MOCID: Motion Context and Displacement Information Learning for Moving Infrared Small Target Detection

Mingjin Zhang[1], Yuanjun Ouyang[1*], Fei Gao[1], Jie Guo[1], Qiming Zhang[2*], Jing Zhang[3]
[1]Xidian University, China
[2]School of Computer Science, The University of Sydney, Australia
[3]School of Computer Science, Wuhan University, China
{mjinzhang, fgao, jguo}@xidian.edu.cn, 23011211078@stu.xidian.edu.cn, {qmzhangzz, jingzhang.cv}@gmail.com

Abstract

In the field of Moving Infrared Small Target Detection (MIRSTD), current methods typically use sequential modeling with two individual modules for spatial and temporal processing. However, such a modeling strategy lacks clear guidance on the motion and displacement difference between moving targets and background noise, thereby limiting the feature discriminability and resulting in error-prone target localization. This paper addresses this issue from clip and frame levels and proposes a novel architecture MOCID for MIRSTD. For clip-level feature fusion, we design a spatio-temporal backbone consisting of several proposed Fourier-inspired Spatio-temporal Attention (FISTA) layers. Each FISTA layer sequentially processes the features from spatial and temporal views to capture clip-level temporal motion context, where Fourier Transformation and Inverse Fourier Transformation are employed for each view. This context is then embedded into dynamic convolutional kernels for subsequent spatial feature extraction, thereby enabling clear motion difference guidance and generating comprehensive features. For frame-level feature fusion, we design a Displacement-aware Mamba Module (DAM) to capture detailed frame-to-frame displacement information. DAM utilizes an innovative Temporal Interpolation and Displacement-aware Scan technique to perform spatio-temporal difference-aware displacement modeling, introducing elaborate temporal indicators into feature extraction. Combining the above improvements, our model captures comprehensive motion and displacement contexts, significantly improving the detection of the small target. Extensive experiments demonstrate that MOCID achieves state-of-the-art detection accuracy on popular IRDST and DAUB datasets. Furthermore, MOCID offers a superior balance between throughput and performance compared to other methods.

Code — https://github.com/TanzanOY/MOCID

Introduction

Moving Infrared Small Target Detection (MIRSTD) involves locating the moving small targets within the target frame by modeling spatio-temporal information across target frame and its reference frames. Figure 1 illustrates the challenges associated with MIRSTD: (1) Small target sizes complicate the extraction of informative features, such as shape and texture, and feature preservation in deep neural network (Zhang et al. 2023). (2) Low contrast due to factors such as complex backgrounds and non-target noise with similar appearance and shape hinders target recognition. While the first challenge can be effectively addressed by integrating semantic-rich high-level features with detail-rich low-level features (Dai et al. 2020; Zhang et al. 2019), the second still remains difficult to overcome (Zhang et al. 2024).

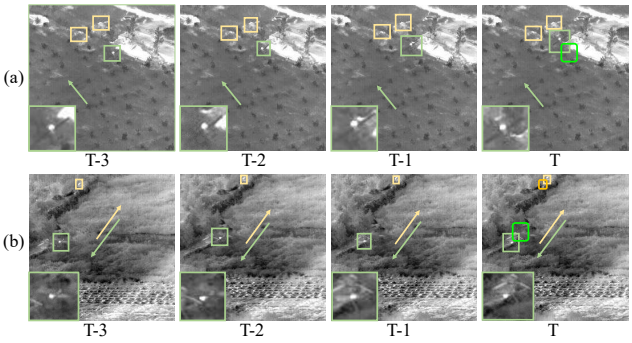

Figure 1: Video clips from two infrared videos. The target and noise regions are highlighted with green and yellow boxes, respectively. Green and yellow arrows denote target and background noise motion, respectively. The lighter green and yellow boxes in time T denote the historical position. The clear clip-level motion difference and frame-level displacement difference illustrate the motivation of MOCID.

As shown in Figure 1, targets typically exhibit distinct motion patterns compared to background noise across multiple frames. Therefore, recent works (Tong et al. 2024; Li et al. 2023) address the challenges in MIRSTD by leveraging these motion and temporal contexts. Most existing methods adopt a spatial-then-temporal framework: a spatial-only feature extractor first independently extracts features from each frame, and then a carefully designed temporal module models the temporal contexts to enhance target features. In terms of temporal contexts utilization, semantic-based methods (Gong et al. 2021; Cui et al. 2021; Zhang et al. 2020) rely on modeling temporal semantic similarity across frames to extract useful target features, while spatio-

*Corresponding author

temporal self-attention-based methods (Zhou et al. 2023; Tong et al. 2024) learn spatio-temporal dependencies within consecutive frames to improve target frame features. Additionally, DTUM (Li et al. 2023) and SSTNet (Chen et al. 2024) explore modeling motion patterns within input clips to enhance target features. Despite achieving excellent performance with robust temporal modeling modules, we argue that such sequential frameworks, which rely on individual modules, rarely incorporate motion and temporal contexts during spatial modeling. The spatial-only feature extractor extracts spatial features without utilizing valuable motion difference information between targets and noise, making it difficult to distinguish target features from noise features and resulting in noise-interfered target representations. This can degrade the performance of temporal modeling in subsequent modules, leading to error-prone target localization.

In addition, frame-to-frame displacement information is crucial for accurate target locating in complex scenarios. As shown in Figure 1 (a), when the target moves into a complex background, it can be easily occluded by noise with similar appearance. An effective approach is to learn the detailed spatial displacement between frames to identify the target. For example, the general video object detection (GVOD) method PTSEformer (Wang et al. 2022) learns the spatial transition information between frames to enhance target frame features using self-attention-based gated correlation. However, the spatial transition method in PTSEformer primarily addresses moderate-range displacement differences in GVOD and is not specifically tailored for perceiving the more subtle pixel changes in the MIRSTD domain, resulting in limited performance improvement for MIRSTD. This underscores the need to develop a specialized frame-to-frame spatio-temporal difference-aware mechanism for the MIRSTD task.

