

Article

TopoRelTrack: A Self-Supervised Visual Tracking Method for Video-Based Seismic Motion Measurement

Honglei Wang ¹, Jinrong Su ², Jilong Li ³, Sirun Niu ⁴, Bo Fan ^{4,*} and Bizheng Luo ^{5,*} 

¹ Hebei Earthquake Agency, No. 262 Huaizhong Road, Yuhua District, Shijiazhuang 050021, China; wanghl@hbdzj.gov.cn

² Sichuan Earthquake Agency, Chengdu 610040, China; sujr@scdzj.gov.cn

³ Institute of Engineering Mechanics, China Earthquake Administration, Harbin 150080, China; lijil@iem.ac.cn

⁴ Shaanxi AI Land Technology Co., Ltd., Xi'an 710065, China; sirun.niu@ai-land.com.cn

⁵ School of Physics and Information Technology, Shaanxi Normal University, Xi'an 710062, China

* Correspondence: bo.fan@ai-land.com.cn (B.F.); luobizheng@snnu.edu.cn (B.L.)

Abstract

Accurate target tracking is essential for video-based seismic-motion measurement, as tracking errors can directly affect displacement reconstruction and subsequent seismic-intensity estimation. To improve tracking accuracy and robustness under vibration scenarios, this paper proposes TopoRelTrack, a visual tracking method based on self-supervised pre-training. In the pre-training stage, a template-guided masked autoencoder (TMAE) uses the target template to guide masked reconstruction in the search region, enabling the model to learn tracking-oriented feature representations from unlabeled video sequences. A topology-preserving scanning Mamba (TPSMamba) is further introduced to efficiently capture long-range spatial dependencies while preserving spatial structural information. In addition, a sparse dynamic relation matcher (SDRM) establishes adaptive and fine-grained correspondences between the template and search regions to enhance target discrimination and suppress background interference. The TPMamba and SDRM are then transferred to the downstream tracking network for continuous target localization and displacement reconstruction. Experiments on the LaSOT and GOT-10K benchmarks show that TopoRelTrack achieves an area under the curve (AUC) of 75.9% and a precision of 81.8% on LaSOT, as well as an AO of 78.5% on GOT-10K, outperforming the compared trackers. In controlled shaking-table video experiments, the proposed method also achieves lower displacement reconstruction errors and improved seismic-intensity prediction performance. The results demonstrate that TopoRelTrack provides accurate and robust target tracking for video-based motion measurement under controlled shaking-table excitation.

Keywords: seismic intensity measurement; visual object tracking; self-supervised; mamba



Academic Editor: Pietro Croce

Received: 15 September 2026

Revised: 15 October 2026

Accepted: 19 October 2026

Published: 22 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Seismic intensity provides a comprehensive measure of the severity of ground shaking and its potential effects on engineering structures and the surrounding environment. It therefore provides important information for rapid post-earthquake assessment and emergency response [1]. Instrumental seismic intensity is conventionally calculated from ground-motion signals recorded by seismographs or strong-motion accelerometers, with peak ground acceleration (PGA) and peak ground velocity (PGV) serving as important indicators for intensity determination [2]. Owing to their explicit physical basis and high

measurement reliability, instrument-based methods remain the primary means of ground-motion monitoring and instrumental seismic-intensity assessment.

Early ground-motion intensity estimation methods generally extracted manually designed features, such as peak amplitude, frequency spectrum, predominant period, and signal energy, from sensor-recorded P waves, S waves, or strong-motion waveforms. Empirical relationships, ground-motion prediction equations, or statistical models were then used to estimate earthquake magnitude, PGA, PGV, and seismic intensity. Although the computational procedures of these methods are relatively straightforward, errors in source-parameter estimation, local site amplification, and differences in seismic-wave propagation paths can alter the amplitude, frequency content, and duration of the recorded signals [3]. Consequently, PGA, PGV, and seismic-intensity estimates derived from fixed empirical relationships may exhibit substantial biases, limiting the generalization of these methods across different regions and earthquake events.

With the development of machine learning, data-driven models have increasingly been employed to learn the relationship between seismic waveforms and ground-motion intensity. Abdalzaher et al. proposed an optimized machine-learning model that rapidly estimates seismic intensity from short-duration waveform features following the arrival of the P wave, thereby improving the efficiency of intensity determination in earthquake early-warning systems [4]. Although this method reduces the dependence on complex manually defined decision rules, it still requires the prior extraction of a limited set of waveform features, and its generalization performance is affected by the distributions of the training regions and station records. Jozinović et al. employed a convolutional neural network to directly process three-component raw waveforms recorded by multiple stations and predict ground-motion intensity measures at different stations using waveform data acquired within a specified time window after an earthquake [5]. This method does not require the source location or magnitude to be specified in advance and can automatically extract features associated with ground-motion intensity from raw waveforms. However, its training and prediction processes still depend on seismic networks with sufficient spatial coverage and extensive historical waveform records. Münchmeyer et al. further proposed the Transformer Earthquake Alerting Model, which takes raw waveforms from multiple seismic stations together with their spatial coordinates as input to predict ground-motion measures, such as PGA, in a target region in real time [6]. This method can accommodate a variable number of stations at different locations and enables information interaction among multiple stations for regional earthquake early warning. Nevertheless, its applicability still depends on the coverage of the seismic network and the availability of reliable real-time data transmission.

Overall, sensor-waveform-based methods can directly provide ground-motion information with explicit physical meaning and have formed a relatively mature framework for seismic monitoring and earthquake early warning. However, their spatial coverage is constrained by station density, equipment deployment costs, communication infrastructure, and sensor operating conditions [7,8]. In regions with sparse strong-motion stations or where conventional instruments cannot be rapidly deployed, sufficiently dense spatial observations may be unavailable. Therefore, developing non-contact seismic-motion measurement methods that use existing imaging devices as a complement to conventional station-based monitoring is of practical significance.

With advances in digital imaging devices and computer vision, video-based motion measurement has increasingly been applied to structural displacement measurement and seismic-motion trajectory reconstruction [9,10]. A typical procedure extracts the positional changes of a target in image coordinates from consecutive video frames, converts pixel displacement into physical displacement through camera calibration or a scale factor, and

subsequently estimates velocity, acceleration, PGA, PGV, and other dynamic-response measures through signal filtering and numerical differentiation [11–13]. Compared with conventional contact sensors, video-based measurement is non-contact, flexible in deployment, and capable of simultaneously observing multiple spatial locations. However, its measurement accuracy depends strongly on the accuracy and temporal stability of the extracted target trajectory [10,11].

To improve the adaptability of visual features to complex measurement environments, deep learning has been introduced into video-based motion trajectory reconstruction. Bai et al. employed Mask R-CNN and HRNet to automatically detect and segment experimental targets and combined them with SIFT feature matching and signal filtering to recover the displacement trajectories and vibration frequencies of structural specimens [14]. This method reduced the manual effort required to select target regions and enabled automatic measurement of multiple targets. However, frame-by-frame instance segmentation incurs considerable computational cost, and temporal fluctuations in the predicted segmentation boundaries may be directly propagated into trajectory noise. Pan et al. proposed a real-time structural vibration measurement method based on YOLOv3-tiny and YOLOv3-tiny-KLT, in which a deep detector locates the target region and KLT estimates its positional changes across consecutive frames [15]. The method combines automatic target recognition with local motion continuity and exhibits favorable robustness under illumination variations and background noise. Nevertheless, its motion estimation still relies on the local brightness-consistency assumption of KLT and may therefore produce incorrect correspondences under rapid motion, severe motion blur, or degradation of local image features. Wen et al. employed the RAFT-GOCor deep optical-flow network to learn dense feature correspondences between consecutive video frames and recover continuous displacement information from different structural regions using the estimated optical-flow field [12]. They further developed a Bayesian RAFT-GOCor model using Monte Carlo dropout to quantify the uncertainty of the resulting visual displacement measurements. However, dense optical flow estimates motion relationships for both target and background regions, resulting in relatively high computational cost, while irrelevant background responses may interfere with target trajectory extraction. Wang et al. proposed LCCTV, a low-computational-cost Transformer-based tracking method that maps the initial target template and the current search region into a deep feature space and establishes relationships between their features through an attention mechanism, thereby continuously predicting the position of a calibration target in vibration videos [16].

Although these deep-learning-based methods have substantially improved automatic target perception and motion estimation, their direct application to seismic video measurement still faces limitations in labeled-data availability, efficient spatial feature encoding, and robust template–search correspondence. The following three challenges therefore remain:

1. In video-based seismic-motion measurement, reliable target tracking is fundamental to accurate motion trajectory reconstruction. Existing deep tracking methods [17,18] generally rely on large amounts of annotated data for supervised training, whereas labeled seismic videos available for model training remain limited. Therefore, learning stable and discriminative target representations from limited labeled data constitutes the first major challenge in seismic video trajectory reconstruction.
2. Accurate target localization requires both the preservation of local spatial structures in the template and search regions and the modeling of long-range dependencies among different spatial positions. Global self-attention provides strong relation-modeling capability but generally incurs high computational cost, whereas state-space models can process long sequences with lower complexity but may alter the original spatial adjacency of two-dimensional features when using one-dimensional

scanning strategies [19,20]. Therefore, achieving a suitable balance between two-dimensional spatial-structure preservation and long-range dependency modeling under a limited computational budget constitutes the second challenge in seismic video target encoding.

3. In seismic video motion measurement, target localization errors directly affect the reconstruction accuracy of displacement trajectories. Previous studies have shown that motion blur and background interference can weaken the target-discrimination capability of search-region features, leading tracking models to produce ambiguous or incorrect responses [21,22]. Moreover, global dense feature interaction may cause irrelevant background information to participate in target-feature aggregation, while localization errors in consecutive frames may gradually accumulate and result in trajectory drift [22]. Therefore, selectively establishing fine-grained, target-relevant correspondences between the template and search regions while suppressing irrelevant background responses constitutes a key challenge in obtaining stable motion trajectories.

Recent studies provide useful methodological foundations for addressing the above challenges. Self-supervised masked image modeling can exploit unlabeled images to learn transferable visual representations and improve label efficiency in downstream tasks [23–25]. Mamba and its visual extensions provide an efficient alternative to conventional self-attention by modeling long-range dependencies with computational complexity that scales linearly with sequence length [19,20,26]. In addition, sparse interaction and foreground–background modeling have demonstrated the potential of selectively establishing template–search correspondences to reduce irrelevant background associations [21,22]. Together, these advances provide the foundations for combining label-efficient pre-training, efficient long-range feature encoding, and selective relation modeling in seismic-video tracking.

Building on these foundations, this study proposes TopoRelTrack, a video-based method for continuous target tracking and displacement reconstruction in seismic-response measurement. The proposed framework follows a two-stage learning strategy consisting of self-supervised pre-training and downstream tracking. During pre-training, a template-guided masked autoencoder (TMAE) uses the target template from the current frame to guide the reconstruction of masked regions in a subsequent search frame, enabling the encoder to learn tracking-oriented representations from unlabeled video sequences. A topology-preserving scanning Mamba (TPSMamba) encoder is developed to model long-range dependencies efficiently while maintaining the spatial structure of two-dimensional visual features. In addition, a sparse dynamic relation matcher (SDRM) dynamically selects target-relevant features and establishes fine-grained template–search correspondences through graph-based feature interaction and cross-attention. After pre-training, the learned encoder and SDRM parameters are transferred to the downstream tracking network, which continuously estimates the target position across video frames for the subsequent reconstruction of structural displacement and dynamic responses.

Accordingly, the objective of this study is to develop and evaluate a self-supervised visual tracking framework for accurate and temporally stable motion-trajectory reconstruction from seismic videos under controlled shaking-table conditions. We hypothesize that the proposed framework improves tracking accuracy and trajectory-reconstruction reliability compared with existing tracking methods under consistent experimental conditions. The main contributions of this work are as follows:

- A self-supervised pre-training method based on a template-guided masked autoencoder is proposed. By using the target template to guide masked reconstruction in

the search region, the model learns tracking-oriented feature representations from unlabeled video sequences, thereby reducing its dependence on annotated data.

- A topology-preserving scanning Mamba encoder is developed to model visual features efficiently. The multi-directional scanning strategy alleviates the loss of spatial continuity caused by one-dimensional token serialization while capturing long-range dependencies with linear computational complexity.
- We propose a sparse dynamic relation matcher to establish adaptive and fine-grained correspondences between the template and search regions. By combining sparse graph-based feature aggregation with cross-attention, SDRM suppresses irrelevant background responses and improves the robustness of target localization.
- Extensive experiments on public tracking benchmarks and a self-built dataset demonstrate the effectiveness and superior performance of the proposed method.

The remainder of this paper is organized as follows. Section 2 reviews signal- and waveform-based seismic intensity estimation and visual tracking for robust seismic trajectory extraction. Section 3 presents the proposed TopoRelTrack framework, including the TMAE pre-training strategy, the TPSMamba encoder, and the SDRM module. Section 4 describes the experimental setup, datasets, and experimental results. Section 5 discusses the findings and limitations. Section 6 concludes the paper.

2. Related Work

2.1. Signal- and Waveform-Based Seismic Intensity Estimation

Instrument-based seismic-intensity estimation derives ground-motion indicators from seismographs and strong-motion accelerometers. Conventional procedures commonly map peak ground acceleration (PGA) and peak ground velocity (PGV) to instrumental intensity scales [2]. In earthquake early-warning systems, regional approaches infer source parameters from measurements collected by multiple triggered stations and subsequently forecast shaking at target sites, whereas on-site approaches estimate local peak motion or intensity directly from the first several seconds of the incoming waveform. These systems provide physically interpretable and operationally reliable measurements and therefore remain the standard means of seismic-intensity estimation.

Recent data-driven approaches have further extended waveform-based estimation by learning the relationship between early seismic signals and subsequent ground shaking [4–6]. Deep learning has been applied to multiple seismic-waveform analysis tasks. Earthquake Transformer performs simultaneous earthquake detection and phase picking from seismic waveforms [27], whereas convolutional and Transformer-based models directly estimate ground-motion intensity measures from single- or multi-station records [5,6]. Multi-station architectures capable of processing different station configurations have also been investigated to improve the applicability of waveform-based prediction across seismic regions [6]. Nevertheless, learned waveform models remain dependent on timely signal acquisition, reliable data transmission, and functioning seismic networks [5,6]. Their performance may also be affected by station geometry, local site conditions, seismic-wave propagation paths, regional differences, and changes in network configuration [3]. Consequently, their generalization to regions or shaking conditions that are insufficiently represented in the training data requires further validation.

Instrumental waveform measurements remain the standard basis for determining ground-motion parameters and instrumental seismic intensity [1,2]. However, their spatial coverage is constrained by sensor availability, station density, telemetry conditions, installation costs, and long-term maintenance requirements [7,8]. Video sensing should therefore be regarded as a complementary modality rather than a replacement for instrumental seismic monitoring [9]. Where calibrated visual observations are available,

cameras provide an alternative means of recovering spatially localized motion from image sequences [10,28,29]. Recent studies on video-based displacement reconstruction and seismic-intensity estimation further demonstrate the potential of this complementary measurement modality [16,28,29].

