Supplementary Materials: Unsupervised Multi-view Pedestrian Detection

Mengyin Liu
blean@live.cn
School of Computer and Communication Engineering,
University of Science and Technology Beijing
Beijing, China

Chao Zhu*
chaozhu@ustb.edu.cn
School of Computer and Communication Engineering,
University of Science and Technology Beijing
Beijing, China

Shiqi Ren
shiqiren@xs.ustb.edu.cn
School of Computer and Communication Engineering,
University of Science and Technology Beijing
Beijing, China

Xu-Cheng Yin
xuchengyin@ustb.edu.cn
School of Computer and Communication Engineering,
University of Science and Technology Beijing
Beijing, China

6 Proof of Lemma 1

LEMMA 1. *The 1st PCA vector $\phi^{(1)}$ of* $\mathbf{X}$*:* (a) *maximizes the global variance* $\mathrm{var}(\phi^{(1)}) = \frac{1}{\mathrm{P}}\sum_{p=1}^{\mathrm{P}}(\phi_p^{(1)})^2$*;* (b) *minimizes the reconstruction loss* $\min_{Z^{(1)}} \|\mathbf{X} - \mathbf{X}Z^{(1)}(Z^{(1)})^\top\|^2$*, where $Z^{(1)}$ is a bi-direction mapping between high-dimensional* $\mathbf{X}$ *and low-dimensional $\phi^{(1)}$.*

PROOF. Given N× images, $\mathbf{X} = (X_1^\top, X_2^\top, \cdots, X_{\mathrm{P}}^\top)^\top \in \mathbb{R}^{\mathrm{P}\times\mathrm{D}}$ are their DINOv2 features. P = N × H × W are total pixels and $X_j \in \mathbb{R}^{\mathrm{D}}$.

After a Zero Standardization, each column $\sum_{p=1}^{\mathrm{P}} x_{pd} = 0$. The 1st PCA vector $\phi^{(1)} = \mathbf{X}Z^{(1)} = (\phi_1^{(1)}, \phi_2^{(1)}, \cdots, \phi_{\mathrm{P}}^{(1)})^\top \in \mathbb{R}^{\mathrm{P}}$ is obtained by a linear mapping $Z^{(1)} = (z_1^{(1)}, z_2^{(1)}, \cdots, z_{\mathrm{D}}^{(1)})^\top \in \mathbb{R}^{\mathrm{D}}$ and $\|Z^{(1)}\|^2 = 1$ as normalization. Then $\phi_p^{(1)} = \sum_{d=1}^{\mathrm{D}} x_{pd} z_d^{(1)}$.

The variance of $\phi^{(1)}$ is $\mathrm{var}(\phi^{(1)}) = \frac{1}{\mathrm{P}}\sum_{p=1}^{\mathrm{P}}(\phi_p^{(1)} - \mathbb{E}(\phi^{(1)}))^2$, then the expectation $\mathbb{E}(\phi^{(1)})$ is calculated as:

$$\mathbb{E}(\phi^{(1)}) = \frac{1}{\mathrm{P}}\sum_{p=1}^{\mathrm{P}}\sum_{d=1}^{\mathrm{D}} x_{pd} z_d^{(1)} = \frac{1}{\mathrm{P}}\sum_{d=1}^{\mathrm{D}} z_d^{(1)} \sum_{p=1}^{\mathrm{P}} x_{pd} = 0. \tag{12}$$

Therefore, $\mathrm{var}(\phi^{(1)}) = \frac{1}{\mathrm{P}}\sum_{p=1}^{\mathrm{P}}(\phi_p^{(1)})^2 = \|\phi^{(1)}\|^2$. Maximal variance $\max_{Z^{(1)}}(\mathrm{var}(\phi^{(1)}))$ is formulated as:

$$\max_{Z^{(1)}} \|\mathbf{X}Z^{(1)}\|^2 = \max_{Z^{(1)}} (Z^{(1)})^\top \mathbf{X}^\top \mathbf{X} Z^{(1)}. \tag{13}$$

