2024 IEEE 27th International Conference on
Intelligent Transportation Systems (ITSC)
September 24- 27, 2024. Edmonton, Canada

MambaST: A Plug-and-Play Cross-Spectral Spatial-Temporal Fuser for Efficient Pedestrian Detection

Xiangbo Gao[1], Asiegbu Miracle Kanu-Asiegbu[2], and Xiaoxiao Du[1]

2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC) | 979-8-3315-0592-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ITSC58415.2024.10920115

***Abstract*— This paper proposes *MambaST*, a plug-and-play cross-spectral spatial-temporal fusion pipeline for efficient pedestrian detection. Several challenges exist for pedestrian detection in autonomous driving applications. First, it is difficult to perform accurate detection using RGB cameras under dark or low-light conditions. Cross-spectral systems must be developed to integrate complementary information from multiple sensor modalities, such as thermal and visible cameras, to improve the robustness of the detections. Second, pedestrian detection models are latency-sensitive. Efficient and easy-to-scale detection models with fewer parameters are highly desirable for real-time applications such as autonomous driving. Third, pedestrian video data provides spatial-temporal correlations of pedestrian movement. It is beneficial to incorporate temporal as well as spatial information to enhance pedestrian detection. This work leverages recent advances in the state space model (Mamba) and proposes a novel Multi-head Hierarchical Patching and Aggregation (MHHPA) structure to extract both fine-grained and coarse-grained information from both RGB and thermal imagery. Experimental results show that the proposed MHHPA is an effective and efficient alternative to a Transformer model for cross-spectral pedestrian detection. Our proposed model also achieves superior performance on small-scale pedestrian detection. The code is available at https://github.com/XiangboGaoBarry/MambaST**

I. INTRODUCTION

Pedestrian detection is an essential task in applications such as autonomous driving. Precise pedestrian detection helps ensure pedestrian safety and helps vehicles to plan paths and avoid collision. Pedestrian detection also has implications in crowd analysis, traffic monitoring and management, and infrastructure planning [1]. In low-illumination scenarios, such as nighttime, it is difficult for visible (RGB) cameras alone to detect moving pedestrians. Cross-spectral fusion methods becomes necessary especially under low-light conditions to take advantage of the complementary information provided by both thermal and visible camera data [2]. Furthermore, pedestrian video data carries sequential movement information. It is beneficial to incorporate both spatial and temporal information from video frames to enhance pedestrian detection performance [3].

While significant progress has been made in multi-modal fusion and spatial-temporal modeling, simultaneous cross-spectral spatial-temporal fusion still lacks exploration. A variety of multi-modal fusion methods have been developed [4]–[8] for single-frame cross-spectral spatial fusion. However, these methods are not easily adapted to temporal fusion due to their reliance of 2D image inductive biases—assumptions about spatial relationships and patterns typical of 2D images. For temporal fusion, 3D convolutions [9]–[12], adaptive 2D convolutions [13], [14], and transformers [15] have been used. However, these methods work for RGB videos and cannot handle multi-modal inputs.

This paper proposes a novel fusion pipeline that addresses spatial-temporal fusion accounting for cross-spectral (RGB and thermal) sensor inputs. The proposed fusion pipeline, named *MambaST*, is based on a state space model (Mamba) [16]. Mamba is a recent state space model architecture that rivals the classic Transformers [17] for sequential data processing and has shown initial promise on computer vision tasks [18]–[20]. Our proposed *MambaST* is the first, to our knowledge, that applies Mamba to cross-spectral fusion accounting for both spatial and temporal information. Within *MambaST*, we propose a novel Multi-head Hierarchical Patching and Aggregation (MHHPA) module, which extracts cross-spectral spatial-temporal features across different hierarchical levels. This module is engineered to balance the extraction of fine-grained details with the removal of noise from coarser-grained information. We show that this module can be easily plug-and-play to perform pedestrian detection with YOLO model architecture [21] and is an effective alternative to transformer-based modules. We also leverage the recurrent capabilities in the visual state space model [22] to enhance the efficiency for *MambaST* in the inference time. We conducted experiments on KAIST, a real-world multispectral pedestrian detection benchmark [23], and we present detailed detection performance evaluation and ablation studies on various parameter choices. Our experimental results show improved pedestrian detection performance and efficiency (e.g., requiring significantly fewer model parameters compared to transformer-based methods).

The contributions of this paper are summarized as follows.

- We propose *MambaST*, a novel cross-spectral spatial-temporal fuser for effective and efficient pedestrian detection. *MambaST* produces superior detection results while requiring less model parameters and GFLOPs.
- We propose a novel plug-and-play MHHPA module for hierarchical spatial-temporal feature extraction.
- We show detailed detection performance evaluation and ablation studies on real-world pedestrian dataset.

This material is based upon work supported by the National Science Foundation under Grant IIS-2153171-CRII: III: Explainable Multi-Source Data Integration with Uncertainty. A. M. Kanu-Asiegbu is supported by a Rackham Merit Fellowship.