2.2. Vision-Based Seismic Motion Measurement

Vision-based measurement provides a non-contact means of recovering structural motion from image sequences. A typical measurement pipeline estimates target positions across consecutive frames, converts pixel displacement into physical displacement through camera calibration or a scale factor, and subsequently derives velocity and acceleration through signal processing and numerical differentiation [9–13]. Detection- and tracking-based methods have demonstrated the feasibility of recovering structural vibration responses from video [12,14–16].

Recent studies have further extended vision-based measurement to earthquake-related displacement reconstruction and seismic-intensity assessment. Choi et al. combined acceleration and computer vision measurements to estimate earthquake-induced permanent displacement by constructing a scale-factor map and tracking stable regions of surrounding natural targets [28]. Although this method exploits the complementary information provided by the two sensing modalities, it still requires acceleration measurements and primarily focuses on permanent rather than continuous dynamic displacement. Yang et al. proposed an image evaluation of seismic intensity method that estimates indoor seismic intensity from the movement percentages of different categories of objects in pre- and post-earthquake images [30]. The method provides a quantitative means of assessing indoor object responses but does not recover continuous displacement, velocity, or acceleration histories. Zhang et al. integrated target detection, scale-factor self-calibration, and sparse optical-flow tracking to recover multi-story structural displacement and identify input ground-motion histories through structural dynamic equations [29]. This method demonstrates the feasibility of vision-based ground-motion identification, but its inversion process depends on an available structural dynamic model and identifiable structural parameters.

Unlike generic visual tracking, seismic-video measurement converts the estimated target position from image coordinates into a physically meaningful displacement through geometric calibration or a scale factor [9–11]. Consequently, frame-wise localization errors and jitter can propagate into the reconstructed displacement trajectory and may be further amplified when velocity and acceleration are calculated through numerical differentiation [12,13]. Therefore, reliable seismic trajectory reconstruction requires not only high frame-wise localization accuracy but also temporal stability and continuous target correspondence across consecutive frames [12,15,16].

2.3. Self-Supervised Pre-Training and Masked Modeling for Visual Tracking

Most deep visual trackers are trained in a supervised manner using large-scale tracking datasets with frame-level target annotations [17,18]. Although this training paradigm provides strong target representations, generating dense annotations for domain-specific seismic videos requires considerable manual effort. Existing video-based structural and seismic-motion measurement studies therefore generally rely on experiment-specific datasets [14–16].

Self-supervised learning provides an alternative means of exploiting unlabeled visual data. He et al. proposed the masked autoencoder (MAE), which randomly masks a high proportion of image patches, encodes only the visible patches, and uses a lightweight decoder to reconstruct the missing pixels [31]. The asymmetric encoder–decoder design reduces the computational cost of pre-training and enables the encoder to learn transferable

representations for downstream visual tasks. Xie et al. subsequently proposed SimMIM, which adopts random patch masking and a lightweight prediction head to recover the raw RGB values of masked regions [32]. Their results showed that strong visual representations can be learned using a relatively simple reconstruction objective without requiring manually assigned semantic labels. More recently, Nandam et al. systematically investigated masked image modeling, contrastive learning, and clustering under low-shot conditions and combined masked reconstruction with multi-level clustering to capture both local context and global semantics [24].

These studies demonstrate that masked image modeling can exploit unlabeled images to learn transferable visual representations and improve model adaptation under limited labeled data. This provides a promising pre-training strategy for seismic-video tracking, where frame-level target annotation is costly. Building on this line of research, the present study incorporates target-template and cross-frame information into masked reconstruction to establish a closer connection between the pre-training objective and downstream template-search matching.

3. Method

3.1. Overview

The overall architecture of the proposed method is illustrated in Figure 1. Given two consecutive frames I_i and I_{i+1} from an image sequence, both frames retain the original scene background. First, the target template is cropped from I_i according to the target state B_i , while a larger search region is extracted from I_{i+1} :

$$Z_i = \mathcal{C}(I_i, B_i), \quad X_{i+1} = \mathcal{S}(I_{i+1}, B_i), \quad (1)$$

where $\mathcal{C}(\cdot)$ and $\mathcal{S}(\cdot)$ denote the template cropping and search-region extraction operations, respectively. Z_i and X_{i+1} denote the template image and the search image.

Based on the constructed template-search pair, the proposed method follows a two-stage learning strategy consisting of template-guided masked autoencoder pre-training and downstream tracking. The two stages share the same TPSMamba-based encoder and sparse dynamic relation matcher. Specifically, the template and search branches take Z_i and X_{i+1} as inputs, respectively, and extract their features using the shared encoder $f_\theta(\cdot)$:

$$\mathbf{F}_z = f_\theta(Z_i), \quad \mathbf{F}_x = f_\theta(X_{i+1}), \quad (2)$$

where $\mathbf{F}_z$ and $\mathbf{F}_x$ denote the template feature and search feature, respectively. By sharing the encoder parameters, both branches are mapped into a unified feature space, which facilitates subsequent relation modeling between the template and search regions.

In the self-supervised pre-training stage, a portion of the patches in the search region is randomly masked, while the template branch remains fully visible. The SDRM then establishes fine-grained cross-frame correspondences between the two branches by treating the search feature as the query and the template feature as the key and value:

$$\tilde{\mathbf{F}}_x = \mathcal{M}_{\text{SDRM}}(\mathbf{F}_x, \mathbf{F}_z, \mathbf{F}_z), \quad (3)$$

where $\mathcal{M}_{\text{SDRM}}(\cdot)$ denotes the dynamic relation matching operation, and $\tilde{\mathbf{F}}_x$ represents the relation-enhanced search feature. The enhanced search feature is then combined with mask tokens and fed into a lightweight decoder to reconstruct the original search image.

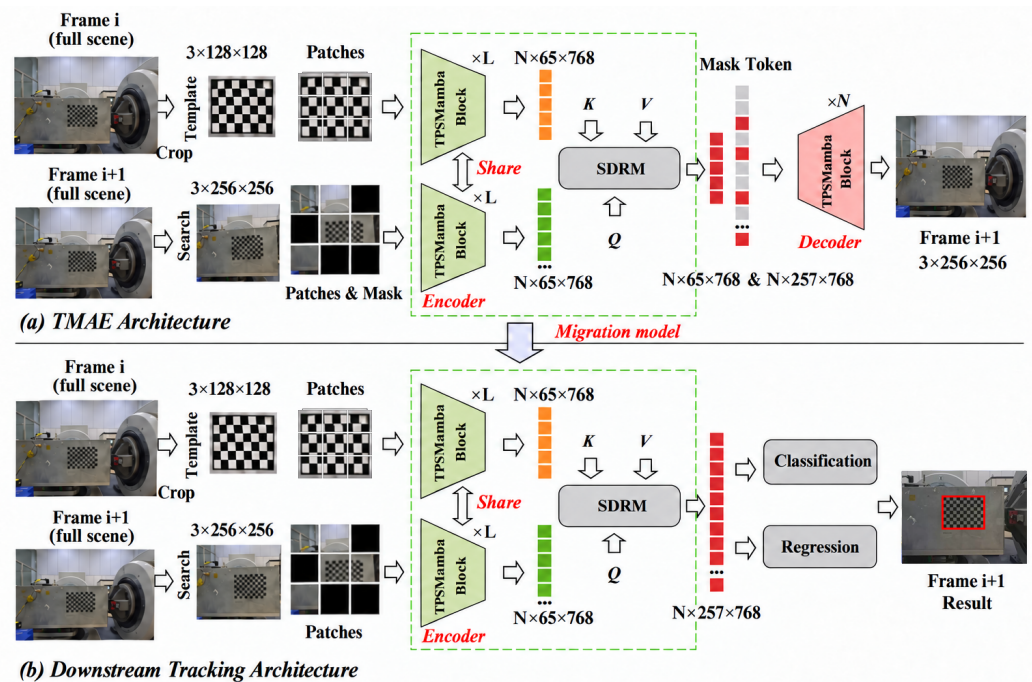


Figure 1. Overall architecture of the proposed method. The upper part shows the template-guided masked autoencoder (TMAE) pre-training stage, while the lower part shows the downstream tracking stage. White arrows indicate the feature-processing flow, bidirectional arrows between the TPSMamba blocks indicate parameter sharing, and the large light-blue downward arrow indicates the transfer of the pretrained modules to the downstream tracking network. Orange, green, red, and gray tokens represent template features, search features, relation-enhanced features, and mask tokens, respectively. The green dashed boxes enclose the encoder and SDRM modules transferred from pre-training to downstream tracking.

After the pre-training stage, the reconstruction decoder is discarded, while the learned parameters of the shared encoder and SDRM are transferred to the downstream tracking network. During tracking, the template and search images are processed by the same TPSMamba encoder and SDRM. The resulting relation-enhanced search feature is subsequently fed into the classification and regression heads. The classification branch predicts the target location, while the regression branch estimates the position and scale of the target bounding box. The final target state in frame I_{i+1} is expressed as

$$\hat{B}_{i+1} = (\hat{x}_{i+1}, \hat{y}_{i+1}, \hat{w}_{i+1}, \hat{h}_{i+1}). \quad (4)$$

Through this two-stage learning strategy, TMAE provides task-relevant initialization for the downstream tracker, while SDRM further models dynamic correspondences between the template and the current search region, thereby improving target localization.

3.2. Topology-Preserving Scanning Mamba Encoder

Accurate target localization requires the feature encoder to preserve local two-dimensional spatial relationships while capturing long-range dependencies between distant image regions. Global self-attention provides effective long-range interaction but incurs quadratic computational complexity with respect to the number of visual tokens, whereas state-space models offer a more efficient alternative with linear scaling in sequence length [19,20]. However, directly serializing two-dimensional template and search features into one-dimensional sequences may alter their original spatial adjacency and weaken spatial continuity. To balance spatial structure preservation, long-range dependency mod-

eling, and computational efficiency, we develop the topology-preserving scanning Mamba encoder (TPSMamba), whose detailed architecture is illustrated in Figure 2.

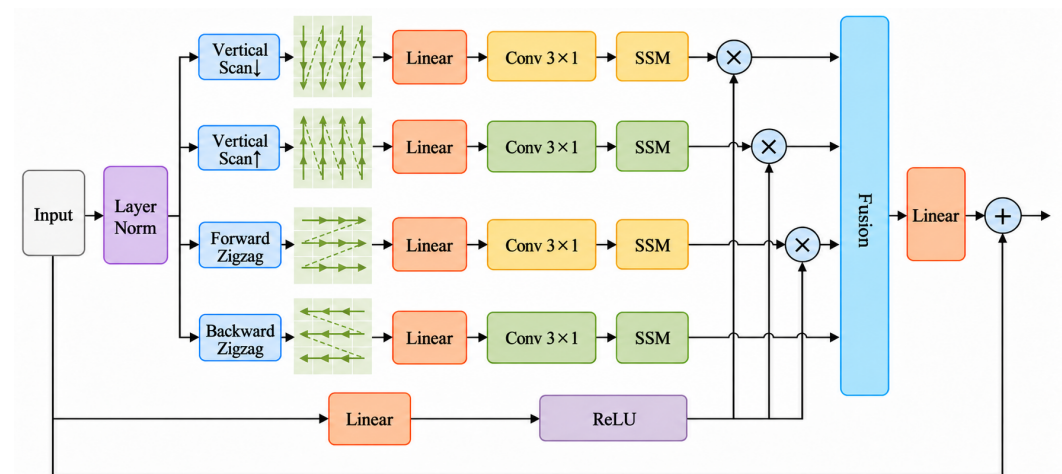


Figure 2. Detailed architecture of the proposed TPMamba block.

Given the input token sequence $\mathbf{L}^{b-1}$ of the b -th TPMamba block, layer normalization is first applied to stabilize feature propagation:

$$\mathbf{L}' = \text{LN}(\mathbf{L}^{b-1}), \quad (5)$$

where $\text{LN}(\cdot)$ denotes the layer normalization operation.

As shown in Figure 2, the normalized features are rearranged into four one-dimensional sequences through two complementary bidirectional scanning strategies. Specifically, the vertical scanning strategy traverses the feature map column by column in opposite directions, generating downward and upward sequences. The zigzag scanning strategy follows an alternating row-wise traversal pattern, in which adjacent rows are scanned in opposite directions, while its reverse branch traverses the same path in the opposite order. By combining vertical and zigzag bidirectional scanning, the proposed mechanism captures contextual dependencies along complementary spatial paths and helps preserve spatial continuity when two-dimensional features are converted into one-dimensional sequences.

For each scanning direction, the resulting sequence is first processed by a linear projection, followed by a one-dimensional convolution and a SiLU activation function. The transformed sequence is subsequently fed into a selective state space model (SSM):

$$\mathbf{s}'_d = \text{SSM}(\text{SiLU}(\text{Conv3x1}(\mathbf{s}_d))), \quad d \in \{1, 2, 3, 4\}, \quad (6)$$

where $\mathbf{s}_d$ denotes the token sequence obtained from the d -th scanning direction. The one-dimensional convolution captures local dependencies along each scanning path, while the selective SSM efficiently propagates long-range contextual information with linear computational complexity.

Since the four scanning sequences correspond to different spatial traversal orders, the features generated by the reverse scanning paths are first restored to the original token arrangement. The aligned multi-directional features are then aggregated as

$$\mathbf{s}_{\text{agg}} = \mathbf{s}'_1 + \text{Reverse}(\mathbf{s}'_2) + \mathbf{s}'_3 + \text{Reverse}(\mathbf{s}'_4), \quad (7)$$

where $\text{Reverse}(\cdot)$ denotes the inverse rearrangement operation that restores the reverse-scanned sequences to the original spatial token order.

The aggregated representation is subsequently modulated by the projected gating feature and fused with the block input through a residual connection:

$$\mathbf{L}^b = \mathbf{L}^{b-1} + \text{Linear}[\text{SiLU}(\mathbf{z}) \odot \mathbf{s}_{\text{agg}}], \quad (8)$$

where $\mathbf{z}$ denotes the projected gating feature and $\odot$ represents element-wise multiplication.

By stacking multiple TPSMamba blocks, the encoder progressively captures both local structural information and long-range spatial dependencies. The multi-directional scanning mechanism preserves the spatial organization of the input features while avoiding the quadratic computational complexity introduced by self-attention. Therefore, the same TPSMamba encoder can be efficiently shared by the template and search branches in both the pre-training and downstream tracking stages.

3.3. Sparse Dynamic Relation Matcher

The detailed structure of the sparse dynamic relation matcher (SDRM) is illustrated in Figure 3. Conventional Siamese trackers usually rely on direct template–search matching, which may introduce irrelevant background information and become less reliable when the target appearance changes. To alleviate this limitation, SDRM represents the template and search features as sparse hypergraphs, where each hyperedge can associate multiple semantically related tokens, and establishes adaptive fine-grained relations between the two branches.