From another perspective, PCA also minimizes the reconstruction loss based on the bi-directional mapping via forward $Z^{(1)}$ and backward $(Z^{(1)})^\top$, which is derived as:

$$\begin{aligned}
&\min_{Z^{(1)}} \|\mathbf{X} - \mathbf{X}Z^{(1)}(Z^{(1)})^\top\|^2 \\
&= \min_{Z^{(1)}} \mathrm{tr}(\mathbf{X}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{X}Z^{(1)}(Z^{(1)})^\top \\
&\quad - Z^{(1)}(Z^{(1)})^\top\mathbf{X}^\top\mathbf{X} + Z^{(1)}(Z^{(1)})^\top\mathbf{X}^\top\mathbf{X}Z^{(1)}(Z^{(1)})^\top) \\
&= \min_{Z^{(1)}} -2\mathrm{tr}(\mathbf{X}^\top\mathbf{X}Z^{(1)}(Z^{(1)})^\top) + \mathrm{tr}(\mathbf{X}^\top\mathbf{X}Z^{(1)}(Z^{(1)})^\top) \\
&= \max_{Z^{(1)}} \mathrm{tr}(\mathbf{X}^\top\mathbf{X}Z^{(1)}(Z^{(1)})^\top) \\
&= \max_{Z^{(1)}} (Z^{(1)})^\top\mathbf{X}^\top\mathbf{X}Z^{(1)} = \text{Eq. 13}
\end{aligned} \tag{14}$$

Here, we can observe that Eq.13 and Eq.14 are equivalent. Let $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$, function $F(Z^{(1)})$ is constructed with a Lagrange Multiplier λ to solve Eq.13 and Eq.14:

$$F(Z^{(1)}) = (Z^{(1)})^\top\mathbf{A}Z^{(1)} + \lambda(1 - (Z^{(1)})^\top Z^{(1)}). \tag{15}$$

Let $\frac{\partial F(Z^{(1)})}{\partial Z^{(1)}} = 0$, thus $\mathbf{A}Z^{(1)} = \lambda Z^{(1)}$ is exactly the form of eigenvalue decomposition. With $\|Z^{(1)}\|^2 = 1$ and Eq. 13 and 14,

$$\max_{Z^{(1)}}(Z^{(1)})^\top\lambda Z^{(1)} = \max_{Z^{(1)}}\lambda(Z^{(1)})^\top Z^{(1)} = \max_{Z^{(1)}}\lambda, \tag{16}$$

i.e., when eigenvalue λ is maximized, its corresponding eigenvector is the solution of $Z^{(1)}$. To find this maximized eigenvalue, Singular Value Decomposition (SVD) can be performed on $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where the items of $\mathbf{U} = (U_1^\top, U_2^\top, \cdots, U_{\mathrm{K}}^\top)^\top$ are eigenvectors correspond to the items of $\mathrm{diag}(\Sigma) = (\Sigma_1, \Sigma_2, \cdots, \Sigma_{\mathrm{K}})^\top$ as eigenvalues.

Therefore, choosing the eigenvector $U_{\max}$ corresponding to the maximized eigenvalue $\Sigma_{\max}$ following Eq. 16, which is denoted as:

$$\Sigma_{\max} = \max\lambda = \max\{\Sigma_1, \Sigma_2, \cdots, \Sigma_{\mathrm{K}}\}, \tag{17}$$

$U_{\max}$ is the solution of mapping $Z^{(1)}$ for the 1st PCA vector $\phi^{(1)}$, which always exists and holds both maximized global variance in Eq.13 and minimized reconstruction loss in Eq.14. □

