Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Lightweight Self-Evolving Framework for Semi-Supervised Video Tampering Localization

Qimeng Zhu
Soochow University
Ke Zhang
kzhang19@suda.edu.cn
Soochow University

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11610145/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Lightweight Self-Evolving Framework for Semi-Supervised Video Tampering Localization

Qimeng Zhu[1] and Ke Zhang[1*]

[1*]School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, China.

*Corresponding author(s). E-mail(s): kzhang19@suda.edu.cn;

Abstract

Temporal localization of video tampering is a fundamental task in multimedia forensics, yet existing approaches typically rely on dense frame-level annotations and computationally intensive architectures, limiting their scalability in real-world applications. In this paper, we propose a lightweight self-evolving framework for semi-supervised video tampering temporal localization, which learns precise temporal boundaries from only a small fraction of labeled data. Unlike prior methods that depend on heavy Transformer-based backbones, our approach adopts a compact temporal modeling strategy with linear complexity. Specifically, we introduce TeLoNet, a lightweight temporal representation operator that operates on pre-extracted features to efficiently generate coarse temporal hypotheses. On top of this backbone, we reformulate localization as a progressive self-evolution process: EvoAlign enforces consistency in temporal evolution dynamics to stabilize pseudo-labels, while EvoRefine performs boundary-aware residual refinement to sharpen temporal boundaries. These components jointly improve localization through iterative prediction refinement rather than model scaling. Experiments on LAV-DF and TVIL show that, using only 10%–20% labeled data, the proposed framework achieves performance comparable to several fully supervised baselines on strict metrics (e.g., AP@0.95), while reducing computational cost by more than an order of magnitude relative to Transformer-based models. These results suggest that stable temporal localization can be achieved through structured evolution in prediction space, offering a practical and efficient solution for real-world video forensics.

Keywords: video tampering localization, semi-supervised learning, self-evolving learning, temporal boundary localization, consistency regularization

1 Introduction

Video tampering temporal localization aims to identify manipulated segments within a video, which plays a critical role in multimedia forensics and content authenticity verification. Despite recent progress, existing methods predominantly rely on fully supervised learning with dense temporal annotations [14, 19, 38]. However, obtaining frame-level annotations for tampered segments is extremely labor-intensive and time-consuming, which severely limits scalability in real-world scenarios.

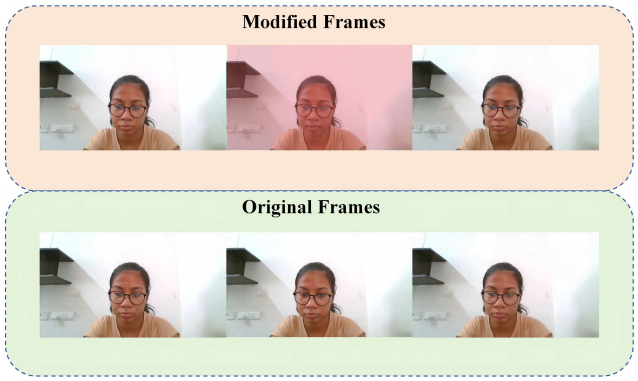

Fig. 1 An example of video face tampering. The video contains multiple frames, of which 24 frames have been tampered (highlighted with red boxes). The tampered face has been replaced with another person's facial features using deepfake technology, but it is difficult to distinguish with the naked eye.

From a security perspective, accurate temporal localization of tampered segments is critical for reliable multimedia forensics. Imprecise boundary estimation may lead to incorrect attribution of manipulated content, thereby undermining the credibility of forensic evidence in high-stakes scenarios such as legal investigations and misinformation control. Moreover, modern content moderation systems require not only detection but also precise localization to enable fine-grained intervention, highlighting the need for efficient and reliable temporal localization methods.

To alleviate the dependency on large-scale annotations, semi-supervised learning provides a promising alternative by leveraging abundant unlabeled data. Compared with weakly supervised approaches that rely on coarse video-level labels, semi-supervised learning enables direct optimization over temporal predictions, offering the potential for higher localization precision. This makes it particularly suitable for video tampering localization, where unlabeled videos are readily available while precise annotations remain scarce.

However, directly applying existing semi-supervised learning frameworks to temporal localization is non-trivial. A fundamental challenge lies in the instability of temporal predictions. Unlike image-level tasks, temporal localization requires precise boundary estimation, where small perturbations can lead to significant deviations in predicted segments. As a result, pseudo-labels generated on unlabeled data are often

highly noisy, especially around boundary regions. When such unreliable signals are used for training, errors accumulate over time, leading to degraded performance.

A common strategy to mitigate this issue is to adopt more powerful architectures, particularly Transformer-based models, to improve representation capacity and temporal modeling. While effective, these approaches suffer from two key limitations. First, their computational complexity grows quadratically with sequence length, making them inefficient for long video sequences. Second, increasing model capacity does not fundamentally address the instability of temporal predictions, but rather compensates for it in a passive manner.

In contrast, we argue that the core challenge of semi-supervised temporal localization is not insufficient model capacity, but the lack of stable prediction dynamics. Instead of relying on heavy architectures, we propose to explicitly model how predictions should evolve over time.

