Article

Not peer-reviewed version

TVEMamba: Enhanced Thermal Video for Better Object

Sargis Hovhannisyan , Sos Agaian , Karen Panetta , Artyom Grigoryan *

Posted Date: 24 December 2024

doi: 10.20944/preprints202412.2099.v1

Keywords: Thermal Video Enhancement; Mamba Model; Motion Deblurring; Video Processing

 doi:10.20944/preprints202412.2099.v1

Article

TVEMamba: Enhanced Thermal Video for Better Object Tracking

Sargis Hovhannisyan [1], Sos Agaian [2], Karen Panetta [3] and Artyom Grigoryan [4,*]

[1] Faculty of Mathematics and Mechanics, Yerevan State University, Yerevan, 0025, Armenia; sargis.hovhannisyan@ysu.am
[2] Department of Computer Science, School of Engineering, City University of New York (CUNY), New York, NY 10031, USA; sos.agaian@csi.cuny.edu
[3] Department of Electrical & Computer Engineering, School of Engineering, Tufts University, Medford, MA 02155, USA; karen@eecs.tufts.edu
[4] Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX 78249, USA; artyom.grigoryan@utsa.edu
* Correspondence: artyom.grigoryan@utsa.edu

Abstract: Object tracking in thermal video is challenging due to noise, blur, and low contrast. We present TVEMamba, a Mamba-based enhancement framework with near-linear complexity that improves tracking in these conditions. Our approach uses a State Space 2D (SS2D) module integrated with Convolutional Neural Networks (CNNs) to filter, sharpen, and highlight important details. Key components include (i) a denoising module to reduce background noise and enhance image clarity, (ii) an optical flow attention module to handle complex motion and reduce blur, and (iii) entropy-based labeling to create a fully labeled thermal dataset for training and evaluation. TVEMamba outperforms existing methods (DCRGC, RLBHE, IECGAN, BBCNN) across multiple datasets (BIRDSAI, FLIR, CAMEL, Autonomous Vehicles, Solar Panels) and achieves higher scores on standard quality metrics (EME, BDIM, DMTE, MDIMTE, LGTA). Extensive tests, including ablation studies and convergence analysis, confirm its robustness. Real-world examples, such as tracking humans, animals, and moving objects for self-driving vehicles and remote sensing, demonstrate the practical value of TVEMamba.

Keywords: thermal video enhancement; mamba model; motion deblurring; video processing

1. Introduction

Visually appealing videos are essential not only for human perception but also for advanced computer vision tasks. Unfortunately, many videos are captured under challenging conditions that cause poor visibility, structural degradation, and unpredictable noise. These issues significantly reduce the performance of automated image analysis systems and object-tracking algorithms used in surveillance [1], monitoring [2], intelligent transportation [3], and remote sensing [4]. Object tracking involves identifying and following objects across visible and thermal video frame sequences. Traditional methods often rely on the Kalman filter, assuming linear motion. However, these methods struggle when objects exhibit complex, nonlinear motions or when videos suffer from uneven illumination, motion blur, and noise.

Many algorithms have been developed to enhance videos in the visible spectrum (VIS) [5], yet they still face challenges in varying illumination and low-light conditions. Improving low-light or nighttime imagery requires a deep understanding of how light interacts with the scene and how images are formed. In response, thermal imaging has emerged as a reliable solution, especially in environments where VIS images perform poorly. For real-world applications, sensors that operate in different spectrums are often necessary to accommodate changing lighting conditions. Studies have shown that thermal infrared data can improve the accuracy of object tracking, semantic segmentation, saliency detection, and object detection. Thermal videos are also widely used in wildlife monitoring

 doi:10.20944/preprints202412.2099.v1

[6], surveillance [7], military operations [8], security [9], and remote sensing [10]. Nevertheless, they pose their own difficulties, such as low contrast, motion blur, and loss of fine details, making it harder to detect targets or enhance infrared imaging technologies, as shown in Figure 1(a).

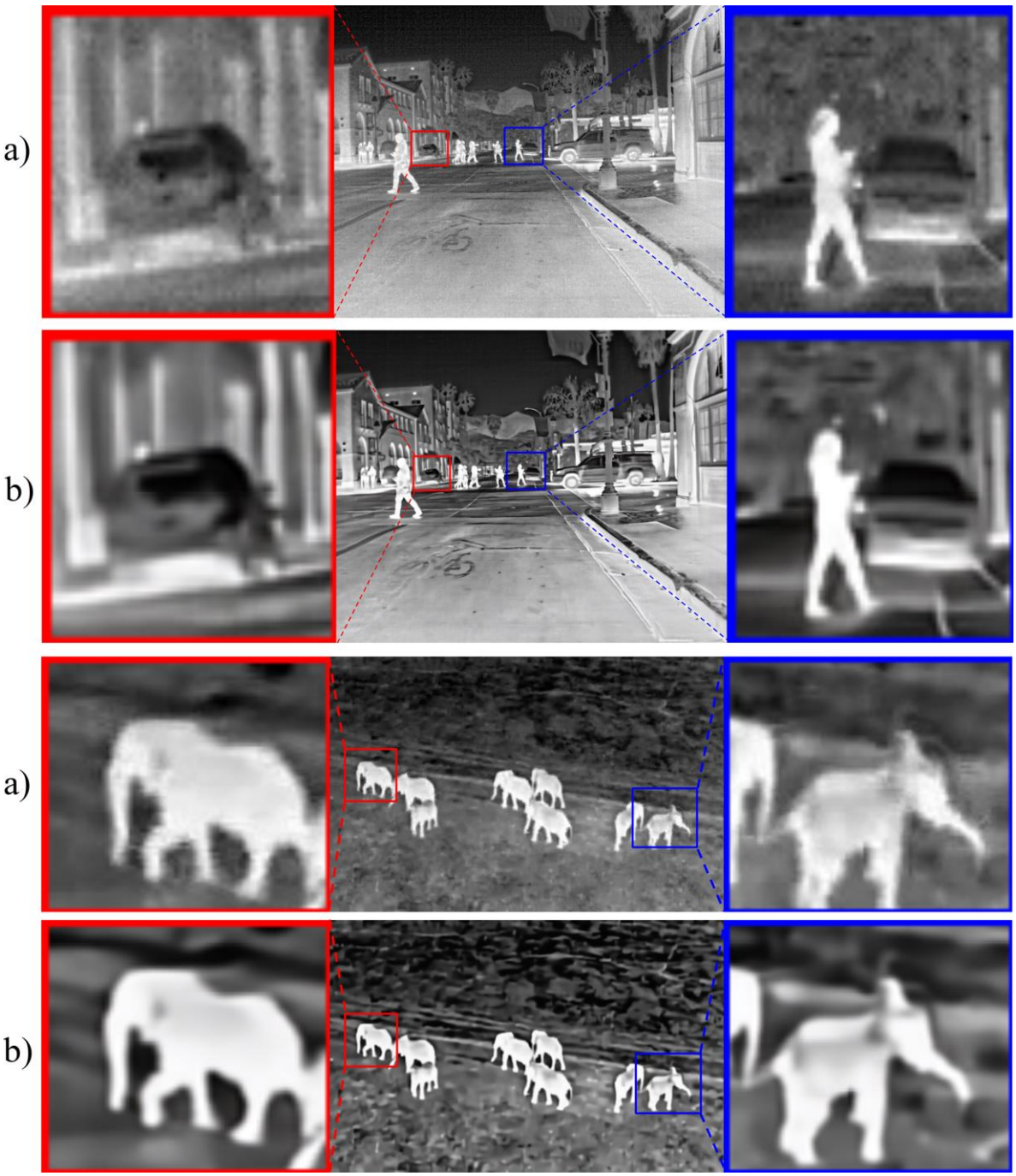

Figure 1. (**a**) Challenging thermal video frames. (**b**) Successful recovery and enhancement by TVEMamba.

Thermal Video Enhancement (TVE) techniques aim to improve the visual quality of thermal footage for automated processing tasks such as analysis, detection, segmentation, and recognition. Recent approaches can be grouped into two categories: methods that use information from neighboring frames to guide enhancement and methods that process each frame independently.

Some properties of the state-of-the-art TVE methods are provided in Table 1. These methods aim to enhance image quality by addressing noise, contrast imbalance, and loss of texture details. However, as shown in the table, many approaches are limited by their inability to handle diverse scenarios effectively, often leading to artifacts, inconsistent brightness, or distorted features in challenging conditions. Overcoming these challenges is crucial to achieve robust performance in real-world

 doi:10.20944/preprints202412.2099.v1

applications. One potential direction involves incorporating blur-resistant motion deblurring methods that leverage the inherent properties of thermal scenes, providing more reliable and adaptive enhancement.