[1]X. Gao and X. Du are with the Robotics Department, University of Michigan, Ann Arbor, MI 48109 USA xiangbog@umich.edu; xiaodu@umich.edu

[2] A. M. Kanu-Asiegbu is with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA akanu@umich.edu

II. RELATED WORK

A. Preliminary on Mamba and Vision Mamba

Mamba [16] is a recent state space model (SSM) proposed for sequence modeling. It maps input $x(t) \in \mathbb{R}$ to output $y(t) \in \mathbb{R}$ through the following translation formulation

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{W}h'(t), \end{aligned} \tag{1}$$

where $h(t) \in \mathbb{R}^{\mathtt{N}}$ is the hidden state. $\mathbf{A} \in \mathbb{R}^{\mathtt{N}\times\mathtt{N}}$ is the evolution parameter. $\mathbf{B} \in \mathbb{R}^{\mathtt{N}\times 1}$ and $\mathbf{W} \in \mathbb{R}^{1\times\mathtt{N}}$ are projection parameters. Originally, Mamba was only used for 1-D sequences. Liu *et al.*[22] and Zhu *et al.*[18] later adapted the SSM-based model to accommodate 2D image data with two slightly different approaches (named *VMamba* and *Vision Mamba* models). Liu *et al.*[22] unfolded image patches into sequences along four distinct traversal paths, processing each patch sequence through Mamba, and then merged their outputs. Zhu *et al.*[18] aligned their model architecture with the transformer [17] architecture and added a positional embedding to each image patch. Both work show the potential of using mamba architecture to extract image features. Dong *et al.*[24], Li *et al.*[25], and Peng *et al.*[26] use Mamba for multi-modal fusion but only address single-frame fusion and does not yet generalize to multi-temporal sequences. VideoMamba[27] focuses on temporal fusion but does not address multi-modal spatial fusion. In this work, we build a novel vision mamba-based pipeline for cross-spectral (RGB and thermal) inputs, accounting for spatial and temporal information in video sequences.

B. Cross-Modality Fusion Methods

Multi-modality sensor data provides complementary information. RGB-Thermal [2], [28], [29], RGB-LiDAR [30]–[32], and RGB-Depth [33], [34] are common cross-modality sensor pairings for pedestrian detection in autonomous driving settings. Thermal cameras, in particular, provide finely detailed grayscale images in a variety of lighting and environmental conditions, and are useful sensor sources for fusion, especially in nighttime and low-light scenarios. A variety of cross-modality (RGB-thermal) fusion methods have been developed based on convolution neural networks [5], [35], [36], probabilistic ensembling methods [37], and transformers [38]–[41]. Feature fusion has also been used for cross-modality pedestrian detection. For example, Network-in-Network (NIN) [4] was used to fuse features from different modalities and reduce feature dimensions; INSANet [42] used intra- and inter-spectral attention blocks to learn mutual spectral relationships; and Guided Attentive Feature Fusion (GAFF) [7] guided the cross-modal feature fusion with an auxiliary pedestrian mask.

C. Temporal Fusion for Video Understanding

Fusion methods including 3D convolutions[9]–[12], adaptive 2D convolutions [13], [14], and Transformers [15] have been specifically designed for temporal fusion only, but these temporal fusion methods lack the capability of utilizing multi-modal inputs. Other approaches [4]–[8] focus on single-frame cross-spectral spatial fusion and cannot directly adapt to temporal fusion. In this work, we propose to extend a Mamba architecture to account for temporal sequences by recurrently connecting patched feature values across frames.

III. METHODOLOGY

We propose a novel Mamba-based Spatial Temporal Fuser named *MambaST* for cross-spectral pedestrian detection. The inputs of *MambaST* are (weakly aligned) multispectral (RGB color and thermal) image pairs containing traffic scenes, including pedestrians. The outputs of *MambaST* are bounding box detections of pedestrians in each frame.

A. Overview of MambaST Model Architecture

Fig. 1 illustrates the backbone and object detection pipeline within *MambaST*. We use YOLOv5 backbone, feature pyramid network (FPN) layer, pyramid attention network (PAN) layer [43], and detector for single-frame RGB and thermal object detection. The RGB and thermal backbone produces $\mathcal{T} \times 5 \times 2$ feature maps. Here, $\mathcal{T}$ represents the temporal duration, with each modality input yielding five layers of feature maps, and the numeral 2 signifies the two modalities–RGB and thermal. For spatial fusion, let $I_R^{(W_i\times H_i\times C_i)}$ and $I_T^{(W_i\times H_i\times C_i)}$ denote the third, fourth, and fifth layers of RGB and thermal feature maps, respectively, where $(W_i, H_i, C_i) \in \{(80, 80, 4D), (40, 40, 8D), (20, 20, 16D)\}$ are selected and input into the MHHPA (denoted as F1, F2, F3, respectively). Here, D is the multiplication factor for channel size; W, H, C represent the feature map width, height, and channel size, respectively. The output from this module is then added back to the original feature maps, enhancing the fused spatial representation. For notation simplicity, we do not differentiate the notations of feature maps between different fusion layers and use W, H, C to notate the width, height, and channel size for each fusion layer. After the last fusion layer, each feature map is passed into the YOLOv5 FPN layer, PAN layer, and detector for final detection outputs.

B. Input Embedding

Consider the RGB feature maps $I_R \in \mathbb{R}^{T\times W\times H\times C}$, and the thermal feature maps $I_T \in \mathbb{R}^{T\times W\times H\times C}$. We add a positional embedding E_{pos} to each feature map to encode the position information of each feature pixel. We also add thermal embedding E_T and RGB embedding E_R for the spectra information of thermal spectrum and RGB spectrum, respectively. All embedding are learnable during training.