In this regard, our framework shares the high-level Mean Teacher spirit [25], but differs in a key aspect: rather than enforcing point-wise output agreement between the student and the teacher, we enforce consistency on the *temporal evolution trajectory* (the first-order difference field) of the prediction. This is motivated by the observation that, for boundary localization, the absolute prediction value is less informative than the shape of its temporal transition; two predictions with similar point-wise values can correspond to very different boundary structures. Constraining the trajectory geometry instead of point-wise outputs therefore makes the model less sensitive to noisy pseudo-labels near boundaries. Semi-supervised temporal action proposal (SSTAP) [23] is the closest prior in spirit, but it targets action localization and relies on proposal-level self-supervised tasks. Our setting differs fundamentally: tampering intervals are typically subtle, short, and weakly cued in appearance, which calls for a forensics-specific evolution constraint and a compact backbone that is deployable in practice. We provide a detailed comparison with these and other recent methods in Section 2.

Specifically, we introduce a lightweight self-evolving framework for semi-supervised video tampering temporal localization. Our method is built upon a compact temporal modeling backbone, termed **TeLoNet**, which operates on pre-extracted features and performs efficient temporal representation learning with linear complexity. Compared with Transformer-based architectures, TeLoNet avoids expensive self-attention operations and instead learns a structured temporal embedding space, significantly reducing computational overhead while preserving temporal ordering information.

On top of this lightweight backbone, we reformulate temporal localization as a progressive evolution process, where predictions are iteratively refined through structured constraints. This self-evolving process consists of three key stages, each addressing a specific limitation in semi-supervised temporal learning:

(1) Evolution Initialization (TeLoNet). TeLoNet generates an initial coarse temporal hypothesis from compact representations. This stage reduces feature redundancy and provides a stable starting point for subsequent refinement.

(2) Evolution Stabilization (EvoAlign). To address the instability of pseudo-labels, we introduce a temporal evolution alignment mechanism that enforces consistency in prediction dynamics rather than raw outputs. By aligning temporal gradients

between student and teacher models, EvoAlign stabilizes boundary transitions and suppresses noisy fluctuations.

(3) Evolution Refinement (EvoRefine). To further improve boundary precision, we design a boundary-aware refinement operator that performs local residual correction on ambiguous regions. This stage enhances boundary sharpness and mitigates fragmented or over-smoothed predictions.

Through this coarse-to-fine evolution process, the model progressively improves prediction quality, effectively suppressing pseudo-label noise and enhancing temporal consistency on unlabeled data.

An important advantage of this formulation is that it decouples temporal reasoning from heavy architectural design. By explicitly modeling prediction evolution, the framework achieves high-precision localization through iterative refinement rather than complex networks, resulting in a lightweight yet effective solution.

Importantly, the lightweight design of the proposed framework enables practical deployment in real-world systems, such as large-scale video screening and online content moderation, where both computational efficiency and localization accuracy are essential. Based on the above, the main contributions of this paper are as follows:

- We formulate video tampering temporal localization under a semi-supervised learning paradigm, addressing a setting that remains largely unexplored in the forensic localization literature. While semi-supervised temporal modeling has been studied for action proposal [23], we adapt and extend the idea to forensics, where tampering boundaries are far more subtle and fragile than action boundaries.
- We propose TeLoNet, a lightweight temporal representation operator that replaces heavy Transformer-based sequential architectures with a compact BiLSTM-based encoding strategy of linear complexity, operating directly on pre-extracted features.
- We introduce a temporal evolution consistency mechanism (EvoAlign) that stabilizes prediction dynamics by aligning first-order temporal gradients, rather than raw outputs, between student and teacher models.
- We propose a boundary-aware refinement operator (EvoRefine) that improves localization accuracy through residual correction in prediction space, and we clarify its role during inference.
- We report a supervised-only baseline and component ablations to isolate the contribution of the semi-supervised machinery and of each proposed component.

2 Related Work

Multimedia Forensics and Deepfake Detection. Multimedia forensics has recently received increasing attention due to the rapid development of realistic content manipulation techniques. Recent studies have summarized representative datasets, forensic clues, and deep learning architectures for multimedia tampering detection, showing that robust manipulation analysis requires both discriminative representation learning and reliable benchmark evaluation [35]. AI-based deepfake generation and detection techniques have also been systematically reviewed, emphasizing the challenges introduced by continuously evolving manipulation methods [1]. Different from

general deepfake detection studies that mainly focus on video-level authenticity classification, our work targets temporal tampering localization, where the objective is to identify precise manipulated segments under limited temporal annotations.

Fully Supervised Video Tampering Localization. Video tampering temporal localization aims to identify manipulated temporal segments within videos, requiring fine-grained boundary prediction. Early methods are largely inspired by temporal action localization and focus on designing temporal feature aggregation architectures [7]. Surveys on autonomous deepfake detection have also systematized the landscape [15], while multimodal analysis surveys highlight the importance of audio-visual cues [32].

Representative works include temporal convolution-based models such as TCN-style frameworks [4], which aggregate multi-scale temporal features for segment-level prediction. Two-stream methods further incorporate RGB and optical flow to capture appearance-motion inconsistency [39]. Attention-based architectures [6, 11] improve temporal dependency modeling by emphasizing discriminative frames. A recent forgery trace-aware Swin Transformer [24] achieves high detection accuracy but heavily relies on full supervision. Hybrid feature fusion strategies also represent a prominent direction [13], yet still require dense annotations. More recently, Transformer-based methods [36, 37] have achieved strong performance by modeling long-range temporal dependencies. Despite their success, these approaches rely heavily on dense frame-level annotations, which are expensive and difficult to obtain in real-world forensic scenarios.