Table 1. Comparative analysis of thermal image enhancement methods across key performance metrics (✓: Fully Performs, ±: Partially Performs).

	DCRGC	RLBHE	IE-CGAN	BBCNN	TVEMamba
Noise reduction			✓		✓
Balanced contrast	✓	✓	±	±	✓
Handles underexposed areas					✓
Handles overexposed areas	±	✓		±	✓
Edge preservation			±	±	✓
Maintains natural brightness				±	✓
Handles complex textures					✓
Artifact-free output					✓

Enhancing thermal videos is inherently more complex than working with visible-light footage. It involves dealing with low image contrast, sensor noise, rapid motion, and limited spatial resolution, which vary with environmental factors, target types, and imaging devices, as shown in Figure 1. Illumination inconsistencies, camera jitter, and atmospheric effects further complicate the task, demanding algorithms that can adapt to different conditions. Addressing these issues can support various applications, from navigation and safety in autonomous systems to reliable object detection in surveillance under unpredictable environments.

This paper focuses on enhancing the quality of thermal videos for practical, real-world applications, including handling complex scenarios such as illumination effects, cluttered backgrounds, noise, or other environmental effects. It presents the Mamba model, a novel approach to thermal video enhancement. The Mamba network uses a SS2D module integrated with a CNN to stabilize and improve thermal video quality. By incorporating a Basic Denoising (BD) module, our method effectively reduces noise and enhances image detail. Meanwhile, the Optical Flow Attention (OFA) module introduces blur-resistant motion deblurring, ensuring that dynamic scenes remain clear and visually coherent. Unlike traditional methods that rely on simplistic assumptions, the Mamba network dynamically adapts to diverse motion patterns and lighting conditions, making it suitable for various practical scenarios. The main contributions are as follows:

1. We introduce a novel Mamba model for thermal video enhancement that integrates the SS2D module with CNNs to handle complex motions and challenging lighting conditions. This model includes:
   a) The Basic Denoising module, which reduces noise and improves image quality.
   b) The Optical Flow Attention module, which provides blur-resistant motion deblurring and preserves scene details even under challenging circumstances.
2. We create a labeled thermal video dataset using entropy-based measures to produce meaningful labels for training and evaluation. This dataset includes over three video sequence pairs, with 4k frame pairs.
3. We evaluate the proposed framework on real-world scenarios like wildlife monitoring and autonomous systems. Our experiments cover diverse thermal video datasets, including BIRDSAI, FLIR, CAMEL, Autonomous Vehicles, and Solar Panel, each presenting unique challenges. Compared to two traditional methods, DCRGC and RLBHE, and two deep learning-based

 doi:10.20944/preprints202412.2099.v1

approaches, IECGAN and BBCNN, the presented Mamba model consistently outperforms existing solutions. This is demonstrated through qualitative improvements and quantitative assessments using state-of-the-art thermal image quality measures such as EME, BDIM, DMTE, MDIMTE, and LGTA.

The integrated design of the Mamba network combines the adaptability of deep learning with the stability of state-space modeling, resulting in enhanced robustness, efficiency, and applicability. This makes it well-suited for complex real-world tasks, such as reliable perception for navigation and safety, effective surveillance under challenging conditions, and improved imaging for security, military, and remote sensing applications.

The rest of the paper is organized as follows. Section 2 provides background information on thermal video enhancement and its challenges and reviews related work on existing thermal video enhancement methods, including traditional and deep learning approaches. Section 3 details the proposed TVEMamba framework, outlining its architecture and describing the data generation process. Section 4 presents experimental results, including qualitative and quantitative comparisons across datasets, with an ablation study and object detection performance. Finally, Section 5 summarizes the contributions of this work and highlights potential future research directions.

2. Related Works

2.1. Thermal Imaging Enhancement Models

Thermal image enhancement algorithms are generally divided into two main categories: traditional methods and learning-based methods. Traditional approaches [11–17] rely exclusively on patterns learned from unlabeled data. A fundamental technique in this category is Histogram Equalization (HE) [11], which], which approximately equalizes the cumulative distribution function of the histogram to map pixel intensities. However, HE often over-enhances image contrast because it doesn't have a way to control the level of enhancement. To address this, Adaptive Histogram Equalization (AHE) [12,17] was developed to preserve more image details compared to standard HE techniques. Despite its advantages, AHE can still cause over-enhancement in certain image regions due to homogeneous blocks. Contrast adjustment techniques [16] aim to enhance visibility by adjusting intensity values and improving overall contrast. These methods are effective but may perform poorly on images with uneven illumination. Wavelet-based methods [13] decompose an image into different frequency components, allowing separate processing to enhance details and reduce noise. Discrete stationary wavelets are often used in this process. However, these methods require careful parameter selection and can be computationally intensive. Improper management may introduce artifacts into the image. Gradient Field Equalization [14] focuses on improving contrast and reducing noise. While effective in some scenarios, these methods often struggle in complex situations and can sometimes over-enhance images, leading to noise amplification and brightness distortion. Frequency-domain-based thermal infrared image enhancement algorithms [15] have also been widely employed. These techniques transform images into the frequency domain, utilize high-pass filters to extract high-frequency components, enhance them, and then convert the images back to the spatial domain.

In contrast, deep learning-based methods [18–21] leverage neural networks to learn and apply enhancement processes. By utilizing large datasets, these methods effectively address issues like low contrast, noise, and blurred details, making thermal images more suitable for analysis. While promising, they come with limitations such as data dependency, high computational resource requirements, training instability, overfitting, interpretability challenges, domain adaptation issues, and noise sensitivity. For instance, a method introduced in [18] uses conditional generative adversarial networks for contrast and detail enhancement in infrared images, preventing background noise amplification and further enhancing image details. Reference [19] presents an end-to-end approach that achieves discriminative enhancement, resulting in superior visual effects in enhanced thermal infrared images. Similarly, Ref. [20] proposes an infrared image enhancement method employing convolutional neural networks to highlight targets and suppress background clutter, effectively improving the contrast between weak targets and the background. The main properties of thermal image

 doi:10.20944/preprints202412.2099.v1

enhancement methods, including their strengths and limitations, are summarized in Table 1. This table provides a systematic comparison across key performance metrics.

2.2. Video Enhancement Models

Video enhancement techniques are crucial for improving the quality of low-quality videos, making them more clear and suitable for applications such as surveillance, identity verification, traffic monitoring, and object recognition [22,23]. The primary goal is to enhance the video's visual appearance or provide better representations for automated processing tasks [24]. Like thermal image enhancement methods, video enhancement algorithms can be broadly classified into two main categories: traditional and context-based fusion methods. Table 2 highlights the benefits and limitations of thermal video technology.

Table 2. Benefits and Limitations of Thermal Video Technology.

Benefits	Limitations
Objects can be observed in no light conditions (dark environments).	Difficulty distinguishing between objects in proximity or of similar temperatures.
High performance in all weather conditions (rain, fog, snow, smoke).	Generally, lower resolution compared to visible light images.
Opportunities for surveillance over large distances and areas, detecting motion over a wide range.	Cannot see through glass or water, as these materials reflect infrared radiation, limiting use cases like capturing images of individuals in cars.
Detection of objects even when partially hidden by vegetation.	More expensive than visible-light cameras.
Promotes early detection of thermal anomalies (e.g. equipment overheating, fire hazards), contributing to preventive safety measures.	Cannot identify detected individuals, as infrared radiation does not create detailed enough images.

Traditional methods include spatial-based domain and transform-based domain techniques [25]. Spatial-based domain methods operate directly on the pixels of video frames and encompass techniques like contrast enhancement, histogram equalization, and tone mapping. Tone mapping is another approach used primarily for high-dynamic-range (HDR) videos. It compresses the luminance levels to displayable ranges on standard devices, enhancing visibility in underexposed areas [26].

Transform-based domain methods modify the frequency components of video frames using techniques like the Discrete Cosine Transform (DCT) and wavelet transforms. Compressed-domain enhancement enhances videos directly in the compressed domain by manipulating transform coefficients, reducing computational complexity and storage requirements [27]. Adjusting DCT coefficients can improve contrast and reduce noise without fully decompressing the video. Wavelet-based methods decompose video frames into different frequency components, allowing separate processing to enhance details and reduce noise [28].