C. Multi-head Hierarchical Patching and Aggregation

We propose a novel Multi-head Hierarchical Patching and Aggregation (MHHPA) structure to extract both fine-grained and coarse-grained information from both RGB and thermal feature maps. Previous work such as VMamba [22] and vision Mamba [18], as well as vision transformer [44], patch and tokenize the feature maps before flattening features, which reduces spatial resolution. This resolution reduction can effectively reduce the time complexity, but can also cause potential information loss and weaken the models' ability to

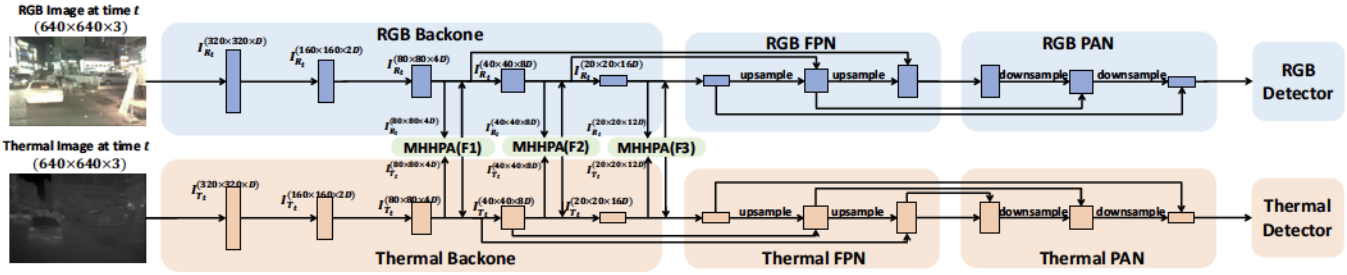

Fig. 1. Visualization of the RGB and thermal object detection network. D denotes the multiplication factor for channel size.

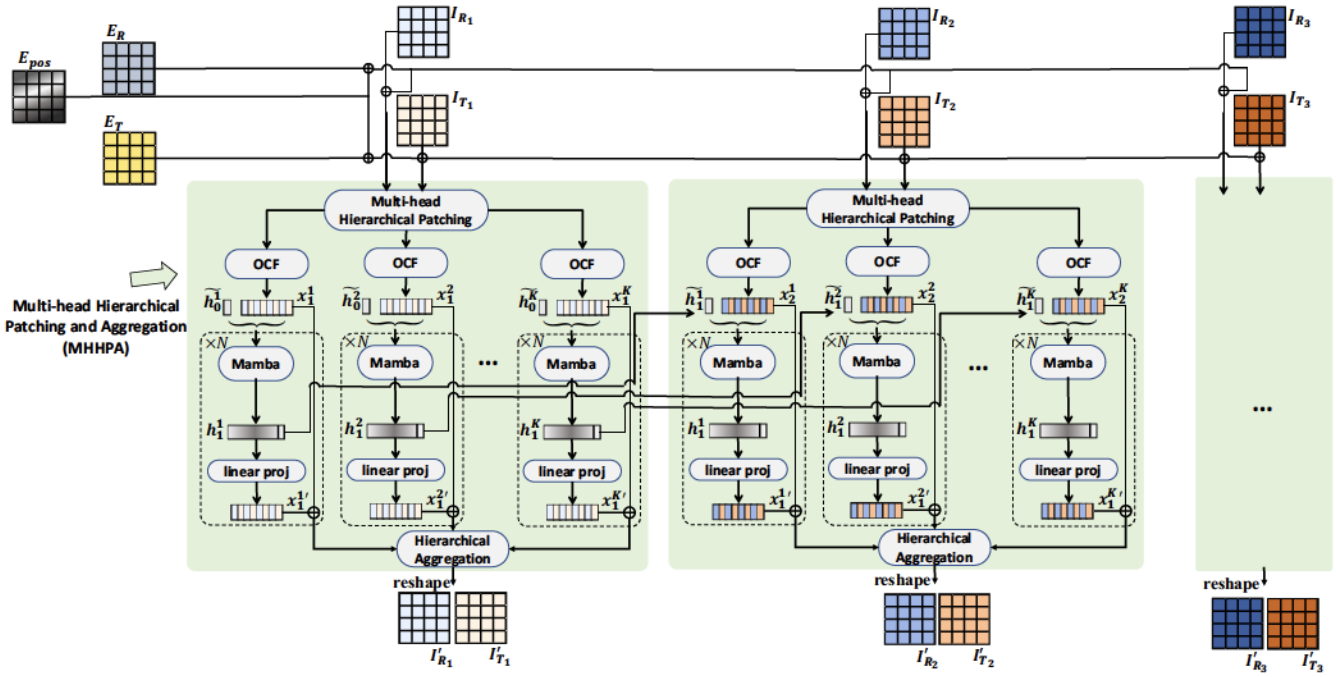

Fig. 2. The proposed *MambaST* pipeline. The input RGB and thermal embeddings are passed through a novel Multi-head Hierarchical Patching and Aggregation (MHHPA) module to extract hierarchical features. An order-aware concatenation and flattening (OCF) procedure is used to concatenate and flatten the patched features. The MHHPA module was applied recurrently to allow for multi-temporal fusion.

extract fine-grained information. On the other hand, prior work in thermal-RGB fusion such as [24] flatten the feature map directly for fusion. We claim this is also sub-optimal as, according to Cheng *et al.* [45], the feature representations of small objects are apt to suffer from noise. In other words, a highly fine-grained information can be useless or even harmful for feature learning. Thus, in this work, we propose a novel MHHPA structure to simultaneously extract both fine-grained and coarse-grained information through hierarchical structures and then combine them. We show later in the Experiments section that the MHHPA module helps improve detection results.

