Learning Time in Static Classifiers

Xi Ding[1*], **Lei Wang**[1, 2*], **Piotr Koniusz**[2, 3, 4, 1] **Yongsheng Gao**[1†]

[1]Griffith University, [2]Data61/CSIRO, [3]University of New South Wales, [4]Australian National University
{x.ding, l.wang4, p.koniusz, yongsheng.gao}@griffith.edu.au

Abstract

Real-world visual data rarely presents as isolated, static instances. Instead, it often evolves gradually over time through variations in pose, lighting, object state, or scene context. However, conventional classifiers are typically trained under the assumption of temporal independence, limiting their ability to capture such *dynamics*. We propose a simple yet effective framework that equips standard feedforward classifiers with temporal reasoning, all without modifying model architectures or introducing recurrent modules. At the heart of our approach is a novel *Support-Exemplar-Query (SEQ) learning paradigm*, which structures training data into temporally coherent trajectories. These trajectories enable the model to learn class-specific temporal prototypes and align prediction sequences via a differentiable soft-DTW loss. A multi-term objective further promotes semantic consistency and temporal smoothness. By interpreting input sequences as *evolving feature trajectories*, our method introduces a strong temporal inductive bias through loss design alone. This proves highly effective in both static and temporal tasks: it enhances performance on fine-grained and ultra-fine-grained image classification, and delivers precise, temporally consistent predictions in video anomaly detection. Despite its simplicity, our approach bridges static and temporal learning in a modular and data-efficient manner, requiring only a simple classifier on top of pre-extracted features.

Code — https://github.com/Darcyddx/time-seq

Introduction

Most classification models are trained under the assumption that data points are independent and identically distributed (i.i.d.). However, in many real-world scenarios such as robotics, surveillance, medical imaging, and video analysis, visual data naturally evolves over time (Wang, Huynh, and Koniusz 2019; Wang 2023; Zhu et al. 2024; Ding and Wang 2025a). A person might turn their head, lighting conditions may shift, or an object's state may gradually change. These temporal variations form coherent, smooth trajectories in feature space. Yet, standard classifiers treat such temporally structured inputs as static, isolated examples, ignoring the rich temporal dynamics inherent in the data.

[*]These authors contributed equally.
[†]Corresponding author.

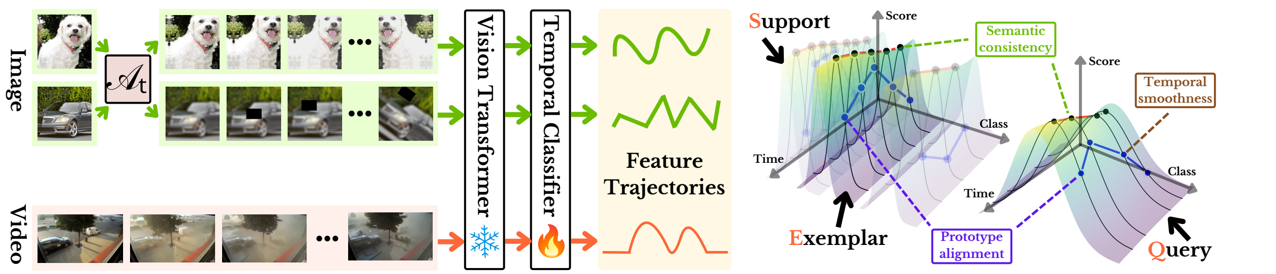
This mismatch between data reality and training assumptions limits the generalization of conventional classifiers, particularly for tasks requiring robustness to structured perturbations or subtle temporal shifts. While sequence models like RNNs, LSTMs, and Transformers can model temporal information (Xu, Zhu, and Clifton 2023), they introduce significant architectural complexity, require temporally annotated data, and are often ill-suited for scenarios with weak or missing frame-level labels (Zhu et al. 2024).

In this work, we ask: *Can standard feedforward classifiers reason over time without modifying their architecture, simply through rethinking how we supervise them?* We show the answer is *yes*. We propose a lightweight, general-purpose training framework that imparts *temporal inductive bias* into static classifiers purely through loss design. Our method operates on smoothly evolving input sequences generated via temporal augmentations that mimic natural transitions such as pose changes or appearance shifts. These sequences pass through a frozen pretrained encoder, followed by a classifier.

At the heart of our framework is a novel *Support-Exemplar-Query (SEQ) learning paradigm* that structures supervision around intra-class temporal patterns. For each query sequence, we align its predictions to class-specific temporal prototypes using a differentiable soft Dynamic Time Warping (soft-DTW) objective. In addition to alignment, we incorporate semantic supervision (via cross-entropy) and a smoothness regularization that penalizes abrupt prediction changes. This yields a key insight: *temporal reasoning can emerge in static feedforward models purely through supervisory signals, without any architectural modifications or explicit sequence modeling*. Our approach enables such models to learn how class semantics evolve temporally, bridging static and dynamic tasks in a unified, modular, and data-efficient manner.

We validate our method on two challenging domains: (i) fine-grained and ultra-fine-grained visual recognition under structured augmentations, where modeling temporal consistency improves generalization, and (ii) frame-level video anomaly detection, where capturing normal temporal behavior enables early and accurate anomaly detection. Our main **contributions** are summarized as follows:

i. We introduce *SEQ learning*, a novel and effective training paradigm that enables static feedforward classifiers to capture and use temporal class-specific prototype tra-



(a) Capturing evolving feature trajectories.

(b) The unified multi-term objective.

Figure 1: Overview of our framework. (a) Temporally smooth sequences are generated via time-indexed transformations $\mathcal{A}_t$ (or sourced from natural videos) and processed by a frozen, image-pretrained vision transformer to extract frame-wise features. A lightweight temporal classifier is then trained to produce feature trajectories. (b) These trajectories are optimized using a multi-term objective with the Support-Exemplar-Query (SEQ) learning framework (see Fig. 3) to (i) align with class-specific prototype trajectories that capture typical temporal patterns (violet block), (ii) achieve accurate classification through semantic supervision (vivid green block), and (iii) ensure smooth and consistent temporal evolution (gray brown block).

jectories, without requiring any architectural changes. This *challenges the common assumption* that temporal reasoning requires specialized sequence models.

ii. We develop *a unified, principled objective* combining soft-DTW temporal alignment, semantic supervision, and smoothness regularization. This framework endows standard classifiers with robust temporal reasoning capabilities purely through loss design and is, to our knowledge, *the first to do so*.

iii. We validate our method on *diverse and challenging tasks*, including fine-grained and ultra-fine-grained image recognition as well as video anomaly detection. Our approach shows significant improvements in generalization, temporal consistency, and anomaly sensitivity while using *only* feedforward architectures.

Related Work

Temporal modeling in classification. Classical approaches for temporal data rely on architectures explicitly designed to capture sequential dependencies, such as recurrent neural networks (RNNs) including LSTMs and GRUs (Hochreiter and Schmidhuber 1997), and more recently, attention-based models like Transformers (Vaswani et al. 2017; Bertasius, Wang, and Torresani 2021; Chen et al. 2024; Raj, Wang, and Gedeon 2025). These methods excel at modeling time series, video, and other sequential data but often require complex architectures, high computational costs, and dense temporal supervision. Their performance degrades when frame-level labels are scarce or when temporal ordering is weak or noisy.