Another significant approach is context-based fusion enhancement, combining information from multiple frames or integrating high-quality background data into low-quality videos [29,30]. This method leverages additional contextual information to enhance video quality, especially under challenging conditions like low light or uneven illumination. Image fusion techniques use Retinex theory to separate illumination and reflectance components, enabling better contrast and detail preservation [31]. By fusing images captured under different lighting conditions, it is possible to enhance the visibility of important scene elements. Motion detection and enhancement utilize algorithms like Gaussian Mixture Models (GMM) to detect moving objects, allowing selective enhancement of these areas [32]. This improves overall visual quality and better detection of important scene elements.

 doi:10.20944/preprints202412.2099.v1

Ensuring temporal coherence between frames is crucial to avoid flickering and maintaining visual consistency [33]. In summary, video enhancement techniques play a vital role in improving the clarity and usability of videos.

2.3. State Space Models

The Mamba [34] model is an advanced dynamic state space model (SSM) featuring efficient selection mechanisms, gaining popularity in computer vision. Its main advantage is handling long-range dependencies in data while maintaining linear computational complexity. This significantly improves over traditional transformers, which suffer from quadratic complexity as image sizes increase. Mamba has shown promising results in various visual tasks such as image classification, feature enhancement, and multimodal fusion [35]. Due to its efficiency, it is poised as a strong candidate to potentially replace CNNs and transformers as the foundational architecture in visual applications.

Recently, Mamba has been successfully applied to various applications, including image enhancement, video analysis, and object detection [36]. These applications highlight its versatility and significant potential to enhance the accuracy and efficiency of computer vision systems. By effectively managing long-range dependencies and keeping computational demands low, Mamba offers a robust solution for modern visual tasks, paving the way for advancements in the field.

Our work integrated the SSM into visual tasks by following the approach outlined in [37]. The SS2D module in our model consists of three primary operations: Scan Expanding, S6 blocks, and Scan Merging. Initially, the input images undergo the Scan Expanding operation, which systematically unfolds the image from its four corners toward the center. This rearrangement of the spatial structure allows the model to capture features from different spatial regions more effectively. Then, the image is flattened, and the sequence is fed into the S6 module responsible for feature extraction. The operations within the S6 module can be expressed as:

$$h_t = A * h_{t-1} + B * x_t \quad (1)$$

$$y_t = C * h_t \quad (2)$$

where, h_t represents the latent state at time t, x_t represents the input variable, y_t is the output, and $A, B, and\ C$ are learnable parameters. The features extracted from the four directions are then summed and merged, and the dimensions of the merged output are adjusted to match the original input size. After processing through the S6 blocks, the Scan Merging operation restores the spatial structure by reorganizing the flattened sequence back into its two-dimensional form. This combination of scan operations enables the SS2D module to effectively capture both local and global features in the image, enhancing feature extraction for our visual tasks.

3. Materials and Methods

3.1. Network Structure

Figure 2 illustrates the presented TVEMamba framework, which consists of three modules: a sharpening and denoising network (SD-Net), a blur-resistant motion estimation network (BRME-Net), and a motion deblurring network (MD-Net). This framework follows three steps to enhance clarity, contrast, and detail in thermal videos.

 doi:10.20944/preprints202412.2099.v1

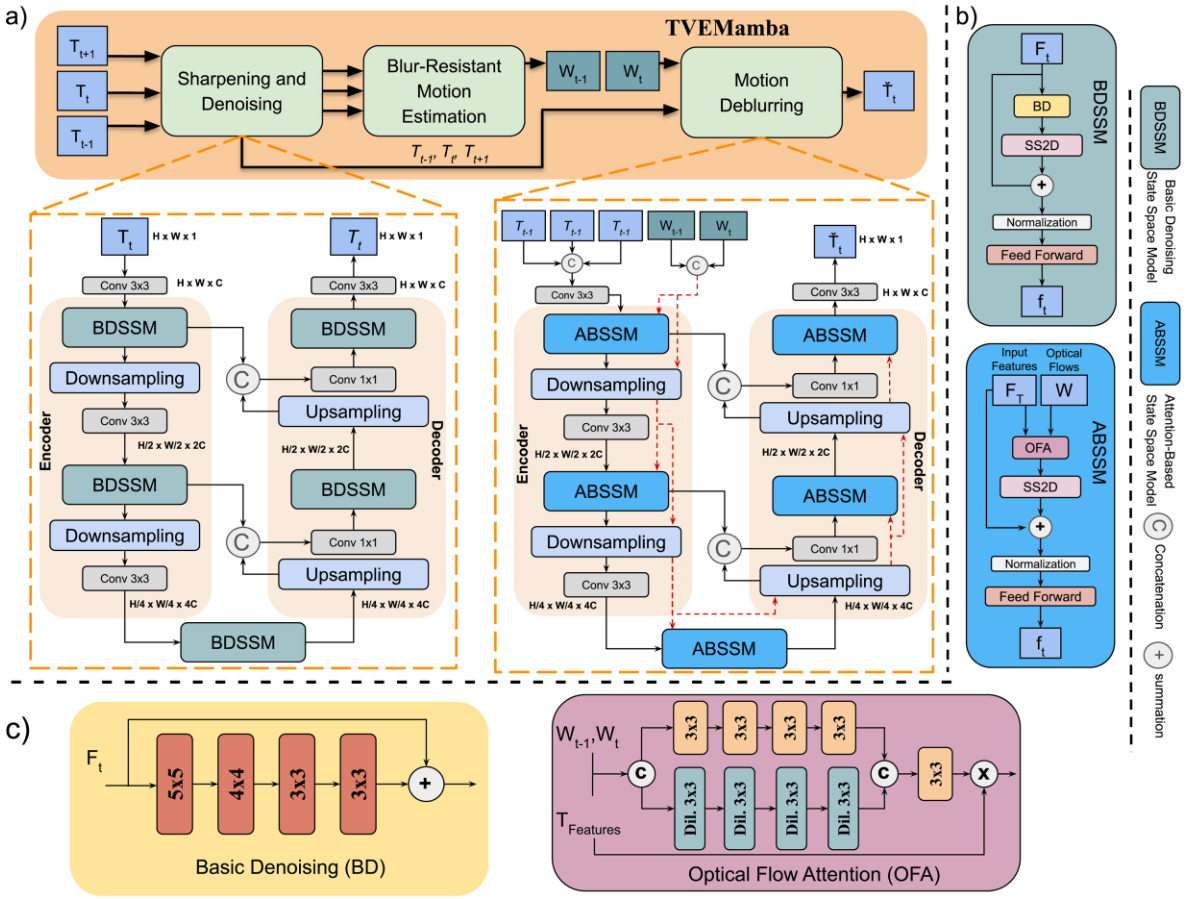

Figure 2. (**a**) Overall architecture of TVEMamba, (**b**) Basic denoising state space model and attention-based state space model, and (**c**) Basic denoising module and optical flow attention module.

Step 1: SD-Net (see Figure 2 (a)) improves the sharpness of thermal images and removes noise using a Mamba-based network with an encoder-decoder structure, capturing local and long-range contextual features. Both processes are symmetric and divided into two levels. Each downsampling level consists of a Basic Denoising State Space Model (BDSSM), downsampling operation, and a convolutional layer with a kernel size of 3 × 3. Similarly, upsampling involves two levels: an upsampling operation, a 1 × 1 convolution applied to the merged features from the corresponding downsampling layer, and a BDSSM. Finally, a 3 × 3 convolution is applied to the image to reduce dimensionality and restore it to grayscale with a single channel. The BDSSM block includes a Basic Denoising (BD) module consisting of four consecutive convolutional layers followed by a residual connection (see Figure 2 (c)), an SS2D module, a normalization layer, and a feed-forward network (FFN), as shown in Figure 2 (b).

Step 2: Then BRM-Net (see Figure 2 (a)) takes three consecutive input frames from the previous step, T_{t-1} , T_t, and T_{t+1}, where T_t is the t-th input frame, then estimates the optical flow W_{t-1} from T_{t-1} to T_t and W_t from T_t to T_{t+1}. The BRM-Net architecture is based on NeuFlow [38], which computes optical flow between two images. We adopt NeuFlow for BRM-Net due to its comparable accuracy to NeuFlowV2 [39] but with significantly fewer parameters. The architecture follows a global-to-local scheme: global matching is performed on a 1/16 resolution to capture large displacements, followed by refinement at 1/8 resolution using lightweight CNN layers for improved accuracy.