In the MHHPA module, at the t^{th} frame, let $I_{R_t} \in \mathbb{R}^{H\times W\times C}$ and $I_{T_t} \in \mathbb{R}^{H\times W\times C}$ denote the RGB and thermal feature maps, respectively. For different patch sizes and given both thermal and RGB feature maps as inputs, the feature maps I_{T_t} and I_{R_t} are firstly patched to $I^k_{T_t}, I^k_{R_t} \in \mathbb{R}^{H/S_k\times W/S_k\times CS_k^2}$, k is the patch size index. Next, they are concatenated and flattened to $z^k_t \in \mathbb{R}^{2HW/{S_k}^2\times CS_k^2}$ following the order-aware concatenation and flattening (OCF) procedure introduced in Sec. III-D. Then, each flattened patches are linearly projected to $x^k_t = z^k_t \mathcal{W}_k$ and passed to the Mamba block to obtain $x^{k\prime}_t = \text{MambaBlock}(x^{k\prime}_t)$. The output from the MambaBlock is reshaped, split, and added back to the patched feature maps to obtain $I^k_{R_t}{}'$ and $I^k_{T_t}{}'$. This procedure will be repeated N times, where N is the number of Mamba layers. Finally, each $I^k_{R_t}$ and $I^k_{T_t}$ are upsampled to their original size and aggregation together by concatenation.

Alg. 1 shows the pseudocode of this procedure, where $\bigcirc$ denotes function aggregation and $\bigoplus$ represents the concatenation operation over all pixel indices.

D. Order-aware Concatenation and Flattening

In the original structured state space sequence model [16], the translation function, as defined in Eq. 1, handles consecutive data sequences. To maintain the spatial continuity within multi-spectral feature maps (I_R for RGB and I_T for thermal), we introduce an order-aware concatenation and flattening (OCF) procedure.

Denote a series of patch sizes as $S_{1:K}$ with corresponding projection matrix $\mathcal{W}_{1:K}$, where $\mathcal{W}_k \in \mathbb{R}^{CS_k^2\times C/K}$ is responsible to map channel size to C/K for all head. Here, K

Algorithm 1 Multi-head Hierarchical Patching and Aggregation (MHHPA)

for $k = 1$ **to** K **do**
  $I_{R_t}^k \leftarrow \text{Patching}_k(I_{R_t})$
  $I_{T_t}^k \leftarrow \text{Patching}_k(I_{T_t})$
  $z_t^k \leftarrow \text{OCF}(I_{R_t}^k, I_{T_t}^k)$
  $x_t^k \leftarrow z_t^k \mathcal{W}_k$
  $h_t^k \leftarrow \bigcirc_i^N \text{MambaBlock}_i(x_t^k)$
  $x_t^{k\prime} \leftarrow \text{Linear}(h_t^k)$
  $\tilde{I}_{R_t}^{k\prime}, \tilde{I}_{T_t}^{k\prime} \leftarrow \text{ReshapeSplit}(x_t^{k\prime})$
  $I_{R_t}^{k\prime} = I_{R_t}^k + \tilde{I}_{R_t}^{k\prime}$
  $I_{T_t}^{k\prime} = I_{T_t}^k + \tilde{I}_{T_t}^{k\prime}$
end for
$I_{T_t}{}' = \bigoplus_k \text{Upsample}_k(I_{T_t}^{k\prime})$
$I_{R_t}{}' = \bigoplus_k \text{Upsample}_k(I_{R_t}^{k\prime})$
Output $I_{T_t}{}', I_{R_t}{}', h_t^{1:K}$

Algorithm 2 Recurrent Structure for Temporal Fusion

Initialize $t = 1, h_0 = 0$
while I_{R_t} and I_{T_t} exist **do**
$I'_{R_t}, I'_{T_t}, h_t^{1:K} = \text{Alg. } 1(I_{R_t}, I_{T_t}, \tilde{h}_{t-1}^{1:K})$
$\tilde{h}_t^{1:K} = \text{Last}(h_t^{1:K})$
end while

is the number of patch sizes. For each frame at time t and patch size index k, let $I_{R_t}^k, I_{T_t}^k \in \mathbb{R}^{H/S_k \times W/S_k \times CS_k^2}$ denote the feature maps for RGB and thermal spectra, respectively. Each pixel within these maps is indexed by its row i and column j. A pixel at position (i, j) in the RGB and thermal maps is denoted as $I_{R_t,(i,j)}^k$ and $I_{T_t,(i,j)}^k$. The OCF constructs a feature vector x_t^k for the t^{th} frame by interleaving pixels from both feature maps, written as

$$x_t^k = \bigoplus_i [J_{\text{even},t}^k(i), J_{\text{odd},t}^k(i)]$$
$$\text{where} \quad J_{\text{even},t}^k(i) = \bigoplus_{j \text{ is even}} [I_{R_t,(i,j)}^k, I_{T_t,(i,j)}^k] \tag{2}$$
$$J_{\text{odd},t}^k(i) = \bigoplus_{j \text{ is odd}} [I_{R_t,(i,W/S_k-j)}^k, I_{T_t,(i,H/S_k-j)}^k]$$

Here, $\bigoplus$ represents the concatenation operation over all pixel indices (i, j). This approach ensures that the structural integrity and the spatial relationships of the multi-spectral data are maintained in the flattened representation.