Based on the discussion above, we propose a novel approach, MOCID, for MIRSTD and improve the performance from clip and frame levels. For the clip level, we introduce a spatio-temporal backbone to effectively leverage the clip-level temporal motion difference between moving targets and background noise. Specifically, the backbone comprises several stacked Fourier-inspired Spatio-temporal Attention (FISTA) layers. Each layer sequentially conducts Fourier Transformation and Inverse Transformation along the spatial and temporal dimensions, aiming to capture comprehensive motion information from the input clip. This information is embedded into dynamic convolution kernels, providing clear guidance of the motion difference between target and background noise, and enabling the ability to dynamically adjust spatial feature extraction for each frame during subsequent feature extraction. By stacking FISTA layers in our backbone, the discriminative feature representations are gradually generated using motion context derived from FISTA.

For the frame level, we propose a Displacement-aware Mamba Module (DAM) to learn displacement information between infrared frames, leveraging the capabilities of Mamba. DAM introduces a novel technique called Temporal Interpolation and Difference-aware Selective Scan (TIDS). Unlike the common linear selection approach in Mamba, TIDS uses 3D central difference convolution (Yu et al. 2021) to capture spatio-temporal context-sensitive system parameters, which are then employed in the subsequent scanning process. Additionally, we propose a unique Temporal Interpolation strategy for frame-to-frame displacement information modeling. It interpolates the target frame with the reference frames in both width and height directions to construct interpolated sequences. Consequently, we develop a displacement-aware scanning approach that uses spatio-temporal-sensitive parameters to scan along these interpolated sequences. This displacement-aware mechanism is rarely considered in other Mamba methods for video tasks (Li et al. 2024; Yang, Xing, and Zhu 2024), thereby leading to inferior performance compared to our DAM.

The contributions of this study has three folds:

- We propose a novel model, MOCID, for MIRSTD. By leveraging clip-level temporal motion context and frame-to-frame displacement information as guidance, MOCID effectively generates discriminative features, enabling the distinction between moving targets and background noise. We demonstrate MOCID's state-of-the-art detection accuracy on public IRDST and DAUB datasets. Furthermore, MOCID achieves a significantly superior balance between throughput and performance compared to other methods. Extensive ablation studies demonstrate the effectiveness of the proposed components.
- We propose Fourier-inspired Spatio-temporal Attention (FISTA) to capture motion context and construct a spatio-temporal backbone with FISTA that dynamically adjusts spatial feature extraction using FISTA-derived temporal motion context.
- We design a Displacement-Aware Mamba Module (DAM) to model the frame-to-frame displacement information. DAM employs novel Temporal Interpolation and Difference-Aware Selective Scan technique to capture subtle spatio-temporal pixel changes between infrared frames. To the best of our knowledge, DAM is the first exploration of spatio-temporal displacement learning using Mamba for MIRSTD.

Related Work

Moving Infrared Small Target Detection

The standard approach for MIRSTD involves initially extracting spatial features from each input frame separately and then enhancing these features using temporal aggregation modules, termed as the spatial-then-temporal framework. Troi (Gong et al. 2021) proposes a method to extract similar temporal contexts using a novel Temporal Align operator. TransVOD++ (Zhou et al. 2023) utilizes a Temporal Deformable Transformer for aggregating temporal contexts. DTUM (Li et al. 2023) models target and clutter motion via motion-to-data mapping and motion encoding. YOlOv (Shi, Wang, and Guo 2023) selects important regions identified by YOlOx (Ge et al. 2021) detection and performs temporal aggregation on these candidates. SSTNet (Chen et al. 2024) proposes a sliced spatio-temporal network that learns cross-slice motion to enhance target features. Despite

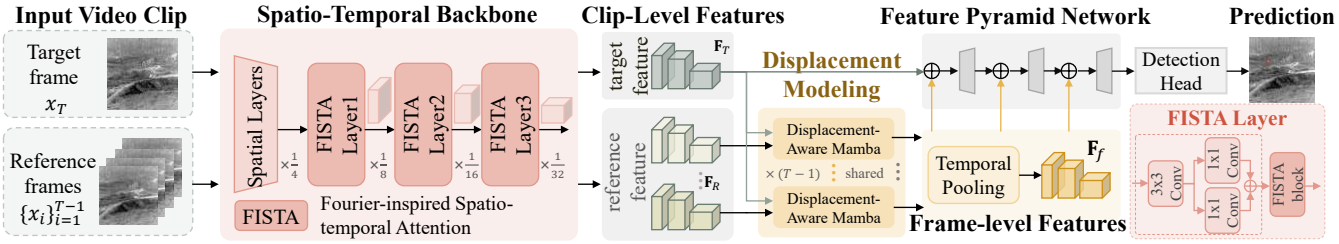

Figure 2: The overall pipeline of MOCID. MOCID comprises a spatio-temporal backbone, a displacement modeling network, a feature pyramid network, and a detection head. The Spatio-Temporal Backbone extracts motion contexts across input frames and dynamically adjusts the spatial feature extraction on each frame with derived contexts. The Displacement Modeling Network models frame-to-frame displacement information between the target frame and each reference frame to enhance representations of small targets. The feature pyramid network fuses features from both the clip and frame levels, while the detection head predicts the final result.

achieving excellent performance with robust temporal aggregation modules, we argue that such sequential spatial-temporal modeling using individual modules rarely considers temporal motion information when performing spatial modeling, which can lead to error-prone target localization.