In contrast, our method injects temporal inductive bias directly into the training objective of standard static classifiers, without architectural modifications or recurrent components. By aligning prediction sequences to learned temporal prototypes via soft-DTW, we enable temporal reasoning within simple feedforward models. This approach reduces complexity and broadens applicability to settings where temporal labels or models are unavailable.

Prototype-based learning. Prototype-based classification methods, central to few-shot and metric learning, represent classes by exemplars or centroids in feature space, facilitating generalization from limited data (Snell, Swersky, and Zemel 2017; Sung et al. 2018; Wang and Koniusz 2022b). Extensions to temporal tasks typically learn prototypes with recurrent or convolutional temporal encoders (Liu, Song, and Qin 2020; Wang and Koniusz 2022a).

Our work introduces a novel perspective by defining prototypes in the prediction space as class-specific softmax trajectories over time. Instead of embedding-level comparisons, we align entire prediction sequences to these temporal prototypes using soft-DTW, enforcing not only correct classification but also coherent temporal evolution of predictions. This shift enables temporal supervision even when only static labels are available, representing a significant departure from prior prototype-based methods.

Temporal and smooth augmentations. Data augmentation techniques improve robustness by exposing models to controlled input variations (Cubuk et al. 2020; Hendrycks et al. 2019). Temporal smoothness regularization and augmentations that mimic natural transitions have been used in video and self-supervised learning to encourage continuity and consistency (Sermanet et al. 2018; Qian et al. 2021; Schiappa, Rawat, and Shah 2023; Chen et al. 2024).

We build upon these ideas by using smooth, structured augmentations to synthesize temporal sequences from static inputs, simulating natural feature trajectories such as pose shifts or illumination changes. Crucially, we use these augmentations not only as regularizers but as core supervisory signals through alignment with temporal prototypes. This enables temporal inductive bias injection even in the absence of real temporal data or frame-level labels.

Learning paradigms for temporal and metric learning. Few-shot and metric learning methods often rely on episodic training paradigms organizing data into support and query sets, promoting generalization from limited exemplars (Snell, Swersky, and Zemel 2017; Sung et al. 2018). Some approaches extend these paradigms to temporal data by incorporating sequential encoding (Liu, Song, and Qin 2020; Wang and Koniusz 2022a,b; Wang et al. 2024b).

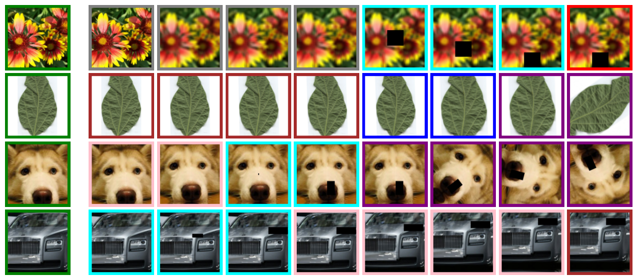


Figure 2: Examples from Flowers-102, SoyAging, Stanford Dogs, and Cars show how augmentations create temporal variations from one image. The first column shows originals (green); others apply augmentations by color: flip (red), zoom (blue), rotation (purple), color jitter (orange), shear (brown), translation (pink), blur (gray), and cutout (cyan), enriching the feature space with varied appearances.

Our proposed SEQ learning paradigm uniquely structures training data as temporally coherent feature trajectories grouped into support, exemplar, and query roles. This design encourages classifiers to internalize intra-class temporal dynamics through alignment with class-specific prediction prototypes. Unlike existing episodic methods, SEQ integrates soft-DTW alignment on prediction trajectories as a central supervision signal, enabling temporal reasoning without architectural or inference-time complexity. To our knowledge, this is the first framework to combine sequence-level prototype alignment, smooth augmentation-driven trajectory generation, and static feedforward classifiers into a lightweight, unified temporal learning paradigm.

Method

Overview

We introduce a novel framework that infuses temporal inductive bias into static classifiers *without requiring architectural changes or recurrent mechanisms*, see Fig. 1 for framework overview. The central insight of our method is to reinterpret static or sequential inputs as temporally coherent *feature trajectories*, which are then aligned with class-specific *temporal prototypes* using a differentiable sequence alignment procedure. This enables conventional feedforward models to exhibit temporal reasoning capabilities, enhancing their performance in scenarios where temporal consistency is crucial. It consists of three key components:

i. **Feature trajectories extraction.** We encode static or video inputs into smoothly evolving feature sequences that reflect temporal coherence.

ii. **Support-Exemplar-Query (SEQ) learning.** We propose a novel SEQ paradigm that uses intra-class temporal structure by organizing data into support, exemplar, and query trajectories, encouraging the model to learn temporally grounded representations.

iii. **Multi-term objective.** We optimize a composite loss function comprising (i) *alignment loss* for matching feature trajectories to temporal prototypes, (ii) *semantic supervision* via class labels, and (iii) a *temporal smoothness* regularization to maintain consistency across time.

The result is a robust, temporally aware classifier that generalizes effectively across both synthetic and real-world temporal variations, all while maintaining compatibility with existing architectures. We begin by describing our notation.

Notation. Let $\mathcal{I}_\tau = \{1, 2, \dots, \tau\}$ denote a time index set of length τ. A stacked vector of elements α_i is written as $[\alpha_i]_{i\in\mathcal{I}_\tau}$, and a matrix formed from elements α_{ij} is denoted $[\alpha_{ij}]_{(i,j)\in\mathcal{I}_I\times\mathcal{I}_J}$. Scalars are represented in standard font (*e.g.*, x), vectors in bold lowercase (*e.g.*, $\boldsymbol{x}$), matrices in bold uppercase (*e.g.*, $\boldsymbol{X}$), and tensors in calligraphic font (*e.g.*, $\mathcal{X}$). The inner product between two matrices $\boldsymbol{\Pi}$ and $\boldsymbol{D}$ is defined as the standard Euclidean inner product between their vectorized forms: $\langle \boldsymbol{\Pi}, \boldsymbol{D}\rangle \equiv \langle \text{vec}(\boldsymbol{\Pi}), \text{vec}(\boldsymbol{D})\rangle$.