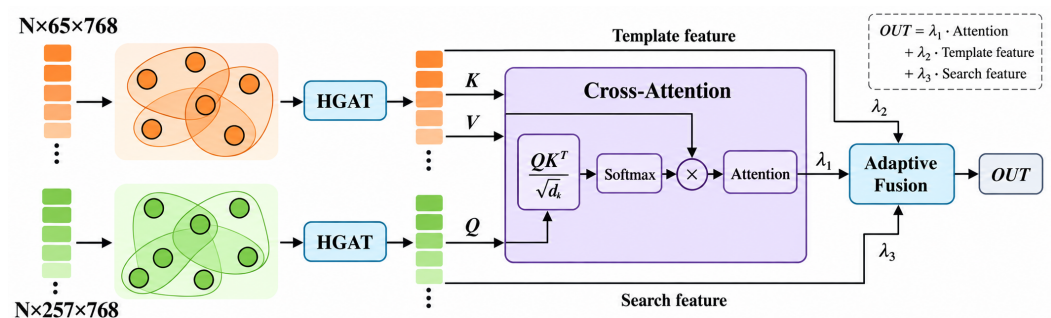


Figure 3. Detailed architecture of the proposed sparse dynamic relation matcher. The upper orange branch processes the template features, while the lower green branch processes the search features. The arrows indicate the feature-processing and fusion flows.

Given the template feature $\mathbf{F}_z$ and search feature $\mathbf{F}_x$ extracted by the shared TPSMamba encoder, two sparse hypergraphs are first constructed:

$$\mathcal{G}_z = (V_z, \mathcal{E}_z, H_z), \quad \mathcal{G}_x = (V_x, \mathcal{E}_x, H_x), \quad (9)$$

where V_z and V_x denote the vertex sets, $\mathcal{E}_z$ and $\mathcal{E}_x$ denote the hyperedge sets, and H_z and H_x are the corresponding incidence matrices. Each feature token is treated as a vertex. For each anchor vertex, its k most relevant tokens are selected according to feature similarity and grouped with the anchor vertex to form a hyperedge. Different from an ordinary graph, in which an edge only describes a pairwise relation between two vertices, each hyperedge can jointly associate multiple related tokens. The resulting sparse hypergraph therefore captures higher-order local relations while avoiding dense interactions with irrelevant background tokens.

After hypergraph construction, a hypergraph attention network (HGAT) is applied independently to the template and search branches. Information is first aggregated within each hyperedge and then propagated back to its associated vertices, allowing each token to adaptively integrate higher-order contextual information from semantically related tokens.

The resulting hypergraph-enhanced representations are denoted by $\mathbf{O}_z$ and $\mathbf{O}_x$. A general hypergraph aggregation for the i -th vertex can be expressed as

$$\mathbf{o}_i = \sum_{e \in \mathcal{E}(i)} \alpha_{i,e} \left(\sum_{j \in e} \beta_{e,j} \mathbf{f}_j \right), \quad (10)$$

where $\mathcal{E}(i)$ denotes the set of hyperedges containing vertex i , $\beta_{e,j}$ represents the contribution of vertex j to hyperedge e , and $\alpha_{i,e}$ denotes the attention weight of hyperedge e when updating vertex i . This vertex–hyperedge–vertex aggregation enables SDRM to model higher-order structural relations before cross-branch matching.

The hypergraph-enhanced features are subsequently used for cross-sequence matching. Specifically, the search representation $\mathbf{O}_x$ is projected as the query, while the template representation $\mathbf{O}_z$ is projected as the key and value:

$$\mathbf{Q} = \mathbf{O}_x \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{O}_z \mathbf{W}_K, \quad \mathbf{V} = \mathbf{O}_z \mathbf{W}_V, \quad (11)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices.

Based on the projected query, key, and value, the cross-attention response is calculated as

$$\mathbf{A} = \text{softmax} \left(\frac{\mathbf{Q} \mathbf{K}^\top}{\sqrt{d_k}} \right) \mathbf{V}, \quad (12)$$

where d_k denotes the feature dimension of the query and key. By using the search feature as the query and the template feature as the key and value, SDRM dynamically propagates target-related information from the template branch to the search branch, thereby establishing flexible template–search correspondences.

Finally, the cross-attention response is adaptively fused with the hypergraph-enhanced template and search representations:

$$\mathbf{F}_{\text{SDRM}} = \lambda_1 \mathbf{A} + \lambda_2 \mathbf{O}_z + \lambda_3 \mathbf{O}_x, \quad (13)$$

where λ_1 , λ_2 , and λ_3 are learnable coefficients. This adaptive fusion enables the network to balance cross-branch matching information with the higher-order structural information preserved in the two hypergraph branches. Meanwhile, the sparse hypergraph construction restricts relation modeling to compact groups of relevant tokens, thereby reducing redundant background interactions.

In the TMAE pre-training stage, SDRM serves as an information interaction bridge between the template tokens and the visible search tokens, providing cross-frame contextual information for masked-region reconstruction. In the downstream tracking stage, the same module is reused to establish dynamic correspondences between the target template and the current search region, and the relation-enhanced features are further fed into the classification and bounding-box regression heads.

3.4. Loss Functions and Optimization Objectives

The proposed method adopts a two-stage learning strategy comprising self-supervised pre-training and downstream tracking fine-tuning. The pre-training stage optimizes the reconstruction of masked search-region patches, whereas the downstream stage jointly optimizes target classification and bounding-box regression.

3.4.1. Self-Supervised Pre-Training Loss

During self-supervised pre-training, TMAE uses the target template extracted from frame i as contextual information and randomly masks a subset of image patches in the

search region of frame $i + 1$. The network is then trained to recover the original pixel content of the masked patches. Let $\mathcal{M}$ denote the index set of masked search-region patches, and let $\mathbf{x}_p$ and $\hat{\mathbf{x}}_p$ denote the ground-truth content and the reconstructed result of the p th patch, respectively. The reconstruction objective is defined as the mean squared error over the masked patches:

$$\mathcal{L}_{\text{TMAE}} = \frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \|\hat{\mathbf{x}}_p - \mathbf{x}_p\|_2^2, \quad (14)$$

where $|\mathcal{M}|$ denotes the number of masked patches. The loss is calculated only over the masked regions, preventing the network from solving the reconstruction task by directly copying visible pixels. Instead, the network is required to recover the missing content by exploiting both the target template and the visible context in the search region. By minimizing Equation (14), the encoder learns target-related representations from unlabeled video frames.

3.4.2. Downstream Tracking Loss

After self-supervised pre-training, the learned TPSMamba encoder and SDRM parameters are transferred to the downstream tracking network. Target classification and bounding-box regression are jointly optimized using a multi-task objective.

For target classification, the network predicts the probability that each candidate location belongs to the target. The binary cross-entropy loss is defined as

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log(1 - p_i)], \quad (15)$$

where p_i is the predicted target probability, $y_i \in \{0, 1\}$ is the corresponding ground-truth label, and N is the number of candidate locations. For bounding-box regression, the L_1 loss is used to constrain the coordinate difference between the predicted and ground-truth boxes:

$$\mathcal{L}_{L1} = \frac{1}{N_+} \sum_{j=1}^{N_+} \|\hat{\mathbf{b}}_j - \mathbf{b}_j\|_1, \quad (16)$$

where $\hat{\mathbf{b}}_j$ and $\mathbf{b}_j$ denote the predicted and ground-truth bounding boxes of the j th positive sample, respectively, and N_+ is the number of positive samples. The GIoU loss is further introduced to constrain the spatial overlap between the predicted and ground-truth boxes:

$$\mathcal{L}_{\text{GIoU}} = \frac{1}{N_+} \sum_{j=1}^{N_+} \left[1 - \frac{|\hat{\mathbf{b}}_j \cap \mathbf{b}_j|}{|\hat{\mathbf{b}}_j \cup \mathbf{b}_j|} + \frac{|\mathbf{C}_j \setminus (\hat{\mathbf{b}}_j \cup \mathbf{b}_j)|}{|\mathbf{C}_j|} \right], \quad (17)$$

where $\mathbf{C}_j$ is the smallest enclosing box covering both $\hat{\mathbf{b}}_j$ and $\mathbf{b}_j$, and $|\cdot|$ denotes the area of a region. The overall downstream tracking objective is defined as

$$\mathcal{L}_{\text{track}} = \gamma_{\text{cls}} \mathcal{L}_{\text{cls}} + \gamma_{L1} \mathcal{L}_{L1} + \gamma_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}, \quad (18)$$

where γ_{cls} , γ_{L1} , and γ_{GIoU} control the contributions of the three loss terms. They are fixed as $\gamma_{\text{cls}} = 1$, $\gamma_{L1} = 5$, and $\gamma_{\text{GIoU}} = 2$ in all experiments. The self-supervised pre-training and downstream tracking stages are optimized separately using $\mathcal{L}_{\text{TMAE}}$ and $\mathcal{L}_{\text{track}}$, respectively.

3.5. Physical Displacement Calibration and Reconstruction

The proposed tracker estimates the target bounding box in each video frame. In the controlled shaking-table experiments, the checkerboard targets were rigidly attached to the

shaking-table platforms. Therefore, the tracking results represent the rigid-body motion of the checkerboard targets and the corresponding shaking-table platforms rather than the deformation or dynamic response of an independent flexible structure.

For the t -th frame, the target state predicted from the k -th camera view is denoted as

$$\hat{\mathbf{B}}_{k,t} = (\hat{x}_{k,t}, \hat{y}_{k,t}, \hat{w}_{k,t}, \hat{h}_{k,t}), \quad (19)$$

where $k \in \{E-W, N-S, U-D\}$ denotes the camera view corresponding to the east–west, north–south, or vertical measurement direction. The variables $\hat{x}_{k,t}$ and $\hat{y}_{k,t}$ denote the horizontal and vertical coordinates of the upper-left corner of the predicted bounding box, while $\hat{w}_{k,t}$ and $\hat{h}_{k,t}$ denote its width and height, respectively. The corresponding target-center coordinate is calculated as

$$\mathbf{p}_{k,t} = \begin{bmatrix} p_{x,k,t} \\ p_{y,k,t} \end{bmatrix} = \begin{bmatrix} \hat{x}_{k,t} + \frac{\hat{w}_{k,t}}{2} \\ \hat{y}_{k,t} + \frac{\hat{h}_{k,t}}{2} \end{bmatrix}. \quad (20)$$

Since the tracker outputs the target position in pixel coordinates, spatial calibration is required to convert the image trajectory into physical displacement. In the present experiment, the optical axes of the three cameras were approximately perpendicular to their corresponding motion planes, the image-coordinate axes were approximately aligned with the principal physical motion directions, and the checkerboard targets remained close to the calibrated planes throughout the experiments. Consequently, the relatively small perspective effects and coupling between the two image-coordinate axes were neglected, and the image-to-physical coordinate relationship was approximated by a scale-and-translation transformation. For the k -th camera, the homogeneous image coordinate of the target center is written as

$$\tilde{\mathbf{p}}_{k,t} = \begin{bmatrix} p_{x,k,t} \\ p_{y,k,t} \\ 1 \end{bmatrix}. \quad (21)$$

The corresponding homogeneous physical coordinate is obtained by

$$\tilde{\mathbf{q}}_{k,t} = \mathbf{H}_k \tilde{\mathbf{p}}_{k,t}, \quad (22)$$

where $\mathbf{H}_k$ is the image-to-physical calibration matrix for the k -th camera and is expressed as

$$\mathbf{H}_k = \begin{bmatrix} s_{x,k} & 0 & t_{x,k} \\ 0 & s_{y,k} & t_{y,k} \\ 0 & 0 & 1 \end{bmatrix}. \quad (23)$$

The parameters $s_{x,k}$ and $s_{y,k}$ denote the physical scale factors along the horizontal and vertical image directions, respectively, and are expressed in centimeters per pixel. The parameters $t_{x,k}$ and $t_{y,k}$ describe the translation between the image-coordinate and physical-coordinate systems. The resulting physical coordinates are

$$\begin{aligned} X_{k,t} &= s_{x,k} p_{x,k,t} + t_{x,k}, \\ Y_{k,t} &= s_{y,k} p_{y,k,t} + t_{y,k}, \end{aligned} \quad (24)$$

The scale parameters in the calibration matrix are determined from the known physical dimensions of the checkerboard target and their corresponding lengths in the image. Let L_x and L_y denote the known physical lengths of the checkerboard in the horizontal and

vertical directions, and let $l_{x,k}$ and $l_{y,k}$ denote their corresponding image lengths in pixels. The scale factors are calculated as

$$s_{x,k} = \frac{L_x}{l_{x,k}}, \quad s_{y,k} = \frac{L_y}{l_{y,k}}. \quad (25)$$

To reduce the influence of individual checkerboard-edge extraction errors, the pixel lengths of multiple checkerboard cells are measured, and their average values are used to calculate the scale factors. The translation parameters are determined from the correspondence between the selected physical-coordinate origin and the image-coordinate origin:

$$t_{x,k} = X_{k,0} - s_{x,k}p_{x,k,0}, \quad t_{y,k} = Y_{k,0} - s_{y,k}p_{y,k,0}, \quad (26)$$

where $(p_{x,k,0}, p_{y,k,0})$ is the target-center coordinate in the reference frame and $(X_{k,0}, Y_{k,0})$ is its corresponding physical coordinate. Taking the target position in the reference frame as the initial position, the physical displacements along the two coordinate directions are expressed as

$$\begin{aligned} \Delta X_{k,t} &= X_{k,t} - X_{k,0} = s_{x,k}(p_{x,k,t} - p_{x,k,0}), \\ \Delta Y_{k,t} &= Y_{k,t} - Y_{k,0} = s_{y,k}(p_{y,k,t} - p_{y,k,0}). \end{aligned} \quad (27)$$

Therefore, the translation terms in the calibration matrix cancel when the displacement is calculated relative to the reference frame. For each camera, only the coordinate component corresponding to its principal observation direction is retained. The physical displacement in the k -th direction can consequently be written as

$$d_k(t) = s_k[u_k(t) - u_k(0)], \quad (28)$$

where $u_k(t)$ is the target-center coordinate projected onto the principal motion direction in the k -th camera view, and s_k is the corresponding physical scale factor.

The three directional displacement components are subsequently combined as

$$\mathbf{d}(t) = \begin{bmatrix} d_{E-W}(t) \\ d_{N-S}(t) \\ d_{U-D}(t) \end{bmatrix}. \quad (29)$$

The calibration matrix for each camera was independently determined before processing the test sequences and remained unchanged throughout the subsequent experiments. Through this calibration procedure, the target-center trajectories in image coordinates were converted into physical displacement histories expressed in centimeters. Under the present controlled experimental conditions, the reconstructed displacement represents the rigid-body motion of the shaking table and its attached checkerboard target.

3.6. Motion Processing and Seismic-Intensity Estimation

The displacement histories reconstructed from the video tracking results may contain target-localization errors, imaging noise, and frame-wise position fluctuations. Direct numerical differentiation of noisy displacement signals can amplify their high-frequency components, particularly when acceleration is obtained through second-order differentiation. Therefore, the three directional displacement histories are filtered before the velocity and acceleration are calculated.

A fourth-order zero-phase Butterworth low-pass filter is independently applied to each directional displacement component:

$$\bar{d}_k(t) = \mathcal{B}_{20}[d_k(t)], \quad k \in \{E-W, N-S, U-D\}, \quad (30)$$

where $\mathcal{B}_{20}(\cdot)$ denotes the Butterworth low-pass filtering operation with a cutoff frequency of 20 Hz, and $\bar{d}_k(t)$ is the filtered displacement in the k -th direction.

