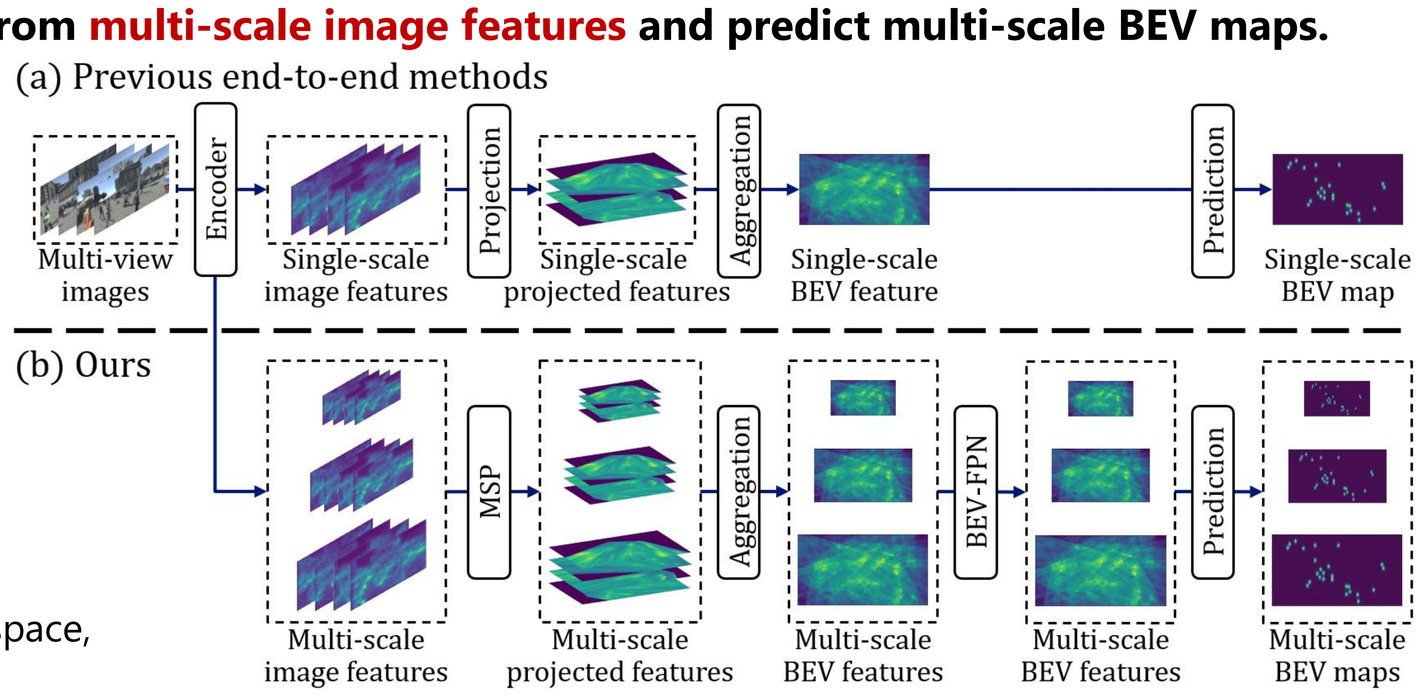
Problem

- Previous methods struggle to detect pedestrians with **consistent small or large scales in views** or **vastly different scales across views**.
  - ➢ This is because they rely solely on **single-scale image features** to generate the BEV feature.