Infrared Small Target Detection in Still Images

Single-frame methods primarily focus on designing spatial feature enhancement and fusion modules to integrate low-level and high-level features, thereby enhancing representations of small targets. ACMNet (Dai et al. 2020) devises an asymmetric contextual modulation module for detecting infrared small targets. RKformer (Zhang et al. 2022a) introduces Runge-Kutta transformer to learn discriminative small target features. ISNet (Zhang et al. 2022c) proposes detecting shapes of infrared small targets using Taylor finite difference-inspired edge blocks. MSHNet (Liu et al. 2024a) designs scale and location sensitive loss and multi-scale head to improve detection accuracy. Single-frame methods fail to capture temporal contexts in infrared video data, thereby limiting their performance in detecting moving small targets.

Vision Mamba

The Mamba architecture represents the latest generation of State Space Models (SSMs) (Gu and Dao 2023; Dao and Gu 2024). Mamba enhances SSMs by dynamically adjusting its parameters based on the input sequence (Li et al. 2024), notable for its capability to model long-range dependencies with linear complexity. Recently, Mamba has demonstrated significant success across various vision tasks. Vision Mamba (Zhu et al. 2024) introduces bidirectional SMMs that perform bidirectional scanning on image tokens. VMamba (Liu et al. 2024b) devises a 2D Selective Scan module for processing visual data. In the domain of video applications, Video Mamba (Li et al. 2024) proposes a forward and backward Spatio-Temporal SMM for processing video data. Vivim (Yang, Xing, and Zhu 2024) introduces spatio-temporal selective scanning to investigate spatio-temporal correlations in video. However, recent video Mamba frameworks rarely incorporate spatio-temporal displacement information into their modeling process.

Our Approach

Overview

Pipeline of MOCID is illustrated at Figure 2. Given an infrared video clip $\{x_i\}_{i=1}^{T}$, x_T is the target frame and preceding frames $\{x_i\}_{i=1}^{T-1}$ are reference frames. MOCID takes $\{x_i\}_{i=1}^{T}$ as input and output results on x_T. It comprises four main components: (1) Spatio-temporal Backbone (STB) that extracts spatial features on each frame with guidance of clip-level motion contexts; (2) Displacement Modeling Network consists of $T-1$ weight-sharing Displacement-aware Mamba modules (DAMs) for frame-level displacement information modeling. A Temporal Pooling function is employed to fuse the results of DAMs; (3) Feature Pyramid Network (Lin et al. 2017) for multi-level feature fusion; and (4) Detection Head from (Ge et al. 2021) for final detection.

Spatio-temporal Backbone

The spatio-temporal backbone (STB) is introduced to leverage the clip-level temporal motion differences between moving targets and background noise to enhance the spatial feature extraction on each input frame. The spatial model CSPDarknet (Bochkovskiy, Wang, and Liao 2020) serves as the baseline for implementing STB. We retain the first two layers of CSPDarknet for low-level spatial feature extraction and replace the last three spatial layers with three cascaded Fourier-inspired Space-time Attention (FISTA) layers. The FISTA layer, illustrated at the bottom-right of Figure 2, contains n convolution blocks for spatial modeling and n proposed FISTA blocks. FISTA block first captures clip-level motion contexts and then dynamically adjust spatial feature extraction with derived contexts. By stacking FISTA layers, the temporally enhanced features from the previous layer are further enhanced with temporal motion contexts in the subsequent layer, delivering comprehensive feature representations in spatial feature for each frame.

Fourier-inspired Spatio-temporal Attention Block

Figure 3 illustrates the details of FISTA block. The FISTA block encompasses motion context capturing using FISTA

and adjustment of spatial feature extraction with the motion context.

Fourier-inspired Spatio-temporal Attention. FISTA follows a spatial-temporal pipeline to capture motion context. We denote the input to FISTA block as $\mathbf{f} \in \mathbb{R}^{T\times C\times H\times W}$. As in GFNet (Rao et al. 2021), FISTA applies 2D spatial Discrete Fourier Transform (DFT) (Duhamel and Vetterli 1990) on each frame to obtain $F_s \in \mathbb{C}^{T\times C\times H\times W}$. F_s is then multiplied by a learnable spatial global filter $\mathcal{K} \in \mathbb{C}^{C\times H\times W}$ to model the spatial global context in frequency domain. The process can be expressed as:

$$F_s = \mathrm{DFT}_{2\mathrm{D}}(\mathbf{f}), \quad \overline{F}_s = F_s \odot \mathcal{K}, \tag{1}$$

where $\odot$, $\mathrm{DFT}_{2\mathrm{D}}$ denote element-wise multiplication and 2D DFT, respectively. An inverse IDFT is applied on $\overline{F}_s$ to transform it to spatial domain, resulting in $\mathbf{f}_s \in \mathbb{R}^{T\times C\times H\times W}$. Since the discriminative features of small targets mainly exist in the high frequency component, e.g. edge, applying 2D DFT and learnable spatial global $\mathcal{K}$ on each image help distinguish the small target features.

The overall motion can be characterized as changes across multiple frames for each spatial pixel. FISTA models such temporal dynamics in the frequency domain. FISTA first applies 1D temporal DFT on each spatial pixel of $\mathbf{f}_s \in \mathbb{R}^{T\times C\times H\times W}$ to convert it into the frequency domain:

$$F_t = \mathrm{DFT}_{1\mathrm{D}}(\mathbf{f}_s). \tag{2}$$

We consider two scenarios: (1) Moving targets with static background: Theoretically, the high-frequency component corresponds to moving targets, while the low-frequency component corresponds to static background (Kothandaraman et al. 2022); (2) Moving targets with moving noise induced by camera motion: The high-frequency component can also correspond to noise. However, the distinction in velocities between targets and noise can lead to a different high-frequency distribution. Therefore, we introduce a learnable temporal global filter $\mathcal{K}_t \in \mathbb{C}^{T\times C\times 1\times 1}$ to uncover the motion of small moving targets in the frequency domain:

$$\overline{F}_t = F_t \odot \mathcal{K}_t. \tag{3}$$

A 1D IDFT is applied to $\overline{F}_t$ to obtain $\widehat{\mathbf{f}} \in \mathbb{R}^{T\times C\times H\times W}$. The amplitude of each pixel in $\widehat{\mathbf{f}}$ represents the temporal dynamics of the signal at each spatio-temporal location. Regions associated with target movement exhibit higher amplitudes. Therefore, we multiply the original features $\mathbf{f}$ by the L_2 normalization of $\widehat{\mathbf{f}}$:

$$\mathbf{M} = \mathbf{f} \odot \|\widehat{\mathbf{f}}\|_2. \tag{4}$$

After applying FISTA, we obtain $\mathbf{M}$, the clip-level motion context, which captures the overall movement of the target.

Motion context guided spatial modeling. To effectively leverage the motion context $\mathbf{M}$ captured by FISTA, and inspired by recent studies (Huang et al. 2021; Cao et al. 2022) on calibrating spatial feature extraction with temporal contexts, we dynamically adjust the spatial convolution weights for each frame using $\mathbf{M}$. Specifically, the output features on each frame from FISTA block can be obtained as:

$$\mathbf{f}_{out} = \mathbf{W_t} * \mathbf{f} = (\alpha_t \cdot \mathbf{W_b}) * \mathbf{f}, \tag{5}$$

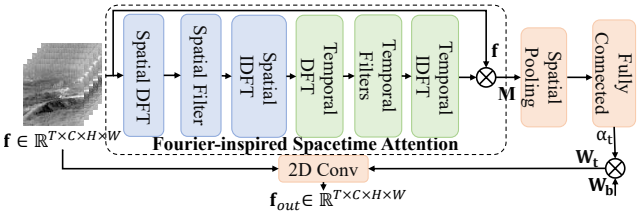

Figure 3: Details of FISTA block. The Fourier-inspired Space-time attention captures motion context across input clip using Discrete Fourier Transform (DFT) and learnable filters. Subsequently, the derived motion context dynamically adjusts spatial feature extraction on each frame. IDFT refers to Inverse Discrete Fourier Transform.

where $\mathbf{f}, \mathbf{f}_{out} \in \mathbb{R}^{T\times C\times H\times W}$, $*$ signifies convolution operation, $\mathbf{W_b} \in \mathbb{R}^{C\times C\times k^2}$ (k is the kernel size) denotes a base weight for each frame, $\alpha_t \in \mathbb{R}^{T\times C\times 1\times 1}$ is the calibration weight derived from motion contexts, and $\mathbf{W_t} \in \mathbb{R}^{T\times C\times C\times k^2}$ is the calibrated weight on each frame. Based on $\mathbf{M}$ derived from FISTA, α_t is calculated as follows:

$$\alpha_t = \mathrm{FC}(\mathrm{GAP_s}(\mathbf{M})), \tag{6}$$

where $\mathrm{GAP_s}$ denotes spatial global average pooling, and FC represents a fully connected layer operating across the temporal dimension for temporal modeling.

Displacement-aware Mamba Module

Mamba Architecture. Mamba models long-range dependencies within an input sequence using input-dependent system parameters with linear complexity (Gu and Dao 2023). Consider the input sequence $X = [x_1, \cdots, x_L] \in \mathbb{R}^{L\times C}$, where L represents sequence length, and C signifies the feature dimension. $x_k \in \mathbb{R}^C$ is a vector from X. Mamba maps x_k to $y_k \in \mathbb{R}^C$ via hidden state $h_k \in \mathbb{R}^N$, utilizing system parameters including the evolution matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, and projection matrices $\mathbf{B} \in \mathbb{R}^{N\times C}$, $\mathbf{C} \in \mathbb{R}^{C\times N}$ as follows:

$$h_k = \overline{\mathbf{A}} h_{k-1} + \overline{\mathbf{B}} x_k, \quad y_k = \mathbf{C} h_{k-1}, \tag{7}$$

$$\overline{\mathbf{A}} = e^{\Delta \mathbf{A}}, \quad \overline{\mathbf{B}} = (e^{\Delta \mathbf{A}} - I)\mathbf{A}^{-1}\mathbf{B}, \tag{8}$$

where N is the hidden dimension, $\overline{\mathbf{A}}, \overline{\mathbf{B}}$ represents discretizations of $\mathbf{A}$ and $\mathbf{B}$ using zero-order hold and timescale $\Delta \in \mathbb{R}^{L\times N}$. The essence of Mamba lies in its selective scan mechanism, where selection involves generating input-dependent system parameters $\overline{\mathbf{A}}, \overline{\mathbf{B}}$ and $\overline{\mathbf{C}}$ with transformation function ϕ (i.e., $\mathbf{C}, \mathbf{B}, \Delta = \phi(x)$) and Eq. (8), and scan refers to utilizing parallel scan algorithms (Liu et al. 2024b) to efficiently compute Eq. (7) in a recurrent manner, adapting to dynamic input-dependent parameters.