Step3: The MD-Net (see Figure 2 (a)) is applied in the final stage of the enhancement process. It takes as input the frames T_{t-1}, T_t, T_{t+1}, and the optical flows W_{t-1} and W_t to generate a blur-free, corrected T_t frame (see Figure 1(b)). Both stages of TVEMamba are built on a U-Net-based encoder-decoder architecture with the integration of vision Mamba. Similar to SD-Net, MD-Net employs a two-level encoder-decoder structure; however, instead of BDSSM, we apply an Attention-Based State

 doi:10.20944/preprints202412.2099.v1

Space Model (ABSSM), which includes an Optical Flow Attention (OFA) module, an SS2D module, a normalization layer, and a feed-forward network (FFN) (see Figure 2 (b)). The OFA module uses two branches to generate attention weights based on both local and global features. The branches share a similar structure, utilizing convolutional layers, but the global branch uses dilated convolutions to capture more global features. The outputs from both branches are concatenated, and the final convolution layer generates a weight map for T_{t-1}, T_t, T_{t+1} frames, as shown in Figure 2 (c).

3.2. Training and Dataset

3.2.1. Dataset Generation

The experiments were conducted on the FLIR dataset [40]. Since no thermal video dataset with ground-truth labels is available, we generated synthetic labels using two methods, as shown in Figure 3. First, we applied a sharpening technique [41] with the following parameters: patch radius r=11, epsilon ε=0.01, scale s=1, and kappa k where k was chosen from a range of 4 to 6, incremented in steps of 0.2, based on the entropy measure for thermal images [42]. However, thermal images are inherently noisy, and the sharpening process amplifies this noise. We applied a denoising method [43] to address this, utilizing a recurrent network with non-local operations for image restoration. Two denoising models were employed, using noise levels of 15 and 25. Images denoised at noise level 15 retained a small amount of noise, while those at noise level 25 were noise-free but lost small details. To balance noise reduction and detail preservation, we merged the two denoised images using the following formula:

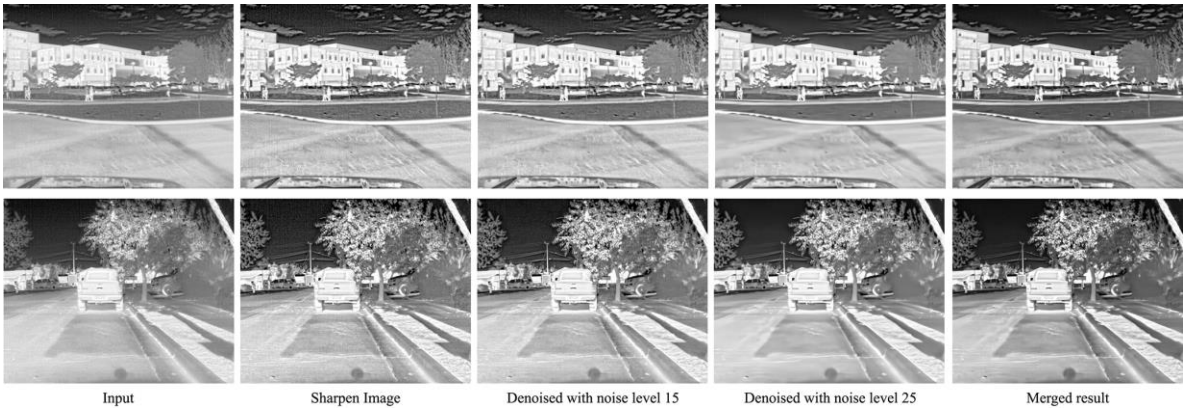

Figure 3. This figure shows the original image from FLIR [40], the corresponding sharpened image, denoised images with different noise levels, and the final merged result.

$$I_{Label} = c * I_{noise=25} + (1 - c) * I_{noise=15} \quad (3)$$

where c=0.6. This approach allowed us to generate high-quality synthetic labels. The data generation process took approximately six days. The dataset contains three videos, totaling 4,224 image frames with a resolution of 640x512.

3.2.2. Sharpening and denoising network

SD-Net was trained for 250 epochs using the Adam optimizer with an initial learning rate 1e−4. The experiments were conducted on the above-described dataset. Simple augmentation techniques were applied to increase generalizability, such as horizontal and vertical flips and random cropping to 256×256-pixel patches. All experiments were performed on an NVIDIA RTX 4090 GPU with 24 GB of memory. The network was trained using the following loss function:

$$L_{SD} = MSE(I_{Input}, I_{Label}) \quad (4)$$

where MSE [44] is the Mean Squared Error, which minimizes the difference between the predicted and labeled images, I_{Input} represents the input image and I_{Label} represents the sharp generated label image.

 doi:10.20944/preprints202412.2099.v1

3.2.3. Blur-resistant motion estimation network

NeuFlow was initially trained using blur-free datasets that provide ground-truth optical flow maps, such as Sintel [45], KITTI [46], and HD1K [47]. However, thermal videos often contain blur, resulting in inaccurate optical flow estimations. Unfortunately, no available datasets offer ground-truth optical flow maps for blurry images. To overcome this limitation, we fine-tuned our BRM-Net using a blurred video dataset [48] that contains pairs of sharp and blurred videos. The blurred video dataset was created by capturing sharp videos at high frame rates and averaging adjacent frames to simulate blur. The dataset contains 71 pairs of blurry videos and corresponding sharp versions, providing 6,708 pairs of 1280×720 blurred and sharp frames. We generated optical flow maps for training using only the sharp frames, employing the NeuFlow model. However, during optical flow generation, the model failed to generate accurate flow for certain videos. The accuracy of optical flow predictions can depend on the dataset on which the optical flow method was initially trained. The failed images may have a different distribution than the training dataset. After filtering out videos with inaccurate flow estimates, the final dataset for training consisted of 46 videos, totaling 4,369 frames. Inspired by [49], we used the same estimated optical flow map for the following pairs:

$$(T_{t-1}^{blur}, T_t^{blur}), (T_{t-1}^{blur}, T_t^{sharp}), (T_{t-1}^{sharp}, T_t^{blur}), (T_{t-1}^{sharp}, T_t^{sharp}). \quad (5)$$

The network was trained for 300 epochs using the Adam optimizer, with an initial learning rate of 1e−4. The input images were cropped to 256×256 pixels. The network was trained using the following loss function:

$$L_{BR} = MSE(W_t^{blur,blur}, W_t) + MSE(W_t^{blur,sharp}, W_t) + MSE(W_t^{sharp,blur}, W_t) + MSE(W_t^{sharp,sharp}, W_t) \quad (6)$$

where, $W_t^{*,*}$, is optical flow predicted by BRM-Net using (T_{t-1}^{*}, T_t^{*}) frames and W_t is the label generated by the same network using only sharp frames, where $* \propto \{blur, sharp\}$.

3.2.4. Motion deblurring network

The MD-Net was trained over 300 epochs using the Adam optimizer, with an initial learning rate of 5×10^{-4}. For training, we utilized the same dataset as in BRM-Net, which consists of 46 pairs of blurry videos and their corresponding sharp videos. We used data augmentation techniques, including horizontal and vertical flips and random cropping to 256×256 pixel patches. The network was trained using the following loss function:

$$L_{MD} = MSE(T_t, T_t^{sharp}) \quad (7)$$

where MSE is the Mean Squared Error, which minimizes the difference between the predicted and target images, T_t represents the network's prediction, and T_t^{sharp} represents the sharp ground-truth frame.

4. Results

This section presents the experimental results of the proposed TVEMamba framework, positioning it alongside several established enhancement methods. For comparison, we selected two representative traditional approaches, DCRGC [16] and RLBHE [17], and two deep learning-based methods, IE-CGAN [18] and BBCNN [21]. These methods were chosen due to their relevance in addressing the common challenges of thermal imaging, such as low contrast and noise, and their demonstrated effectiveness in various thermal image enhancement tasks.