Capturing Evolving Feature Trajectories

Smooth temporal augmentations from images. Static images inherently lack temporal structure, limiting a model's capacity to learn temporal dynamics or develop temporal reasoning. To overcome this limitation, we synthesize *virtual temporal sequences* from a single image $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ by applying *smooth, time-varying augmentations* over a virtual time index $t \in \mathcal{I}_\tau$. Formally, we construct a sequence:

$$\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_\tau], \quad \text{with} \quad \mathcal{X}_t = \mathcal{A}_t(\mathcal{X}), \tag{1}$$

where $\mathcal{A}_t$ denotes a transformation with parameters $\boldsymbol{\theta}_t$ that vary smoothly over time. Each augmentation parameter $p \in \boldsymbol{\theta}$ (*e.g.*, rotation angle, brightness, translation, *etc.*) evolves linearly over the sequence length τ:

$$p_t = p_{\text{start}} + \frac{t-1}{\tau-1}(p_{\text{end}} - p_{\text{start}}), \tag{2}$$

where p_{start} and p_{end} are randomly sampled endpoints. This linear interpolation ensures that the transformations evolve continuously across time, mimicking realistic temporal transitions. The operator $\mathcal{A}_t$ thus combines spatial and photometric effects such as rotation, translation, scaling, brightness, contrast, and blur into a time-indexed transformation:

$$\mathcal{A}_t = \mathcal{T}(\boldsymbol{\theta}_t). \tag{3}$$

These augmentations emulate plausible temporal changes, such as gradual pose shifts, or camera zooms, without requiring access to video data or temporal annotations. Fig. 2 shows visualizations of temporal augmentations on images.

Natural temporal sequences from videos. In contrast, video data naturally provides temporal continuity, capturing authentic dynamics such as object motion, scene evolution, and environmental changes. A video clip can be represented as: $\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_\tau]$, where each frame $\mathcal{X}_t$ is temporally correlated with its neighbors, forming a coherent sequence. This inherent structure encodes rich temporal information that can be directly exploited during training.

Extracting frame-wise features. We adopt a frozen image-pretrained backbone $\mathcal{M}_{\text{Img}}$ to extract frame-wise features from both synthetic and natural sequences. Each frame $\mathcal{X}_t$ is independently processed:

$$\boldsymbol{z}_t = \mathcal{M}_{\text{Img}}(\mathcal{X}_t), \tag{4}$$

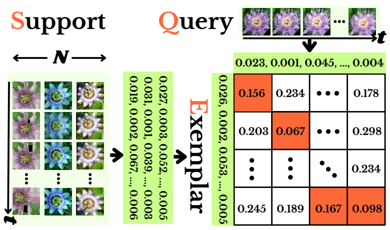


Figure 3: Support-Exemplar-Query (SEQ) models class-consistent temporal dynamics by constructing a *support set* of sequences to form a class-specific *exemplar* that captures typical prediction trajectories over time. A *query sequence* is then aligned against this exemplar to enforce temporal consistency and reveal deviations from expected class behavior.

yielding a sequence of feature vectors:

$$\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_\tau] \in \mathbb{R}^{\tau \times d}, \tag{5}$$

where d denotes the dimensionality of the extracted features. We adopt image-pretrained backbones for their rich, transferable visual representations, stability across domains, and efficiency benefits, enabling our classifier to focus solely on learning temporal relationships from strong, frozen features.

We then train a classifier f, typically a fully connected layer followed by a softmax activation:

$$\boldsymbol{\phi}_t = f(\boldsymbol{z}_t; \boldsymbol{W}), \quad \boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\tau] \in \mathbb{R}^{\tau \times C}, \tag{6}$$

where C denotes the number of classes. The output $\boldsymbol{\Phi}$ is used in classification tasks guided by dedicated loss objectives. This clean separation between feature extraction and temporal modeling maintains architectural simplicity, while enabling our framework to process both synthetic and real temporal sequences in a unified and scalable manner.

Support-Exemplar-Query (SEQ) Learning

We propose *Support-Exemplar-Query (SEQ) learning*, a novel framework for modeling class-consistent temporal dynamics and detecting structural deviations within sequential data. SEQ is built on three key components: (i) a *support set* consisting of class-consistent sequences, (ii) a class-conditioned *exemplar* that summarizes temporal regularities, and (iii) a *query set*, containing sequences evaluated against their corresponding class exemplars for consistency.

The SEQ framework operates in two stages (see Fig. 3). First, a *support-query matching* phase selects relevant support sequences for a given query. Second, an *exemplar-query alignment* phase measures the temporal similarity between the query and a synthesized class exemplar via differentiable alignment. The exemplar acts as a dynamic reference that encodes intra-class temporal coherence, facilitating interpretable matching and anomaly detection.

By explicitly capturing the temporal structure within each class and comparing incoming sequences against these learned exemplars, SEQ enables both fine-grained classification and structural deviation detection. Importantly, SEQ uses an *episodic training paradigm*, inspired by few-shot learning, which promotes robust generalization to novel classes and distribution shifts.

Support-query matching. In each training episode, we sample two disjoint subsets from the training data: a *query set* and a *support set*. The query set, denoted as $\mathcal{S}^{*}$, consists of sequences from various classes (*e.g.*, a batch of training samples), simulating real-world inputs that may be ambiguous or noisy. Given a query sequence $\boldsymbol{\Phi}^{*} \in \mathcal{S}^{*}$ with known class label c, we construct the corresponding support set $\mathcal{S}^{\bullet} = \{\boldsymbol{\Phi}^{\bullet}_n\}_{n \in \mathcal{I}_N}$ by sampling N additional sequences from the same class c. These support sequences are used to synthesize a class exemplar that represents typical temporal score evolution for class c.

To compare score sequences of variable lengths, we use the γ-Soft Dynamic Time Warping (Soft-DTW) distance, a differentiable relaxation of classical DTW. It enables smooth, gradient-based optimization and aggregates alignment costs over multiple plausible warping paths.

Let $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\tau] \in \mathbb{R}^{\tau \times C}$ and $\boldsymbol{\Phi}' = [\boldsymbol{\phi}'_1, \ldots, \boldsymbol{\phi}'_{\tau'}] \in \mathbb{R}^{\tau' \times C}$ denote two sequences of softmax prediction scores. The Soft-DTW distance is computed as:

$$d^2_{\text{DTW}}(\boldsymbol{\Phi}, \boldsymbol{\Phi}') = \text{SoftMin}_\gamma \left(\{ \langle \boldsymbol{\Pi}, \boldsymbol{D}(\boldsymbol{\Phi}, \boldsymbol{\Phi}') \rangle \mid \boldsymbol{\Pi} \in \mathcal{P}_{\tau, \tau'} \} \right), \tag{7}$$

where $\mathcal{P}_{\tau,\tau'}$ is the set of valid alignment paths between the two sequences, and the alignment cost $\langle \boldsymbol{\Pi}, \boldsymbol{D} \rangle$ is computed over the distance matrix $\boldsymbol{D} \in \mathbb{R}_+^{\tau \times \tau'}$, defined by:

$$\boldsymbol{D} = \left[d^2_{\text{base}}(\boldsymbol{\phi}_m, \boldsymbol{\phi}'_n) \right]_{(m,n) \in \mathcal{I}_\tau \times \mathcal{I}_{\tau'}}. \tag{8}$$

Here, $d^2_{\text{base}}(\cdot, \cdot)$ is typically the squared Euclidean distance. The SoftMin operator is given by:

$$\text{SoftMin}_\gamma(\boldsymbol{\alpha}) = -\gamma \log \sum_i \exp(-\alpha_i / \gamma), \tag{9}$$

where $\gamma \geq 0$ controls the softness of the alignment. As $\gamma \to 0$, it converges to standard DTW; larger γ values yield smoother, more flexible alignments.

