HiddenObject: Modality-Agnostic Fusion for Multimodal Hidden Object Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Detecting hidden or partially concealed objects remains a fundamental challenge in multimodal environments, where factors like occlusion, camouflage, and lighting variations significantly hinder performance. Traditional RGB-based detection methods often *fail under such adverse conditions*, motivating the need for more robust, modality-agnostic approaches. In this work, we present **HiddenObject**, a novel fusion framework that integrates RGB, thermal, and depth data using a Mamba-based fusion mechanism. Our method *captures complementary signals across modalities*, enabling enhanced detection of obscured or camouflaged targets. Specifically, the proposed approach identifies modality-specific features and fuses them in a unified representation that generalizes well across challenging scenarios. We validate **HiddenObject** across multiple benchmark datasets, demonstrating *state-of-the-art or competitive performance* compared to existing methods. These results highlight the efficacy of our fusion design and expose key limitations in current unimodal and naïve fusion strategies. More broadly, our findings suggest that Mamba-based fusion architectures can significantly advance the field of multimodal object detection, especially under visually degraded or complex conditions.

1 Introduction

Rapid advancements in autonomous systems and robotic technologies have been motivated by the need to enhance operational efficiency, mitigate labor shortages, and address complex real-world challenges across diverse domains, such as industrial automation, agriculture, surveillance, search-and-rescue operations, and security applications. As these scenarios grow increasingly sophisticated, accurately detecting hidden or partially obscured objects becomes critical for effective decision-making and operational success. Traditional object detection methods primarily depend on single-modality imaging, notably RGB cameras, which are significantly limited by factors like lighting variations, occlusions, camouflaged objects, and adverse environmental conditions [1].

Real-world objects often appear partially or fully concealed due to foliage, structural elements, environmental clutter, or intentional concealment, posing substantial detection challenges for conventional RGB imaging, which relies heavily on ideal lighting and unobstructed visibility. Conversely, multimodal imaging techniques—such as thermal [2] and depth [3] imaging—offer robust solutions by capturing complementary information even under challenging visual conditions. Thermal imaging, for instance, detects infrared radiation independent of visible light, demonstrating strong resilience against lighting variability and providing reliable performance in nighttime [4] or obscured environments [5].

Integrating multiple imaging modalities thus presents significant advantages for handling complex detection tasks, including partial occlusions [6], dense object clustering [7], and camouflage [8]. Objects exhibiting minimal color or texture contrast in RGB modalities often possess distinct signa-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

tures in thermal, depth, or near-infrared (NIR) modalities, thereby greatly enhancing detectability. Specifically, thermal imaging differentiates objects based on thermal contrast, while depth imaging captures structural three-dimensional information, further facilitating the detection of hidden or obscured objects [9].

However, fusing multiple modalities into a cohesive detection framework introduces considerable technical challenges, such as modality misalignment, information redundancy, and computational efficiency. Existing approaches typically rely on static fusion strategies [10, 11] or emphasize single modalities [12], limiting their ability to interpret critical information effectively in complex or unfamiliar environments. To overcome these limitations, we introduce **HiddenObject**, a modality-agnostic detection framework utilizing a novel Mamba-based fusion mechanism coupled with a channel-aware decoder. Our approach efficiently extracts and seamlessly integrates critical information across diverse imaging modalities (e.g., RGB, thermal, depth), enabling robust and accurate detection of concealed objects in challenging scenarios.