Our design. Building on Mamba's properties of input sensitivity and efficiency, we propose DAM to model frame-to-frame displacement information between the target frame and each reference frame. The details of DAM are illustrated in Figure 4. Since pixel changes between frames in MIRSTD scenarios are subtle, a system sensitive to fine-grained spatio-temporal differences can more effectively

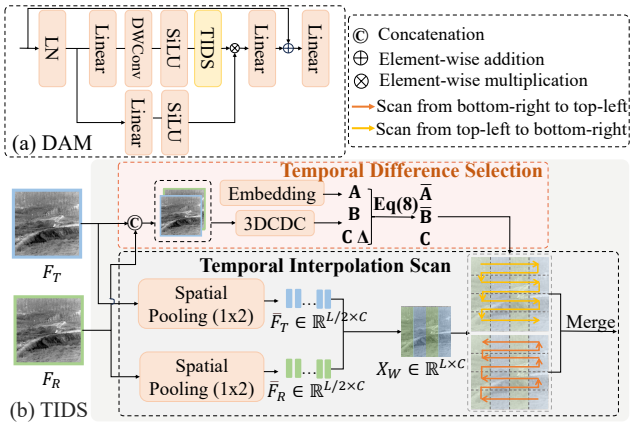

Figure 4: (a) Details of DAM. (b) Details of TIDS. We illustrate the scanning on X_W in top-left to bottom-right and bottom-right to top-left directions. X_W represents width-interpolated sequence. The other similar operation is conducted along the height dimension and omitted here for a clear illustration.

capture this information. To achieve this, DAM devises the novel Temporal Interpolation and Difference-aware Selective Scan (TIDS), which consists of two main processes: Spatio-temporal Difference Selection (SDS) and Temporal Interpolation Scan (TIS).

SDS. Given target frame $F_T \in \mathbb{R}^{C\times H\times W}$ and one reference frame $F_R \in \mathbb{R}^{C\times H\times W}$, they are concatenated along the temporal dimension, resulting in $x \in \mathbb{R}^{2\times C\times H\times W}$. Existing methods (Gu and Dao 2023; Li et al. 2024) employ linear transformation for the selection function ϕ, which cannot effectively model fine-grained spatio-temporal differences in x. In contrast, we adopt ϕ with 3D Central Difference Convolution (Yu et al. 2021) (3DCDC), which captures detailed local spatio-temporal features for modeling subtle pixels changes. We derive system parameters that are responsive to the spatio-temporal differences of the input by computing $\mathbf{B}, \mathbf{C}, \Delta = \text{3DCDC}(x)$. Subsequently, $\overline{\mathbf{A}}, \overline{\mathbf{B}}$, and $\overline{\mathbf{C}}$ are computed based on $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and Δ with Eq. (8).

TIS. After SDS, we initially perform spatial pooling (denoted as SP in Eq. (9) and (10)) along the width dimension using a kernel size of (1×2) on F_T and F_R, and reshape them into $\mathbb{R}^{L\times C}(L = HW)$, respectively:

$$\overline{F}_T = \text{SP}(F_T) = \left[f_T^1, f_T^2, \ldots, f_T^{L/2}\right] \in \mathbb{R}^{L/2\times C}, \quad (9)$$

$$\overline{F}_R = \text{SP}(F_R) = \left[f_R^1, f_R^2, \ldots, f_R^{L/2}\right] \in \mathbb{R}^{L/2\times C}. \quad (10)$$

Subsequently, $\overline{F}_T$ is interpolated with $\overline{F}_R$ to create a width-interpolated input sequence X_W,

$$\begin{aligned} X_W &= \text{Interpolation}(\overline{F}_T, \overline{F}_R) \\ &= \left[f_R^1, f_T^1, f_R^2, f_T^2, \ldots, f_R^{L/2}, f_T^{L/2}\right] \in \mathbb{R}^{L\times C}. \end{aligned} \quad (11)$$

A similar operation is conducted along the height dimension with pooling kernel size of (2×1), resulting in $X_H \in \mathbb{R}^{L\times C}$.

Then, we employ the expanding scan method from (Liu et al. 2024b) to perform scans from top-left to bottom-right and bottom-right to top-left on X_W, resulting in two scanned sequences. Theses two sequences are merged via addition and reshaped into one feature map. A similar procedure is also applied on X_H, and the results from X_W and X_H are also merged via addition to obtain the final enhanced output features. Scanning from multiple directions is crucial for aggregating representations of small targets due to their sizes.

Analysis and Discussion. Let reformulate Eq. (7) in matrix form with X_W:

$$\mathbf{C} = diag(\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_L) \quad (12)$$

$$Y = [y_1, y_2, y_3, ..., y_L]^T = \mathbf{CH}X_W = \quad (13)$$

$$\mathbf{C}\begin{bmatrix} \overline{\mathbf{B}}_1 & 0 & \cdots & 0 \\ \overline{\mathbf{B}}_1\overline{\mathbf{A}}_2 & \overline{\mathbf{B}}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mathbf{B}}_1\prod_{i=2}^{L}\overline{\mathbf{A}}_i & \overline{\mathbf{B}}_2\prod_{i=3}^{L}\overline{\mathbf{A}}_i & \cdots & \overline{\mathbf{B}}_L \end{bmatrix}\begin{bmatrix} f_R^1 \\ f_T^1 \\ \vdots \\ f_T^{L/2} \end{bmatrix},$$

where X_W captures displacement information between two frames along width. For instance, given $\overline{F}_T = [b, t, b, b]$ and $\overline{F}_R = [b, b, t, b]$, where b and t represent background and target tokens respectively, X_W is formed as $[b, b+t, b+t, b]$ by interpolating $\overline{F}_T$ with $\overline{F}_R$. Elements in the matrix $\mathbf{CH}$ encompass rich spatio-temporal difference contexts relevant to X_W. Consequently, following matrix computation, DAM can model subtle pixel changes between frames in MIRSTD scenarios and learn the detailed frame-to-frame displacement information for MIRSTD accordingly. The same analysis can also be conducted on X_H.

Experiments

Experimental Details

Datasets. We evaluate MOCID on two MIRSDT datasets, DAUB (Hui et al. 2019) and IRDST (Sun et al. 2023). DAUB is a widely used MIRSDT dataset comprising 17 video sequences totaling 13,778 frames. IRDST consiss of 92 video sequences and 40,656 frames captured from real-world scenarios. Following (Chen et al. 2024), for DAUB, we select 10 videos totaling 8,983 frames for training and 7 videos totaling 4,795 frames for validation, and for IRDST, we use 42 videos comprising 20,398 frames for training and 43 videos comprising 20,258 frames for validation.