Query-exemplar alignment. To represent the temporal dynamics of each class, we synthesize an *exemplar* sequence by computing the Fréchet mean (or barycenter) of the support set under Soft-DTW. This exemplar captures the average temporal evolution of softmax scores for class c, acting as a dynamic prototype for alignment.

Given support set $\mathcal{S}^{\bullet} = \{\boldsymbol{\Phi}^{\bullet}_n\}_{n \in \mathcal{I}_N}$ with possibly varying sequence lengths τ_n, the exemplar $\boldsymbol{M}^{\bullet} \in \mathbb{R}^{\bar{\tau} \times C}$ (where $\bar{\tau}$ is the average length of the sequences in $\mathcal{S}^{\bullet}$) is defined as:

$$\boldsymbol{M}^{\bullet} = \underset{\boldsymbol{M}^{\bullet} \in \mathbb{R}^{\bar{\tau} \times C}}{\arg\min} \sum_{n=1}^{N} \frac{w_n}{\tau_n} d^2_{\text{DTW}}(\boldsymbol{\Phi}^{\bullet}_n, \boldsymbol{M}^{\bullet}), \tag{10}$$

where $w_n \in \mathbb{R}_+$ are normalized weights satisfying $\sum_{n=1}^{N} w_n = 1$. This formulation jointly aligns and averages the support sequences, yielding a smooth, representative trajectory of class-consistent score dynamics.

Episodic training paradigm. In each episode, we select a query sequence $\boldsymbol{\Phi}^{*}$, sample a support set $\mathcal{S}^{\bullet}$ of size N, and compute the corresponding class exemplar $\boldsymbol{M}^{\bullet}$. We then align the query to the exemplar using Soft-DTW, obtaining a class-conditioned similarity score. Note that for generated virtual sequences, we ensure that both the query and its corresponding support sequences undergo identical temporal augmentations. This consistency preserves alignment integrity and allows the model to *focus on class-specific dynamics rather than artificial temporal discrepancies*. This alignment-based approach equips the model to capture temporal consistency, detect deviations from class patterns, and generalize to new temporal dynamics. Training across diverse episodes encourages abstraction of temporal class structure and adaptability to unseen scenarios.

Temporal Classifier with Multi-term Objective

We propose a lightweight yet expressive classifier that operates over temporal sequences of features using a single fully connected layer followed by softmax. The objective is to train this model to produce temporally coherent, semantically accurate, and class-consistent prediction trajectories. Specifically, the output sequence $\mathbf{\Phi} = [\phi_1, \ldots, \phi_\tau]$ should: (i) align with a class-specific prototype trajectory that encodes the typical temporal prediction pattern, (ii) accurately classify each timestep or sequence via semantic supervision, (iii) evolve smoothly over time, avoiding abrupt output changes. We now present the unified multi-term objective.

Temporal prototype alignment. For each class $c \in \{1,\ldots,C\}$ in an episode, we construct a class-specific prototype sequence $\boldsymbol{M}^{\bullet} \in \mathbb{R}^{\bar{\tau} \times C}$ using the Soft-DTW barycenter method (see equation 10). This prototype captures the characteristic temporal evolution of predictions for class c. To align each training query sequence $\mathbf{\Phi}^{*}$ with its corresponding prototype $\boldsymbol{M}^{\bullet}$, we minimize their Soft-DTW distance:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{S}^{*}|} \sum_{i=1}^{|\mathcal{S}^{*}|} d^2_{\text{DTW}} \left(\mathbf{\Phi}^{*}_i, \boldsymbol{M}^{\bullet}_i \right). \tag{11}$$

Here, $\mathcal{S}^{*}$ denotes the set of query sequences, with $|\mathcal{S}^{*}|$ as its cardinality. This alignment encourages the model to produce temporally structured and class-consistent prediction sequences, even when the input dynamics are nonlinear.

Cross-entropy supervision. To ensure semantic accuracy, we apply standard cross-entropy (CE) loss at either the frame or sequence level, depending on the task:

i. For sequence tasks (*e.g.*, anomaly detection) with frame-level labels y_t, cross-entropy is applied at each timestep:

$$\mathcal{L}_{\text{CE}} = \frac{1}{\tau |\mathcal{S}^{*}|} \sum_{i=1}^{|\mathcal{S}^{*}|} \sum_{t=1}^{\tau} \text{CE}(\phi^{*}_{i,t}, y_{i,t}). \tag{12}$$

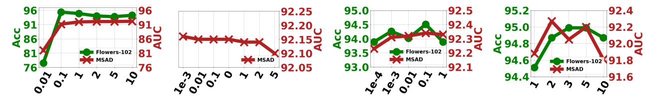
ii. For static tasks (*e.g.*, image classification), predictions are averaged over time to capture visual variations:

$$\mathcal{L}_{\text{CE}} = \frac{1}{|\mathcal{S}^{*}|} \sum_{i=1}^{|\mathcal{S}^{*}|} \text{CE} \left(\frac{1}{\tau} \sum_{t=1}^{\tau} \phi^{*}_{i,t}, y_i \right). \tag{13}$$

This term ensures that predictions convey the correct semantic labels at the *appropriate temporal granularity*.

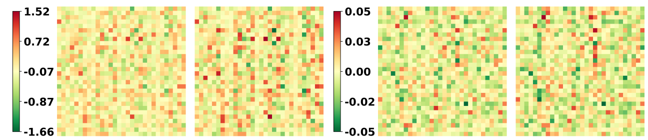
Temporal smoothness regularization. We add a smoothness loss to enforce temporal stability by penalizing abrupt changes between consecutive predictions:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{(\tau - 1)|\mathcal{S}^{*}|} \sum_{i=1}^{|\mathcal{S}^{*}|} \sum_{t=2}^{\tau} \left\| \phi^{*}_{i,t} - \phi^{*}_{i,t-1} \right\|_2^2. \tag{14}$$



(a) α (b) β (c) γ (d) N

Figure 4: Evaluation of key hyperparameters.



(a) Flowers-102 (b) SoyAging

Figure 5: Visualization of selected FC weight regions shows a clear comparison between the baseline (left) and our temporal modeling (right). Temporal modeling yields stronger, more distinct patterns, enhancing feature discrimination. Even on the ultra-fine-grained SoyAging, our approach produces clearer, more structured weights, demonstrating the advantages of temporal supervision in feature learning.

This encourages the model to produce gradual, interpretable prediction changes that mirror natural dynamics such as motion, progression, or transitions.

Final multi-term objective. The complete training loss combines these three components:

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \alpha \, \mathcal{L}_{\text{CE}} + \beta \, \mathcal{L}_{\text{smooth}}, \tag{15}$$

where α and β are hyperparameters that balance semantic supervision and temporal regularity. In our experiments, we also incorporate exemplars into both CE loss and temporal smoothness regularization to enhance robustness against semantic variations, perturbations, and class prototype shifts. For sequence tasks, exemplars help capture fine-grained temporal variations within sequences. For image classification, they assist in addressing shifts in class prototypes, thereby improving generalization across diverse conditions.