Weakly Supervised Temporal Localization. To reduce annotation cost, weakly supervised approaches formulate temporal localization under video-level supervision. Most methods follow the Multiple Instance Learning (MIL) paradigm [16], where a video is treated as a bag of temporal segments. RTFM [27] introduces feature magnitude learning to distinguish abnormal segments, while subsequent methods improve temporal smoothness and attention aggregation to enhance localization quality. Recently, multimodal deviation-based frameworks such as MDP [33] and multitask learning based approaches like WMMT [34] have been proposed for temporal forgery localization, further narrowing the gap between weak and full supervision. EM-based frameworks such as GEM-TFL [10] attempt to bridge weak supervision and latent temporal decomposition. A generalized taxonomy of video anomaly detection also provides context for such methods [18]. However, weakly supervised methods suffer from inherent ambiguity: supervision is indirect, leading to unstable and fragmented temporal predictions.

Semi-Supervised Learning for Video Understanding. Semi-supervised learning (SSL) provides a more flexible paradigm by leveraging both labeled and unlabeled data. Representative frameworks such as Mean Teacher [25] enforce consistency between teacher and student models via exponential moving average (EMA), and have been widely adopted in visual understanding tasks. Recent extensions introduce stronger regularization strategies such as FixMatch [21], which combines confidence-based pseudo-labeling with consistency learning. For temporal tasks, SSTAP [23] proposes a self-supervised learning approach for semi-supervised temporal action proposal, where unlabeled frames are exploited through proposal-level pretext tasks

within a pseudo-label framework. However, action proposal and video tampering localization differ in important ways: action boundaries are typically visually salient and relatively long, whereas tampering intervals are often short, subtle, and visually inconspicuous. Directly transferring SSL-for-action methods to forensics is therefore non-trivial.

Recent Advances in Video Tampering Localization. The last two years have seen rapid progress in fully supervised video tampering localization. VideoFACT [30] detects video forgeries by jointly exploiting attention, scene context, and forensic traces, but relies on full frame-level supervision and multimodal encoding. A spatio-temporal co-attention fusion network [17] targets video splicing localization by modeling cross-region consistency. For inpainting-type tampering, a trusted video inpainting localization method [26] learns deep attentive noise representations, contrastive-learning-based inpainting localization [22] improves boundary discrimination via representation contrast, and a high-low frequency decoupling network [8] separates low- and high-frequency clues for partial inpainting forgery detection. On the audio side, CFPRF [5] targets audio temporal forgery localization; on the audio-visual side, EVAS [9], UniSkip-Mamba [29], and UniCaCLF [28] substantially advance LAV-DF AP@0.95 using multimodal and state-space architectures. Collectively, these methods improve accuracy by scaling model capacity and exploiting multimodal or frequency-domain cues, but they share two assumptions that limit their practicality: (i) dense frame-level annotations are available for all training videos, and (ii) computational budgets are sufficient to support Transformer- or Mamba-scale backbones. Our work targets the complementary setting in which annotations are scarce and computation is constrained.

Discussion. Existing approaches can be categorized into three paradigms: fully supervised methods relying on dense annotations, weakly supervised methods relying on coarse video-level labels, and semi-supervised methods relying on consistency learning. However, none of these paradigms explicitly model temporal evolution dynamics. Fully supervised methods focus on static boundary fitting, weakly supervised methods suffer from ambiguous supervision, and existing SSL methods primarily enforce output-level consistency. Meanwhile, the robustness of forensic methods against adversarial manipulations has recently attracted attention [2], raising additional requirements for localization frameworks. In contrast, our method reformulates temporal localization as a self-evolving prediction process, where temporal boundaries are progressively refined through evolution-aware consistency and structure-aware correction.

3 Method

We propose a lightweight semi-supervised framework for temporal video tampering localization under a Mean Teacher paradigm. Instead of relying on heavy spatiotemporal architectures, we reformulate temporal localization as a self-evolving prediction process, where coarse temporal predictions are progressively refined through structured evolution constraints. The framework consists of three components: a lightweight

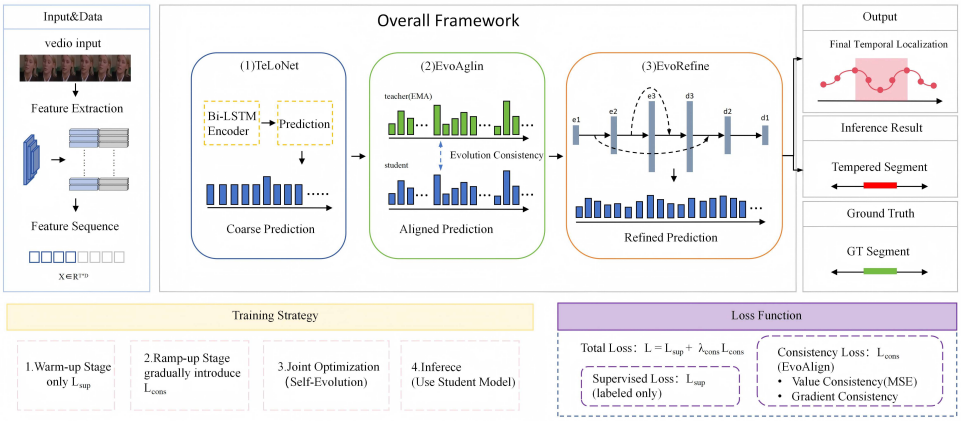

Fig. 2 Given pre-extracted features, TeLoNet generates an initial temporal hypothesis, which is progressively refined through the self-evolution process. EvoAlign enforces temporal evolution consistency in the gradient space, while EvoRefine performs structure-aware local refinement for precise boundary localization.

temporal representation operator (TeLoNet), a temporal evolution alignment mechanism (EvoAlign), and a boundary-aware refinement module (EvoRefine). These components are jointly optimized to enable stable and efficient temporal localization under limited supervision.