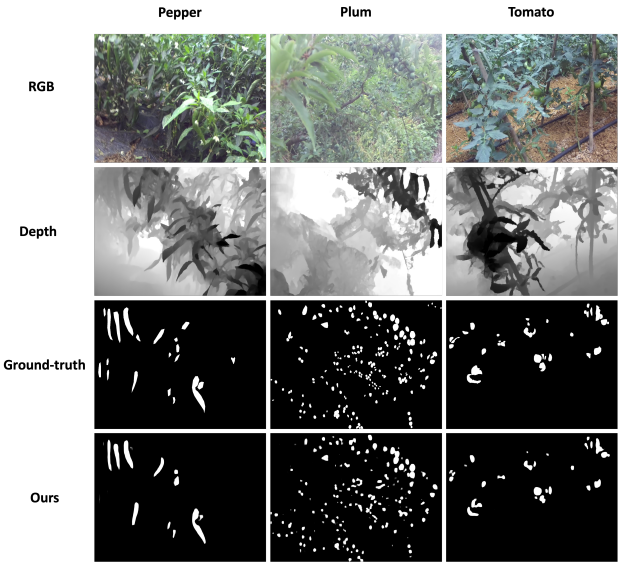

Figure 1: Our proposed method effectively addresses severe occlusion and densely distributed small objects, enhancing detection performance. The examples utilized are sourced from the ACOD-K12 dataset [6].

As demonstrated in Figure 1, our framework maintains reliable detection performance even under severe occlusion and complex clutter conditions, a capability especially crucial in applications such as agriculture [13–15] and robotics [10], where target objects frequently remain hidden or obscured. The key contributions of our work are summarized as follows:

- We propose a Mamba-based fusion mechanism integrated with a channel-aware decoder to effectively extract and merge multimodal information from RGB, thermal, depth, and other modalities.
- Our detection pipeline robustly handles a wide range of hidden objects, including lightly obscured, partially occluded, heavily occluded, concealed, or camouflaged targets.
- Through extensive experimentation and analysis across multiple datasets, we validate the effectiveness and efficiency of our proposed framework, establishing a benchmark for future research into the capabilities of Mamba in multimodal learning.

In the following sections, we provide detailed descriptions of our approach, including the fusion mechanism, dataset selection, experimental protocols, and comprehensive performance evaluations, further emphasizing the practical applicability and versatility of our proposed **HiddenObject** framework.

2 Related Work

2.1 Object Detection and Segmentation Methods

In recent advancements, Carion et al. [16] pioneered the development of an end-to-end object detection framework utilizing transformers, termed DEtection TRansformer (DETR). This approach redefines object detection as a set prediction task, employing binary matching to generate one-to-one object predictions during training. Such a method streamlines the detection process by removing the necessity for manually crafted anchor boxes and non-maximum suppression (NMS) post-processing. Despite its benefits, DETR suffers from prolonged training convergence times, prompting the emergence of various enhanced DETR variants. Deformable DETR [17] improves training speed by predicting 2D anchor points and incorporating a deformable cross-attention mechanism to selectively sample features near reference points. Conditional DETR [18] separates content

and positional operations, introducing conditional cross-attention to expedite convergence. Efficient DETR [19] enhances the pipeline's efficiency by integrating dense detection with sparse set prediction. DAB-DETR [20] employs 4D reference points to refine bounding box predictions iteratively. DN-DETR [21] boosts training efficiency and label assignment by incorporating query denoising during training. DINO [22] consolidates these advancements into a robust DETR-based detection framework. For real-time performance, RT-DETR [23, 24] develops an end-to-end detector with an Efficient Hybrid Encoder and an IoU-aware Query Selection strategy.

Recent advancements in State Space Models (SSMs), particularly Mamba, offer efficient linear-time complexity for sequence modeling tasks, addressing limitations of Transformers in handling long sequences [25]. Extensions such as Vision Mamba (Vim) have adapted Mamba to visual tasks, improving speed and reducing memory consumption compared to traditional Vision Transformers like DeiT [26]. Further integration into object detection, including Mamba YOLO [27] and Fusion-Mamba [28], demonstrated significant gains in detection accuracy for multimodal data, crucial for identifying concealed objects. Multimodal fusion methods have also benefited from Mamba, particularly through Coupled Mamba, which enhances cross-modal representation consistency and inference speed, suggesting strong potential for applications in hidden object detection scenarios [29].