Experiment

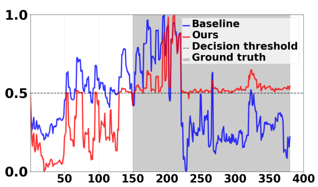
Experimental Setup

Datasets. We evaluate fine-grained recognition on Stanford Cars (Krause et al. 2013), Dogs (Khosla et al. 2011), Flowers-102 (Nilsback and Zisserman 2008), and the ultra-fine-grained SoyAging dataset (Yu et al. 2021b). Video anomaly detection uses MSAD (Zhu et al. 2024) with Protocol ii, covering various anomaly types and scenarios. All evaluations follow standard protocols for fair comparison.

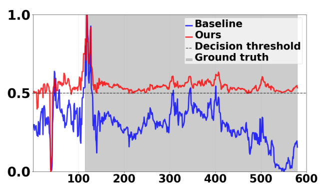
Models. Our method trains a single FC on frozen vision transformer features, without fine-tuning the backbone. We extract features using CLIP-ViT-L/14/224 (ImageNet-1K) for Stanford Cars, CLIP-ViT-B/16/224 (ImageNet-1K) for Oxford Flowers-102, and ViT-B/16/224 (ImageNet-1K) for Stanford Dogs. For the specialized SoyAging dataset, we use CLE-ViT (Swin-B/448) pretrained on ImageNet-21K (Yu, Wang, and Gao 2023) to use its larger corpus for ultra-fine-

Stanford Cars		Stanford Dogs		Oxford Flowers-102		SoyAging	
Method	Acc	Method	Acc	Method	Acc	Method	Acc
AP-CNN (Ding et al. 2021)	95.4	RAMS-Trans (Hu et al. 2021)	92.4	MGE-CNN (Zhang et al. 2019)	95.9	Cutmix (Yun et al. 2019)	62.3
P2P-Net (Yang et al. 2022)	95.4	PMG-V2 (Du et al. 2022)	90.7	SJFT (Ge and Yu 2017)	97.0	DCL (Chen et al. 2019)	73.2
CP-CNN (Liu et al. 2022)	95.4	ViT-NeT (Kim, Nam, and Ko 2022)	93.6	OPAM (Peng, He, and Zhao 2017)	97.1	ViT (Dosovitskiy et al. 2021)	67.0
TransFG (He et al. 2022)	94.8	TransFG (He et al. 2022)	92.3	Cosine (Barz and Denzler 2020)	97.2	DeiT (Touvron et al. 2021)	69.5
ViT-NeT (Kim, Nam, and Ko 2022)	95.0	IELT (Xu et al. 2023)	91.8	PMA (Song et al. 2020)	97.4	MaskCOV (Yu et al. 2021a)	75.9
DCAL (Zhu et al. 2022)	95.3	LGTF (Zhu et al. 2023)	92.1	DSTL (Cui et al. 2018)	97.6	TransFG (He et al. 2022)	72.2
PMG-V2 (Du et al. 2022)	95.4	ACC-ViT (Zhang et al. 2024)	92.9	MC-Loss (Chang et al. 2020)	97.7	SPARE (Yu, Zhao, and Gao 2022)	75.7
GDSMP-Net (Ke et al. 2023)	95.3	MP-FGVC (Jiang et al. 2024)	91.0	CAP (Behera et al. 2021)	97.7	Mix-ViT (Yu et al. 2023)	76.3
MPSA (Wang et al. 2024a)	95.4	MPSA (Wang et al. 2024a)	95.4	SR-GNN (Bera et al. 2022)	97.9	CLE-ViT (Yu, Wang, and Gao 2023)	79.0*
Baseline	94.7	Baseline	93.5	Baseline	97.6	Baseline	79.6
w/ feat. traj.	95.6	*w/ feat. traj.*	96.0	*w/ feat. traj.*	**98.4**	*w/ feat. traj.*	79.8
w/ feat. traj.&SEQ	**96.1**	*w/ feat. traj.&SEQ*	**96.3**	*w/ feat. traj.&SEQ*	**98.4**	*w/ feat. traj.&SEQ*	**80.0**

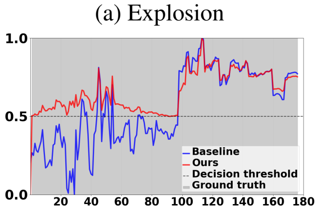
Table 1: Performance on fine-grained and ultra-fine-grained recognition datasets. Baseline and our methods use a single-layer classifier. *w/ feat. traj.* applies smooth temporal augmentations to produce feature trajectories, while *w/ feat. traj.&SEQ* is our full model with both sequence and temporal modeling. **Bold** marks the best. Temporal classifier improves performance across image datasets, highlighting the value of temporal cues in fine-grained recognition. * indicates reproduced results.



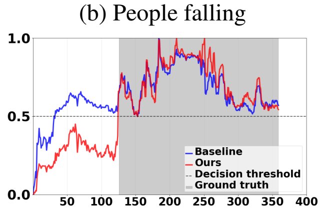
(a) Explosion



(b) People falling



(c) Shooting



(d) Traffic accident

Figure 6: Anomaly prediction comparison. Grey regions indicate ground-truth anomalies. Blue and red curves show the baseline and our method. Our approach detects anomalies more accurately and earlier, with scores crossing the 0.5 threshold in closer alignment with the ground truth.

grained tasks. For MSAD, frame-level features are extracted using CLIP-ViT-B/16/224 pretrained on WebImgText.

Setups. In all experiments, the baseline is a static classifier with a single FC layer and softmax. We extend this by exploring feature trajectories, which represent smooth feature evolution through synthetic temporal augmentations (for static images) or natural temporal changes (for videos like MSAD). Our full model integrates feature trajectories with SEQ to effectively capture dynamic visual patterns over time. We benchmark against recent state-of-the-art methods on each dataset to validate the competitiveness of our results.

Quantitative and Qualitative Evaluation

Hyperparameter evaluation. Fig. 4 shows the impact of key hyperparameters. The weight α controls the classification loss, with performance improving as α increases and stabilizing beyond $\alpha \geq 1$, underscoring the need to balance alignment and classification. The smoothness regularizer β exhibits stable performance across a wide range, indicating robustness on MSAD. The softness parameter γ in soft-DTW benefits from moderate values, *e.g.*, 0.1, particularly on Flowers-102. Lastly, N defines the number of sequences in the support set for computing class exemplars. Larger N leads to more reliable temporal estimates, with gains saturating at $N = 3$ on Flowers-102.

Analysis of learned weights. As shown in Fig. 5, temporal modeling produces stronger, more structured, and more distinct weight patterns compared to the baseline. These clearer weight structures suggest enhanced feature discrimination, as the model learns to better separate meaningful variations through temporal supervision. Even on challenging ultra-fine-grained datasets such as SoyAging, where differences are inherently subtle, our method leads to more pronounced and organized weight patterns. This demonstrates that our approach introduces an implicit temporal inductive bias into otherwise static classifiers, without requiring architectural modifications. Instead, this bias is induced through simple temporal augmentations and sequence-based training objectives, providing a lightweight yet effective alternative to traditional heavy temporal models.