3.1 Problem Formulation

Given a video V with T temporal segments, we aim to predict a frame-level tampering score sequence:

$$\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_T\}, \quad \hat{y}_t \in \mathbb{R} \tag{1}$$

where $\hat{y}_t$ denotes the raw logit for each temporal segment. The corresponding probability is obtained via sigmoid:

$$p_t = \sigma(\hat{y}_t) \tag{2}$$

In the semi-supervised setting, we are given a labeled dataset $\mathcal{D}_L = \{(X_i, Y_i)\}$ and an unlabeled dataset $\mathcal{D}_U = \{X_j\}$, where $X \in \mathbb{R}^{T \times D}$ denotes pre-extracted TSN features. We consider a realistic forensic scenario where only a small portion of videos are temporally annotated, while large-scale unlabeled videos are available.

3.2 Semi-Supervised Training Framework

We adopt a Mean Teacher framework. The student network is updated via gradient descent, while the teacher network is updated via exponential moving average (EMA):

$$\theta' \leftarrow \alpha\theta' + (1 - \alpha)\theta \tag{3}$$

The student produces logits $\hat{Y}^{(s)}$, and the teacher produces logits $\hat{Y}^{(t)}$. We convert them into probabilities:

$$p^{(s)} = \sigma(\hat{Y}^{(s)}), \quad p^{(t)} = \sigma(\hat{Y}^{(t)}) \tag{4}$$

To stabilize early training, we apply a ramp-up weighting function:

$$w(e) = \exp\left(-5\left(1 - \frac{e}{e_{\text{ramp}}}\right)^2\right) \tag{5}$$

3.3 Temporal Representation Operator (TeLoNet)

TeLoNet is a lightweight temporal encoder operating on pre-extracted features:

$$H_0 = W_p X, \quad H = \text{BiLSTM}(H_0) \tag{6}$$

It outputs logits:

$$\hat{Y} = f_\phi(X) \tag{7}$$

Compared with Transformer-based models, TeLoNet has linear complexity $\mathcal{O}(T \cdot d^2)$ and avoids quadratic self-attention, making it suitable for long video sequences.

3.4 Temporal Evolution Alignment (EvoAlign)

Instead of enforcing consistency on raw outputs, we define the temporal evolution (first-order difference) of the prediction in probability space:

$$\mathcal{E}(p_t) = p_{t+1} - p_t \tag{8}$$

We then enforce consistency between the student and teacher evolution trajectories by decomposing the objective into a *gradient-consistency* term and a *point-wise* term:

$$\mathcal{L}_{evo} = \underbrace{\|\nabla_t p^{(s)} - \nabla_t p^{(t)}\|_2^2}_{\mathcal{L}_{\text{grad}}} + \underbrace{\|p^{(s)} - p^{(t)}\|_2^2}_{\mathcal{L}_{\text{out}}} \tag{9}$$

The gradient term $\mathcal{L}_{\text{grad}}$ is the core of EvoAlign: it constrains the *shape* of the temporal transition rather than its absolute value, so that the student and the teacher agree on where boundaries rise and fall even when their point-wise outputs differ. The output term $\mathcal{L}_{\text{out}}$ acts as a regularizer that prevents the two trajectories from drifting apart globally. This differs from standard Mean Teacher consistency [25], which constrains point-wise outputs only.

3.5 Boundary-Aware Refinement (EvoRefine)

To further refine localization results, we introduce a residual correction module that operates in logit space:

$$\hat{Y}^* = \hat{Y} + \Delta_\psi(\hat{Y}) \tag{10}$$

where $\Delta_\psi(\cdot)$ is a lightweight 1D convolutional network with a bounded residual output, implemented as a small stack of 1D convolutions over the temporal axis. We define *ambiguous temporal regions* as the temporal positions whose predicted probability lies in an intermediate band around the decision threshold, i.e., $|p_t - 0.5| \leq \epsilon$ for a small ϵ. Because the base prediction $\hat{Y}$ is already confident away from the boundaries, the residual correction $\Delta_\psi(\hat{Y})$ naturally concentrates its effect on these ambiguous transition zones, sharpening the boundary without disturbing confident predictions. The corresponding consistency loss is defined in probability space:

$$\mathcal{L}_{ref} = \|\sigma(\hat{Y}^{*(s)}) - p^{(t)}\|_2^2 \tag{11}$$

3.6 Optimization Objective

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{evo} + \lambda_2 \mathcal{L}_{ref} \tag{12}$$

The supervised loss is applied on labeled data using logits:

$$\mathcal{L}_{sup} = \text{BCEWithLogitsLoss}(\hat{Y}, Y) \tag{13}$$

3.7 Inference

During inference, only the student network is used, and the boundary-aware refinement is applied to the student logits:

$$\hat{Y}^* = \hat{Y}^{(s)} + \Delta_\psi(\hat{Y}^{(s)}) \tag{14}$$

The per-frame tampering probability is then computed from the refined logits:

$$p_t = \sigma(\hat{y}_t^*) \tag{15}$$

where $\hat{y}_t^*$ denotes the t-th element of the refined sequence $\hat{Y}^*$. In other words, the refined prediction $\hat{Y}^*$ of EvoRefine, rather than the raw logits $\hat{Y}^{(s)}$, is used for inference. Binary localization results are obtained via thresholding:

$$\hat{y}_t^{bin} = \mathbb{I}[p_t > \tau] \tag{16}$$

3.8 Computational Complexity

TeLoNet has complexity $\mathcal{O}(T \cdot d^2)$, EvoAlign operates in $\mathcal{O}(T)$, and EvoRefine in $\mathcal{O}(T \cdot k)$, giving an overall complexity of $\mathcal{O}(T \cdot d^2)$. Compared with Transformer-based methods $\mathcal{O}(T^2 \cdot d)$, our method achieves linear scaling with respect to sequence length.