Earlier studies on infrared-visible object detection have largely relied on frameworks designed for single-modality object detection, typically categorized into one-stage detectors like Faster R-CNN [30] and two-stage detectors such as YOLO [31–33]. To integrate the complementary data from infrared and visible imagery, König et al. [34] proposed a fully convolutional fusion Region Proposal Network (RPN) that merges features from both modalities through concatenation, finding that mid-level fusion yields superior outcomes. Building on this, subsequent research developed CNN-based attention mechanisms to enhance the synergy between infrared and visible images [35–37]. Furthermore, some approaches incorporated transformer-based fusion modules to capture broader complementary relationships across these modalities [38–41]. Beyond direct feature fusion, certain methods utilized illumination data as global weights to combine infrared and visible features or to refine multi-branch detection outputs, mitigating the effects of noise [42, 43]. Recognizing that complementary traits vary by region, some studies employed semantic segmentation at the bounding box level [44, 45] or Region of Interest (ROI) prediction [46, 47] to direct regional fusion. Others leveraged confidence or uncertainty metrics to refine multi-branch predictions post-fusion [48, 49]. To tackle modality misalignment, Zhang et al. [46] introduced the AR-CNN framework, aligning features across modalities using paired bounding box annotations to address object displacement. Similarly, Kim et al. [50] applied a multi-label learning strategy to adapt detection in misaligned scenarios. Recently, many multimodal methods have been introduced [51, 52]. Guo et al. [51] present a method for infrared-visible object detection that dynamically focuses on dominant modality objects and adaptively fuses complementary information, while Zhang et al. [52] proposed a novel end-to-end algorithm for multimodal fusion detection.

2.2 Datasets and Benchmarks

The KAIST Multispectral Pedestrian Dataset [53] consists of manually annotated RGB Color and Thermal imaging from the heat signatures of pedestrians from the perspective of a motor vehicle, with 103,128 manual annotations. A multi-modality Neural Network-based fusion detection algorithm paper [54] collected RGB and Thermal data from a high-end FLIR T650sc infrared camera, capturing 454 pairs of images containing construction defects from twenty years of wear. RGBT-1k [55] captured 1000 RGB-Thermal paired images from indoor and outdoor scenes with post-processed transformation steps for pixel-to-pixel alignment, which is an underlying concern for all RGB-X modality datasets. Osroosh et. al [56] developed an RGB-Thermal image collection technique, which in their methodology uniquely included a microclimate monitoring system with a rainfall simulator system to simulate different weather. The CitDet dataset [7] features 579 high-resolution images and 32,000 bounding box annotations of citrus trees, specifically targeting trees affected by the Huanglongbing bacterial infection disease, capable of significantly impacting citrus crop yield and quality through early detection and disease monitoring.

Thermal cameras, which capture infrared radiation emitted from object surfaces, have become increasingly popular due to their robustness under harsh conditions, widespread availability, and ability to provide unique information such as temperature. The proliferation of thermal cameras has led to the introduction of numerous thermal imaging datasets. The MultiSpectralMotion dataset [57] contains

indoor and outdoor thermal images collected through handheld devices along with ground-truth depth maps. The OdomBeyondVision dataset [58], in contrast, focuses on indoor environments using various platforms including handheld devices, unmanned ground vehicles (UGVs), and unmanned aerial systems (UAS). ViViD++ [59] provides outdoor thermal imagery captured using vehicle-mounted and handheld systems. The SubT-MRS dataset [60] comprises outdoor scenes collected from vehicles, UAS, legged robots, and handheld setups under varying degrees of visual degradation; it is most similar to our requirements but lacks imagery specific to wildland forests. Recent outdoor thermal stereo datasets include STheReO [61], MS2 [62], and FIReStereo [14]. MS2 offers thermal images captured in diverse environmental conditions, including visually challenging scenarios like rainy conditions. However, these datasets primarily focus on urban driving environments, inheriting limitations typical of driving datasets, such as constrained motion patterns and large baseline separation between cameras. FIReStereo [14] specifically targets small UAS applications aimed at developing depth estimation algorithms suitable for visually degraded environments.