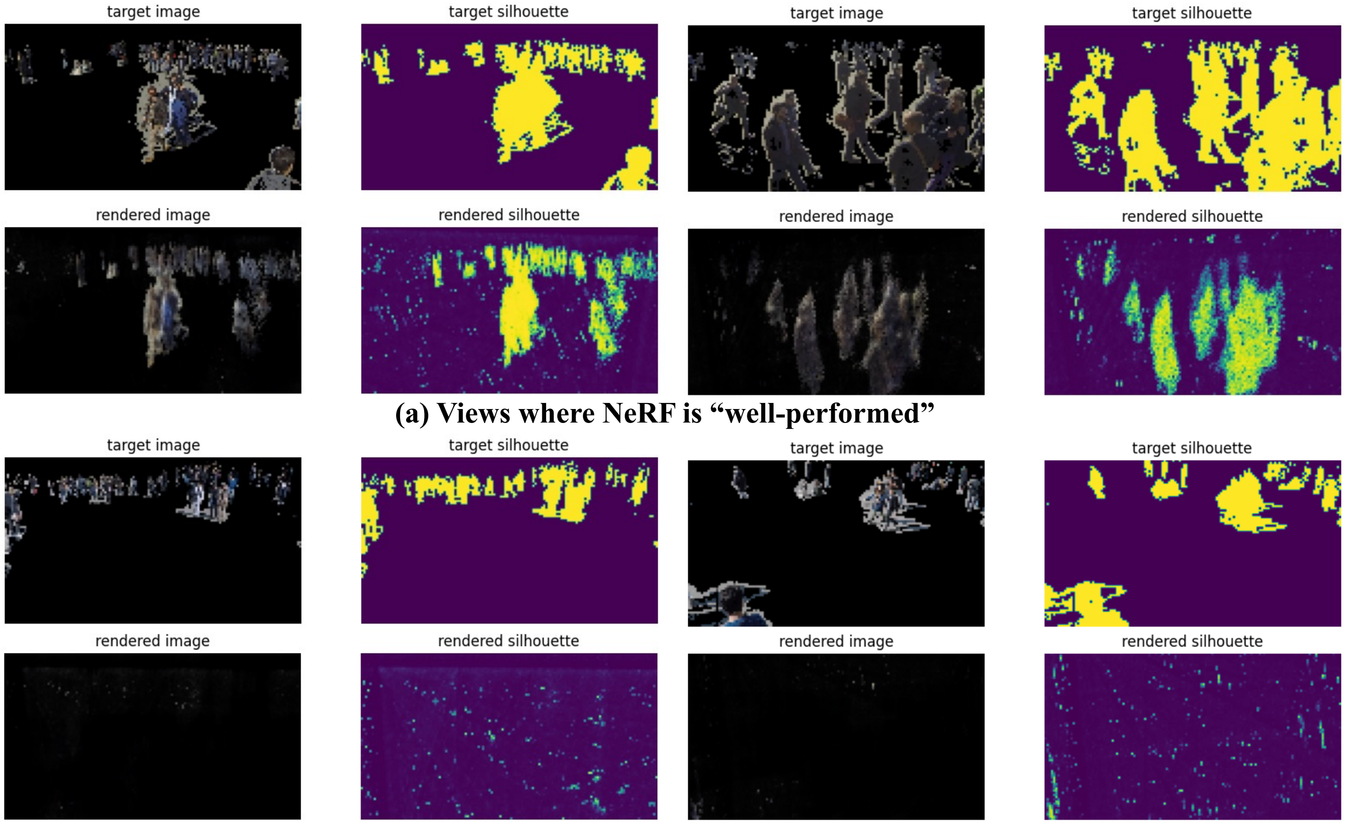
COROLLARY 1. *If the k^{th} maximum eigenvalue $\Sigma_{\max-\mathrm{k}}$ in Eq. 17 is chosen, its corresponding $U_{\max-\mathrm{k}}$ is denoted as k^{th} PCA vector $\phi^{(k)}$, which represents the k^{th} principal component that more loosely holds the Eq.13 and in Eq.14 with the linear mapping $Z^{(k)}$ than the 1st one.*

COROLLARY 2. *According to the bi-direction mapping ensured by a minimized reconstruction loss in Lemma 1, the 1st PCA values $\phi^{(1)}$ of features* X *are also similar and distinguished from the opposite parts. Then, $\phi_p^{(1)} \in \mathbb{R}$ as scalars can be more easily divided by a threshold value than comparing the complicated feature vectors $X_p \in \mathbb{R}^{\mathrm{D}}$.*

Based on Lemma 1 and Corollary 2 above, DINOv2 features are bi-directionally mapped to 1st PCA values. As is shown in Figure 3 of the main paper, these features represent the ground surrounded by the multiple cameras as the most distinguished background, and the remaining parts are pedestrians and their contexts.