3.9 Discussion

The key design principle of our framework is to decouple representation learning from temporal evolution modeling. Instead of relying on heavy architectures, we explicitly

model prediction dynamics in probability space while maintaining stable logit-based optimization. It is worth clarifying how the three components relate to, and differ from, their closest established counterparts. TeLoNet is a BiLSTM over pre-extracted features; its contribution is not the recurrence itself but the choice of a linear-complexity, lightweight temporal encoder that removes the quadratic self-attention of Transformer backbones. EvoAlign extends Mean Teacher consistency [25] from point-wise output agreement to trajectory (first-order difference) agreement; the gradient-consistency term $\mathcal{L}_{\text{grad}}$ is the distinguishing ingredient. EvoRefine is a residual 1D-conv correction in logit space, and differs from generic boundary-refinement networks by operating entirely in prediction space rather than feature space, which keeps its cost negligible. The novelty of the framework therefore lies in the specific combination of these three ingredients and their joint optimization under label scarcity, rather than in any single module in isolation.

4 Experiments

4.1 Datasets and Implementation Details

LAV-DF. LAV-DF (Localized Audio-Visual DeepFake) [4] is a large-scale audio-visual deepfake detection and temporal forgery localization dataset. It contains authentic videos together with forged videos produced by multiple manipulation approaches, where either the face or the audio (or both) can be manipulated. Each tampered video is annotated with the precise temporal intervals of the manipulated segments, which provides the frame-level supervision needed for temporal localization. The dataset contains approximately 36,000 video clips. We follow the official split, in which the test set contains 26,100 videos (12,821 tampered and 13,279 authentic) and the remaining clips are used for training. Since our focus is visual tampering, we use only the video stream and ignore the audio tampering annotations.

TVIL. TVIL (Temporal Video Inpainting Localization) [37] focuses on inpainting-based tampering, in which a temporal region of a video is removed and inpainted. Each video is annotated with the temporal range of the inpainted region. The official split provides 3,471 videos for training, 474 for validation, and 508 for testing.

Preprocessing. Both datasets are converted into fixed-length temporal feature sequences before being fed to the model. Following [4, 37], we use pre-extracted TSN [31] features (RGB and optical flow) instead of raw frames. Each stream produces a 2048-dimensional feature vector per temporal segment; when both streams are used, the RGB and flow vectors are concatenated into a 4096-dimensional input.

We follow the evaluation protocols in [4, 37] and use average precision (AP) and average recall (AR) as the localization metrics. By convention, we set the tIoU thresholds to $\{0.5, 0.75, 0.95\}$ and the average number of proposals (AN) to $\{10, 20, 50, 100\}$. For video-level classification, we infer the video label from the localization output: if any frame's predicted probability exceeds 0.5, the video is classified as tampered.

Semi-Supervised Setting. We simulate label scarcity by keeping only a small labeled subset and treating the remaining training videos as unlabeled. For LAV-DF we use 10% labeled data and for TVIL 20%. The label ratios were selected on the validation split only and fixed before any test evaluation. For each dataset, the labeled

subset is sampled while preserving the tampered/authentic class balance, and the remaining training videos form the unlabeled subset with no annotations used.

Implementation Details. We train the model for 100 epochs with batch size 64 for LAV-DF and 16 for TVIL, where each batch contains an equal number of labeled and unlabeled samples. The teacher is updated by EMA with decay $\alpha = 0.99$, consistent with [25]. The consistency-loss weight λ_{cons} (i.e., λ_1 in Eq. (12)) is set to 0.5, and the warm-up length is $e_{\text{ramp}} = 40$.

4.2 Main Results

Table 1 Performance comparison on the LAV-DF full set.

Method	AP@0.5	AP@0.75	AP@0.95	AR@10	AR@20	AR@50	AR@100
MDS	12.78	1.62	0.00	37.88	36.71	34.39	32.15
AGT	17.85	9.42	0.11	43.15	34.23	24.59	16.71
BMN	24.01	7.61	0.07	53.26	41.24	31.60	26.93
BMN (I3D)	10.56	1.66	0.00	48.49	44.39	37.13	31.55
BA-TFD	58.55	28.60	0.16	62.49	58.77	53.86	50.29
UMMAFormer	**97.30**	**92.96**	25.68	90.19	90.85	91.14	91.18
LSEF (Ours)	96.57	88.84	**42.33**	**92.14**	**92.14**	**92.14**	**92.14**

Table 2 Performance comparison on the TVIL full set.