References

[1] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

[2] Mritunjay Rai, Tanmoy Maity, and RK Yadav. Thermal imaging system and its real time applications: a survey. *Journal of Engineering Technology*, 6(2):290–303, 2017.

[3] Alexandre Lopes, Roberto Souza, and Helio Pedrini. A survey on rgb-d datasets. *Computer Vision and Image Understanding*, 222:103489, 2022.

[4] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Hanzhe Teng, Yipeng Wang, Xiaoao Song, and Konstantinos Karydis. Multimodal dataset for localization, mapping and crop monitoring in citrus tree farms. In *Advances in Visual Computing*, pages 571–582, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47969-4.

[6] Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, and Feng Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17201–17211, June 2024.

[7] Jordan James, Heather Manching, Matthew Mattia, Kim Bowman, Amanda Hulse-Kemp, and William Beksi. Citdet: A benchmark dataset for citrus fruit detection. *IEEE Robotics and Automation Letters*, PP:1–8, 12 2024. doi: 10.1109/LRA.2024.3474473.

[8] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021.

[9] Zhangyong Tang, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler. Multi-level fusion for robust rgbt tracking via enhanced thermal representation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(10):1–24, 2024.

[10] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. doi: 10.1109/IROS.2017.8206396.

[11] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

[12] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network, 2019.

[13] Mahmoud Abdulsalam, Zakaria Chekakta, Nabil Aouf, and Maxwell Hogan. Fruity: A multi-modal dataset for fruit recognition and 6D-Pose estimation in precision agriculture. In *2023 31st Mediterranean Conference on Control and Automation (MED)*, pages 144–149. IEEE, June 2023.

[14] Devansh Dhrafani, Yifei Liu, Andrew Jong, Ukcheol Shin, Yao He, Tyler Harp, Yaoyu Hu, Jean Oh, and Sebastian Scherer. Firestereo: Forest infrared stereo dataset for uas depth perception in visually degraded environments. *IEEE Robotics and Automation Letters*, 10(4):3302–3309, 2025. doi: 10.1109/LRA.2025.3536278.

[15] Connor Lee, Matthew Anderson, Nikhil Raganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Cart: Caltech aerial rgb-thermal dataset in the wild. *arXiv preprint arXiv:2403.08997*, 2024.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

[18] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021.

[19] Zhi Yao, Jiang Ai, Boxun Li, and Chi Zhang. Efficient detr: Improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

[20] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

[21] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.

[22] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[24] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer, 2024. URL `https://arxiv.org/abs/2407.17140`.

[25] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[26] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

[27] Zeyu Wang, Chen Li, Huiying Xu, and Xinzhong Zhu. Mamba yolo: A simple baseline for object detection with state space model. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025.

[28] Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Xuhui Liu, Juan Zhang, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. *arXiv preprint arXiv:2402.16853*, 2024.

[29] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[32] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.

[33] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[34] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 49–56, 2017.

[35] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022.

[36] Kamil Roszyk, Michał R Nowicki, and Piotr Skrzypczyński. Adopting the yolov4 architecture for low-latency multispectral pedestrian detection in autonomous driving. *Sensors*, 22(3):1082, 2022.

[37] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–411, 2023.

[38] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

[39] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible-infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[40] Yaohui Zhu, Xiaoyu Sun, Miao Wang, and Hua Huang. Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[41] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024.

[42] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 787–803. Springer, 2020.

[43] Xiaoxiao Yang, Yeqiang Qian, Huijie Zhu, Chunxiang Wang, and Ming Yang. Baanet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2920–2926. IEEE, 2022.

[44] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280. IEEE, 2020.

[45] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 72–80, 2021.