Therefore, we propose more iterations of PCA as Semantic-aware Iterative Segmentation (SIS) to further segment the pedestrians and

[*] Corresponding author.



(a) Views where NeRF is "well-performed"

(b) Views where NeRF fails to render

Figure 10: Visualizations of how NeRF [7] performs on the 1st frame of Wildtrack dataset. (a) In some views with large scale pedestrians, NeRF fits color and silhouette (i.e., masks). (b) However, NeRF fails to render small scale pedestrians in other views. Instead, the sparse dots might come from the leaning pedestrians from other views. For example, the left-most 1st view of (a) and (b) is physically overlapping. If the pedestrians are rendered to be standing, they should also be observed in the view inside (b). However, there are no pedestrians in this view, thus they are rendered to be leaning and sparse in the global 3D volume.

Table 7: Comparison of different methods on more popular Wildtrack [1] and larger-scale CityStreet [17] datasets.

Datasets	UMPD	MVDet	MVDeTr	3DROM	SHOT
Wildtrack	61.2	75.7	82.1	75.9	76.5
CityStreet	45.1	65.7	74.1	70.1	72.4

non-human background inside these coarse-grained "foreground" features, equipped with the zero-shot semantic capability of powerful vision-language model CLIP to identify the pedestrian parts and the proof of Lemma 1 to segment the foreground and background.

7 Experiments on Larger-scale Dataset

For a long time, CityStreet [17] is used to evaluate summed coarse-grained counting number of tiny pedestrians (MAE/NAE), then adopted by [18] recently for accurate 3D localization (MODA/MODP). In Table 7, most popular supervised detectors, including MVDet [5], MVDeTr [4], 3DROM [10] and SHOT [13], are worse on the larger-scale scenes with tinier pedestrians in CityStreet than Wildtrack, reported in this recent paper [18]. But our UMPD also performs competitively on such a larger-scale and challenging dataset.

8 Qualitative Analysis on Using NeRF

The typical input views in multi-view pedestrian detection datasets are 6~7 [1, 5] or even 4 [3]. However, the normal number of views for NeRF [7] ranges in 10~10^3. In practice, more views (>100) are needed to render large-scale crowded scenes than the simple scenes with single or a few objects. Thus, qualitative analysis are performed about using NeRF on multi-view pedestrian detection dataset.

As is illustrated in Figure 10, NeRF fails to render in some views where there are too few foreground in Figure 10(b), compared to the other views with more large scale pedestrians in Figure 10(a). Moreover, the sparse dots in the failing views might come from the non-vertical pedestrians from other views. For instance, both the 1st left views of Figure 10(a) and (b) are overlapping observations. If these pedestrians are rendered, they should also be observed in Figure 10(b). However, no pedestrians are rendered in this view. Thus, they are non-vertical in the other views of the global volume, which hinders the correct multi-view pedestrian detection on BEV.

In addition to the limited view numbers, most of the NeRF-based methods [7, 8, 14] utilize a neural network to implicitly represent the 3D volume, which adopts a 3D voxel location and an observation ray as its inputs and the color and density as its outputs, i.e., the information of 3D volume is “memorized” in neural network parameters. Therefore, it is difficult for our vertical-aware loss $\mathcal{L}_{VBR}$ designed for a whole explicit 3D volume to address the issue above.

Instead, based on the 2D-3D cross modal mapping, our 3D-to-2D rendering losses based on PyTorch3D [6] directly learn the colors and densities in the explicit 3D volume predicted by our proposed Geometric-aware Volume-based Detector (GVD), then our proposed Vertical-aware BEV Regularization (VBR) $\mathcal{L}_{VBR}$ can regularize the volume following the vertical characteristics of pedestrians.