Method	AP@0.5	AP@0.75	AP@0.95	AR@10	AR@20	AR@50	AR@100
TAGS	18.40	12.68	0.09	24.41	25.05	25.56	25.56
DCAN	82.75	75.00	3.22	64.73	66.02	68.82	69.97
ActionFormer	86.27	83.03	28.17	–	–	–	–
UMMAFormer	**88.68**	**84.70**	**62.43**	**87.09**	**88.21**	**90.43**	**91.16**
LSEF (Ours)	72.49	66.32	58.87	77.93	77.93	77.93	77.93

LAV-DF Full Set. As shown in Table 1, our method achieves competitive performance against fully supervised baselines despite using only 10% labeled data. On AP@0.5 it reaches 96.57% (vs. UMMAformer's 97.30%), and on AP@0.75 it reaches 88.84% (vs. 92.96%). On the strict AP@0.95 metric our method reaches 42.33%, higher than UMMAformer (25.68%) in this comparison. Importantly, this AP@0.95 advantage should not be attributed to a single mechanism. The ablation in Table 4 shows that the base TeLoNet backbone already attains 41.31% AP@0.95, and the gain over UMMAformer comes mainly from the lightweight backbone together with

EvoRefine's boundary correction, with EvoAlign contributing a smaller but consistent improvement. Furthermore, recent fully supervised methods that additionally exploit multimodal or frequency-domain cues report substantially higher AP@0.95. We therefore do not claim state-of-the-art accuracy; our contribution is the label- and compute-efficient setting.

TVIL Full Set. Table 2 presents results on TVIL, which focuses on inpainting-based tampering. Our method achieves 58.87% AP@0.95, approaching UMMAformer (62.43%) while using only 20% labeled data. Compared with the classical proposal-free method TAGS [20], the improvement is substantial across all metrics. Compared with DCAN [12], our method is stronger on AP@0.95 but weaker on AP@0.5 and AP@0.75. We also include ActionFormer [36], a stronger fully supervised baseline (86.27% AP@0.5, 83.03% AP@0.75, 28.17% AP@0.95), against which our method trails on AP@0.5/AP@0.75 but exceeds on AP@0.95. A performance gap remains on AP@0.5 and AP@0.75 relative to the strongest fully supervised methods, indicating that coarse-level temporal coverage is still limited by the amount of labeled data.

Semi-Supervised Baseline. To isolate the contribution of the semi-supervised machinery itself, we report a supervised-only baseline in Table 3: the same TeLoNet backbone trained exclusively on the labeled subset (10% for LAV-DF and 20% for TVIL), with no unlabeled data and no consistency or refinement terms.

Table 3 Supervised-only baseline versus the full framework.

Dataset	Method	Labeled	AP@0.5	AP@0.75	AP@0.95
LAV-DF	TeLoNet (supervised-only)	10%	96.02	86.00	35.80
	LSEF (full framework)	10%	96.57	88.84	42.33
TVIL	TeLoNet (supervised-only)	20%	71.47	65.30	58.10
	LSEF (full framework)	20%	72.49	66.32	58.87

Two observations follow from Table 3. First, on both datasets the supervised-only backbone already attains AP@0.5 close to the full framework (within roughly 0.5–1.0 point), indicating that most of the coarse-level localization ability comes from the lightweight backbone itself rather than from the semi-supervised machinery. Second, the full framework yields a more substantial gain on the strict AP@0.95 metric on LAV-DF (35.80% → 42.33%, +6.53 points), consistent with the intended role of EvoAlign and EvoRefine in sharpening precise temporal boundaries; on TVIL the corresponding gain is smaller (58.10% → 58.87%). These results suggest that the self-evolving components primarily improve boundary precision under label scarcity rather than overall coarse coverage.

4.3 Ablation Studies

To verify the effectiveness of each self-evolution component, we conduct ablation experiments on the LAV-DF dataset under a fixed 10% label ratio. Introducing EvoAlign consistently improves performance across multiple metrics: TeLoNet+EvoAlign

Table 4 Component ablation experiment (10% labels).

Dataset	Method	EA	ER	AP@0.5	AP@0.75	AP@0.95	Mean IoU
LAV-DF	TeLoNet	×	×	95.35	87.15	41.31	87.83
	TeLoNet+EA	✓	×	95.86	87.72	41.46	88.47
	TeLoNet+ER	×	✓	96.25	88.76	42.09	89.05
	Full Model	✓	✓	96.57	88.84	42.33	89.67

improves AP@0.5 by 0.51 points (95.35% → 95.86%), AP@0.75 by 0.57 points (87.15% → 87.72%), and Mean IoU by 0.64 points (87.83% → 88.47%). The Evolution Refinement Operator (EvoRefine) brings more significant improvements: compared with TeLoNet, TeLoNet+EvoRefine improves AP@0.5 by 0.90 points (95.35% → 96.25%), AP@0.75 by 1.61 points (87.15% → 88.76%), AP@0.95 by 0.78 points (41.31% → 42.09%), and Mean IoU by 1.22 points (87.83% → 89.05%). The full model achieves the best overall performance, reaching 96.57% AP@0.5, 88.84% AP@0.75, 42.33% AP@0.95, and 89.67% Mean IoU, indicating that EvoAlign and EvoRefine are complementary.

Table 5 Label ratio ablation experiment.

Dataset	Label Ratio	AP@0.5	AP@0.75	AP@0.95	Mean IoU
LAV-DF	10%	96.57	88.84	42.33	89.67
	20%	96.97	89.73	44.19	89.87
	30%	97.33	90.65	45.99	90.30
	50%	97.73	91.21	47.65	90.77

As shown in Table 5, performance consistently improves with increasing label ratio. However, the improvement margin gradually diminishes as the label ratio increases from 20% to 50%, indicating that the model can already effectively exploit unlabeled data under low-supervision settings. Notably, even with only 10% labeled data, the model achieves 96.57% AP@0.5 and 42.33% AP@0.95, which are only marginally lower than the 50% setting.