E. Recurrent Structure for Temporal Fusion

The structured state space sequential model states that the Mamba translation function (1) resembles a recurrent neural network structure with an input-variant translation function [16]. To perform temporal fusion, we model recurrent connections between temporal frames on top of the MHHPA module. Suppose our MambaST performed fusion for the first t frames and produced hidden vector $h_t \in \mathbb{R}^{WH \times C}$. We take the last hidden output, $\tilde{h}_t \in \mathbb{R}^{1 \times C}$, concatenate it with the flattened feature map of the $t + 1^{\text{th}}$ frame x_{t+1}, and input them into N layers of MambaBlock. This results in the updated outputs $I'_{T_{t+1}}$ and $I'_{R_{t+1}}$ by Alg. 1, with $\tilde{h}_{t+1}$ ready to concatenate to the order-aware flattened feature map of the $(t+2)^{\text{th}}$ frame. The procedure is depicted in Fig. 2, and formulated in Alg. 2.

IV. EXPERIMENTAL RESULTS

A. Dataset and Evaluation Metric

We evaluate our proposed *MambaST* approach on the KAIST Multispectral Pedestrian Detection Benchmark dataset [23]. For training, the sanitized annotations provided by Li. *et. al.* [5] which includes 41 video series with 7,601 images pairs are used. While testing is performed on 25 video series with 2,252 images featuring (nearly) aligned thermal-RGB pairs that capture traffic scenes in both day and night/low-light environments.

We provide evaluation results on two settings from the KAIST benchmark, reasonable and reasonable small. The "reasonable" setting includes non-occluded and partially-occluded pedestrians taller than 55 pixels, and the "reasonable small" setting includes pedestrians between 50 to 75 pixels in height. Both settings use the log-average miss rate (LAMR) over the range of $[10^{-2}, 10^0]$ false positives per image (FPPI). Lower LAMR corresponds to better performance. We also report recall values, where higher recall is desirable (reduces false negative rate). Additionally, to evaluate the efficiency of the algorithm, we report the number of parameters and giga floating point operations (GFLOPs) during inference, where lower number of parameters and lower GFLOPs correspond to smaller number of parameters and floating point operations required for processing the image sequences (lower is regarded as more efficient).

B. Implementation Details

The Multi-head Hierarchical Patching and Aggregation (MHHPA) module employs patch sizes $S_{1:4}^{(1)} = \{1, 2, 4, 8\}$ for the first MHHPA block and omits patching for the subsequent blocks. Patch sizes are constrained to powers of two for dimensional consistency. The number of MambaBlock layer $N = 8$. For the backbone, we follow the standard YOLOv5L setting and set $D = 64$. The number of frames (temporal duration) $\mathcal{T} = 3$, unless otherwise specified (in ablation studies). The KAIST images are of size 640×512, and we pad it to 640×640 (i.e., $H = W = 640$) during training. The original KAIST dataset was captured at 20Hz. To avoid redundancy from consecutive frames, we applied a temporal stride of three, i.e., skipping every two frames. Our proposed network was implemented using Python 3.10.13 and Pytorch 2.1.2, and executed on NVIDIA A100 GPUs.

C. Comparison with Other Cross-Modal Fusion Methods

We evaluate our proposed *MambaST* fusion module against the fusion sources (RGB only and Thermal only), as well as a basic feature addition strategy and an advanced Cross-Modality Fusion Transformer (CFT) [38]. In the basic feature addition approach, the RGB and thermal features were simply added and the resulting feature maps were broadcasted across modalities. This serves as a baseline comparison. For a more advanced cross-modality fusion approach, we

TABLE I

PEDESTRIAN DETECTION RESULTS ON THE KAIST DATASET (FULL AND SMALL).

Fusion Method	LAMR(%)↓ All	Day	Night	Recall(%)↑ All	LAMR(%)↓ All-Small	Day-Small	Night-Small	Recall(%)↑ All-Small
RGB only	13.96	16.72	12.47	99.61	18.11	18.48	19.31	99.01
Thermal only	15.54	19.64	8.28	99.52	20.67	22.19	18.59	99.20
Feat. Add.	12.47	15.31	7.48	99.18	16.61	19.02	11.80	98.01
CFT [38]	11.34	13.37	7.26	98.42	16.76	18.88	12.43	96.87
T-CFT	10.38	12.16	7.15	97.73	16.11	17.68	12.06	96.49
D-CFT	8.68	11.45	4.53	98.69	15.21	16.51	12.97	96.59
MambaST (Ours)	**6.67**	**8.67**	**3.12**	**99.86**	**11.37**	**13.56**	**7.32**	**99.34**

TABLE II

EFFICIENCY COMPARISON BETWEEN OUR MAMBAST AND CFT VARIANTS ON THE KAIST DATASET.

	Param.(M)↓ F1	F2	F3	Total	GFLOPs↓ F1	F2	F3	Total	Latency (ms)↓ F1	F2	F3	Total
CFT [38]	5.12	25.29	100.89	131.3	4.86	19.35	77.37	101.61	**8.7**	8.2	8.8	**25.6**
T-CFT	6.42	25.41	101.17	133.01	4.87	19.37	77.39	101.63	8.9	8.3	8.8	26.0
T-CFT (w/o DS)	16.15	30.13	103.23	149.51	2907.72	2903.4	2901.24	8712.36	570.0	532.0	562.8	1664.8
D-CFT	**2.86**	9.62	34.94	47.42	90.6	84.93	82.1	257.63	20.9	20.7	20.4	62.0
MambaST (Ours)	3.07	**3.80**	**15.64**	**22.52**	**1.83**	**1.82**	**1.79**	**5.43**	25.8	**6.8**	**6.5**	39.1

TABLE III

ABLATION STUDY ON THE KAIST DATASET WHEN VARYING $\mathcal{K}$ (DIFFERENT PATCH SIZES), OCF, AND THE NUMBER OF FRAMES $\mathcal{T}$ IN TRAINING. THE MEDIAN, MAX, AND MIN VALUES OF MISS RATE WERE REPORTED ACROSS FIVE RUNS.