By conducting a comparative analysis, we aim to identify each technique's unique contributions and potential limitations, including their ability to maintain image integrity, enhance critical features, and avoid common artifacts. While some existing methods may offer marginal improvements in contrast, they often come at the cost of increased noise or halo effects that undermine the overall image quality. In contrast, the TVEMamba achieves a favorable balance between visual clarity and structural fidelity, providing a more stable and versatile foundation for tasks like object detection. This combination of enhancement quality and robustness under varying lighting conditions highlights the practical value of TVEMamba in real-world thermal imaging scenarios.

 doi:10.20944/preprints202412.2099.v1

4.1. Qualitative Comparision

Figure 4 presents a qualitative comparison of the TVEMamba framework against several established thermal image enhancement methods applied to a variety of datasets, including BIRDSAI [50], FLIR [40], CAMEL [51], Autonomous Vehicles [52] and Solar Panels [53]. In some cases, DCRGC can achieve balanced contrast, but it frequently introduces halo artifacts and amplifies noise (Figure 4 (a,c)). These issues become especially noticeable in images that are too dark or bright, ultimately reducing their overall visual quality and usefulness. RLBHE applies smaller contrast adjustments but is highly sensitive to bright areas, making already bright regions even more intense while failing to reduce existing noise. IE-CGAN, while aiming to reduce noise in underexposed scenes, often produces overly dark images that obscure subtle details critical for further analysis (Figure 4 (c,d)). Conversely, when dealing with overexposed images, it tends to over-brighten them, causing the loss of important features (Figure 4 (b)). BBCNN, on the other hand, does not sufficiently highlight dark defects, resulting in the loss of essential information. It also struggles to maintain consistent brightness and contrast across different scenes, occasionally introducing artifacts. Furthermore, its tendency to add excessive sharpness can distort the natural appearance of the images, potentially leading to misinterpretation. In comparison, TVEMamba achieves a more balanced enhancement. It preserves structural details, maintains natural brightness and contrast levels, and reduces noise without introducing distracting artifacts. Even under difficult scenarios such as low-light conditions, moving elements, or complex textures TVEMamba consistently produces stable, clear, and visually coherent frames. Additionally, Figure 5 provides a detailed view of how our TVEMamba framework preserves and refines subtle features in the image. The improvement in edge sharpness, textural fidelity, and balance of contrast and brightness is visible, underscoring the method's ability to recover important scene details. Moreover, a simple colorization technique enhances the image interpretability, allowing distinct elements to stand out more clearly.

Figure 4. Visual comparison on (a) BIRDSAI [50], (b) FLIR [40], (c) CAMEL [51], and (d) Autonomous Vehicles [52] datasets.

 doi:10.20944/preprints202412.2099.v1

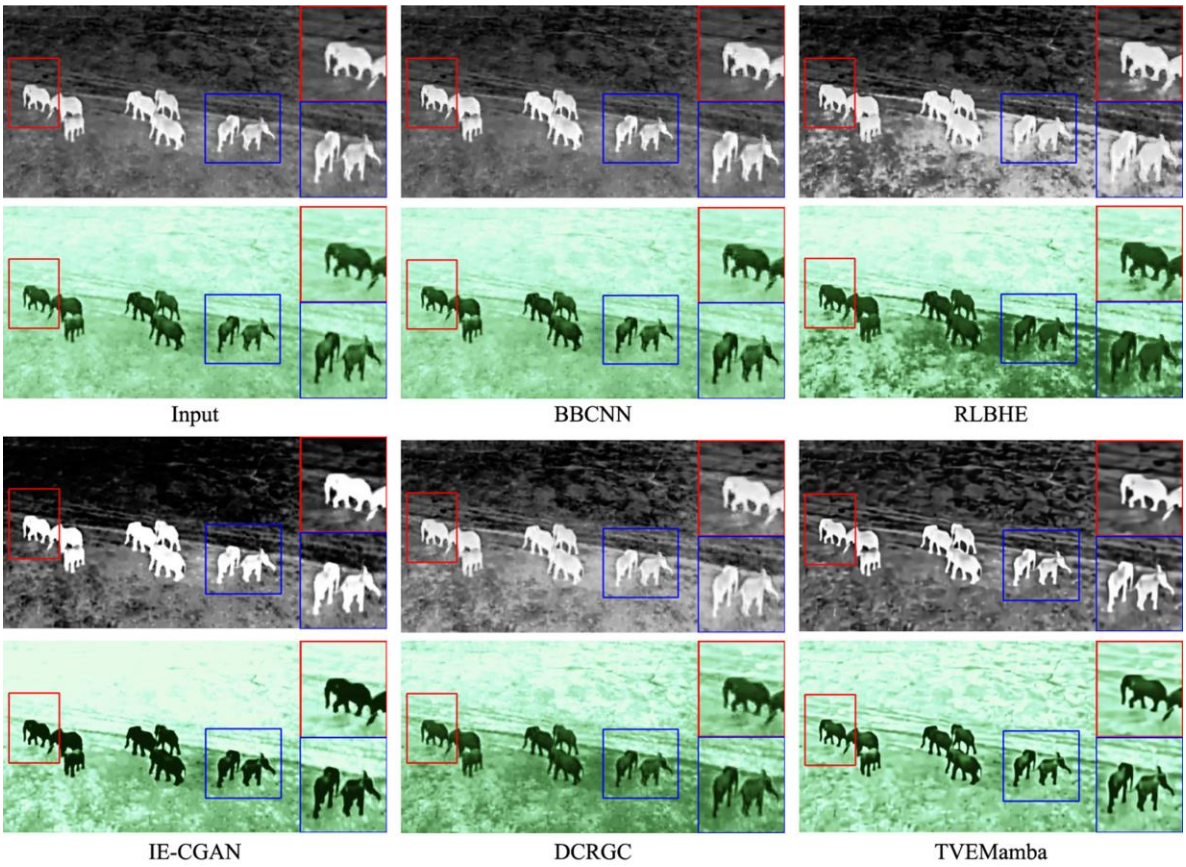

Figure 5. Performance of TVEMamba on BIRDSAI dataset.

Finally, Figure 6 illustrates the effectiveness of our TVEMamba framework on a low-quality solar panel video dataset. This figure displays five sequential frames from the original video (top row) and their enhanced counterparts generated by TVEMamba (bottom row). The input frames suffer from poor visibility and a lack of detail, making it difficult to identify subtle structural features. In contrast, the enhanced frames significantly improve clarity, contrast, and edge definition. TVEMamba highlights critical details and ensures smooth temporal consistency across frames, an essential factor for video analysis applications. This example further demonstrates the robustness and adaptability of our model in handling challenging real-world scenarios.

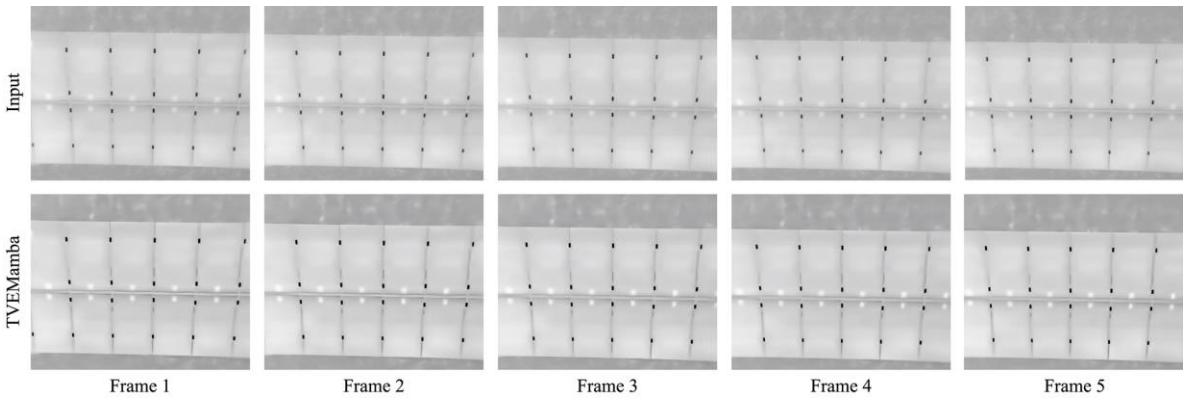

Figure 6. Qualitative results of TVEMamba framework on solar panel video frames.

4.2. Quantitative Comparision

To assess the effectiveness of the TVEMamba framework, we used five non-reference image quality metrics: (i) Measure of Enhancement (EME) [54], which evaluates image contrast entropy on a block basis rather than individual pixels. This metric is essential for assessing enhanced images and

 doi:10.20944/preprints202412.2099.v1

highlighting contrast variations within blocks. (ii) Block Distribution-Based Information Measure (BDIM) [55], which quantifies the information in image blocks by examining local contrast and structural details. It ensures that fine details are preserved and enhanced effectively. (iii) Density-based Measure of Thermal-image Enhancement (DMTE) [56], which incorporates elements of the human visual system with density-based analysis. (iv) Global Contrast Measure of Enhancement (MDIMTE) [56], which combines features related to human vision, information theory, and distribution-based metrics. This measure provides a holistic assessment of enhancement quality by focusing on overall contrast improvements that align with human perception and effective information distribution. (v) Local and Global Thermal Assessment (LGTA) [57], which integrates both local and global features to evaluate thermal image quality comprehensively. By combining block-level analysis with global intensity distribution, it closely aligns with human perception, offering nuanced insights into image clarity and enhancement.