4.4 Computational Efficiency Analysis

We conduct an efficiency analysis of the proposed framework on a single NVIDIA A100 GPU with batch size 1. Table 6 presents a comparison between our method and the Transformer-based UMMAFormer.

The results reveal a substantial efficiency gap, although we emphasize that the two methods are not measured under identical conditions: UMMAFormer is evaluated at sequence length $T = 768$, whereas our model uses $T = 128$, which partially inflates the gap. Our full RGB model contains 3.23M parameters (roughly an order of magnitude smaller than UMMAFormer's 44.93M), requires 301.29M FLOPs versus 20.35G, and

Table 6 Model efficiency analysis.

Model	Feat. Dim	Params	FLOPs	Infer. (ms)	FPS
UMMAFormer (RGB+Flow)	4096	44.929M	20.345G‡	799.33	1.25
TeLoNet (RGB)	2048	2.60M	271.75M	6.33	158.0
TeLoNet (RGB+Flow)	4096	3.65M	405.96M	10.60	94.3
EvoRefine	–	0.63M	29.54M	2.41	415.8
Full System (RGB)	2048	3.23M	301.29M	10.40	96.2

‡ UMMAFormer is measured at sequence length $T = 768$, whereas our model is measured at $T = 128$; the two settings are not matched.

runs at 96.2 FPS under the measured setting. We note two caveats. First, both methods share the same TSN feature-extraction front-end, which dominates the end-to-end cost and is excluded from the figures above. Second, because the sequence lengths differ, the exact speed-up factor should be interpreted with caution. We therefore describe our method as lightweight and fast rather than claiming a strict real-time end-to-end guarantee.

The efficiency of the proposed framework stems from a fundamental design shift: temporal modeling is performed in a compact representation space, and refinement is conducted in the prediction space rather than the feature space. TeLoNet performs temporal modeling on pre-extracted TSN features with linear complexity $\mathcal{O}(T \cdot d^2)$. EvoAlign is a parameter-free operator with negligible complexity $\mathcal{O}(T)$. EvoRefine performs boundary correction via lightweight 1D convolutions with complexity $\mathcal{O}(T \cdot k)$, operating entirely in prediction space. This shifts the computational burden from feature learning to structured prediction refinement, enabling amortized computation and scalable inference.

5 Conclusion

This paper presents a lightweight self-evolving paradigm for semi-supervised video tampering temporal localization. Different from conventional pipeline-based or module-centric designs, we reformulate temporal localization as a unified evolution process in prediction space, where coarse temporal hypotheses are progressively refined through structural consistency and localized correction. At the core of the proposed framework is the idea of self-evolving temporal reasoning: a compact temporal representation operator (TeLoNet) provides an initial coarse hypothesis, an evolution alignment operator (EvoAlign) enforces trajectory-level temporal consistency, and a refinement operator (EvoRefine) further corrects local boundary ambiguities.

Extensive experiments on LAV-DF and TVIL demonstrate that the proposed framework achieves competitive performance under limited supervision. With only 10%–20% labeled data, our method attains localization accuracy comparable to several fully supervised baselines while using roughly an order of magnitude fewer parameters.

On the strict AP@0.95 metric of LAV-DF, our method achieves 42.33%, exceeding UMMAFormer [37] in that specific comparison, although recent multimodal fully supervised methods report higher absolute accuracy. We therefore position the contribution as label- and compute-efficient localization rather than absolute state of the art.

Despite its effectiveness, the current framework still has limitations. It remains less robust for extremely short or highly fragmented tampering intervals, and does not yet incorporate multimodal cues such as audio signals. Future work will explore multi-scale evolution dynamics, cross-modal self-evolution, and more adaptive evolution scheduling strategies to further enhance the generalization ability of the framework. Improving robustness against adversarial manipulations and unseen tampering patterns also remains an important direction for future research. Overall, this work demonstrates that lightweight self-evolving modeling is a promising direction for scalable and high-precision video tampering localization under limited annotation budgets.

Acknowledgements. Not applicable.

Declarations

Funding Not applicable.

Conflict of interest The authors declare no competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Data availability The two benchmark datasets used in this study, LAV-DF and TVIL, are publicly available from their original sources [4, 37].

Code availability The source code of the proposed framework will be made publicly available on GitHub upon acceptance.

Author contribution Q.Z. and K.Z. conceived the study and designed the method. Q.Z. conducted the experiments, and K.Z. supervised the project. All authors reviewed and approved the final manuscript.

References

[1] Abbas, F., Taeihagh, A.: Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications **252**, 124260 (2024).

[2] Abdullah, S.M., Cheruvu, A., Kanchi, S., Chung, T., Gao, P., Jadliwala, M., Viswanath, B.: An analysis of recent advances in deepfake image detection in an evolving threat landscape. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 91–109. IEEE (2024).

[3] Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: BMN: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV), pp. 3889–3898. IEEE (2019).

[4] Cai, Z., Stefanov, K., Dhall, A., Hayat, M.: Do you really mean that? Content-driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In: Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE (2022).

[5] Wu, J., Lu, W., Luo, X., Yang, R., Wang, Q., Cao, X.: Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). ACM (2024).

[6] Chugh, K., Gupta, P., Dhall, A., Subramanian, R.: Not made for each other—Audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM International Conference on Multimedia (MM '20), pp. 439–447. ACM (2020).