Forward–backward filtering is employed to eliminate the phase delay introduced by the filter and to prevent temporal shifts in the displacement peaks. Different candidate cutoff frequencies are compared using the validation split, with the error between the video-reconstructed displacement and the reference displacement serving as the primary selection criterion. The results show that a cutoff frequency of 20 Hz produces the lowest trajectory-reconstruction error. This parameter is therefore fixed and applied unchanged to all test sequences.

The videos were recorded at a frame rate of 120 fps. Accordingly, the sampling frequency of the displacement sequence is

$$f_s = 120 \text{ Hz}, \quad (31)$$

and the corresponding Nyquist frequency is

$$f_N = \frac{f_s}{2} = 60 \text{ Hz}. \quad (32)$$

The time interval between two consecutive video frames is

$$\Delta t = \frac{1}{120} \text{ s}. \quad (33)$$

For the internal samples of each filtered displacement sequence, the velocity and acceleration are calculated using central finite differences. The velocity component in the k -th direction at time t_i is calculated as

$$v_k(t_i) = \frac{\bar{d}_k(t_{i+1}) - \bar{d}_k(t_{i-1}))}{2\Delta t}. \quad (34)$$

The corresponding acceleration component is calculated as

$$a_k(t_i) = \frac{\bar{d}_k(t_{i+1}) - 2\bar{d}_k(t_i) + \bar{d}_k(t_{i-1}))}{\Delta t^2}. \quad (35)$$

A fourth-order Butterworth low-pass filter with a cutoff frequency of 20 Hz is applied to the reconstructed displacement signal. Forward–backward filtering is employed to obtain a zero-phase response and avoid phase distortion. Symmetric reflection padding is applied at both ends of each sequence before filtering, and the padding is removed afterward to retain the original signal length. Filtering is performed before numerical differentiation to suppress the amplification of target-localization noise in the reconstructed velocity and acceleration histories. Because central finite differences cannot be directly calculated for the first and last samples of a sequence, these two endpoints are excluded from the subsequent peak-value calculation.

According to the Chinese Seismic Intensity Scale [2], the acceleration components in the east–west, north–south, and vertical directions are combined to obtain the resultant acceleration:

$$a(t_i) = \sqrt{a_{E-W}^2(t_i) + a_{N-S}^2(t_i) + a_{U-D}^2(t_i)}. \quad (36)$$

Similarly, the three directional velocity components are combined to obtain the resultant velocity:

$$v(t_i) = \sqrt{v_{E-W}^2(t_i) + v_{N-S}^2(t_i) + v_{U-D}^2(t_i)}. \quad (37)$$

The peak ground acceleration and peak ground velocity are obtained from the resultant motion histories:

$$\text{PGA} = \max_i a(t_i), \quad (38)$$

and

$$\text{PGV} = \max_i v(t_i). \quad (39)$$

PGA and PGV are expressed in m/s^2 and m/s , respectively. Because the calibrated displacement is initially expressed in millimeters, the displacement, velocity, and acceleration are converted into SI units before PGA and PGV are substituted into the intensity equations.

The acceleration-based instrumental-intensity estimate is calculated as

$$I_A = 3.17 \log_{10}(\text{PGA}) + 6.59, \quad (40)$$

while the velocity-based instrumental-intensity estimate is calculated as

$$I_V = 3.00 \log_{10}(\text{PGV}) + 9.77, \quad (41)$$

where I_A and I_V denote the intensity values derived from the reconstructed peak acceleration and peak velocity, respectively.

The final instrumental-intensity estimate is determined as

$$\hat{I} = \begin{cases} I_V, & I_A \geq 6.0 \text{ and } I_V \geq 6.0, \\ \frac{I_A + I_V}{2}, & I_A < 6.0 \text{ or } I_V < 6.0. \end{cases} \quad (42)$$

Through the above processing pipeline, the target locations estimated from consecutive video frames are sequentially converted into physical displacement, velocity, acceleration, PGA, PGV, and instrumental-intensity estimates. The present pipeline validates video-based rigid-motion measurement and seismic-intensity estimation under controlled shaking-table excitation.

4. Experiments

4.1. Implementation Details

The proposed method was implemented using PyTorch 2.2 and trained on a Linux workstation equipped with an NVIDIA A100 GPU. All input frames were converted to RGB format. For each pair of consecutive frames, the target template was cropped from the current frame and resized to 128×128 pixels, while the corresponding search region in the subsequent frame was resized to 256×256 pixels.

The network was trained using a two-stage strategy. In the self-supervised pre-training stage, only the training split of GOT-10K was used for TMAE pre-training. The TMAE was optimized using the Adam optimizer with a batch size of 64 and an initial learning rate of 1.5×10^{-4} . The masking ratio of the search-region patches was set to 0.75, and the model was trained for 100 epochs. The template and search branches shared the same TPSMamba encoder parameters, and the reconstruction objective was optimized using the masked-patch mean squared error $\mathcal{L}_{\text{TMAE}}$.

Following self-supervised pre-training, the learned TPSMamba encoder and SDRM parameters were transferred to the downstream tracking network. The network was fine-tuned for 20 epochs using the Adam optimizer with a batch size of 8 and an initial learning rate of 1×10^{-5} . Binary cross-entropy was adopted for target classification, while the L_1 and GloU losses were used for bounding-box regression. The corresponding loss weights

were fixed as $\gamma_{\text{cls}} = 1$, $\gamma_{L1} = 5$, and $\gamma_{\text{GIoU}} = 2$. These settings were kept unchanged in all downstream tracking experiments.

To evaluate the robustness of the reported results to training randomness and provide a basis for statistical comparison, ten repeated training-and-evaluation runs were conducted for the public-benchmark tracking experiment and the seismic-video trajectory-reconstruction experiment using the same random seeds, ranging from 2017 to 2026. Specifically, repeated experiments were performed on LaSOT and GOT-10K to evaluate tracking performance using AUC, Prec, NormPrec, AO, $\text{SR}_{0.5}$, and $\text{SR}_{0.75}$. Repeated evaluations were also conducted on the self-built Shaking-Table Seismic-Video dataset using displacement RMSE, displacement MAE, EPE, and seismic-intensity-estimation metrics. The ten consecutive integers from 2017 to 2026 were specified before the experiments solely to ensure reproducibility. These values served only as random-seed identifiers and controlled model initialization, data shuffling, and other randomized training operations.

4.2. Datasets

Three complementary datasets were selected according to the two evaluation objectives of this study: assessing generic visual tracking performance and validating video-based seismic-intensity prediction. LaSOT was selected because its long sequences and diverse tracking challenges are suitable for evaluating long-term localization accuracy and temporal tracking stability. GOT-10K complements LaSOT through its diverse object categories and motion patterns and its one-shot evaluation protocol, providing an assessment of generalization to unseen targets. However, the ultimate objective of this study is not generic target localization, but seismic-intensity prediction based on video-reconstructed motion trajectories and the subsequent calculation of velocity, acceleration, PGA, and PGV. Therefore, a self-built Shaking-Table Seismic-Video dataset containing controlled seismic excitations, three orthogonal visual observations, synchronized laser displacement references, and instrumental-intensity levels II–VI was used to evaluate the complete pipeline from target tracking to displacement reconstruction and intensity estimation.

LaSOT [18] is a large-scale long-term visual tracking benchmark containing 280 test sequences with an average length of approximately 2500 frames. It covers a wide range of object categories and challenging tracking conditions and is therefore suitable for evaluating long-term tracking accuracy and robustness. Following the standard evaluation protocol, area under the success curve, precision, and normalized precision are adopted as the evaluation metrics. Since LaSOT does not provide an official validation split, two sequences were randomly selected from the 16 training sequences of each target category for validation, while the remaining 14 sequences were used for training. Consequently, the adopted LaSOT partition consisted of 980 training sequences, 140 validation sequences, and 280 official test sequences. The validation set was used for hyperparameter selection and checkpoint determination, whereas the official test set was used exclusively for final performance evaluation.

GOT-10K [18] is a large-scale one-shot tracking benchmark comprising 9335 training sequences, 180 validation sequences, and 420 test sequences, with the test set covering 84 object categories. Following the official protocol, only the GOT-10K training split was used for GOT-10K-specific model training. The official validation split was used for hyperparameter selection and checkpoint determination. Average overlap, success rate at an IoU threshold of 0.5 ($\text{SR}_{0.5}$), and success rate at an IoU threshold of 0.75 ($\text{SR}_{0.75}$) are reported.

Shaking-Table Seismic-Video Dataset is a self-built dataset collected through controlled shaking-table experiments at the Institute of Engineering Mechanics, China Earthquake Administration, for evaluating video-based displacement reconstruction and seismic-intensity

estimation. The experimental system was centered on an earthquake-simulation shaking table, with three cameras and three laser displacement sensors arranged along the X-, Y-, and Z-directions, respectively. Three checkerboard calibration targets were rigidly mounted on the three platforms of the shaking table and served as visual tracking targets. The experimental setup and the checkerboard target mounted on the shaking table are illustrated in Figure 4. No independent structural specimen, such as a building or bridge model, was included in the experiments. Therefore, the measured motion corresponds to the motion of the shaking table and the checkerboard targets rigidly attached to it, rather than to the dynamic response of a flexible structure. This configuration was adopted to isolate the accuracy of the video-based motion measurement and seismic-intensity estimation pipeline from structure-dependent dynamic effects. The video and laser displacement signals were recorded independently by their respective acquisition devices. For each complete excitation record, the onset of the shaking-table excitation was first used as a common temporal reference to perform an initial alignment between the video frames and the laser displacement signal. The residual time offset was then determined by maximizing the cross-correlation between the reconstructed visual displacement and the laser displacement signal. For each complete excitation record, only one temporal offset was estimated and applied uniformly to the displacement trajectories obtained by all tracking methods. After correcting this offset, the laser measurements were resampled at the video timestamps, ensuring that the visual and reference displacements used for error calculation corresponded to the same sampling instants.

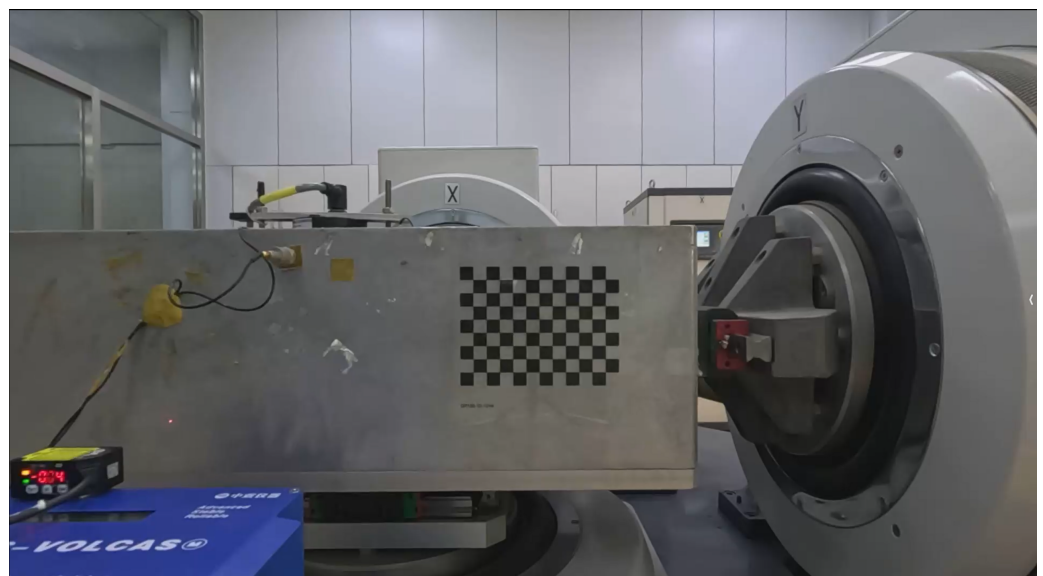


Figure 4. Experimental setup of the shaking-table seismic-video data acquisition system. The checkerboard target is rigidly mounted on the shaking table and serves as the visual tracking target.

The earthquake-simulation shaking table used in the experiments was a 3ES-20 system with a platform size of 600×600 mm. Its rated uniaxial sinusoidal excitation force is 20 kN, and its maximum uniaxial random-excitation frequency is 1200 Hz. Under no-load conditions, the maximum acceleration, velocity, and displacement are 120 m/s^2 , 1.2 m/s , and $\pm 19 \text{ mm}$, respectively. Three GoPro cameras were used to record the checkerboard targets, while HG-C1100 laser displacement sensors (Panasonic Industry, Osaka, Japan) were employed to acquire reference displacement measurements. Each checkerboard target had physical dimensions of 12×12 cm, with an individual checker size of 10 mm. The cameras were positioned approximately 1 m from the corresponding shaking-table surfaces

and directly faced the checkerboard targets, whereas the laser displacement sensors were positioned approximately 10 cm from the table surfaces.

The videos were recorded at a spatial resolution of 1920×1080 pixels and a frame rate of 120 fps. The dataset contains 84 seismic-video sequences, including 16 sine-sweep signals and 68 real earthquake-wave records. The real earthquake records include motions originating from events such as the Wenchuan, Wushi, and Maerkang earthquakes. According to the reference instrumental measurements, the dataset covers instrumental-intensity levels II to VI. Specifically, the dataset contains 3, 15, 26, 20, and 20 records corresponding to intensity levels II, III, IV, V, and VI, respectively. The recorded videos were segmented according to the prescribed shaking intervals of the experimental platform, resulting in effective vibration clips of approximately 60–80 s. Each video segment was further decomposed into individual frames, and the checkerboard targets were annotated to construct the seismic-video dataset.

In total, the dataset corresponds to approximately 1.63 h of effective seismic video and 705,600 image frames. The dataset was divided into mutually exclusive development, validation, and test subsets containing 58, 16, and 10 complete excitation records, respectively. Partitioning was performed at the complete excitation-record level so that different temporal segments originating from the same input motion could not appear in multiple subsets. The held-out test set was used exclusively for the quantitative evaluation of displacement reconstruction and seismic-intensity estimation.

Laser displacement measurements were used as reference ground truth for evaluating the reconstructed displacement trajectories. The camera views along the three mutually orthogonal directions provide the corresponding motion components required for subsequent velocity, acceleration, and instrumental-intensity estimation. Accordingly, the reference signals in this study represent the controlled shaking-table motion rather than the response of an independent structural specimen.

4.3. Data Partitioning and Model-Development Protocol

During self-supervised pre-training, TMAE was trained exclusively using raw video frames from the GOT-10K training split without using manual annotations. Through the template-guided masked-reconstruction task, the model learned tracking-oriented representations from unlabeled video frames. The pre-training checkpoint was selected according to the reconstruction loss on the GOT-10K validation split.

During downstream tracking, the model was fine-tuned separately for each dataset. For LaSOT and GOT-10K, the dataset partitions were kept fixed across all evaluations. For LaSOT, 980 sequences were used for training, a fixed subset of 140 sequences from the official training set was used for validation, and the 280 official test sequences were used for final evaluation. For GOT-10K, the 9335 official training sequences, 180 validation sequences, and 420 test sequences were used. Ten repeated training runs were conducted using the random seeds 2017–2026. For these runs, the random seeds controlled stochastic training factors, including model initialization and data shuffling, while the dataset partitions remained unchanged. Therefore, the variability reported for LaSOT and GOT-10K reflects stochastic variation across repeated training runs under fixed dataset partitions.