High scores across these metrics indicate superior enhancement and a more natural visual appearance. Table 3 presents the comparative analysis, showcasing TVEMamba's performance against existing methods. Our approach outperforms both traditional and CNN-based methods, achieving the highest average scores across all metrics. These results demonstrate TVEMamba's outstanding ability to enhance thermal images while preserving critical details and maintaining a realistic appearance.

Table 3. Quantitative comparison of TVEMamba with the state-of-the-art methods.

	BBCNN	**DCRGC**	**IE-**	**RLBHE**	**TVEMamba**
BIRDSAI					
EME	10.060	20.264	17.748	18.377	**22.942**
DMTE	0.297	0.297	0.296	0.297	**0.299**
MDIMTE	45.060	42.620	31.620	46.001	**47.132**
BDIM	0.974	0.986	0.988	0.986	**0.991**
LGTA	1.158	1.167	1.423	1.154	**1.172**
CAMEL					
EME	14.633	24.214	24.010	23.796	**25.371**
DMTE	0.293	0.292	0.290	0.294	**0.296**
MDIMTE	39.833	41.309	32.004	40.747	**42.786**
BDIM	0.990	0.988	0.992	0.990	**0.994**
LGTA	1.239	1.235	1.381	1.089	**1.548**
FLIR					
EME	10.743	13.424	10.560	11.185	**14.152**
DMTE	0.295	0.296	0.295	0.294	**0.298**
MDIMTE	43.801	40.627	41.024	42.486	**50.146**
BDIM	0.972	0.977	0.965	0.971	**0.982**
LGTA	1.137	1.146	1.105	1.080	**1.167**
Autonomous Vehicles					
EME	2.929	3.088	7.260	8.130	**12.513**
DMTE	0.299	0.298	0.297	0.297	**0.310**
MDIMTE	51.517	48.326	53.659	47.925	**57.369**
BDIM	0.937	0.959	0.943	0.957	**0.963**
LGTA	1.180	1.393	1.189	1.411	**1.499**

4.3. Evaluation Metrics for Object Detection

To validate our approach for thermal video enhancement, we employed object detection methods and evaluated their performance using standard metrics: $mAP_{0.5}$ and $mAP_{0.5:0.95}$. These metrics are widely used to assess object detection models' localization and classification accuracy. Precision

 doi:10.20944/preprints202412.2099.v1

and recall are fundamental components of these evaluations, where precision measures the proportion of true positive samples in all the predicted positive samples, and recall is used to measure the proportion of true positive samples in all the predicted positive samples. These can be mathematically expressed as:

$$precision = \frac{TP}{TP + FP} \tag{8}$$

$$recall = \frac{TP}{TP + FN} \tag{9}$$

where, TP (True Positive) represents the number of objects that are correctly recognized as belonging to the target class, FP (False Positives) refers to the number of instances where non-target objects are incorrectly identified as belonging to the target class, and FN (False Negatives) indicates the number of instances where target objects are incorrectly classified as non-target objects.

We use Intersection over Union (IoU), which measures the overlap between the predicted bounding box and the ground-truth bounding box to assess the accuracy of the predicted bounding boxes. Mathematically, IoU is defined as the ratio of the intersection of the two bounding boxes to their union:

$$IoU = \frac{|B_g \cap B_p|}{|B_g \cup B_p|} \tag{10}$$

where, B_g and B_p represent the ground truth and predicted bounding boxes, respectively. A prediction is a true positive if the IoU exceeds a predefined threshold, such as 0.5.

Building on these concepts, the mean average precision (mAP) is the primary evaluation metric for object detection. The mAP calculates the average precision (AP) for each class and then averages the values across all classes. For $mAP_{0.5}$, the IoU threshold is fixed at 0.5, meaning that predicted bounding boxes are required to have at least 50% overlap with the ground truth. For a more rigorous evaluation, $mAP_{0.5:0.95}$ averages the precision over multiple IoU thresholds, ranging from 0.5 to 0.95 in steps of 0.05, providing a more comprehensive measure of detection accuracy. This approach ensures that both the localization quality of the bounding boxes and the ability to classify objects correctly are accounted for. A higher mAP value indicates better overall performance of the object detection model. Using these metrics, we can effectively measure the improvements in object detection accuracy achieved through our thermal video enhancement method.

4.4. Ablation Study

We conducted a series of ablation experiments to assess the contribution of the BD and OFA blocks to the thermal image enhancement task. Specifically, we trained TVEMamba with and without these modules to understand their impact on performance. Only the BD block was retained in one variant, while the OFA block was removed with the optical flow estimation module. In another variant, only the OFA block was retained while the BD block was removed. As shown in Figure 7, the highest values of each measure were achieved when both blocks were integrated, indicating that their combination significantly improves the model ability to enhance overall image quality. Figure 8 provides qualitative comparisons, illustrating the network performance for each module. Furthermore, to evaluate the effectiveness of our TVEMamba framework on downstream computer vision tasks, we utilized object detection experiments using two datasets: BIRDSAI and FLIR. The BIRDSAI dataset contains thermal videos of elephants captured for wildlife monitoring. To assess our enhancement method under different scenarios, we designed two labeling schemes: first, with two classes, "Elephant" and "Unknown" (including all non-elephant objects), and second, with three classes, "Elephant," "Human," and "Unknown." This allowed us to evaluate the model ability to distinguish between multiple object categories in enhanced thermal videos. The FLIR dataset, commonly used for autonomous driving and surveillance applications, focuses on two classes: "Pedestrian" and "Car". We employed two object detection architectures: YOLOR [58], which combines implicit and explicit knowledge within a single model for efficient detection, and Hyper-YOLO [59], incorporating hyperparameter optimization and architectural improvements. For each dataset and model, we trained on both the original thermal images and the enhanced images produced by the TVEMamba framework. As shown in Table 4, the notable improvements observed on the BIRDSAI dataset suggest that our enhancement method particularly benefits datasets with lower initial image quality due to low

 doi:10.20944/preprints202412.2099.v1

contrast and noise. In addition, Figure 9 illustrates YOLOR predictions on both original and enhanced BIRDSAI frames, clearly demonstrating improved detection accuracy after enhancement. There was no significant improvement in object detection performance for the FLIR dataset. However, the lack of deterioration confirms that our method does not introduce artifacts or distortions that could negatively impact detection. These findings indicate that the proposed TVEMamba framework benefits applications requiring reliable object detection in challenging thermal environments, such as wildlife monitoring and surveillance under adverse conditions.

Table 4. Object detection performance on the BIRDSAI and FLIR datasets. YOLOR$_1$ and Hyper-YOLO$_1$ models are trained on original datasets, and YOLOR$_2$ and Hyper-YOLO$_2$ models are trained on enhanced datasets produced by the TVEMamba framework.

Dataset	**BIRDSAI**		**BIRDSAI**		**BIRDSAI**		**FLIR**	
Classes	**2**		**3**		**2**		**2**	
Architecture	YOLOR$_1$	YOLOR$_2$	YOLOR$_1$	YOLOR$_2$	Hyper-YOLO$_1$	Hyper-YOLO$_2$	Hyper-YOLO$_1$	Hyper-YOLO$_2$
mAP$_{0.5}$	38.1	**44.2**	25.0	**29.7**	38.0	**43.9**	89.8	**89.9**
mAP$_{0.5:0.9}$	13.2	**16.8**	9.3	**10.9**	12.9	**16.4**	56.6	**56.7**

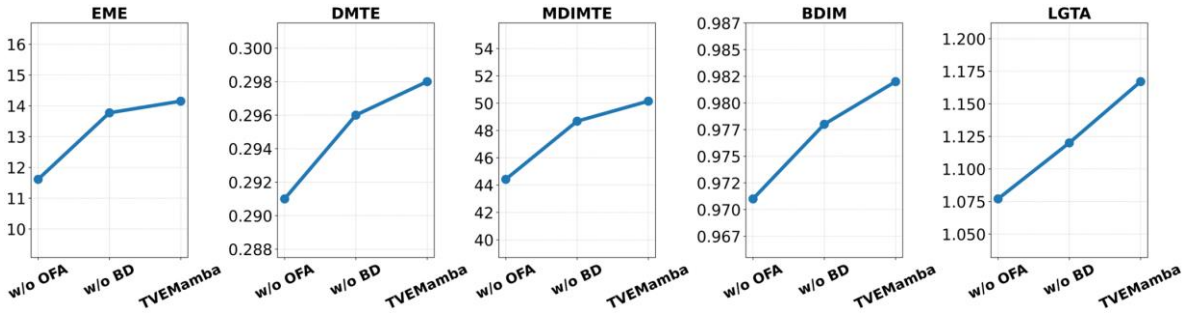

Figure 7. Evaluating the Contribution of OFA and BD Blocks in TVEMamba.