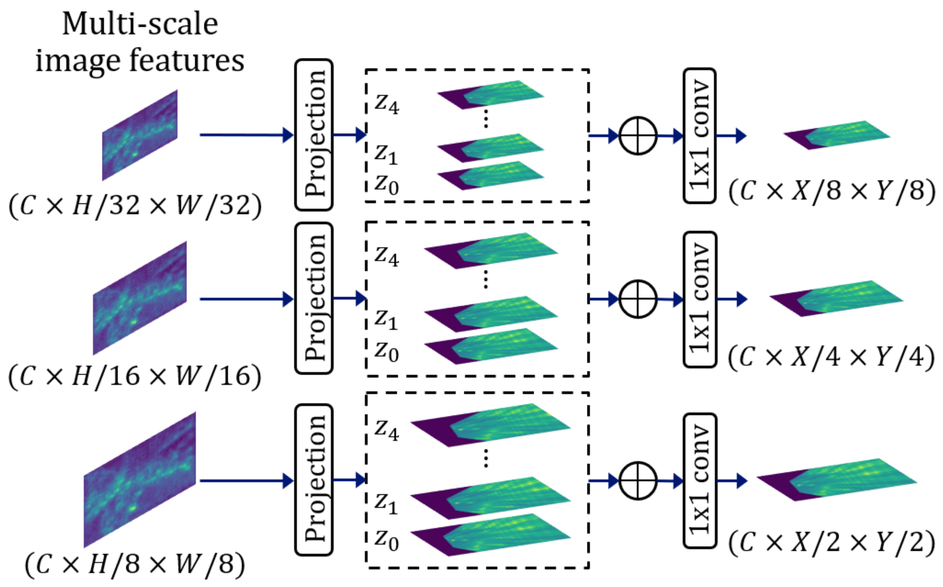


Our proposed method

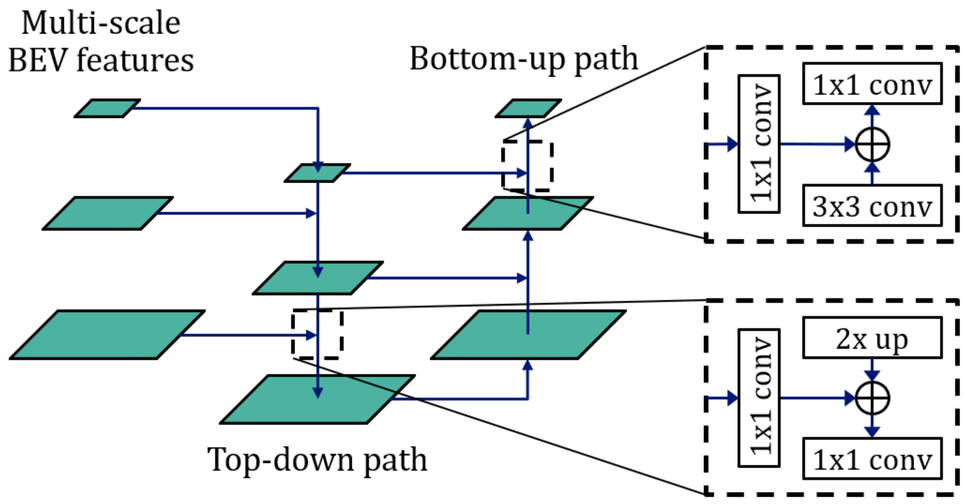
- **Our method generates multi-scale BEV features from multi-scale image features and predict multi-scale BEV maps.**
  - ➢ Core components are **MSP** and **BEV-FPN**.
  - ➢ **Multi-scale image features have the ability to effectively represent pedestrians at various scales.**
  - ➢ Training: We calculate the loss for each of multi-scale BEV maps.
  - ➢ Inference: We merge multi-scale BEV maps into one BEV map by up-sampling and averaging.



- **Multi-scale projection (MSP)**
  - ➢ This projects multi-scale image features onto the BEV space, scale-by-scale, to generate multi-scale BEV features.
  - ➢ This improves the detection of pedestrians <u>with consistently small or large scales in views</u>.



- **BEV feature pyramid network (BEV-FPN)**
  - ➢ This combines information at different scales of multiple views to detect pedestrians <u>with vastly different scales between views</u>.



Experimental results on GMVD dataset

- Quantitative results: MSP and BEV-FPN improved the detection performance, and incorporating both outperformed previous methods that only use single-scale image features. → **Effectiveness of utilizing multi-scale BEV features**
  - ➢ Experiments are conducted on GMVD dataset [Vora+, WACV'23] which includes various camera layouts.
  - ➢ Baseline: The model only uses single-scale image features (corresponding to (a) in the above figure).

	MODA	MODP	Precision	Recall
SHOT [Song+, ICCV'21]	71.5	77.9	93.7	76.7
3DROM [Qiu+, ECCV'22]	73.7	77.3	92.2	80.5
MVAug [Engilberge+, WACV'23]	73.8	77.9	91.9	80.8
MVFP [Aung+, WACV'24]	75.7	78.2	94.3	80.5
Baseline	71.8	80.0	93.6	76.9
+MSP (+multi-scale prediction)	**74.9**	**80.5**	**94.3**	**80.3**
+MSP & BEV-FPN	**80.2**	**81.3**	**95.7**	**83.9**

- ✓ $MODA = 1 - \frac{FP+FN}{GT}$
- ✓ $MODP = \frac{\sum(1-d[TP]/0.5)}{TP}$
- ✓ GT: The number of Ground Truth.
- ✓ TP: The number of True Positive.
- ✓ FP: The number of False Positive.
- ✓ FN: The number of False Negative.
- ✓ A detected pedestrian is classified as a TP if its distance from GT is within 0.5 meters.
- ✓ $d[TP]$ is the distance of TP from GT.

- Qualitative results: see the figure at the top right.
  - ➢ Detection of pedestrians consistently small or large scales in views (yellow, pink, and purple rectangles) → **Effectiveness of MSP**
  - ➢ Detection of pedestrians with vastly different scales between views (red and blue rectangles) → **Effectiveness of BEV-FPN**