For the self-built Shaking-Table Seismic-Video dataset, the random seeds 2017–2026 were used exclusively to generate ten different dataset partitions. For each seed, the 84 complete excitation records were randomly divided into 58 training records, 16 validation records, and 10 test records. All compared methods used the same seed-specific partition, enabling paired comparisons under identical data conditions. The training, validation, and test subsets were mutually exclusive within each evaluation. Therefore, the variability reported for the self-built dataset primarily reflects sensitivity to dataset partitioning rather

than variability across repeated training runs. The fixed seed set 2017–2026 was selected solely to ensure the reproducibility of the repeated experiments.

Model-specific and signal-processing hyperparameters, including the TMAE masking ratio, hypergraph neighborhood size, number of training epochs, learning rate, and filter cutoff frequency, were determined exclusively according to performance on the corresponding validation sets. Accordingly, the parameter-sensitivity experiments report validation-set performance rather than test-set performance. After the optimal configuration had been selected, the model architecture, training settings, signal-processing parameters, and checkpoint-selection criteria were fixed before test-set evaluation.

The official LaSOT and GOT-10K test sets were used to evaluate the models from the ten repeated training runs. For the self-built dataset, the corresponding 10 test records in each seed-specific partition were used for final evaluation. No test result was used for model selection, hyperparameter optimization, or subsequent configuration adjustment.

For the statistical comparisons reported in the following sections, the null hypothesis (H_0) assumes no mean paired performance difference between TopoRelTrack and the selected strongest competing method, whereas the alternative hypothesis (H_1) assumes a nonzero mean paired difference. Two-sided paired t -tests were conducted at a significance level of $\alpha = 0.05$. For LaSOT and GOT-10K, the compared methods were paired using the results obtained with the same training seed. For the self-built dataset, the methods were paired using the results obtained from the same seed-specific dataset partition.

Because the self-built dataset was repartitioned for each seed, partitions generated by different seeds could contain overlapping records. Consequently, although no record was shared among the training, validation, and test subsets within an individual evaluation, the ten evaluations were not completely independent. The statistical results for the self-built dataset were therefore interpreted with this cross-partition dependence in mind.

4.4. Training and Validation Convergence Analysis

To evaluate the training stability and convergence of the proposed two-stage learning method, the training and validation losses were monitored throughout the optimization process, as shown in Figure 5. During TMAE pre-training on GOT-10K, the reconstruction loss decreases rapidly from approximately 1.70 during the early epochs and falls below 0.60 after about 50 epochs. The rate of reduction subsequently becomes slower, and the loss stabilizes at approximately 0.42 after 100 epochs, with only minor fluctuations observed during the remaining epochs. This trend indicates that the masked-reconstruction objective reaches a stable optimization state and that extending pre-training beyond 100 epochs produces only limited additional improvement.

During downstream fine-tuning, similar convergence patterns can be observed on LaSOT, GOT-10K, and the self-built Shaking-Table Seismic-Video dataset. Most of the loss reduction occurs within the first 15–20 epochs, after which both the training and validation curves gradually enter stable plateaus. On LaSOT, the training and validation losses converge to approximately 0.08 and 0.095, respectively. On GOT-10K, the corresponding losses stabilize at approximately 0.09 and 0.108, while those on the self-built seismic-video dataset converge to approximately 0.09 and 0.109.

Overall, these results demonstrate that the proposed two-stage learning method achieves stable convergence on all datasets. Since further training after the observed plateau produces only negligible reductions in loss, 100 epochs are considered sufficient for TMAE pre-training, while 20 epochs provide an appropriate training duration for downstream fine-tuning on LaSOT, GOT-10K, and the self-built seismic-video dataset.

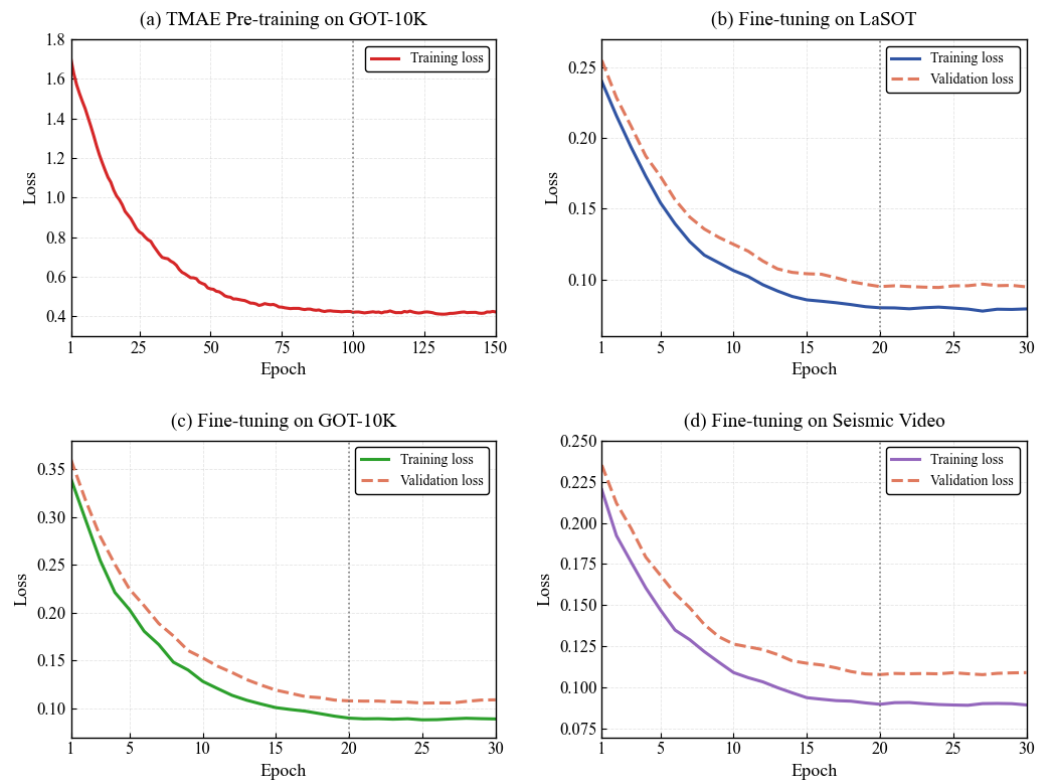


Figure 5. Training and validation loss curves of the proposed two-stage learning method. (a) TMAE pre-training loss on GOT-10K; downstream training and validation losses during fine-tuning on (b) LaSOT, (c) GOT-10K, and (d) the self-built Shaking-Table Seismic-Video dataset.

4.5. Tracking Performance on Public Benchmarks

To evaluate the general tracking performance of the proposed method, experiments were conducted on two widely used single-object tracking benchmarks, LaSOT and GOT-10K. For LaSOT, the area under the success curve (AUC), precision (Prec), and normalized precision (NormPrec) are reported. For GOT-10K, average overlap (AO), success rate at an intersection-over-union threshold of 0.5 ($SR_{0.5}$), and success rate at an intersection-over-union threshold of 0.75 ($SR_{0.75}$) are reported. All results in Tables 1 and 2 are reported as the mean and sample standard deviation over ten repeated training runs using random seeds 2017–2026.

As shown in Table 1, TopoRelTrack achieves an average AUC of 76.1%, a precision of 81.3%, and a normalized precision of 84.6% on LaSOT. The proposed method obtains the highest AUC and NormPrec among the compared trackers while matching HIPTrack in mean precision. Compared with HIPTrack, TopoRelTrack improves the mean AUC and NormPrec by 1.3 and 1.1 percentage points, respectively. Compared with LCCTV, which achieves the second-best NormPrec, the proposed method improves this metric by 1.0 percentage point. These results indicate that TopoRelTrack provides accurate target localization and robust overlap estimation in long-term tracking sequences.

On GOT-10K, TopoRelTrack achieves the best mean results across all three evaluation metrics, with an AO of 78.3%, an $SR_{0.5}$ of 88.5%, and an $SR_{0.75}$ of 75.8%, as shown in Table 2. Compared with the strongest competing results, the proposed method improves AO by 0.5 percentage points over OTrack-384 and HIPTrack and improves $SR_{0.75}$ by 0.9 percentage points over OTrack-384. The performance gains on both LaSOT and GOT-10K demonstrate that TopoRelTrack maintains competitive localization accuracy across long-term and one-shot tracking scenarios.

To further assess the robustness of the results to training randomness, Tables 3 and 4 report the detailed results of ten independent training runs using matched random seeds from 2017 to 2026. The results are summarized using the mean and sample standard deviation, and two-sided paired *t*-tests are performed using the matched results of HIPTrack and TopoRelTrack. On LaSOT, TopoRelTrack improves the mean AUC from 74.80% to 76.10% and the mean NormPrec from 83.50% to 84.60%, corresponding to improvements of 1.30 and 1.10 percentage points, respectively. These improvements are statistically significant at the 0.05 level, with $p = 0.0156$ for AUC and $p = 0.0169$ for NormPrec. HIPTrack and TopoRelTrack obtain essentially identical mean precision values, and their difference is not statistically significant ($p = 0.9995$). Nevertheless, TopoRelTrack exhibits lower standard deviations for all three LaSOT metrics, indicating more stable performance across independent training runs.

Table 1. Comparison with state-of-the-art trackers on LaSOT. Results are reported as mean $\pm$ standard deviation over ten runs with seeds 2017–2026. Best results are bold; second-best are underlined.

Method	AUC (%)	Prec (%)	NormPrec (%)
OSTrack-384 [18]	72.9 $\pm$ 1.72	79.8 $\pm$ 1.85	83.4 $\pm$ 1.65
ODTrack-B256 [33]	73.7 $\pm$ 1.58	80.4 $\pm$ 1.78	83.3 $\pm$ 1.62
LCCTV [16]	74.3 $\pm$ 1.65	81.0 $\pm$ 1.92	83.6 $\pm$ 1.70
HIPTrack [34]	<u>74.8 $\pm$ 1.60</u>	<u>81.3 $\pm$ 1.80</u>	83.5 $\pm$ 1.68
TopoRelTrack (ours)	76.1 $\pm$ 1.48	81.3 $\pm$ 1.35	84.6 $\pm$ 1.57

Table 2. Comparison with state-of-the-art trackers on GOT-10K. Results are reported as mean $\pm$ standard deviation over ten runs with seeds 2017–2026. Best results are bold; second-best are underlined.

Method	AO (%)	SR _{0.5} (%)	SR _{0.75} (%)
OSTrack-384 [18]	77.8 $\pm$ 1.85	86.7 $\pm$ 1.62	<u>74.9 $\pm$ 1.57</u>
ODTrack-B256 [33]	77.3 $\pm$ 1.70	85.3 $\pm$ 1.55	<u>74.4 $\pm$ 1.48</u>
LCCTV [16]	77.5 $\pm$ 1.68	<u>88.4 $\pm$ 1.68</u>	74.1 $\pm$ 1.32
HIPTrack [34]	<u>77.8 $\pm$ 1.75</u>	<u>87.6 $\pm$ 1.60</u>	74.3 $\pm$ 1.60
TopoRelTrack (ours)	78.3 $\pm$ 1.62	88.5 $\pm$ 1.55	75.8 $\pm$ 1.38

Table 3. Comparison of HIPTrack and TopoRelTrack on LaSOT over ten repeated training runs using random seeds 2017–2026. AUC, Prec, and NormPrec are reported in %. The *p*-values are obtained using two-sided paired *t*-tests: * $p < 0.05$.

Seed	AUC (%)		Prec (%)		NormPrec (%)	
	HIPTrack	TopoRelTrack	HIPTrack	TopoRelTrack	HIPTrack	TopoRelTrack
2017	74.884	75.923	81.395	81.143	83.588	84.459
2018	74.356	76.399	80.801	81.556	83.034	84.771
2019	75.938	74.399	82.580	79.807	84.695	83.321
2020	74.841	76.497	81.347	81.656	83.543	84.966
2021	73.530	73.590	79.871	79.020	82.166	82.060
2022	75.367	76.249	81.938	81.448	84.095	84.860
2023	77.297	78.505	84.109	83.518	86.122	87.317
2024	76.566	78.108	83.287	83.143	85.354	86.798
2025	73.185	75.770	79.484	80.963	81.805	83.946
2026	72.035	75.560	78.190	80.745	80.597	83.503
Mean $\pm$ SD	74.80 $\pm$ 1.60	76.10 $\pm$ 1.48	81.30 $\pm$ 1.80	81.30 $\pm$ 1.35	83.50 $\pm$ 1.68	84.60 $\pm$ 1.57
Paired <i>t</i> -test <i>p</i> -value	0.0156 *		0.9995		0.0169 *	

Note: Bold values indicate the best results.

Table 4. Comparison of HIPTrack and TopoRelTrack on GOT-10K over ten independent training runs using random seeds 2017–2026. AO, SR_{0.5}, and SR_{0.75} are reported in %. The *p*-values are obtained using two-sided paired *t*-tests: **p* < 0.05.

Seed	AO (%)		SR _{0.5} (%)		SR _{0.75} (%)	
	HIPTrack	TopoRelTrack	HIPTrack	TopoRelTrack	HIPTrack	TopoRelTrack
2017	77.892	78.270	90.141	90.588	74.384	75.595
2018	77.314	78.147	88.384	90.341	73.856	76.211
2019	79.045	78.236	85.759	87.369	75.438	73.761
2020	77.845	78.522	87.460	87.163	74.341	76.209
2021	76.411	76.216	87.194	90.392	73.030	73.430
2022	78.420	78.755	87.389	88.594	74.867	75.834
2023	80.531	81.160	88.826	88.170	76.797	77.794
2024	79.732	80.473	85.556	86.282	76.066	77.534
2025	76.034	77.050	89.471	88.997	72.685	75.802
2026	74.776	76.172	85.820	87.105	71.535	75.830
Mean ± SD	77.80 ± 1.75	78.30 ± 1.62	87.60 ± 1.60	88.50 ± 1.55	74.30 ± 1.60	75.80 ± 1.38
Paired <i>t</i> -test <i>p</i> -value		0.0324 *		0.0428 *		0.0161 *

Note: Bold values indicate the best results.

On GOT-10K, TopoRelTrack improves the mean AO, SR_{0.5}, and SR_{0.75} over HIPTrack by 0.50, 0.90, and 1.50 percentage points, respectively. The corresponding paired *t*-test results are *p* = 0.0324, *p* = 0.0428, and *p* = 0.0161, indicating that all three improvements are statistically significant at the 0.05 level. TopoRelTrack also produces lower standard deviations for all three metrics. These repeated-run results indicate that the performance gains on GOT-10K are consistent across different training seeds rather than being determined by a single favorable initialization.

4.6. Parameter Sensitivity Analysis

4.6.1. Sensitivity to Camera–Target Distance

To evaluate the influence of camera–target distance on displacement reconstruction, three error metrics are adopted: root mean square error (RMSE), mean absolute error (MAE), and endpoint error (EPE). RMSE measures the overall deviation between the reconstructed displacement trajectory and the reference measurement, MAE represents the average absolute displacement error over the entire sequence, and EPE quantifies the displacement error at the final frame.