9 About the Insights of Our Proposed UMPD

As is introduced in the main paper, our proposed UMPD is mostly motivated by the difficulty of heavy burden to annotate the BEV pedestrian labels on real data. Even cross-domain methods [15] cannot perform as ideally as in-domain ones. Therefore, we have found a solution to learn an annotation-free detector:

- Given the 3D existence (i.e., density) of pedestrians, the BEV labels mean a top-down observation, and their 2D masks from multiple cameras mean the surrounding observations. Without laborious BEV labels, the latter one is the key.
- Hence, the 3D density becomes a “bridge”, to be predicted by our proposed volume-based detector GVD from 2D multi-view images (2D-to-3D geometric projection), learned from unsupervised 2D masks (3D-to-2D rendering losses), and vertically projected on BEV as the detection results.
- For such a multi-view task, all-view information should be considered by segmentation rather than just single-view. Therefore, we propose a new unsupervised method SIS with Iterative PCA based on multi-image DINOv2 features.
- Moreover, to better identify the pedestrians, powerful vision-language model CLIP is fundamental for our SIS, and to constraint the predicted 3D density with no leaning or laying down, we further proposes VBR to regularize the prediction.

Please note that these insights are logically step-by-step for such a novel and challenging unsupervised task. To our best knowledge, these are never discussed before in multi-view pedestrian detection. Thus, UMPD is effective on real-world and simulated datasets.

For future researches, we have also discussed more insights about issues that remain to solve in Section 4 of the main paper:

- In Section 4.4, the experiments show that the quality of 2D masks greatly affects the performance of our UMPD, even some supervised models like Grounded-SAM [12] fail in generalization to this task, which is worth future studies for better masks in supervised or unsupervised manner.
- In Section 4.5, the visualizations show some wrong results near the edges of region, where the information from less overlapped camera views is insufficient for accurate detection, especially in more populated MultiviewX dataset [5].

The first item above also shows the capability of supervised models are limited. Once the new domain of real world or realistic simulation [16] is far from their capability of generalization, the 2D masks are essential as supervisions for in-domain fine-tuning. However, pixel-wise 2D masks are even harder to annotate than pedestrian BEV positions to directly supervise multi-view pedestrian detectors. Based on the powerful capability of single-modal as well as multi-modal unsupervised models [9, 11], our UMPD is applicable in a plug-and-play manner for both real-world and simulated scenes without any source domain labels.

Finally, we hope this work, as the first fully-unsupervised multi-view pedestrian detection method, could be a start and inspire more interesting future works in this field and beyond.

References

[1] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. 2018. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5030–5039.

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1290–1299.

[3] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. 2007. Multi-camera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence* 30, 2 (2007), 267–282.

[4] Yunzhong Hou and Liang Zheng. 2021. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*. 1673–1682.

[5] Yunzhong Hou, Liang Zheng, and Stephen Gould. 2020. Multiview detection with feature perspective transformation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 1–18.

[6] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. 2020. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 Courses*. 1–1.

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 405–421.

[8] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5480–5490.

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193* (2023).

[10] Rui Qiu, Ming Xu, Yuyao Yan, Jeremy S Smith, and Xi Yang. 2022. 3D Random Occlusion and Multi-layer Projection for Deep Multi-camera Pedestrian Localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Springer, 695–710.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[12] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

[13] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. 2021. Stacked homography transformations for multi-view pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6049–6057.

[14] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4190–4200.

[15] Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. 2023. Bringing generalization to deep multi-view pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 110–119.

[16] Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. 2024. ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object. *arXiv preprint arXiv:2403.18775* (2024).

[17] Qi Zhang and Antoni B Chan. 2019. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8297–8306.

[18] Qi Zhang, Yunfei Gong, Daijie Chen, Antoni B Chan, and Hui Huang. 2024. Multi-View People Detection in Large Scenes via Supervised View-Wise Contribution Weighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7242–7250.