Evaluation Metrics. We assess MOCID's performance using Average Precision with IoU threshold 0.5 (AP_{50}), precision (*Pr*), recall (*Re*), and F_1. AP_{50} measures the accuracy of target detection by considering detections as true positives if they overlap with ground truth by at least 50%. *Pr* evaluates the model's accuracy in positive predictions, while *Re* measures its ability to correctly identify positives among all actual positives. F_1 is the harmonic mean of Pr and Re, offering a balanced metric for evaluating both *Pr* and *Re*.

Implementation. The algorithm is implemented in PyTorch using the SGD optimizer with specific configurations: weight decay and momentum are set to 5×10^{-4} and 0.937.

Method	DAUB				IRDST			
	AP_{50}	Pr	Re	F_1	AP_{50}	Pr	Re	F_1
DNANet (Li et al. 2022) $*$	89.93	92.49	98.27	95.29	63.61	82.92	77.48	80.11
ISNet (Zhou et al. 2023) $*$	83.43	89.36	94.99	92.09	59.78	80.24	75.08	77.58
UIU-Net(Wu, Hong, and Chanussot 2022) $*$	86.41	94.46	92.03	93.23	56.38	80.95	70.29	77.58
FC3-Net (Zhang et al. 2022b)	81.92	93.07	88.83	90.90	82.44	88.37	89.87	89.36
MTU-Net (Wu et al. 2023)	84.43	96.23	92.83	94.50	83.04	89.30	90.07	89.68
YOLOx (Ge et al. 2021)	83.59	94.27	89.34	91.74	87.25	95.55	94.52	95.03
TransVOD++ (Zhou et al. 2023)	76.69	94.35	83.4	93.90	82.25	89.37	83.89	86.54
YOLOv (Shi, Wang, and Guo 2023)	86.76	94.97	88.92	91.85	89.37	93.19	92.55	92.87
SSTNet (Chen et al. 2024) $*$	95.59	98.08	**98.10**	98.09	71.55	88.56	81.92	85.11
LSTFE-Net (Xiao et al. 2023)	87.23	97.54	94.41	95.95	88.30	97.14	93.41	95.24
CSPDarknet+DTUM (Li et al. 2023)	88.72	97.58	91.39	94.38	92.14	97.95	94.16	96.02
MOCID (Ours)	**95.93**	**99.12**	97.34	**98.22**	**94.74**	**98.92**	**96.86**	**97.88**

Table 1: Performance comparisons on DAUB and IRDST. $*$ represents that the results are cited from (Chen et al. 2024).

The initial learning rate and learning rate reduction coefficient are set to 0.01 and 0.1, respectively. Input frames are resized to 512×512, and the length of the input video clip T is set to 5. For data augmentation, random flipping of video clips is employed. To supervise the training, we employ Binary Cross Entropy loss for classification and IoU loss for regression: $\mathcal{L} = \mathcal{L}_{reg} + \mathcal{L}_{cls}$. The spatio-temporal backbone is initially trained for 100 epochs, after which it is frozen while we proceed to train the DAM for an additional 100 epochs. Training utilizes two RTX3090 GPUs, while testing is performed using a single RTX3090 GPU.

Main Results

Comparisons with State-of-the-art. Comparisons on DAUB and IRDST are shown in Table 1. We compare our method primarily with the top-performing single-image and video methods. For video approaches, we employ a multi-stage training pipeline. The model is pre-trained using still images in both datasets and then equipped with temporal modules to process video data. Specifically, for DTUM (Li et al. 2023), we select our baseline spatial backbone CSPDarknet (Bochkovskiy, Wang, and Liao 2020) as DTUM's spatial feature extractor. As shown in the Table 1, our method achieves SOTA performance on AP_{50}, Pr and F_1 on DAUB, and AP_{50}, Pr, Re and F_1 on IRDST.

Speed-accuracy Tradeoff. Figure 5 visualizes the speed-accuracy comparisons on DAUB with 5 top-performing methods on MIRSTD (SSTNet, DTUM), video small object detection (LSTFE-Net) and video object detection (YOLOv, TransVOD++). Speed is measured with FPS. MOCID achieves a tradeoff between speed and accuracy.

Ablation Study

Impact of Each Module. The ablation study of FISTA and DAM is presented in Table 2. The base backbone is the spatial-only CSPDarknet. Table 2 illustrates the positive effects of both FISTA and DAM, with their combination yielding the best results. This is because FISTA captures motion contexts to enhance the features of small targets, while

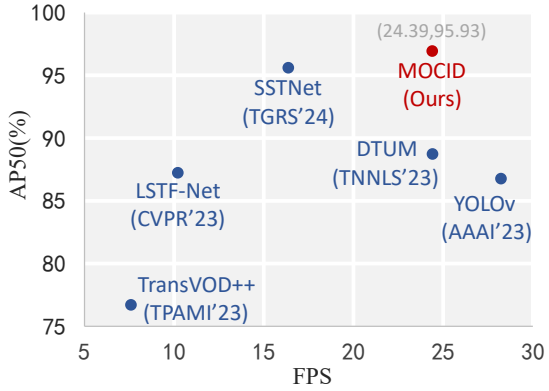

Figure 5: Speed-accuracy comparisons on DAUB.

	AP_{50}	F_1	Para. (M)	Time (s)
Base	83.59	91.74	8.94	0.003
.+Convs	85.43	93.22	13.51	0.003
.+FISTA	92.42	96.40	9.45	0.022
.+DAM	90.22	95.14	12.54	0.032
.+FISTA+DAM (MOCID)	95.93	98.22	13.05	0.041

Table 2: Ablations on the proposed components on DUAB.