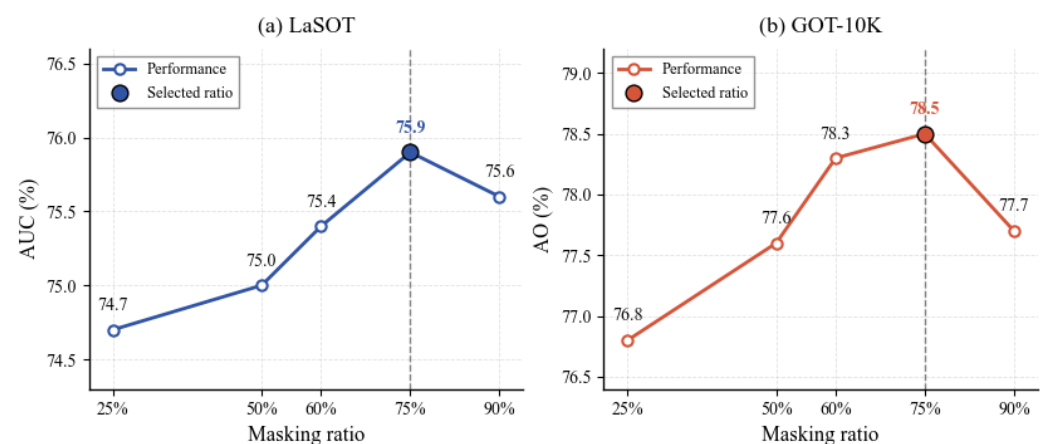
As shown in Table 5, the displacement-reconstruction error increases gradually as the camera–target distance increases from 0.5 to 2.0 m. Specifically, the RMSE increases from 0.326 to 0.411 mm, the MAE from 0.267 to 0.356 mm, and the EPE from 0.465 to 0.548 mm. This trend is reasonable because a larger observation distance reduces both the apparent target size and the image-plane displacement corresponding to the same physical motion, thereby increasing the influence of pixel-level localization errors after geometric calibration. Although the shortest distance provides slightly lower reconstruction errors, an excessively close camera position reduces the effective field of view and increases the risk that the target approaches the image boundary during large-amplitude motion. Therefore, a camera–target distance of 1.0 m is adopted as the reference configuration, providing a practical balance between target image resolution, displacement-measurement accuracy, and sufficient field-of-view coverage.

Table 5. Sensitivity of displacement reconstruction and seismic-intensity estimation to camera–target distance. All other acquisition, calibration, and filtering settings are kept unchanged.

Camera–Target Distance (m)	Displacement RMSE (mm)	Displacement MAE (mm)	EPE (mm)
0.5	0.326	0.267	0.465
1.0	0.342	0.279	0.483
1.5	0.371	0.314	0.519
2.0	0.411	0.356	0.548

4.6.2. Sensitivity to the Masking Ratio

To evaluate the sensitivity of the proposed method to the masking ratio used during TMAE self-supervised pre-training, experiments were conducted with masking ratios of 0.25, 0.50, 0.60, 0.75, and 0.90. The corresponding tracking performance on LaSOT and GOT-10K is presented in Figure 6.

**Figure 6.** Sensitivity of tracking performance to the masking ratio used during TMAE self-supervised pre-training: (a) AUC on LaSOT and (b) AO on GOT-10K. The vertical dashed lines and filled markers indicate the selected masking ratio of 0.75.

As the masking ratio increases from 0.25 to 0.75, the performance improves consistently on both benchmarks. Specifically, the LaSOT AUC increases from 74.7% to 75.9%, while the GOT-10K AO increases from 76.8% to 78.5%, corresponding to improvements of 1.2 and 1.7 percentage points, respectively. This trend indicates that an appropriately increased masking ratio strengthens the reconstruction task and encourages the encoder to learn more informative visual representations from the visible template and search-region patches.

When the masking ratio is further increased to 0.90, the LaSOT AUC decreases to 75.6%, and the GOT-10K AO decreases more noticeably to 77.7%. An excessively high masking ratio may remove too much spatial and appearance information from the search region, thereby increasing reconstruction difficulty and weakening the quality of the learned tracking representations. In contrast, a relatively low masking ratio provides an easier reconstruction task but may offer insufficient self-supervised learning difficulty. Therefore, a masking ratio of 0.75 provides the best balance between the amount of visible context and the difficulty of masked reconstruction and is adopted in the subsequent experiments.

4.6.3. Sensitivity to the Hypergraph Neighborhood Size

To evaluate the influence of the hypergraph neighborhood size on tracking performance, experiments were conducted with $k \in \{4, 6, 8, 10, 12\}$, while keeping all other components and training settings unchanged. As shown in Figure 7, both benchmarks exhibit an initial improvement followed by a decline as k increases. Increasing k from 4 to 8

improves the LaSOT AUC from 74.3% to 75.9% and the GOT-10K AO from 77.2% to 78.5%, corresponding to gains of 1.6 and 1.3 percentage points, respectively. Further increasing k to 12 reduces these metrics to 75.0% and 77.7%.

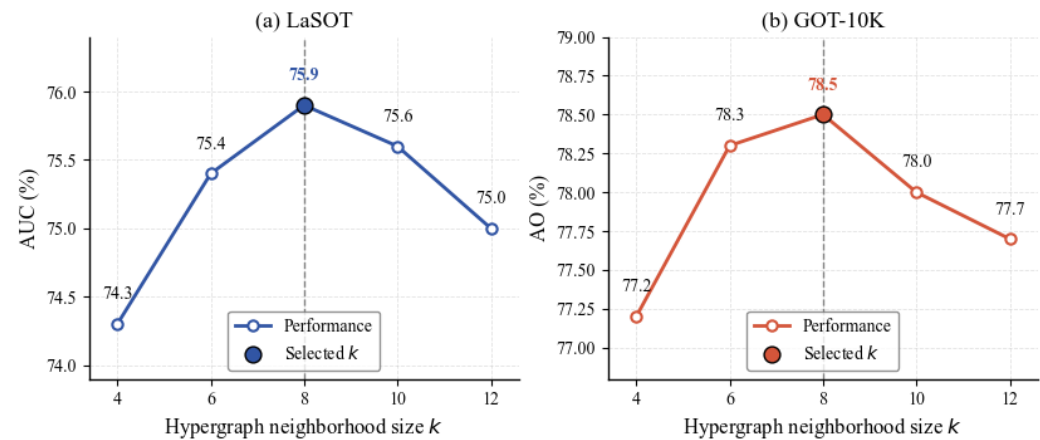


Figure 7. Sensitivity of tracking performance to the hypergraph neighborhood size k : (a) AUC on LaSOT and (b) AO on GOT-10K. The vertical dashed lines and filled markers indicate the default setting $k = 8$. All other components and training settings remain unchanged.

These results suggest that a small neighborhood may limit the aggregation of useful related features, whereas an excessively large neighborhood may introduce redundant or less relevant feature associations. Among the evaluated settings, $k = 8$ achieves the highest performance on both benchmarks and is therefore adopted as the default hypergraph neighborhood size.

4.6.4. Sensitivity to the Filter Cutoff Frequency

To evaluate the influence of the filter cutoff frequency on displacement trajectory reconstruction, four settings ranging from 10 to 25 Hz were compared while keeping all other processing settings unchanged. As shown in Table 6, all three errors decrease as the cutoff frequency increases from 10 to 20 Hz. At 20 Hz, the displacement RMSE, MAE, and EPE reach their lowest values of 0.349, 0.294, and 0.496 mm, respectively. Compared with the 10 Hz setting, these errors decrease by 0.063, 0.057, and 0.082 mm, respectively. Increasing the cutoff frequency further to 25 Hz raises the corresponding errors to 0.366, 0.307, and 0.518 mm.

Table 6. Effect of the filter cutoff frequency on displacement trajectory reconstruction.

Cutoff Frequency (Hz)	Displacement RMSE (mm)	Displacement MAE (mm)	EPE (mm)
10	0.412	0.351	0.578
15	0.375	0.318	0.539
20	0.349	0.294	0.496
25	0.366	0.307	0.518

This trend suggests that a lower cutoff frequency may attenuate useful motion components, whereas a higher cutoff frequency may retain more localization noise. Among the evaluated settings, 20 Hz provides the lowest displacement-reconstruction errors and is therefore adopted for the subsequent experiments.

4.7. Displacement Measurement Performance

To evaluate the accuracy of the reconstructed displacement trajectories, RMSE, MAE, and EPE are used as the evaluation metrics. To reduce the dependence of the results on a sin-

gle data partition, ten repeated record-level evaluations were conducted using consecutive random seeds ranging from 2017 to 2026. For each random seed, the self-built seismic-video dataset was repartitioned at the complete-excitation-record level, with all frames from the same excitation record assigned to the same subset. All compared trackers were evaluated using the same five data partitions, calibration parameters, and filtering protocol. Table 7 reports the mean and standard deviation over the five repeated evaluations.

Table 7. Trajectory-reconstruction performance on the seismic-video test set. All methods use the same calibration and filtering protocol.

Tracker	Displacement RMSE (mm)	Displacement MAE (mm)	EPE (mm)
OSTrack-256 [18]	0.539 ± 0.047	0.443 ± 0.064	0.781 ± 0.043
ODTrack-B256 [33]	0.446 ± 0.039	0.358 ± 0.052	0.582 ± 0.044
LCCTV [16]	0.412 ± 0.034	0.310 ± 0.041	0.547 ± 0.038
TopoRelTrack (ours)	0.349 ± 0.028	0.294 ± 0.035	0.496 ± 0.032

Note: Bold values indicate the best results.

As shown in Table 7, TopoRelTrack achieves the lowest mean trajectory-reconstruction errors across all three evaluation metrics. Specifically, the proposed method obtains a displacement RMSE of 0.349 ± 0.028 mm, a displacement MAE of 0.294 ± 0.035 mm, and an EPE of 0.496 ± 0.032 mm. Compared with LCCTV, which achieves the best mean performance among the competing methods, TopoRelTrack reduces the RMSE, MAE, and EPE by 0.063 mm, 0.016 mm, and 0.051 mm, respectively. Compared with OSTrack-256, the corresponding reductions are 0.19 mm, 0.149 mm, and 0.285 mm, respectively. TopoRelTrack also exhibits the lowest standard deviations for all three metrics, indicating that its trajectory-reconstruction performance is less sensitive to variations in the record-level data partitions.

To further assess whether the improvements observed in the mean trajectory-reconstruction errors are statistically significant, LCCTV, which achieves the lowest mean errors among the competing methods, was selected as the strongest baseline for paired comparison. A two-sided exact paired permutation test was conducted separately for each evaluation metric using the matched results obtained by LCCTV and TopoRelTrack under identical record-level data partitions for the ten random seeds, as reported in Table 8.

The resulting p -values are 0.0078, 0.0098, and 0.0117 for RMSE, MAE, and EPE, respectively. The differences in all three metrics are statistically significant at the 0.05 level, while the improvements in RMSE and MAE are also significant at the 0.01 level.

Although TopoRelTrack does not outperform LCCTV in every individual data partition, it achieves lower errors in the majority of the repeated evaluations and produces lower mean errors for all three metrics. These results indicate that the observed improvements are not restricted to a particular random data partition and provide statistical evidence for the improved trajectory-reconstruction accuracy of TopoRelTrack under the evaluated controlled shaking-table conditions.

Figures 8–10 further compare the reconstructed X-direction displacement histories under three representative seismic-motion patterns. The reference displacement is provided by the laser displacement sensor, while the results of LCCTV, ODTrack-B256, and TopoRelTrack are obtained from the corresponding visual tracking trajectories.

For the low-intensity micro-motion case, the vibration amplitude is relatively small and is therefore more sensitive to frame-wise localization fluctuations. As shown in Figure 8, both LCCTV and ODTrack-B256 exhibit noticeable high-frequency jitter around the reference trajectory, whereas TopoRelTrack follows the reference waveform more closely

and preserves the main oscillation trend with less fluctuation. This indicates that the proposed method is more effective in recovering weak structural motion.

Table 8. Trajectory-reconstruction results of LCCTV and TopoRelTrack over ten repeated record-level evaluations. RMSE, MAE, and EPE are expressed in mm. The p -values are obtained using two-sided paired t -tests: * $p < 0.05$ and ** $p < 0.01$.

Seed	Displacement RMSE (mm)		Displacement MAE (mm)		EPE (mm)	
	LCCTV	TopoRelTrack	LCCTV	TopoRelTrack	LCCTV	TopoRelTrack
2017	0.373	0.354	0.330	0.300	0.541	0.501
2018	0.395	0.307	0.263	0.242	0.528	0.448
2019	0.429	0.381	0.357	0.334	0.603	0.533
2020	0.440	0.326	0.290	0.265	0.591	0.470
2021	0.384	0.372	0.344	0.323	0.553	0.522
2022	0.418	0.317	0.249	0.254	0.578	0.459
2023	0.361	0.391	0.371	0.346	0.516	0.544
2024	0.406	0.344	0.303	0.288	0.491	0.491
2025	0.463	0.363	0.317	0.311	0.566	0.512
2026	0.451	0.335	0.276	0.277	0.503	0.480
Mean $\pm$ SD	0.412 $\pm$ 0.034	0.349 $\pm$ 0.028	0.310 $\pm$ 0.041	0.294 $\pm$ 0.035	0.547 $\pm$ 0.038	0.496 $\pm$ 0.032
Paired t -test		0.0078 **		0.0098 **		0.0117 *
p -value						

Note: Bold values indicate the best results.

For the pulse-type event, the comparison mainly focuses on the preservation of rapid displacement changes and peak responses. As shown in Figure 9, TopoRelTrack maintains better agreement with the reference displacement around the dominant positive and negative peaks. In contrast, LCCTV and ODTrack-B256 show larger deviations near several rapid transitions and peak regions. The result demonstrates that the proposed method provides improved peak preservation and temporal alignment under strong transient motion.