Fine-grained image recognition. In Table 1, when smooth temporal augmentations are applied, performance improves across all datasets. Adding SEQ learning on top of feature trajectories yields further improvements. This suggests that modeling temporal dependencies, even in static images, can enhance feature discrimination by implicitly learning consistent patterns and relationships across augmented views. Particularly in ultra-fine-grained recognition like SoyAging, the incremental gains indicate that temporal cues help address extreme subtlety in class differences where traditional spatial cues alone may be insufficient. These findings reveal that temporal consistency, typically associated with video or time-series data, can be used as a powerful inductive bias to improve static image classification, especially in challenging fine-grained domains, without increasing architectural complexity (Ding et al. 2025b). This insight opens new avenues for bridging temporal modeling techniques with static image

Method	Assault		Explosion		Fighting		Fire		Obj. Fall		People Fall		Robbery		Shooting		Traffic Acc.		Vandalism		Water Inc.		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
RTFM (I3D)	53.9	<u>**66.4**</u>	66.0	76.6	79.8	88.6	44.9	71.1	84.6	89.3	45.7	<u>**52.6**</u>	**70.2**	<u>**88.0**</u>	<u>**87.5**</u>	<u>**89.2**</u>	64.1	**57.7**	74.9	73.0	98.1	99.6	86.6	68.4
MGFN (SwinT)	50.2	49.6	50.9	58.1	57.2	67.1	51.4	74.2	41.3	51.6	44.4	40.3	40.1	68.5	51.4	63.9	50.4	42.3	42.6	40.9	58.6	87.2	69.3	33.6
MGFN (I3D)	53.9	60.2	59.1	66.5	**80.6**	**89.5**	66.1	82.9	89.9	94.6	**53.6**	44.9	<u>**72.2**</u>	85.4	68.3	80.6	**66.9**	54.7	84.4	78.5	81.9	96.1	81.2	59.3
UR-DMU	56.9	**64.5**	67.9	74.5	<u>**83.9**</u>	<u>**90.4**</u>	61.2	82.9	**92.1**	<u>**95.8**</u>	42.5	43.7	63.5	79.3	**81.4**	87.8	62.0	55.6	84.7	77.0	98.5	99.5	85.0	68.3
EGO	52.2	57.5	57.6	74.4	66.5	72.8	62.9	**86.7**	<u>**92.3**</u>	**94.8**	35.4	43.8	64.8	<u>**87.5**</u>	68.6	78.4	69.9	<u>**64.3**</u>	<u>**88.1**</u>	<u>**81.4**</u>	81.9	95.4	87.3	64.4
IEF-VAD	<u>**66.0**</u>	–	66.3	–	79.8	–	49.4	–	75.9	–	42.5	–	66.9	–	**86.9**	–	<u>**70.1**</u>	–	75.8	–	**88.9**	–	82.1	–
Baseline	48.2	51.5	**77.3**	84.6	73.1	83.9	78.2	**94.7**	83.7	89.7	49.8	46.3	65.5	86.4	79.8	88.1	63.0	55.2	76.2	75.0	**99.6**	**99.9**	86.7	72.2
w/ feat. traj.	50.6	51.1	76.1	**85.3**	70.5	82.8	76.7	94.5	**85.6**	90.7	<u>**56.2**</u>	**50.7**	67.0	86.2	79.1	88.3	61.4	52.6	84.3	78.3	99.3	99.8	<u>**92.1**</u>	**77.3**
w/ feat. traj.&SEQ	**59.3**	60.2	<u>**84.9**</u>	<u>**88.9**</u>	79.8	<u>**89.9**</u>	<u>**81.4**</u>	<u>**95.6**</u>	85.2	90.8	52.3	48.7	68.1	87.4	79.5	**88.6**	60.6	50.4	**85.0**	**80.3**	<u>**99.6**</u>	<u>**99.9**</u>	**90.5**	<u>**77.5**</u>

Table 2: Performance by anomaly type on MSAD. The best result is marked with <u>**bold and underline**</u>, and the second-best is shown in **bold**. We compare against recent methods, including RTFM (Tian et al. 2021), MGFN (Chen et al. 2023), UR-DMU (Zhou, Yu, and Yang 2023), EGO (Ding et al. 2025a), and IEF-VAD (Jeong, Park, and Imani 2025), which rely on complex architectures and spatio-temporal feature extraction. In contrast, our method, despite using only a temporal classifier on top of frozen features, consistently achieves strong and often superior results across anomaly types. This highlights the effectiveness of simple temporal modeling in capturing temporal dynamics without heavy architectural complexity.

Method	Frontdoor		Mall		Office		Parkinglot		Pedestr. st.		Restaurant		Road		Shop		Sidewalk		St. highview		Train		Warehouse		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
RTFM (I3D)	81.8	79.3	88.1	76.6	76.6	72.8	80.7	45.8	94.0	**48.5**	88.3	79.1	**84.3**	57.9	85.3	75.6	**88.3**	68.8	72.0	28.5	51.4	3.3	82.7	57.0	86.6	68.4
MGFN (SwinT)	59.5	51.7	18.5	20.1	64.1	52.3	67.9	19.0	75.9	9.7	67.9	44.0	70.6	26.3	62.7	43.0	69.0	25.9	75.3	23.3	65.4	5.2	70.1	30.1	69.3	33.6
MGFN (I3D)	82.5	80.8	73.8	71.3	71.5	58.2	68.9	14.8	**94.8**	36.2	<u>**95.1**</u>	<u>**91.3**</u>	76.5	35.8	85.6	78.4	78.5	57.2	77.9	29.3	40.3	2.1	58.3	24.2	81.2	59.3
UR-DMU	84.8	82.8	<u>**91.0**</u>	**83.8**	77.8	67.3	91.4	53.9	81.9	11.5	93.1	**87.4**	83.0	**64.4**	81.3	64.5	86.5	64.1	85.0	37.7	59.0	3.1	81.2	59.1	85.0	68.3
EGO	**85.2**	81.6	82.3	73.4	80.0	71.7	**96.8**	75.2	<u>**97.5**</u>	<u>**52.0**</u>	**94.3**	73.9	<u>**89.8**</u>	<u>**64.6**</u>	83.4	72.2	87.1	45.0	28.2	10.1	80.8	7.8	84.7	46.6	87.3	64.4
IEF-VAD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.1	-
Baseline	83.9	**83.3**	**90.1**	82.0	79.5	**76.0**	96.3	**83.9**	49.9	25.4	85.9	77.3	63.1	43.5	<u>**92.1**</u>	<u>**84.7**</u>	85.4	67.8	99.7	98.8	91.7	24.2	79.0	45.9	86.7	72.2
w/ feat. traj.	**85.2**	82.4	**90.1**	82.0	**83.0**	75.0	<u>**97.0**</u>	<u>**86.3**</u>	46.8	10.7	90.8	81.5	81.4	60.4	**90.9**	**82.0**	<u>**93.2**</u>	<u>**80.9**</u>	<u>**99.9**</u>	<u>**99.4**</u>	<u>**95.7**</u>	<u>**42.9**</u>	<u>**94.2**</u>	**69.9**	<u>**92.1**</u>	**77.3**
w/ feat. traj.&SEQ	<u>**86.0**</u>	<u>**84.6**</u>	89.6	<u>**84.2**</u>	<u>**84.5**</u>	<u>**79.2**</u>	92.9	64.4	39.1	10.0	92.2	84.0	79.5	62.4	88.0	79.6	87.3	**71.4**	**99.8**	**99.2**	**95.3**	**36.7**	**93.5**	<u>**70.9**</u>	**90.5**	<u>**77.5**</u>

Table 3: Performance by Scenario on MSAD. We report results on 12 test scenarios, excluding *Highway* and *Park*, which do not contain anomalous events. Our methods, *w/ feat. traj.* and *w/ feat. traj.&SEQ*, achieve strong performance across all scenarios.

tasks to push the limits of visual recognition accuracy.