 doi:10.20944/preprints202412.2099.v1

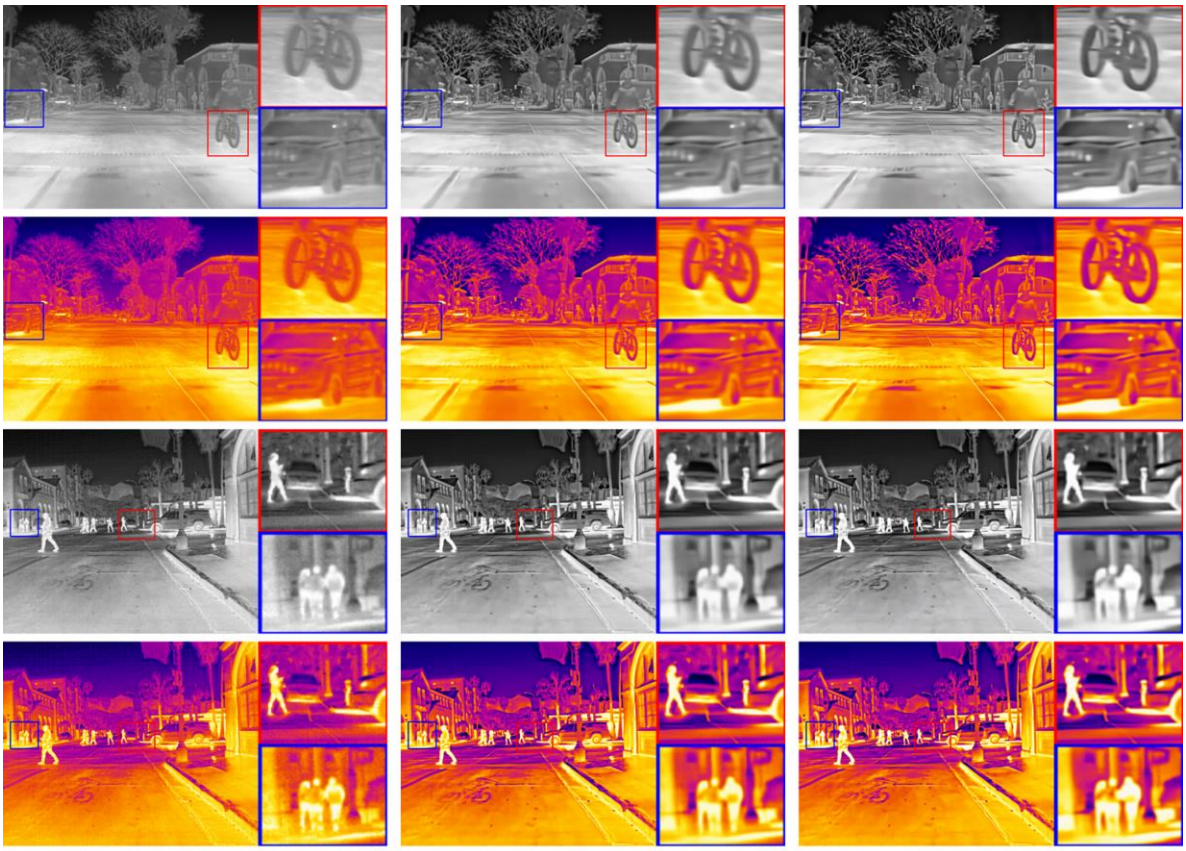

Figure 8. Details of intermediate results from the proposed TVEMamba, showing the effects of each module.

Figure 9. Object detection results on original and enhanced Images using the YOLOR method.

5. Discussion

This paper introduces the Mamba model, a novel thermal video enhancement method that leverages a State Space 2D module integrated with a Convolutional Neural Network. The Mamba model addresses major challenges such as low contrast, motion blur, and noise by incorporating the Basic Denoising and Optical Flow Attention modules. Simulation results demonstrated across multiple datasets, including BIRDSAI, FLIR, CAMEL, Autonomous Vehicles, and Solar Panel, highlight the Mamba model's ability to outperform both traditional and deep learning-based methods, resulting in higher-quality thermal videos suitable for a wide range of applications. Through this integration of state-space modeling and deep learning, the Mamba network adapts to diverse lighting conditions and varying motion patterns, making it well-suited for practical use cases. Applications include surveillance, where robust detection and tracking under challenging conditions are essential; autonomous systems, where reliable perception ensures safety and navigation; and remote sensing, where improved thermal imaging can aid critical monitoring tasks.

 doi:10.20944/preprints202412.2099.v1

Future research directions involve extending the Mamba model's capabilities. First, we aim to enhance data association in trackers by leveraging richer thermal information. Second, we will explore the fusion of visible and non-visible spectral data to improve tracking accuracy under different lighting and environmental conditions. Finally, optimizing the model for real-time performance will facilitate its deployment in time-sensitive applications. Advancing state-of-the-art thermal video enhancement will unlock the full potential of thermal imaging technologies, ultimately improving the performance and reliability of a broad spectrum of computer vision tasks. Finally, we will develop a GUI-based image enhancement, object detection, and tracking framework that can run through the cloud environment.

Author Contributions: Conceptualization, S.A.; methodology, S.H.; software, S.H.; validation, S.H., A.G. and S.A.; formal analysis, S.H, A.G. and K.P.; investigation, S.H., S.A, A.G and K.P.; resources, S.H.; data curation, S.H. and K.P.; writing—original draft preparation, S.H.; writing—review and editing, S.H., S.A, A.G and K.P.; visualization, S.H.; supervision, S.A., A.G and K.P.; project administration, S.A.; funding acquisition, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Advance Research Grants provided by the Foundation for Armenian Science and Technology.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Due to the nature of this research, the participants did not agree to share their data publicly, so supporting data is not available.

Acknowledgments: The first author thanks Sarkis and Nune Sepetjians for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shidik, G.; Noersasongko, E.; Nugraha, A.; Andono, P. N.; Jumanto, J.; Kusuma, E.J. A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and datasets. *IEEE Access* **2019**, *7*, 170457–170473.
2. Karpuzov, S.; Petkov, G.; Ilieva, S.; Petkov, A.; Kalitzin, S. Object tracking based on optical flow reconstruction of motion-group parameters. *Information* **2024**, *15*, 296.
3. Alsrehin, N.; Klaib, A.; Magableh, A. Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study. *IEEE Access* **2019**, *7*, 49830–49857.
4. Zhang, L.; Xiong, N.; Gao, W.; Wu, P. Improved detection method for micro-targets in remote sensing images. *Information* **2024**, *15*, 108.
5. Dong, Y.; Pan, W.D. A survey on compression domain image and video data processing and analysis techniques. *Information* **2023**, *14*, 184.
6. **Yoshida, E.; Kato, S.; Sato, T.; Suzuki, T.; Koyama, H.; Kato, S.** Proposal and prototyping on wildlife tracking system using infrared sensors. In Proceedings of the *International Conference on Information Networking (ICOIN)*, Jeju-si, Korea, 7–10 January 2022; pp. 292–297.
7. **Kim, E.; Kim, W.; Park, J.; Yeo, K.** Human detection in infrared image using daytime model-based transfer learning for military surveillance system. In Proceedings of the *14th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, 11–13 October 2023; pp. 1306–1308.
8. Yuan, C.-J.; Lan, S.-J.; Yan, G.-D.; Wang, D.; Lu, J.-H.; Meng, Q.-F. Application of near-infrared spectroscopy in rapid determination of adenosine and polysaccharide in Cordyceps militaris. In Proceedings of the *Fifth International Conference on Natural Computation*, Tianjin, China, 14–16 August 2009; pp. 578–582.
9. **Alheeti, K.; McDonald-Maier, K.** An intelligent security system for autonomous cars based on infrared sensors. In Proceedings of the *23rd Internat. Conference on Automation and Computing (ICAC)*, Huddersfield, UK, 7–8 September 2017; pp. 1–5.
10. **Mo, F.; Li, H.; Yao, X.; Wang, Q.; Jing, Q.; Zhang,** L. Intelligent onboard processing and multichannel transmission technology for infrared remote sensing data. In Proceedings of the *IGARSS 2019—2019 IEEE*

 doi:10.20944/preprints202412.2099.v1

International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 9063–9066.