DAM learns displacement information to address noise occlusion issues. The second row in Table 2 indicates that we replace FISTA blocks with an equivalent number of convolution blocks. The results demonstrate that the improvement with FISTA is attributed to the incorporation of motion contexts rather than simply increasing model depth. To further investigate the effectiveness of FISTA and DAM, we visualize the heatmaps from different modules in MOCID in Figure 6. Compared to the baseline, FISTA obtain more distinguished target representations. After DAM, the activation in the noise and background gets reduced.

Cost Analysis. Table 2 presents an analysis of model size and inference time for each component of MOCID. FISTA

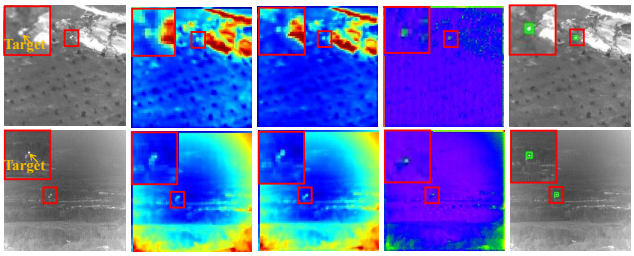

Figure 6: Visualization of heatmaps. (a) The original target frame. (b) The heatmap from spatial-only base model. (c) The heatmap after FISTA. (d) The featmap after FISTA and DAM (MOCID). (f) The detection result.

and DAM introduce only a minor increase in parameters compared to the baseline. The baseline operates as a spatial-only model processing one frame at a time, whereas our method processes T frames simultaneously, resulting in increased inference time. Considering the trade-offs between detection accuracy and computational cost, our approach demonstrates significant performance improvements.

Impact of FISTA. To ablate FISTA, we compare it with recent methods that dynamically adjust spatial feature extraction using temporal information: Tada (Huang et al. 2021) and TadaCNN (Cao et al. 2022). We replace the FISTA block with these methods. Results are shown in Figure 7. FISTA delivers the best performance because it captures more comprehensive motion information from the input clip in the frequency domain compared to other methods.

Impact of DAM. To ablate DAM, we compare it with recent video applications using Mamba architecture on DAUB: Video Mamba (Li et al. 2024) and Vivim (Yang, Xing, and Zhu 2024). DAM consists of two primary processes: SDS and TIS. As shown in Figure 8, applying SDS leads to performance improvements across all methods. When DAM uses only TIS, it still achieves the best results. Combining SDS and TIS allows DAM to deliver optimal performance, because they complement each other: SDS provides spatio-temporal difference-sensitive parameters, while TIS constructs sequences containing displacement information between frames. In addition, DAM shares the concept of modeling frame-to-frame displacement information with PTSEFormer (Wang et al. 2022). Building on FISTA, we compare our DAM with self-attention-based Spatial Transition Awareness Module (STAM) from PTSEFormer. Figure 9 presents the results. DAM achieves superior performance. because it can perceive more subtle pixel changes in MIRSTD compared to STAM, owing to its spatio-temporal difference-aware mechanism.

Impact of Input Video Length. Table 3 illustrates the impact of the input video clip length T on accuracy and inference speed. An increase in T within a certain range provides additional temporal information. A sequence of 5 or 6 frames contains sufficient temporal details, while sequences exceeding 6 frames may encounter significant motion changes, resulting in a decline in detection accuracy and speed.

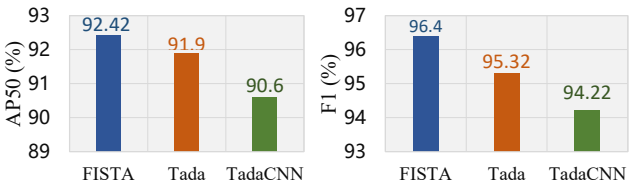

Figure 7: Comparisons on FISTA, with Tada and TadaCNN on DAUB.

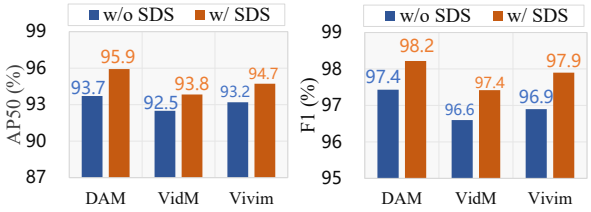

Figure 8: Impact of DAM and SDS on different video Mamba models, including Video Mamba (VidM) and Vivim.

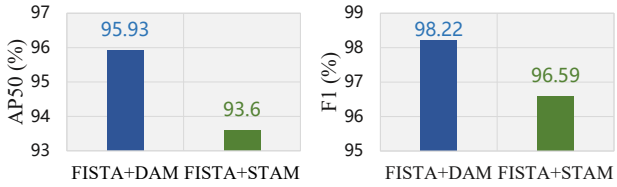

Figure 9: Comparisons on proposed DAM and self-attention-based STAM from PTSEFormer on DUAB.

T	3	4	**5**	6	7	8	9
AP_{50}	87.57	92.31	**95.93**	95.91	94.47	95.10	93.32
Time (s)	0.023	0.034	**0.041**	0.048	0.060	0.068	0.074

Table 3: Impact on the length of input video clip on DAUB.