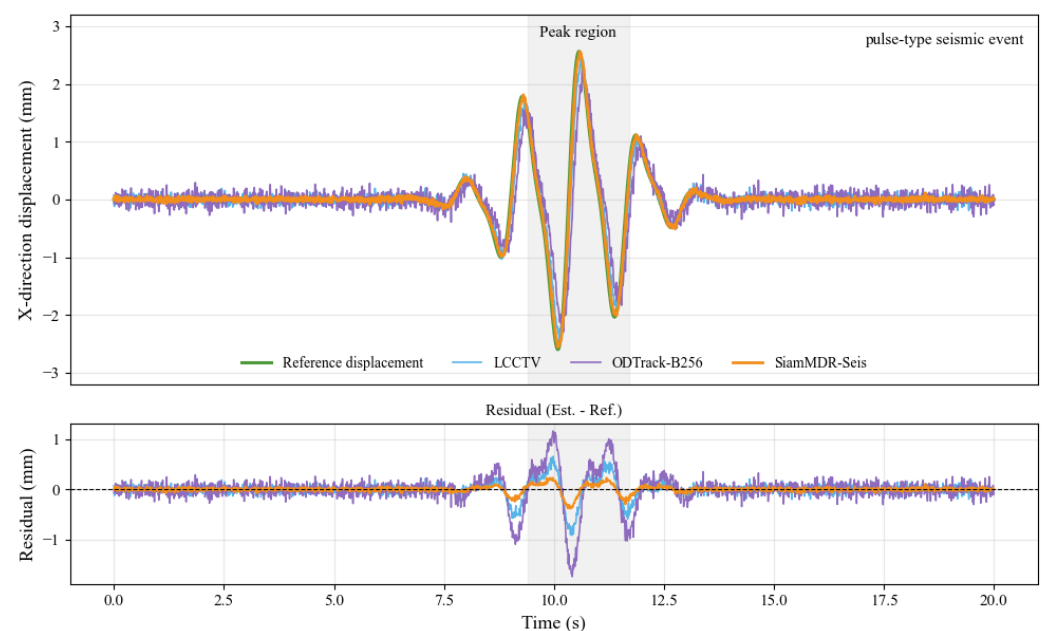


Figure 8. X-direction displacement-time histories and corresponding residuals for a representative pulse-type seismic event. The upper panel compares the displacement trajectories reconstructed by different tracking methods with the reference displacement, while the lower panel presents the corresponding residuals, defined as the estimated displacement minus the reference displacement. The shaded region indicates the peak-motion interval.

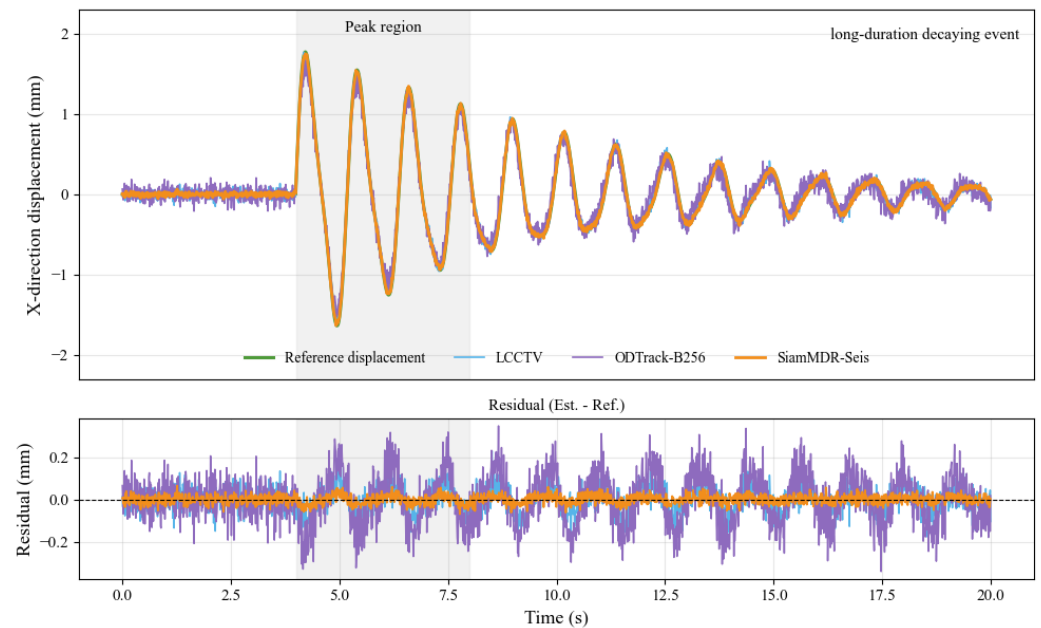


Figure 9. X-direction displacement reconstruction for a representative pulse-type seismic event. The upper panel compares the reconstructed and reference displacement–time histories, while the lower panel shows the corresponding residuals calculated as the estimated displacement minus the reference displacement. The shaded region indicates the peak-motion interval.

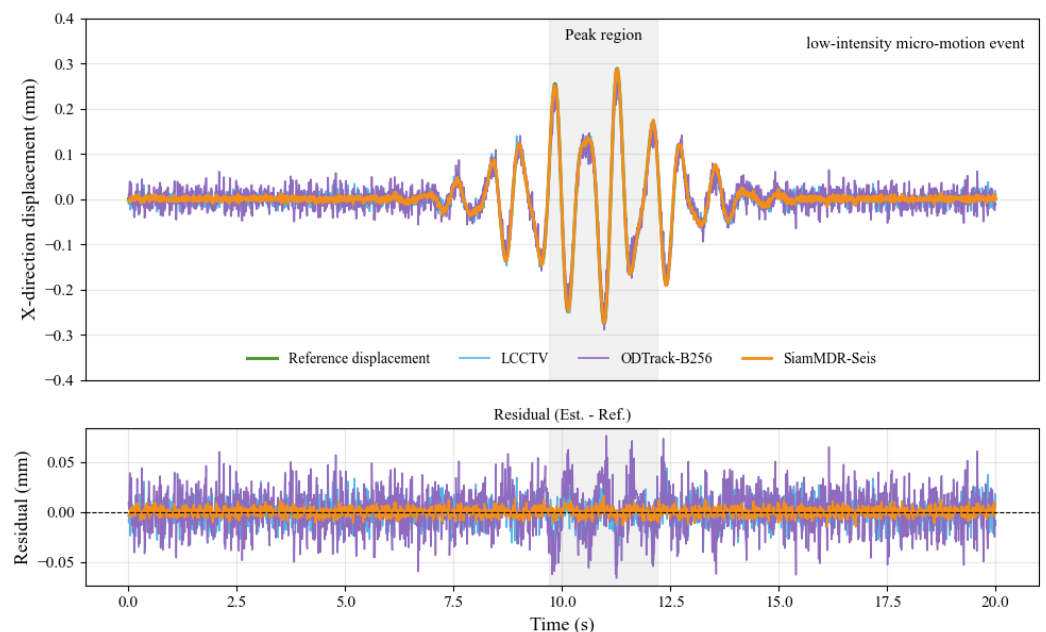


Figure 10. X-direction displacement reconstruction for a representative long-duration decaying seismic event. The upper panel compares the reconstructed and reference displacement–time histories, while the lower panel shows the corresponding residuals calculated as the estimated displacement minus the reference displacement. The shaded region indicates the selected high-amplitude motion interval.

For the long-duration decaying event, the tracking method is required to maintain stable localization over an extended oscillatory interval. As shown in Figure 10, TopoRelTrack remains closely aligned with the reference displacement throughout the decay process, while LCCTV and ODTrack-B256 exhibit larger local fluctuations and gradually accumulated deviations in the later stage. The improved waveform consistency indicates that

TopoRelTrack can maintain more stable long-term tracking and better preserve the phase and amplitude characteristics of decaying structural vibration.

Overall, the three representative cases demonstrate that TopoRelTrack achieves more accurate displacement reconstruction under weak motion, transient pulse excitation, and long-duration decaying vibration. The closer agreement with the laser reference in waveform shape, peak response, and temporal consistency further confirms the reliability of the proposed method for structural seismic-response measurement.

4.8. Seismic Intensity Estimation

As shown in Table 9, the seismic-intensity prediction experiment conducted on the Shaking-Table Seismic-Video Dataset shows that TopoRelTrack achieves the best performance among all compared trackers. Specifically, the proposed method obtains an accuracy of 86.7%, a macro precision of 85.0%, a macro recall of 84.7%, and a macro F1-score of 84.5%, while achieving the lowest intensity MAE of 0.402.

Table 9. Seismic-intensity prediction on the held-out seismic-video test set. All trackers use the same coordinate transformation, filter, and intensity mapping.

Tracker	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)	Intensity MAE
OSTrack-256 [18]	73.3	75.8	71.5	72.5	0.574
ODTrack-B256 [33]	76.7	80.3	74.0	75.6	0.485
LCCTV [16]	83.3	82.5	82.2	82.1	0.421
TopoRelTrack (ours)	86.7	85.0	84.7	84.5	0.402

Note: Bold values indicate the best results.

Figure 11 further presents the row-normalized confusion matrices of the compared trackers. Each row corresponds to a reference seismic-intensity class, while each column denotes the predicted class. The values in each cell indicate the number of samples and the corresponding row-normalized percentage. Compared with the baseline trackers, TopoRelTrack exhibits a more concentrated diagonal distribution, indicating more consistent predictions across different intensity levels. In particular, the proposed method correctly classifies all samples in intensity classes II and VI, while maintaining relatively high recognition accuracy for classes III and V. The remaining errors mainly occur between adjacent intensity levels, suggesting that the prediction deviations are generally limited in magnitude rather than involving severe cross-class confusion.

Compared with LCCTV, which provides the strongest competing results, TopoRelTrack improves the accuracy, macro precision, macro recall, and macro F1-score by 3.4, 2.5, 2.5, and 2.4 percentage points, respectively, while reducing the intensity MAE by 0.019. Compared with ODTrack-B256, the corresponding improvements reach 10.0, 4.7, 10.7, and 8.9 percentage points, together with a reduction of 0.172 in intensity MAE. These consistent improvements across both classification and error metrics demonstrate the effectiveness of TopoRelTrack for seismic-intensity estimation.

The results further show that the more accurate and stable trajectories produced by TopoRelTrack on the Shaking-Table Seismic-Video dataset can effectively improve downstream seismic-response analysis. Since all trackers employ the same coordinate transformation, filtering, numerical differentiation, and intensity-mapping procedure, the differences in intensity-prediction performance mainly originate from the quality of the upstream visual tracking and displacement reconstruction. This confirms that improved target localization accuracy and trajectory stability contribute directly to more reliable seismic-intensity estimation.

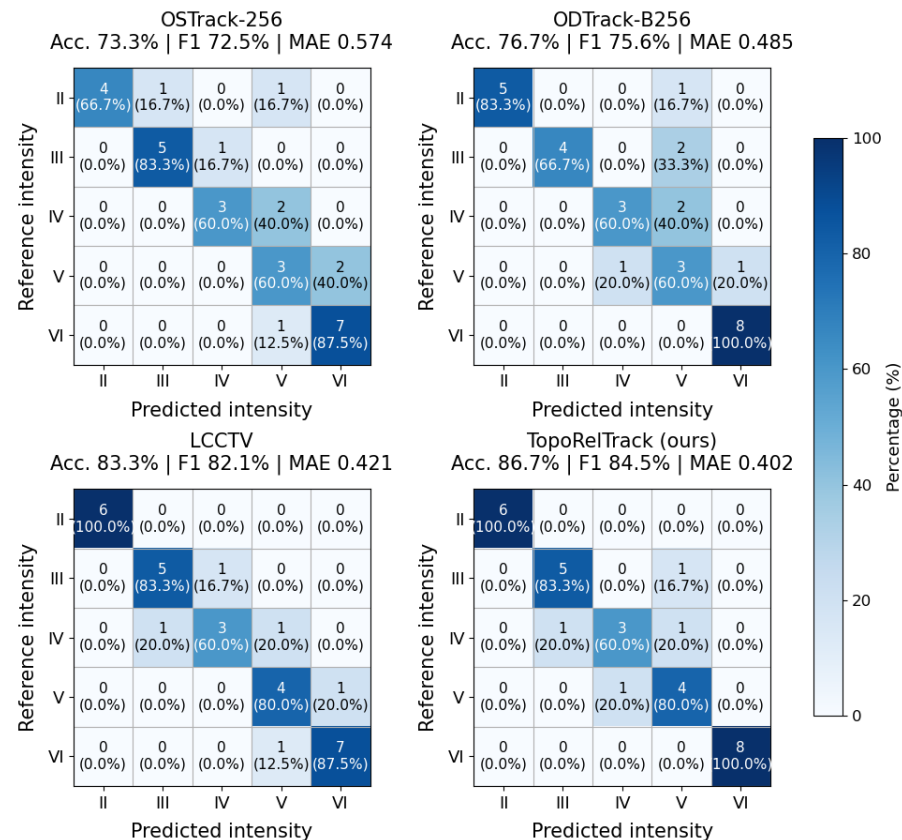


Figure 11. Row-normalized confusion matrices of different trackers for seismic-intensity prediction on the Shaking-Table Seismic-Video Dataset. Acc., F1, and MAE denote accuracy, F1-score, and mean absolute error, respectively. Each cell reports the number of samples and the corresponding percentage within the reference-intensity class.

4.9. Ablation Study

As shown in Table 10, each component of TopoRelTrack contributes positively to the tracking performance. Starting from the baseline model, introducing TPSMamba improves the LaSOT AUC and precision from 68.7% and 72.1% to 70.4% and 74.5%, respectively, while the GOT-10K AO increases from 70.3% to 73.6%. This demonstrates the effectiveness of the multi-directional scanning strategy in enhancing feature representation.

Table 10. Component ablation study of TopoRelTrack on LaSOT and GOT-10K.

Config.	TPSMamba	SDRM	TMAE	LaSOT AUC (%)	LaSOT Prec (%)	GOT-10K AO (%)
A (baseline)	×	×	×	68.7	72.1	70.3
B	✓	×	×	70.4	74.5	73.6
C	✓	✓	×	72.3	77.8	75.2
D (TopoRelTrack)	✓	✓	✓	75.9	81.8	78.5

Note: ✓ indicates that the corresponding component is included, whereas × indicates that it is excluded.

After further introducing SDRM, the LaSOT AUC and precision increase to 72.3% and 77.8%, and the GOT-10K AO reaches 75.2%. The consistent improvements indicate that adaptive relation modeling between the template and search regions can further enhance target discrimination and suppress irrelevant background responses.

Finally, incorporating TMAE pre-training yields the best performance, achieving 75.9% AUC and 81.8% precision on LaSOT and 78.5% AO on GOT-10K. Compared with the configuration without TMAE, the three metrics are improved by 3.6, 4.0, and 3.3 percentage points, respectively. These results demonstrate that the task-oriented self-supervised pre-

training strategy provides more effective feature initialization and further improves the tracking accuracy and generalization capability of the model.

To further investigate the contribution of the scanning paths within TPSMamba, Table 11 compares the four individual directions and their combinations. Among the single-direction configurations, the upward vertical scan achieves the highest LaSOT AUC and precision of 73.9% and 80.2%, respectively, whereas the forward zigzag scan obtains the highest GOT-10K AO of 76.5%. The relatively small differences among Configurations A–D indicate that each direction can independently capture useful spatial dependencies. Combining the two vertical directions increases the three metrics to 74.5%, 80.7%, and 77.1%, while bidirectional zigzag scanning further improves them to 74.7%, 80.9%, and 77.6%, respectively. The full four-direction configuration achieves the best results, outperforming the stronger bidirectional zigzag configuration by 1.2, 0.9, and 0.9 percentage points in LaSOT AUC, LaSOT precision, and GOT-10K AO, respectively. These results indicate that the vertical and zigzag scanning strategies capture complementary spatial dependencies, and their combination improves spatial continuity preservation and long-range contextual modeling in TPSMamba.

Table 11. Ablation study of the four scanning directions in TPSMamba. All other components and training settings remain unchanged.