11. **Gonzalez, R.; Woods, R.** *Digital Image Processing*; 2nd ed.; Prentice-Hall: Upper Saddle River, NJ, USA, 2002; Volume 793.
12. **Dhariwal, S.** Comparative analysis of various image enhancement techniques. *Int. J. Electron. Commun. Technol.* **2011**, *2*, 91–95.
13. **Mudavath, T.; Niranjan, V.** Thermal image enhancement for adverse weather scenarios: A wavelet transform and histogram clipping approach. *Signal Image Video Process.* **2024**. (In Press)
14. **Grigoryan, A.; Agaian, S.** Asymmetric and symmetric gradient operators with application in face recognition in Renaissance portrait art. In *Proceedings of the SPIE, Defense + Commercial Sensing, Mobile Multimedia/Image Processing, Security, and Applications*, Baltimore, MD, USA, 4–18 May 2019; Volume 10993, p. 12.
15. **Agaian, S.; Panetta, K.; Grigoryan, A.** Transform-based image enhancement algorithms with performance measure. *IEEE Trans. Image Process.* **2001**, *10*, 367–382.
16. **Wang, Z.; Liang, Z.; Liu, C.** A real-time image processor with combining dynamic contrast ratio enhancement and inverse gamma correction for PDP. *Displays* **2009**, *30*, 133–139.
17. **Zuo, C.; Chen, Q.; Sui, X.** Range limited bi-histogram equalization for image contrast enhancement. *Optik* **2013**, *124*, 425–431.
18. **Kuang, X.; Sui, X.; Liu, Y.; Chen, Q.; Gu, G.** Single infrared image enhancement using a deep convolutional neural network. *Neurocomputing* **2019**, *332*, 119–128.
19. **Wang, D.; Lai,** R.; Juntao. G. Target attention deep neural network for infrared image enhancement. *Infrared Phys. Technol.* **2021**, *115*, 103690.
20. **Fan, Z.; Zhang, Y.; Li, X.; Zhao, L.; Wang, H.** Dim infrared image enhancement based on convolutional neural network. *Neurocomputing* **2018**, *272*, 396–404.
21. **Lee, K.; Lee, J.; Lee, J.; Hwang, S.; Lee, S.** Brightness-based convolutional neural network for thermal image enhancement. *IEEE Access* **2017**, *5*, 26867–26879.
22. **Kastrinaki, V.; Zervakis, M.; Kalaitzakis, K.** A survey of video processing techniques for traffic applications. *Image Vis. Comput.* **2003**, *21*, 359–381.
23. **González-Cepeda, J.; Ramajo, Á.; Armingol, J.M.** Intelligent video surveillance systems for vehicle identification based on multinet architecture. *Information* **2022**, *13*, 325.
24. **Rao, Y.; Lin, W.; Chen, L.** Image-based fusion for video enhancement of nighttime surveillance. *Opt. Eng.* **2010**, *49*, 120501.
25. **Agaian, S.; Blair, S.; Panetta, K.** Transform coefficient histogram-based image enhancement algorithms using contrast entropy. *IEEE Transactions on Image Processing* **2007**, *16*, 741–758.
26. **Reinhard, E.; Ward, G.; Pattanaik, S.; Debevec, P.** *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*; Morgan San Francisco, CA, USA, 2005.
27. **Lee, S.** An efficient content-based image enhancement in the compressed domain using Retinex theory. *IEEE Trans. Circuits Syst. Video Tech.* ***2007***, 17, 199–213.
28. **Balster, E.; Zheng, Y.F.; Ewing, R.L.** Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising. *IEEE Transactions on Circuits and Systems for Video Technology* **2006**, *16*, 220–230.
29. **Wan, T.;** Tzagkarakis, G.; Tsakalides, P.; Canagarajah, C.N.; Achim, A. Context enhancement through image fusion: A multi-resolution approach based on convolution of Cauchy distributions, *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 1309–1312.
30. **Li, J.; Li, S.Z.; Pan, Q.; Yang, T.** Illumination and motion-based video enhancement for night surveillance. In Proceedings of the *IEEE workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, China, 15–16 October 2005; pp. 169–175.
31. **Land, E.H.; McCann, J.J.** Lightness and Retinex theory. *Journal of the Optical Society of America* **1971**, *61*, 1–11.
32. **Stauffer, C.; Grimson, W.E.L.** Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2000**, *22*, 747–757.

33. **Cho, Y.; Lee, H.; Park, D.; Kim, C.Y.** Enhancement for temporal resolution of video based on multi-frame feature trajectory and occlusion compensation. In Proceedings of the *IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, 7–10 November 2009; pp. 389–392.
34. **Gu, A.; Dao, T.** Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
35. **Yuan, C.; Zhao, D.; Agaian, S.** MUCM-Net: A Mamba powered UCM-Net for skin lesion segmentation. *Explor Med*. **2024**, *5*, 694–708.
36. **Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; Ye, Z.** A survey on visual Mamba. *arXiv* **2024**, arXiv:2404.15956v2.
37. **Xu, X.; Wang, R.; Fu, C.; Jia, J.** SNR-aware low-light image enhancement. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022; pp. 17693–17703.
38. **Zhang, Z.; Jiang, H.; Singh, H.** NeuFlow: Real-time, high-accuracy optical flow estimation on robots using edge devices. *arXiv* **2024**, arXiv:2403.10425.
39. **Zhang, Z.; Gupta, A.; Jiang, H.; Singh, H.** NeuFlow v2: High-efficiency optical flow estimation on edge devices. *arXiv* **2024**, arXiv:2408.10161.
40. **FLIR Thermal Dataset.** Available online: https://www.kaggle.com/datasets/deepnewbie/flir-thermal-images-dataset (accessed on 21 September 2024).
41. **Deng, G.; Galetto, F.; Al–nasrawi, M.; Waheed, W.** A guided edge-aware smoothing-sharpening filter based on patch interpolation model and generalized gamma distribution. *IEEE Open Journal of Signal Processing* **2021**, *2*, 119–135.
42. **Ayunts, H.; Grigoryan, A.; Agaian, S.** Novel entropy for enhanced thermal imaging and uncertainty quantification. *Entropy* **2024**, *26*, 374.
43. **Liu, D.; Wen, B.; Fan, Y.; Loy, C.C.; Huang, T.S.** Non-local recurrent network for image restoration. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Montréal, QC, Canada, 3–8 December 2018; Volume 31.
44. **Murphy, A.H.** Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Weather Rev.* **1988**, *116*, 2417–2424.
45. **Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J.** A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Florence, Italy, 7–13 October 2012.
46. **Geiger, A.; Lenz, P.; Urtasun, R.** Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
47. **Kondermann, D.; et al.** The HCI benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the CVPR Workshops*, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 19–28.
48. **Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; Wang, O.** Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 237-246.
49. **Son, H.; Lee, J.; Lee, J.; Cho, S.; Lee, S.** Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *Association for Computing Machinery* **2021**, *40*, 5.
50. **Bondi, E.; et al.** BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1736–1745.
51. **Gebhardt, E.; Wolf, M.** CAMEL dataset for visual and thermal infrared multiple object detection and tracking. In Proceedings of the *15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
52. **Takumi, K.; et al.** Multispectral object detection for autonomous vehicles. In Proceedings of the *Thematic Workshops of ACM Multimedia*, Mountain View, CA, USA, 23–27 October 2017; pp. 35–43.
53. **Bommes, L.; et al.** Georeferencing of photovoltaic modules from aerial infrared videos using structure-from-motion. *Progress in Photovoltaics* **2022**, *30*, 1122–1135.

54. **Agaian, S.;** Panetta**, K.; Grigoryan, A.** A new measure of image enhancement. In Proceedings of the *IASTED International Conference on Signal Processing and Communications (SPC)*, Marbella, Spain, 19–21 September 2000.
55. Trongtirakul, T; Agaian, S. Unsupervised and optimized thermal image quality enhancement and visual surveillance application. *Signal Process. Image Commun.* **2022**, *105*, 116714.
56. **Agaian, S.; Roopaei, M.; Akopian, D.** Thermal-image quality measurements. In Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 4–9 May 2014; pp. 1–5.
57. **Agaian, S.; Ayunts, H.; Trongtirakul, T.; Hovhannisyan, S.** A new method for judging thermal image quality with applications. *Signal Process.* **2024**, *229*, 109769.
58. **Wang, C.; Yeh, I.; Liao, H.** You only learn one representation: Unified network for multiple tasks. *Journal of Information Science and Engineering* **2023**, *39*, 691–709.
59. **Feng, Y.; et al.** Hyper-YOLO: When visual object detection meets hypergraph computation. *arXiv* **2024**, arXiv:2408.04804.