[46] Lu Zhang, Zhiyong Liu, Xiangyu Zhu, Zhan Song, Xu Yang, Zhen Lei, and Hong Qiao. Weakly aligned feature fusion for multimodal object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[47] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1510–1523, 2021.

[48] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2022.

[49] Qing Li, Changqing Zhang, Qinghua Hu, Pengfei Zhu, Huazhu Fu, and Lei Chen. Stabilizing multispectral pedestrian detection with evidential hybrid fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[50] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi. Mlpd: Multi-label pedestrian detector in multispectral domain. *IEEE Robotics and Automation Letters*, 6(4): 7846–7853, 2021.

[51] Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *European Conference on Computer Vision*, pages 464–481. Springer, 2024.

[52] Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, Daixun Li, Wenbo Huang, Yunsong Li, and Xue Yang. E2E-MFD: Towards end-to-end synchronous multimodal fusion detection. *Advances in Neural Information Processing Systems*, 37:52296–52322, 2024.

[53] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[54] Xincong Yang, Runhao Guo, and Heng Li. Comparison of multimodal rgb-thermal fusion techniques for exterior wall multi-defect detection. *Journal of Infrastructure Intelligence and Resilience*, 2(2):100029, 2023. ISSN 2772-9915. doi: https://doi.org/10.1016/j.iintel.2023.100029.

[55] Santosh Kumar Panda and Pankaj Kumar Sa. Rgbt-1k: An rgb-thermal paired dataset, 2024.

[56] Yasin Osroosh and R. Troy Peters. Detecting fruit surface wetness using a custom-built low-resolution thermal-rgb imager. *Computers and Electronics in Agriculture*, 157:509–517, 2019. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2019.01.023.

[57] W. Dai, Y. Zhang, S. Chen, D. Sun, and D. Kong. A multi-spectral dataset for evaluating motion estimation systems. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5560–5566. IEEE, 2021.

[58] P. Li, K. Cai, M. R. U. Saputra, Z. Dai, and C. X. Lu. Odombeyondvision: An indoor multi-modal multi-platform odometry dataset beyond the visible spectrum. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3845–3850. IEEE, 2022.

[59] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung. Vivid++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022.

[60] S. Zhao, Y. Gao, T. Wu, D. Singh, R. Jiang, H. Sun, M. Sarawata, Y. Qiu, W. Whittaker, I. Higgins, Y. Du, S. Su, C. Xu, J. Keller, J. Karhade, L. Nogueira, S. Saha, J. Zhang, W. Wang, C. Wang, and S. Scherer. Subt-mrs dataset: Pushing slam towards all-weather environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22647–22657, 2024.

[61] S. Yun, M. Jung, J. Kim, S. Jung, Y. Cho, M.-H. Jeon, G. Kim, and A. Kim. Sthereo: Stereo thermal dataset for research in odometry and mapping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3857–3864. IEEE, 2022.

[62] U. Shin, J. Park, and I. S. Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1043–1053, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper proposes a new fusion mechanism, which is discussed in the Proposed Method and experiementation is performed and detailed in Section 4,

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the algorithm are covered in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The work produced in this paper is empirical, focusing on a new Mamba-based fusion architecture and experimentally evaluating the results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Formulas, along with the specific datasets, compute, and description of the Mamba architecture are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: There is no provided open access to the data and code for now, but it is in the future intent to open source the code within the next calendar year.

   Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details on a training and test details are specified in the experimentation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our paper does not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper includes information on the bare metal that the Mamba architecture, along with the experiments that are performed upon.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and checked that the paper confroms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the Broader Impacts section discusses the posiitve and negative societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper is not a pretrained language model, image generator, scraped dataset, or anything that may have high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used published datasets that were correctly referenced and followed their terms of use in accordance with usage and citations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The paper evalutes on several public datasets, and describes a framework of the approach.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Core research methods do not involve usage of LLM's.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.