Evaluation on MSAD. Analyzing performance by anomaly type (Table 2), our method consistently achieves competitive or superior results compared to recent state-of-the-art approaches, despite using a simple classifier on frozen features. This demonstrates the effectiveness of lightweight temporal modeling in capturing complex anomaly dynamics without relying on heavy spatio-temporal architectures. Anomalies such as explosions, fires, and vandalism benefit notably from integrating feature trajectories and SEQ learning, leading to substantial performance gains (see Fig. 6 for prediction comparison). The strong performance achieved without complex architectures further suggests that well-extracted frozen features, when paired with effective temporal modeling, provide a robust foundation for anomaly recognition. Across scenarios (Table 3), the model maintains stable and reliable detection across diverse environments, from indoor settings like malls and offices to outdoor scenes such as sidewalks and parking lots. This demonstrates that temporal modeling via feature trajectories and SEQ learning adapts well to varied spatial contexts and activity patterns.

On the role of *time*. Our findings offer new insights into the importance of time in both visual and video recognition. Temporal information is not only crucial for video analysis but also enhances static image recognition when properly used. With simple temporal augmentations and SEQ-based supervision, standard feedforward classifiers, without any architectural modifications, can effectively reason over time. This demonstrates that temporal inductive bias can be introduced through training strategies rather than architectural complexity. Our experiments consistently show performance gains across diverse datasets, from fine-grained image classification to video anomaly detection. These improvements stem from modeling feature trajectories and enforcing temporal consistency, enabling models to capture subtle dynamics and patterns over time. This lightweight approach *challenges the common belief* that temporal reasoning requires heavy sequential models like transformers. Instead, our results emphasize that the way we supervise temporal information, by aligning it with feature evolution, can be equally or even more impactful.

Conclusion

We introduced a simple yet effective training framework that equips standard feedforward classifiers with temporal reasoning through a carefully designed loss function, without altering model architecture. At its core is our SEQ learning, which aligns prediction sequences with temporal prototypes via soft-DTW, further guided by semantic consistency and temporal smoothness objectives. This framework enables lightweight classifiers to model temporal dynamics, yielding robust, temporally consistent predictions for both fine-grained visual recognition and video anomaly detection. By bridging static classifiers and temporal modeling through supervision alone, our method offers an efficient alternative to specialized sequence models.

Acknowledgments

Xi Ding, a visiting scholar at the ARC Research Hub for Driving Farming Productivity and Disease Prevention, Griffith University, conducted this work under the supervision of Lei Wang. Lei Wang proposed the algorithm and developed the theoretical framework, while Xi Ding implemented the code and performed the experiments.

We thank the anonymous reviewers for their invaluable insights and constructive feedback, which have contributed to improving our work.

This work was supported by the Australian Research Council (ARC) under Industrial Transformation Research Hub Grant IH180100002.

This work was also supported by the National Computational Merit Allocation Scheme 2025 (NCMAS 2025; Lead CI: Lei Wang) and the ANU Merit Allocation Scheme (ANUMAS 2025; Lead CI: Lei Wang), with computational resources provided by NCI Australia, an NCRIS-enabled capability supported by the Australian Government.

References

Barz, B.; and Denzler, J. 2020. Deep learning on small datasets without pre-training using cosine loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1371–1380.

Behera, A.; Wharton, Z.; Hewage, P. R.; and Bera, A. 2021. Context-aware attentional pooling (cap) for fine-grained visual classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 929–937.

Bera, A.; Wharton, Z.; Liu, Y.; Bessis, N.; and Behera, A. 2022. SR-GNN: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, 31: 6017–6031.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Chang, D.; Ding, Y.; Xie, J.; Bhunia, A. K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; and Song, Y.-Z. 2020. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29: 4683–4695.

Chen, Q.; Wang, L.; Koniusz, P.; and Gedeon, T. 2024. Motion meets attention: Video motion prompts. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Y.; Bai, Y.; Zhang, W.; and Mei, T. 2019. Destruction and Construction Learning for Fine-Grained Image Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; and Wu, Y.-C. 2023. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(1), 387–395.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Cui, Y.; Song, Y.; Sun, C.; Howard, A.; and Belongie, S. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4109–4118.

Cuturi, M.; and Blondel, M. 2017. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, 894–903. PMLR.

Ding, D.; Wang, L.; Zhu, L.; Gedeon, T.; and Koniusz, P. 2025a. Learnable Expansion of Graph Operators for Multi-Modal Feature Fusion. In *The Thirteenth International Conference on Learning Representations*.

Ding, X.; and Wang, L. 2024. Quo Vadis, Anomaly Detection? LLMs and VLMs in the Spotlight. *arXiv preprint arXiv:2412.18298*.

Ding, X.; and Wang, L. 2025a. Do language models understand time? In *Companion Proceedings of the ACM on Web Conference 2025*, 1855–1868.

Ding, X.; and Wang, L. 2025b. The Journey of Action Recognition. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, 1869–1884. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713316.

Ding, X.; Wang, L.; Koniusz, P.; and Gao, Y. 2025b. Graph Your Own Prompt. *Advances in Neural Information Processing Systems*.

Ding, Y.; Ma, Z.; Wen, S.; Xie, J.; Chang, D.; Si, Z.; Wu, M.; and Ling, H. 2021. AP-CNN: Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-Grained Visual Classification. *IEEE Transactions on Image Processing*, 30: 2826–2836.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Du, R.; Xie, J.; Ma, Z.; Chang, D.; Song, Y.-Z.; and Guo, J. 2022. Progressive Learning of Category-Consistent Multi-Granularity Features for Fine-Grained Visual Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9521–9535.

Ge, W.; and Yu, Y. 2017. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1086–1095.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1705–1714.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning Temporal Regularity in Video Sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 733–742.

He, J.; Chen, J.-N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; and Wang, C. 2022. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 852–860.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Hu, Y.; Jin, X.; Zhang, Y.; Hong, H.; Zhang, J.; He, Y.; and Xue, H. 2021. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the 29th ACM international conference on multimedia*, 4239–4248.

Huang, C.; Kang, Z.; and Wu, H. 2024. A prototype-based neural network for image anomaly detection and localization. *Neural Processing Letters*, 56(3): 169.

Jeong, S.; Park, J.; and Imani, M. 2025. Uncertainty-Weighted Image-Event Multimodal Fusion for Video Anomaly Detection. *arXiv preprint arXiv:2505.02393*.