$\mathcal{K}$ F1	F2	F3	OCF	$\mathcal{T}$	LAMR (Reasonable) ↓ median	max	min	LAMR (Reasonable small) ↓ median	max	min	Para. (M) ↓ F1	F2	F3	total	GFLOPs ↓ F1	F2	F3	total
1	1	1		3	7.22	7.86	6.85	12.5	13.08	12.19	8.5	16.3	55.6	80.4	2.12	1.90	1.79	5.81
2	1	1		3	7.30	7.70	7.10	13.68	14.3	12.68	5.1	16.3	55.6	77.0	1.96	1.90	1.79	5.65
2	2	1		3	7.19	7.54	7.06	12.39	12.61	12.02	5.1	9.2	55.6	69.9	1.96	1.82	1.79	5.57
4	1	1		3	**6.78**	**6.95**	**6.6**	**11.94**	**12.84**	**11.3**	4.1	16.3	55.6	75.9	1.84	1.90	1.79	5.53
4	2	1		3	6.87	7.58	6.8	12.8	13.15	11.68	4.1	9.2	55.6	68.9	1.84	1.82	1.79	5.45
4	2	2		3	7.48	7.88	7.14	13.72	14.13	13.38	4.1	9.2	30.5	43.8	1.84	1.82	1.75	5.40
4	4	2		3	7.81	8.16	7.38	14.21	14.61	13.48	4.1	10.1	30.5	44.7	1.84	1.76	1.75	5.34
4	1	1	✓	1	7.44	7.82	6.89	11.87	13.02	11.60	4.1	16.3	55.6	75.9	1.84	1.90	1.79	5.53
4	1	1	✓	3	6.73	7.2	6.62	11.37	12.33	10.92	4.1	16.3	55.6	75.9	1.84	1.90	1.79	5.53
4	1	1	✓	7	**6.32**	**6.82**	**6.20**	**11.11**	**11.68**	**10.25**	4.1	16.3	55.6	75.9	1.84	1.90	1.79	5.53

compare to CFT [38], a top-ranking cross modality fuser for pedestrian detection. Note that the original (vanilla) CFT model only works for a single frame. To account for temporal fusion, we implemented three variations of CFT for comprehensive comparison. 1) *CFT* model, where the original CFT was applied frame-to-frame; 2) *T-CFT* model, where the temporal information was integrated by concatenating feature maps from all timesteps, written as

$$x'_{i,k} = x_{i,k} + \text{CFT}(\text{Concat}_i^N(x_{i,k})); \tag{3}$$

and 3) *D-CFT* model, a deformable variant that replaces standard self-attention in the transformers with deformable attention [46] to handle temporal data more efficiently, written as

$$x'_{i,k} = x_{i,k} + \text{DCFT}(\text{Concat}_i^N(x_{i,k})). \tag{4}$$

Table I shows the pedestrian detection results on the KAIST dataset. Overall, the fusion methods outperform the single-modality sources (first two rows), which indicates the necessity of cross-spectral fusion. Thermal-only produced lower miss rate at Night than RGB-only, while RGB-only performed better during the Day. Among the CFT variants, D-CFT (CFT with deformable attention) performed the best compared with temporal concatenation (T-CFT) and original CFT. Our proposed *MambaST* outperforms the single-modality baselines as well as all other cross-modal fusion models for both day and night settings on the KAIST dataset, with the lowest log-average miss rate of 6.67% overall, 8.67% during the day, 3.12% at night, and a highest 99.86% recall.

D. Evaluation on Small Object Detection

Following the KAIST benchmark settings, pedestrians beteween 50 to 75 pixels in height are considered small-sized objects. The last four columns in Table I reports the detection results specifically the small-scaled pedestrians. As shown, the CFT and T-CFT, which use arithmetic addition as a fusion strategy, performed poorly. This is likely due to resolution reduction via average pooling, which removed the fine-grained information for small-scaled objects. Our *MambaST* kept full resolution before the fusion step and produced superior performance across all settings, with a LAMR of 11.37% overall, 13.56% during the day, and 7.32% at night, and achieving the highest recall rate of 99.56%. This demonstrates the effectiveness of our *MambaST* approach in detecting small-scale pedestrians.

E. Efficiency Evaluation

Table II reports the number of parameters, gigafloating point operations (GFLOPs), and latency (ms) during inference. Note that in our T-CFT experiments, the feature maps were