[7] Diwan, A., Dixit, S., Subbiah, R., Mahadeva, R.: Systematic analysis of video tampering and detection techniques. Cogent Engineering **11**(1), 2347217 (2024).

[8] Xie, X., et al.: Detecting the imperceptible: A high-low frequency decoupling network for partial video inpainting forgery detection. Expert Systems with Applications (2026).

[9] Shen, S., Zhang, Q., Jiang, D., Zhang, K.: EVAS: Efficient multimodal temporal forgery localization via audio-visual synergy and steered boundary calibration. arXiv preprint arXiv:2607.04472 (2026).

[10] Zhu, X., Zheng, Y., Wang, S., Yang, J., Yang, Y., Tu, W., Wang, Z.: GEM-TFL: Bridging weak and full supervision for forgery localization through EM-guided decomposition and temporal refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026).

[11] Guan, J., Zhou, H., Hong, Z., Ding, E., Wang, J., Quan, C., Zhao, Y.: Delving into sequential patches for deepfake detection. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 4517–4530. Curran Associates, Inc. (2022).

[12] Su, H., Gan, W., Wu, W., Qiao, Y., Yan, J.: DCAN: Improving temporal action detection via dual context aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 248–257. AAAI Press (2022).

[13] Jayashre, K., Amsaprabhaa, M.: Safeguarding media integrity: A hybrid optimized deep feature fusion based deepfake detection in videos. Computers & Security **142**, 103860 (2024).

[14] Jegaveerapandian, L.M., Rani, A.J., Periyaswamy, P., Velusamy, S.: A survey on passive digital video forgery detection techniques. International Journal of Electrical and Computer Engineering **13**(6), 6324–6334 (2023).

[15] Khichi, M., Kumar, P., Sharma, S.: Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. Heliyon **11**(3), e42273 (2025).

[16] Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., Lu, Q.: Sharp multiple instance learning for deepfake video detection. In: Proceedings of the ACM International Conference on Multimedia (MM), pp. 1644–1653. ACM (2020).

[17] Lin, M., et al.: Spatio-temporal co-attention fusion network for video splicing localization. Journal of Electronic Imaging (2024).

[18] Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., Sun, P., Song, L.: Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. ACM Computing Surveys **56**(7), 189 (2024).

[19] Mohiuddin, S., Malakar, S., Kumar, M., Sarkar, R.: A comprehensive survey on state-of-the-art video forgery detection techniques. Multimedia Tools and Applications **82**(22), 33499–33539 (2023).

[20] Nag, S., Zhu, X., Song, Y.-Z., Xiang, T.: Proposal-free temporal action detection via global segmentation mask learning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 645–662. Springer (2022).

[21] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.-L.: FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 596–608. Curran Associates, Inc. (2020).

[22] Lin, M., et al.: Video inpainting localization with contrastive learning. IEEE Signal Processing Letters (2025).

[23] Wang, X., Zhang, S., Qing, Z., Shao, Y., Gao, C., Sang, N.: Self-supervised learning for semi-supervised temporal action proposal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2021).

[24] Hu, Z., Zhang, S., Liu, F.: Video tampering detection with forgery trace-aware Swin Transformer. Neurocomputing **675**, 132995 (2026).

[25] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised learning results. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1195–1204. Curran Associates, Inc. (2017).

[26] Lou, J., Cao, Y., et al.: Trusted video inpainting localization via deep attentive noise learning. IEEE Transactions on Dependable and Secure Computing (2025).

[27] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4975–4985. IEEE (2021).

[28] Yin, Q., Lu, W., Luo, X., Cao, X.: Context-aware TFL: A universal context-aware contrastive learning framework for temporal forgery localization. IEEE Transactions on Dependable and Secure Computing (2026).

[29] Yu, C., Zhang, Q., Jiang, D., Zhang, K.: UniSkip-Mamba: A frequency-aware state space model for audio-visual temporal forgery localization. arXiv preprint arXiv:2607.04498 (2026).

[30] Nguyen, T.D., Fang, S., Stamm, M.C.: VideoFACT: Detecting video forgeries using attention, scene context, and forensic traces. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (2024).

[31] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 20–36. Springer (2016).

[32] Wang, M.: Deepfake detection: A multimodal survey. ITM Web of Conferences **78**, 02027 (2025).

[33] Xu, W., Wu, J., Lu, W., Luo, X., Wang, Q.: A multimodal deviation perceiving framework for weakly-supervised temporal forgery localization. In: Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), pp. 11581–11589. ACM (2025).

[34] Xu, W., Lu, W., Luo, X.: Weakly supervised multimodal temporal forgery localization via multitask learning. arXiv preprint arXiv:2508.02179 (2025).

[35] Yadav, A., Vishwakarma, D.K.: Datasets, clues and state-of-the-arts for multimedia forensics: An extensive review. Expert Systems with Applications **249**, 123756 (2024).

[36] Zhang, C.-L., Wu, J., Li, Y.: ActionFormer: Localizing moments of actions with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 492–510. Springer (2022).

[37] Zhang, R., Wang, H., Du, M., Liu, H., Zhou, Y., Zeng, Q.: UMMAFormer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In: Proceedings of the 31st ACM International Conference on Multimedia (MM '23), pp. 8712–8721. ACM (2023).

[38] Zhang, Y., Pang, Z., Huang, S., Wang, C., Zhou, X.: Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. Journal of King Saud University – Computer and Information Sciences **37**(6), 102148 (2025).

[39] Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15044–15054. IEEE (2021).