Config.	Vertical ↓	Vertical ↑	Zigzag-F	Zigzag-R	LaSOT AUC (%)	LaSOT Prec (%)	GOT-10K AO (%)
A	✓	×	×	×	73.4	79.5	76.2
B	×	✓	×	×	73.9	80.2	76.4
C	×	×	✓	×	73.8	80.0	76.5
D	×	×	×	✓	73.7	79.9	76.3
E (Vertical bidirectional)	✓	✓	×	×	74.5	80.7	77.1
F (Zigzag bidirectional)	×	×	✓	✓	74.7	80.9	77.6
G (Full TPSMamba)	✓	✓	✓	✓	75.9	81.8	78.5

Note: ↓ and ↑ denote downward and upward vertical scanning, respectively. ✓ indicates that the corresponding scanning direction is included, whereas × indicates that it is excluded. Bold values indicate the best results.

4.10. Computational Cost and Throughput

As shown in Table 12, TopoRelTrack introduces a higher computational cost than the compared methods, with 103.4 M parameters and 98.7 GFLOPs. Its inference speed on a single NVIDIA A100 GPU is 58 FPS, which is lower than that of the competing trackers but indicates the potential for real-time tracking of a single video stream acquired at up to 58 FPS. The reported FPS measures only the inference stage of the tracking model. Its peak GPU memory usage is 37.7 GB, slightly higher than the 37.1 GB required by ARTrack-256, indicating a substantial memory requirement for deployment. Despite the increased computational complexity, the proposed method achieves improvements in tracking accuracy, displacement reconstruction, and seismic-intensity estimation. Therefore, TopoRelTrack provides a trade-off between computational cost and measurement accuracy for video-based seismic-motion measurement under controlled shaking-table conditions.

Table 12. Efficiency comparison. Speed is measured on a single NVIDIA GPU.

Method	Params (M)	GFLOPs	Speed (FPS)	GPU Memory (GB)
OSTrack-256 [18]	92.1	49.5	105	32.4
ODTrack-B256 [33]	92.1	53.0	92	33.9
ARTrack-256 [35]	92.8	57.3	62	37.1
LCCTV [16]	97.9	78.5	63	38.6
TopoRelTrack (ours)	103.4	98.7	58	37.7

Note: Bold values indicate the best results.

5. Discussion

The experimental results demonstrate that TopoRelTrack provides reliable target tracking and displacement reconstruction for video-based seismic-motion measurement under controlled shaking-table conditions. The combination of TPSPMamba, SDRM, and task-oriented self-supervised pre-training improves the modeling of spatial dependencies and template-search correspondences, enabling the reconstructed displacement histories to maintain better agreement with the laser reference under different vibration patterns. From a practical perspective, the reduction in frame-wise trajectory fluctuations improves the preservation of waveform shape, motion peaks, and temporal consistency. This is particularly important because localization noise can be amplified during numerical differentiation and subsequently affect the estimation of velocity, acceleration, PGA, PGV, and instrumental intensity. Therefore, the benefits of TopoRelTrack extend beyond improvements in tracking metrics and contribute to the stability of the overall video-based seismic-motion measurement pipeline.

Despite these advantages, several limitations remain. First, the introduction of multi-directional feature modeling and dynamic relation matching increases the computational complexity of the network. TopoRelTrack achieves a tracking inference speed of 58 FPS on a single NVIDIA A100 GPU, indicating its potential for real-time tracking of a single video stream acquired at up to 58 FPS. However, its parameter count, computational cost, and GPU memory consumption are higher than those of the compared trackers. The current implementation is therefore more suitable for GPU-equipped workstations or edge servers than for resource-constrained embedded devices. Its deployment in field-monitoring systems may also be constrained by available memory, power consumption, and computing capacity. The reported speed corresponds only to the inference stage of the tracking model.

Second, the current shaking-table experiments were conducted under a fixed-camera condition. In real earthquake scenarios, the camera itself may undergo translational or rotational motion, introducing additional apparent motion into the observed target trajectory and thereby affecting the accuracy of physical displacement reconstruction.

Third, the laser displacement sensors provide direct validation only for the reconstructed displacement trajectories. Velocity, acceleration, PGA, and PGV are derived from the filtered displacement signals through numerical differentiation and have not been independently validated using dedicated velocity sensors or accelerometers. Instrumental intensity is subsequently estimated from the derived PGA and PGV according to the adopted intensity-mapping criterion. Consequently, errors introduced by spatial calibration, filtering, temporal alignment, and numerical differentiation may propagate into these derived quantities.

Future work will focus on developing a more lightweight TPSPMamba architecture and reducing the computational overhead of dynamic relation matching while preserving tracking accuracy. Model compression, knowledge distillation, and hardware-aware opti-

mization will also be investigated to reduce memory usage and power consumption for embedded deployment. In addition, camera-motion compensation will be investigated by exploiting stationary background features or fixed reference targets to estimate and remove global camera motion before target-trajectory reconstruction. Synchronized accelerometers will also be incorporated to independently validate the reconstructed acceleration, PGA, PGV, and instrumental-intensity estimates.

6. Conclusions

This study proposed TopoRelTrack, a video-based tracking and displacement reconstruction framework for video-based seismic-motion measurement. The method adopts a two-stage learning strategy consisting of self-supervised pre-training and downstream tracking. A template-guided masked autoencoder is first employed to learn tracking-oriented representations from unlabeled video sequences. Subsequently, the topology-preserving scanning Mamba encoder is used to capture long-range spatial dependencies efficiently, while the sparse dynamic relation matcher establishes adaptive and fine-grained correspondences between the template and search regions to improve target localization robustness.

Experiments on the LaSOT and GOT-10K benchmarks demonstrate that TopoRelTrack achieves competitive tracking performance and strong generalization capability. In the controlled shaking-table seismic-video experiments, the proposed method also provides more accurate displacement reconstruction and improved seismic-intensity prediction compared with the competing trackers. The ablation results further verify the effectiveness and complementarity of TMAE, TPSMamba, and SDRM. Although the proposed method introduces additional computational cost, it achieves a tracking inference speed of 58 FPS on a single NVIDIA A100 GPU, indicating its potential for real-time tracking of a single video stream acquired at up to 58 FPS. Overall, TopoRelTrack provides a favorable balance between tracking accuracy and computational efficiency for video-based seismic-motion measurement under controlled shaking-table conditions.

Future work will focus on further reducing the computational complexity of the model and improving its robustness under extreme visual conditions, such as severe target occlusion and substantial image-quality degradation, while extending validation to more diverse seismic scenarios.

Author Contributions: H.W.: Writing–review & editing, Writing–original draft, Visualization, Investigation, Validation, Supervision, Software, Methodology, Funding acquisition. J.S.: Writing–review & editing, Project administration, Formal analysis, Resources. J.L.: Formal analysis, Conceptualization, Validation. S.N.: Investigation, Conceptualization. B.F.: Project administration, Conceptualization, Supervision. B.L.: Project administration, Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project “Demonstration Application of an Earthquake Early-Warning System Based on Ubiquitous Vibration Sensing”, grant number 2024YFC3012805.

Acknowledgments: The authors gratefully acknowledge Jie Su for assistance with manuscript polishing. During the preparation of this manuscript, the authors used ChatGPT-5 for language editing and color rendering of module-architecture figures. All AI-assisted outputs were reviewed by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Authors Sirun Niu and Bo Fan were employed by Shaanxi AI Land Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

- Allen, R.M.; Melgar, D. Earthquake Early Warning: Advances, Scientific Challenges, and Societal Needs. *Annu. Rev. Earth Planet. Sci.* **2019**, *47*, 361–388. <https://doi.org/10.1146/annurev-earth-053018-060457>.
- GB/T 17742–2020; The Chinese Seismic Intensity Scale. State Administration for Market Regulation and Standardization Administration of the People's Republic of China: Beijing, China, 2020.
- Douglas, J. Earthquake Ground Motion Estimation Using Strong-Motion Records: A Review of Equations for the Estimation of Peak Ground Acceleration and Response Spectral Ordinates. *Earth-Sci. Rev.* **2003**, *61*, 43–104. [https://doi.org/10.1016/S0012-8252\(02\)00112-5](https://doi.org/10.1016/S0012-8252(02)00112-5).
- Abdalzaher, M.S.; Soliman, M.S.; El-Hady, S.M. Seismic Intensity Estimation for Earthquake Early Warning Using Optimized Machine Learning Model. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5914211. <https://doi.org/10.1109/TGRS.2023.3296520>.
- Jozinović, D.; Lomax, A.; Štajduhar, I.; Michelini, A. Rapid Prediction of Earthquake Ground Shaking Intensity Using Raw Waveform Data and a Convolutional Neural Network. *Geophys. J. Int.* **2020**, *222*, 1379–1389. <https://doi.org/10.1093/gji/ggaa233>.
- Münchmeyer, J.; Bindi, D.; Leser, U.; Tilmann, F. The Transformer Earthquake Alerting Model: A New Versatile Approach to Earthquake Early Warning. *Geophys. J. Int.* **2021**, *225*, 646–656. <https://doi.org/10.1093/gji/ggaa609>.
- Nolte, K.A.; Tsoflias, G.P.; Holubnyak, Y.; Raney, J.; Wreath, D. Designing monitoring networks for local earthquakes. *J. Geophys. Eng.* **2022**, *19*, 75–84. <https://doi.org/10.1093/jge/gxab075>.
- Ringler, A.T.; Anthony, R.E.; Aster, R.C.; Ammon, C.J.; Arrowsmith, S.; Benz, H.; Ebeling, C.; Frassetto, A.; Kim, W.Y.; Koelemeijer, P.; et al. Achievements and Prospects of Global Broadband Seismographic Networks After 30 Years of Continuous Geophysical Observations. *Rev. Geophys.* **2022**, *60*, e2021RG000749. <https://doi.org/10.1029/2021RG000749>.
- Dong, C.Z.; Catbas, F.N. A Review of Computer Vision-Based Structural Health Monitoring at Local and Global Levels. *Struct. Health Monit.* **2021**, *20*, 692–743. <https://doi.org/10.1177/1475921720935585>.
- Zhuang, Y.; Chen, W.; Jin, T.; Chen, B.; Zhang, H.; Zhang, W. A Review of Computer Vision-Based Structural Deformation Monitoring in Field Environments. *Sensors* **2022**, *22*, 3789. <https://doi.org/10.3390/s22103789>.
- Weng, Y.; Quek, S.T.; Yeoh, J.K.W. Robust Vision-Based Sub-Pixel Level Displacement Measurement Using a Complementary Strategy. *Mech. Syst. Signal Process.* **2025**, *223*, 111898. <https://doi.org/10.1016/j.ymssp.2024.111898>.
- Wen, H.; Dong, R.; Dong, P. Structural Displacement Measurement Using Deep Optical Flow and Uncertainty Analysis. *Opt. Lasers Eng.* **2024**, *181*, 108364. <https://doi.org/10.1016/j.optlaseng.2024.108364>.
- Won, J.; Park, J.W.; Song, M.H.; Kim, Y.S.; Moon, D. Robust Vision-Based Displacement Measurement and Acceleration Estimation Using RANSAC and Kalman Filter. *Earthq. Eng. Eng. Vib.* **2023**, *22*, 347–358. <https://doi.org/10.1007/s11803-023-2173-0>.
- Bai, Y.; Abdullah, R.M.; Sezen, H.; Yilmaz, A. Automatic Displacement and Vibration Measurement in Laboratory Experiments with a Deep Learning Method. In Proceedings of the 2021 IEEE Sensors, Sydney, Australia, 31 October–3 November 2021; pp. 1–4. Available online: <http://arxiv.org/abs/2109.04960> (accessed on 22 October 2026).
- Pan, X.; Yang, T.Y.; Xiao, Y.; Yao, H.; Adeli, H. Vision-Based Real-Time Structural Vibration Measurement Through Deep-Learning-Based Detection and Tracking Methods. *Eng. Struct.* **2023**, *281*, 115676. <https://doi.org/10.1016/j.engstruct.2023.115676>.
- Wang, H.; Bai, Z.; Bai, R.; Zhao, L.; Lin, M.; Han, Y. Seismic Intensity Prediction with a Low-Computational-Cost Transformer-Based Tracking Method. *Sensors* **2025**, *25*, 6269. <https://doi.org/10.3390/s25206269>.
- Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
- Ye, B.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. In *Proceedings of the Computer Vision—ECCV 2022; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13682*, pp. 341–357.
- Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
- Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
- Yang, D.; He, J.; Ma, Y.; Yu, Q.; Zhang, T. Foreground-Background Distribution Modeling Transformer for Visual Object Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 10117–10127. <https://doi.org/10.1109/ICCV51070.2023.00928>.
- Fu, Z.; Fu, Z.; Liu, Q.; Cai, W.; Wang, Y. SparseTT: Visual Tracking with Sparse Transformers. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; pp. 905–912. <https://doi.org/10.24963/ijcai.2022/127>.

23. Hondru, V.; Croitoru, F.A.; Minaee, S.; Ionescu, R.T.; Sebe, N. Masked Image Modeling: A Survey. *Int. J. Comput. Vis.* **2025**, *133*, 7154–7200. <https://doi.org/10.1007/s11263-025-02524-1>.
24. Nandam, S.R.; Atito, S.; Feng, Z.; Kittler, J.; Awais, M. Investigating Self-Supervised Methods for Label-Efficient Learning. *Int. J. Comput. Vis.* **2025**, *133*, 4522–4537. <https://doi.org/10.1007/s11263-025-02397-4>.
25. Qian, Y.; Wang, Y.; Zou, J.; Lin, J.; Pan, Y.; Yao, T.; Sun, Q.; Mei, T. Kernel Masked Image Modeling Through the Lens of Theoretical Understanding. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 13512–13526. <https://doi.org/10.1109/TNNLS.2024.3443088>.
26. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. VideoMamba: State Space Model for Efficient Video Understanding. In *Proceedings of the Computer Vision–ECCV 2024*; Springer: Cham, Switzerland, 2024.
27. Mousavi, S.M.; Ellsworth, W.L.; Zhu, W.; Chuang, L.Y.; Beroza, G.C. Earthquake Transformer: An Attentive Deep-Learning Model for Simultaneous Earthquake Detection and Phase Picking. *Nat. Commun.* **2020**, *11*, 3952. <https://doi.org/10.1038/s41467-020-17591-w>.
28. Choi, J.; Chung, J.; Kim, K.; Sohn, H. Seismic-Induced Permanent Displacement Estimation Combining Acceleration and Computer Vision Measurements. *Mech. Syst. Signal Process.* **2023**, *200*, 110504.
29. Zhang, S.; Ni, P.; Han, Q.; Wen, J.; Du, X.; Li, J. Vision-Based Identification Method of Earthquake Ground Motions. *Eng. Struct.* **2026**, *346*, 121642.
30. Yang, J.; Lu, G.; Wu, Y.; Peng, F. Method for Indoor Seismic Intensity Assessment Based on Image Processing Techniques. *J. Imaging* **2025**, *11*, 129. <https://doi.org/10.3390/jimaging11050129>.
31. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
32. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 9653–9663.
33. Zheng, Y.; Zhong, B.; Liang, Q.; Mo, Z.; Zhang, S.; Li, X. ODTrack: Online Dense Temporal Token Learning for Visual Tracking. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 7588–7596.
34. Cai, W.; Liu, Q.; Wang, Y. HIPTrack: Visual Tracking with Historical Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024; pp. 19258–19267.
35. Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; Gong, Y. Autoregressive Visual Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17–24 June 2023; pp. 9697–9706.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.