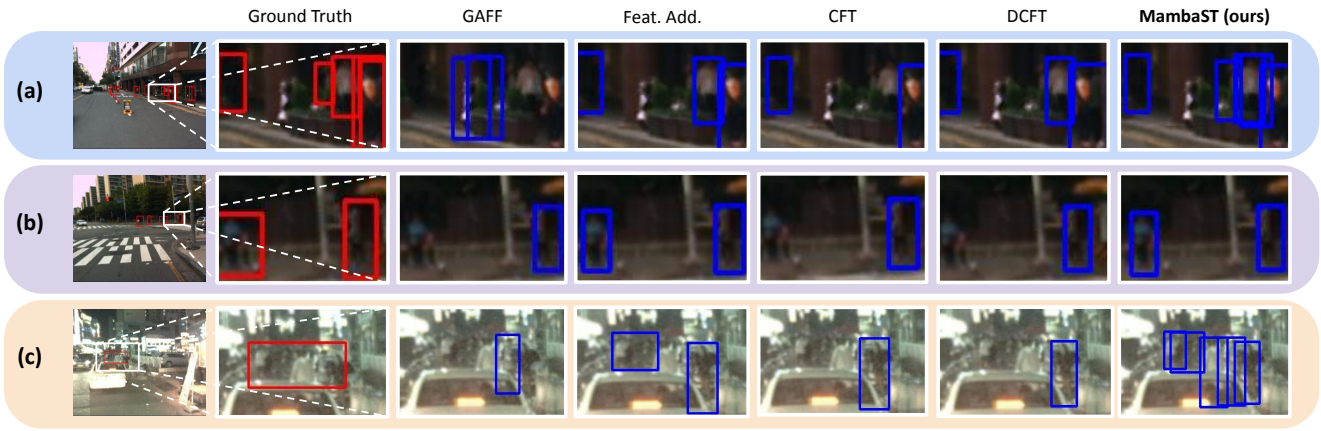

Fig. 3. Visual examples of detection results on the KAIST dataset. All bounding boxes are filtered by confidence ≥ 0.5.

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART OF THE KAIST DATASET WITH "SANITIZED" ANNOTATION. **BEST**; SECOND BEST IN MAMBAST PERFORMANCE.

Method	LAMR(%↓)		
	All	Day	Night
MSDS-RCNN [5]	7.49	8.09	5.92
GAFF [7]	6.48	8.35	3.46
MS-DETR [40]	6.13	7.78	3.18
CFR [47]	6.13	7.68	3.19
MambaST (1 frame)	7.44	9.65	3.42
MambaST (3 frames)	6.67	8.67	3.12
MambaST (7 frames)	6.32	8.22	**2.88**

first downsampled to 8×8 before fusion. This reduces the number of parameters and produced better results empirically. The T-CFT module, without downsampling (w/o DS), has 149.51M parameters and very high GFLOPs (8712.36). The downscaled T-CFT and the original CFT had significantly lower GFLOPs (101.63 and 33.87, respectively). The deformable variant of CFT (D-CFT) requires fewer parameters but has high GFLOPs. In contrast, our *MambaST* was able to achieve superior detection performance while requiring the lowest number of parameters (22.52M) and GFLOPs (5.43). Our *MambaST* also has a relatively low inference latency (less than 40ms).

F. Comparison with State-of-the-art Methods

Table IV shows the comparison results against the state-of-the-art fusion methods on the KAIST dataset with "sanitized" annotations. MSDS-RCNN [5] combines a convolution neural network-based multispectral proposal network and a multispectral classification network to perform fusion. GAFF [7] proposes inter- and intra-modality attention modules to dynamically weigh and fuse the multispectral features. MS-DETR [40] fuses RGB and thermal features in a multi-modal Transformer decoder and adaptively learns the attention weights. CFR [47] cyclically fuses and refines spectral feature from each modality to achieve cross-spectral complementary/consistency balance. Our *MambaST* achieved competitive detection performance compared with the state-of-the-art methods, and achieves superior detection performance at Night. As the number of input frames increased, our detection performance also improves, achieving a low 2.88% LAMR at Night given seven frames as input.

G. Ablation Studies

We conducted multiple sets of ablation studies to evaluate the effect of parameter choices. To reduce the result variance and ensure fairness, we trained using the entire set and selected the checkpoint from the 10th epoch for testing. We also trained the model five times with different seeds for each hyperparameter setup and reported the median overall LAMR.

First, we varied tested different numbers of patch sizes ($\mathcal{K}$) across MHHPA blocks, as outlined in Table III. The patch sizes range from one size ($\mathcal{K} = 1$) to four sizes ($\mathcal{K} = 4$) per block, tailored to maintain powers of two for consistent embedding dimensions. The first seven rows in Table III show that the (4,1,1) setting, i.e., using four patch sizes in the first MHHPA block and omitting patching for the subsequent blocks, achieves lowest Log-average Miss Rate (LAMR) in both "reasonable" and "reasonable small" settings of the KAIST dataset without excessive computational overhead.

Second, we evaluated the impact of the Order-aware Concatenation and Flattening module (OCF). We observed based on row 4 and row 9 in Table III that incorporating OCF further enhanced detection performance, reducing the median LAMR from 6.78% to 6.73% in the "reasonable" setting and from 11.94% to 11.37% in the "reasonable small" setting.

Third, we performed further experiments varying the number of frames (temporal duration $\mathcal{T}$ =1, 3, and 7). The last three rows in Table III show that our model's performance improves with the number of frames used, achieving the lowest LAMR with seven frames as input. This makes sense as one of the advantage of a Mamba-based model is its strength in modeling longer sequences. Future work will include evaluation on longer sequences and other datasets.

H. Visual Results

Fig. 3 shows example visual results for pedestrian detection on the KAIST dataset. We present our proposed *MambaST* model results, compared to feature addition, CFT, D-CFT, and GAFF models. In row (a), we observed that the CFT model failed to detect some small pedestrians near the center of the scene, while other methods, including the naive feature addition strategy was able to correctly detect more pedestrians. This implies the importance of avoiding resolution reduction in cross-spectral spatial temporal fusion. Similarly, in row (b), our MambaST successfully detected the pedestrians in the scene. Row (c) presents an interesting case, where some annotations in the KAIST dataset were noisy. It loosely labeled several pedestrians in the same bounding box. As shown, our MambaST model was still able to correctly detect multiple pedestrians in the scene.