Conclusion

This paper introduces MOCID, a novel model for MIRSTD. MOCID leverages clip-level motion context and frame-level displacement information in infrared videos through two key designs: the Spatio-temporal Backbone (STB) and the Displacement-aware Mamba Module (DAM). Within STB, we capture motion context using the proposed Fourier-inspired Spatio-temporal Attention (FISTA) and dynamically adjust spatial feature extraction for each frame based on derived context. DAM incorporates a novel Temporal Interpolation and Difference-aware Scan to model subtle pixel changes between frames. These components synergistically enhance spatio-temporal feature modeling, leading MOCID to outperform existing methods on the public IRDST and DAUB datasets. Additionally, ablation studies validate the efficacy of MOCID's key modules.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272363, Grant 92470108; in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2021QNRC001; in part by the Joint Laboratory for Innovation in Satellite-Borne Computers and Electronics Technology Open Fund 2023 under Grant 2024KFKT001-1; in part by the Proof of Concept Foundation of Xidian University Hangzhou Institute of Technology under Grant No. GNYZ2023YL0301; in part by the Fundamental Research Funds for the Central Universities under Grant No. ZYTS24012.

References

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934.

Cao, Z.; Huang, Z.; Pan, L.; Zhang, S.; Liu, Z.; and Fu, C. 2022. TCTrack: Temporal Contexts for Aerial Tracking. arXiv:2203.01885.

Chen, S.; Ji, L.; Zhu, J.; Ye, M.; and Yao, X. 2024. SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Cui, Y.; Yan, L.; Cao, Z.; and Liu, D. 2021. TF-Blender: Temporal Feature Blender for Video Object Detection. arXiv:2108.05821.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2020. Asymmetric Contextual Modulation for Infrared Small Target Detection. arXiv:2009.14530.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Duhamel, P.; and Vetterli, M. 1990. Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing*, 19(4): 259–299.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

Gong, T.; Chen, K.; Wang, X.; Chu, Q.; Zhu, F.; Lin, D.; Yu, N.; and Feng, H. 2021. Temporal RoI Align for Video Object Recognition. arXiv:2109.03495.

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

Huang, Z.; Zhang, S.; Pan, L.; Qing, Z.; Tang, M.; Liu, Z.; and Ang Jr, M. H. 2021. Tada! temporally-adaptive convolutions for video understanding. *arXiv preprint arXiv:2110.06178*.

Hui, B.; Song, Z.; Fan, H.; Zhong, P.; Hu, W.; Zhang, X.; Lin, J.; Su, H.; Jin, W.; Zhang, Y.; et al. 2019. A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background. *Sci. Data Bank*, 5(12): 4.

Kothandaraman, D.; Guan, T.; Wang, X.; Hu, S.; Lin, M.; and Manocha, D. 2022. FAR: Fourier Aerial Video Recognition. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, 657–676. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-19835-9.

Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; and Guo, Y. 2022. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32: 1745–1758.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*.

Li, R.; An, W.; Xiao, C.; Li, B.; Wang, Y.; Li, M.; and Guo, Y. 2023. Direction-coded temporal U-shape module for multiframe infrared small target detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. arXiv:1612.03144.

Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; and FU, Y. 2024a. Infrared Small Target Detection with Scale and Location Sensitivity. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17490–17499.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024b. VMamba: Visual State Space Model. arXiv:2401.10166.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global Filter Networks for Image Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shi, Y.; Wang, N.; and Guo, X. 2023. YOLOV: Making Still Image Object Detectors Great at Video Object Detection. arXiv:2208.09686.

Sun, H.; Bai, J.; Yang, F.; and Bai, X. 2023. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Tong, X.; Zuo, Z.; Su, S.; Wei, J.; Sun, X.; Wu, P.; and Zhao, Z. 2024. ST-Trans: Spatial-Temporal Transformer for Infrared Small Target Detection in Sequential Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, H.; Tang, J.; Liu, X.; Guan, S.; Xie, R.; and Song, L. 2022. PTSEFormer: Progressive Temporal-Spatial Enhanced TransFormer Towards Video Object Detection. arXiv:2209.02242.

Wu, T.; Li, B.; Luo, Y.; Wang, Y.; Xiao, C.; Liu, T.; Yang, J.; An, W.; and Guo, Y. 2023. MTU-Net: Multilevel TransUNet for Space-Based Infrared Tiny Ship Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Wu, X.; Hong, D.; and Chanussot, J. 2022. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32: 364–376.

Xiao, J.; Wu, Y.; Chen, Y.; Wang, S.; Wang, Z.; and Ma, J. 2023. LSTFE-net: Long short-term feature enhancement network for video small object detection. In *the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 14613–14622.

Yang, Y.; Xing, Z.; and Zhu, L. 2024. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*.

Yu, Z.; Zhou, B.; Wan, J.; Wang, P.; Chen, H.; Liu, X.; Li, S. Z.; and Zhao, G. 2021. Searching multi-rate and multimodal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 30: 5626–5640.

Zhang, M.; Bai, H.; Zhang, J.; Zhang, R.; Wang, C.; Guo, J.; and Gao, X. 2022a. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1730–1738.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2019. Deep Latent Low-Rank Representation for Face Sketch Synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10): 3109–3123.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2020. Neural Probabilistic Graphical Model for Face Sketch Synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7): 2623–2637.

Zhang, M.; Yang, H.; Guo, J.; Li, Y.; Gao, X.; and Zhang, J. 2024. IRPruneDet: efficient infrared small target detection via wavelet structure-regularized soft channel pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7224–7232.

Zhang, M.; Yue, K.; Zhang, J.; Li, Y.; and Gao, X. 2022b. Exploring Feature Compensation and Cross-level Correlation for Infrared Small Target Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1857–1865. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; and Guo, J. 2022c. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 877–886.

Zhang, M.; Zhang, R.; Zhang, J.; Guo, J.; Li, Y.; and Gao, X. 2023. Dim2Clear Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Zhou, Q.; Li, X.; He, L.; Yang, Y.; Cheng, G.; Tong, Y.; Ma, L.; and Tao, D. 2023. TransVOD: End-to-End Video Object Detection With Spatial-Temporal Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7853–7869.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv:2401.09417.