V. CONCLUSION

We propose *MambaST*, a Mamba-based spatial-temporal fusion pipeline for effective and efficient cross-spectral pedestrian detection. By utilizing the novel Multi-head Hierarchical Patching and Aggregation (MHHPA) module, *MambaST* efficiently handles the complexities of cross-spectral data without the excessive computational overhead commonly associated with similar models. The MHHPA module can be easily swapped (e.g, with a CFT) and can plug-and-play with a variety of detectors. Our experiments on the KAIST dataset show superior detection performance, particularly in low-light conditions and for small pedestrian detection.

REFERENCES

[1] A. Baul, W. Kuang, J. Zhang, H. Yu, and L. Wu, "Learning to detect pedestrian flow in traffic intersections from synthetic data," in *IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2639–2645.

[2] H. Zhang, Z. Li, Z. Wu, and D. Wang, "A lightweight rgb-t fusion network for practical semantic segmentation," in *IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 4233–4238.

[3] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 221–230, 2022.

[4] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[5] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," *arXiv preprint arXiv:1808.04818*, 2018.

[6] L. Zhang, Z. Liu, S. Zhang, *et al.*, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, pp. 20–29, 2019.

[7] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 72–80.

[8] Y.-T. Chen, J. Shi, C. Mertz, S. Kong, and D. Ramanan, "Multimodal object detection via bayesian fusion," *arXiv preprint arXiv:2104.02904*, vol. 3, no. 6, 2021.

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[11] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5552–5561.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[13] Z. Huang, S. Zhang, L. Pan, *et al.*, "Tada! temporally-adaptive convolutions for video understanding," in *International Conference on Learning Representations*, 2021.

[14] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 708–13 718.

[15] S. Yang, X. Wang, Y. Li, *et al.*, "Temporally efficient vision transformer for video instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2885–2895.

[16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[19] H. Zhang, Y. Zhu, D. Wang, L. Zhang, T. Chen, and Z. Ye, "A survey on visual mamba," *arXiv preprint arXiv:2404.15956*, 2024.

[20] R. Xu, S. Yang, Y. Wang, B. Du, and H. Chen, "A survey on vision mamba: Models, applications and challenges," *arXiv preprint arXiv:2404.18861*, 2024.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object de-

tection," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[22] Y. Liu, Y. Tian, Y. Zhao, *et al.*, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[23] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[24] W. Dong, H. Zhu, S. Lin, *et al.*, "Fusion-mamba for cross-modality object detection," *arXiv preprint arXiv:2404.09146*, 2024.

[25] H. Li, Q. Hu, Y. Yao, K. Yang, and P. Chen, "Cfmw: Cross-modality fusion mamba for multispectral object detection under adverse weather conditions," *arXiv preprint arXiv:2404.16302*, 2024.

[26] S. Peng, X. Zhu, H. Deng, Z. Lei, and L.-J. Deng, "Fusionmamba: Efficient image fusion with state space model," *arXiv preprint arXiv:2404.07932*, 2024.

[27] K. Li, X. Li, Y. Wang, *et al.*, "Videomamba: State space model for efficient video understanding," *arXiv preprint arXiv:2403.06977*, 2024.

[28] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 9, pp. 15 940–15 950, 2022.

[29] Q. Li, C. Zhang, Q. Hu, P. Zhu, H. Fu, and L. Chen, "Stabilizing multispectral pedestrian detection with evidential hybrid fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[30] A. González, G. Villalonga, J. Xu, D. Vázquez, J. Amores, and A. M. López, "Multiview random forest of local experts combining rgb and lidar data for pedestrian detection," in *IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 356–361.

[31] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 7–12.

[32] M. Fürst, O. Wasenmüller, and D. Stricker, "Lrpd: Long range 3d pedestrian detection leveraging specific strengths of lidar and rgb," in *IEEE 23rd international conference on intelligent transportation systems (ITSC)*, 2020, pp. 1–7.

[33] F. Farahnakian and J. Heikkonen, "Rgb-depth fusion framework for object detection in autonomous vehicles," in *14th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2020, pp. 1–6.

[34] H. Rashed, A. El Sallab, S. Yogamani, and M. ElHelw, "Motion and depth augmented semantic segmentation for autonomous navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[35] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5127–5137.

[36] Y. Zhang, Z. Yin, L. Nie, and S. Huang, "Attention based multi-layer fusion of multispectral images for pedestrian detection," *IEEE Access*, vol. 8, pp. 165 071–165 084, 2020.

[37] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *European Conference on Computer Vision*, Springer, 2022, pp. 139–158.

[38] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

[39] W.-Y. Lee, L. Jovanov, and W. Philips, "Cross-modality attention and multimodal fusion transformer for pedestrian detection," in *European Conference on Computer Vision*, Springer, 2022, pp. 608–623.

[40] Y. Xing, S. Wang, S. Zhang, G. Liang, X. Zhang, and Y. Zhang, "Ms-detr: Multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization," *arXiv preprint arXiv:2302.00290*, 2023.

[41] Y. Zhu, X. Sun, M. Wang, and H. Huang, "Multi-modal feature pyramid transformer for rgb-infrared object detection," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[42] S. Lee, T. Kim, J. Shin, N. Kim, and Y. Choi, "Insanet: Intra-inter spectral attention network for effective feature fusion of multispectral pedestrian detection," *Sensors*, vol. 24, no. 4, p. 1168, 2024.

[43] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[45] G. Cheng, X. Yuan, X. Yao, *et al.*, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[46] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[47] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *2020 IEEE International conference on image processing (ICIP)*, 2020, pp. 276–280.