Jiang, X.; Tang, H.; Gao, J.; Du, X.; He, S.; and Li, Z. 2024. Delving into multimodal prompting for fine-grained visual classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 2570–2578.

Ke, X.; Cai, Y.; Chen, B.; Liu, H.; and Guo, W. 2023. Granularity-aware distillation and structure modeling region proposal network for fine-grained image classification. *Pattern Recognition*, 137: 109305.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO.

Kim, S.; Nam, J.; and Ko, B. C. 2022. ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 11162–11172. PMLR.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, 554–561.

Kuang, H.; Zhu, Y.; Zhang, Z.; Li, X.; Tighe, J.; Schwertfeger, S.; Stachniss, C.; and Li, M. 2021. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3195–3204.

Liu, J.; Song, L.; and Qin, Y. 2020. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 741–756. Springer.

Liu, M.; Zhang, C.; Bai, H.; Zhang, R.; and Zhao, Y. 2022. Cross-Part Learning for Fine-Grained Image Classification. *IEEE Transactions on Image Processing*, 31: 748–758.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, H.; Noh, J.; and Ham, B. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14372–14381.

Peng, Y.; He, X.; and Zhao, J. 2017. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3): 1487–1500.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; and Cui, Y. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6964–6974.

Raj, A.; Wang, L.; and Gedeon, T. 2025. Tracknetv4: Enhancing fast sports object tracking with motion attention maps. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Schiappa, M. C.; Rawat, Y. S.; and Shah, M. 2023. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s): 1–37.

Sermanet, P.; Lynch, C.; Chebotar, Y.; Hsu, J.; Jang, E.; Schaal, S.; Levine, S.; and Brain, G. 2018. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, 1134–1141. IEEE.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Song, K.; Wei, X.-S.; Shu, X.; Song, R.-J.; and Lu, J. 2020. Bi-Modal Progressive Mask Attention for Fine-Grained Recognition. *IEEE Transactions on Image Processing*, 29: 7006–7018.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-Supervised Video Anomaly Detection With Robust Temporal Feature Magnitude Learning.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4975–4986.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Xu, Q.; Jiang, B.; Luo, B.; and Tang, J. 2024a. Multi-Granularity Part Sampling Attention for Fine-Grained Visual Classification. *IEEE Transactions on Image Processing*, 33: 4529–4542.

Wang, L. 2023. *Robust human action modelling*. Ph.D. thesis, The Australian National University (Australia).

Wang, L.; Huynh, D. Q.; and Koniusz, P. 2019. A comparative review of recent kinect-based action recognition algorithms. *IEEE Transactions on Image Processing*, 29: 15–28.

Wang, L.; Huynh, D. Q.; and Mansour, M. R. 2019. Loss switching fusion with similarity search for video classification. In *2019 IEEE international conference on image processing (ICIP)*, 974–978. IEEE.

Wang, L.; and Koniusz, P. 2022a. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Proceedings of the Asian Conference on Computer Vision*, 4176–4193.

Wang, L.; and Koniusz, P. 2022b. Uncertainty-dtw for time series and sequences. In *European Conference on Computer Vision*, 176–195. Springer.

Wang, L.; Liu, J.; Zheng, L.; Gedeon, T.; and Koniusz, P. 2024b. Meet jeanie: a similarity measure for 3d skeleton sequences via temporal-viewpoint alignment. *International Journal of Computer Vision*, 132(9): 4091–4122.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132.

Xu, Q.; Wang, J.; Jiang, B.; and Luo, B. 2023. Fine-Grained Visual Classification via Internal Ensemble Learning Transformer. *IEEE Transactions on Multimedia*, 25: 9015–9028.

Yang, X.; Wang, Y.; Chen, K.; Xu, Y.; and Tian, Y. 2022. Fine-grained object classification via self-supervised pose alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7399–7408.

Yu, X.; Wang, J.; and Gao, Y. 2023. CLE-ViT: contrastive learning encoded transformer for ultra-fine-grained visual categorization. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4531–4539.

Yu, X.; Wang, J.; Zhao, Y.; and Gao, Y. 2023. Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *Pattern Recognition*, 135: 109131.

Yu, X.; Zhao, Y.; and Gao, Y. 2022. SPARE: Self-supervised part erasing for ultra-fine-grained visual categorization. *Pattern Recognition*, 128: 108691.

Yu, X.; Zhao, Y.; Gao, Y.; and Xiong, S. 2021a. MaskCOV: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition*, 119: 108067.

Yu, X.; Zhao, Y.; Gao, Y.; Yuan, X.; and Xiong, S. 2021b. Benchmark Platform for Ultra-Fine-Grained Visual Categorization Beyond Human Performance. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10265–10275.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, L.; Huang, S.; Liu, W.; and Tao, D. 2019. Learning a Mixture of Granularity-Specific Experts for Fine-Grained Categorization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8330–8339.

Zhang, Z.-C.; Chen, Z.-D.; Wang, Y.; Luo, X.; and Xu, X.-S. 2024. A vision transformer for fine-grained classification by reducing noise and enhancing discriminative information. *Pattern Recognition*, 145: 109979.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3769–3777.

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4692–4702.

Zhu, L.; Chen, T.; Yin, J.; See, S.; and Liu, J. 2023. Learning gabor texture features for fine-grained recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1621–1631.

Zhu, L.; Wang, L.; Raj, A.; Gedeon, T.; and Chen, C. 2024. Advancing Video Anomaly Detection: A Concise Review and a New Dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Appendices

A. Video Anomaly Detection

Video anomaly detection aims to identify abnormal events in temporal visual data. Conventional methods include reconstruction-based models (Hasan et al. 2016; Ding and Wang 2024; Liu et al. 2018; Gong et al. 2019) and temporal embedding learning (Ding and Wang 2025b; Park, Noh, and Ham 2020), which often require specialized architectures and are sensitive to irrelevant changes like background motion or lighting (Wang, Huynh, and Mansour 2019; Zhu et al. 2024).

Prototype-based anomaly detection has gained traction by modeling normality via latent prototypes (Huang, Kang, and Wu 2024). Our approach redefines normality in terms of prediction-space dynamics: deviations from class-consistent prediction trajectories indicate anomalies. This results in a lightweight, interpretable model capable of frame-level anomaly detection without complex reconstruction losses or custom temporal encoders.

We select image-pretrained backbones for several compelling reasons. First, they encode rich, generalizable visual representations transferable across diverse tasks. Second, unlike video-pretrained models that may overfit to motion-specific artifacts, ImageNet embeddings remain stable and domain-agnostic. Third, freezing the backbone yields substantial efficiency gains, reducing training time, memory footprint, and energy consumption, while mitigating overfitting risks. This modularity allows the classifier to concentrate exclusively on learning temporal relationships from high-quality visual features.

B. Temporal Inductive Bias in Action

We present a principled and versatile strategy for embedding temporal inductive biases into standard feedforward classifiers. By encouraging models to learn not only class semantics but also the expected temporal evolution of predictions under smooth input variations, our method enables temporal reasoning without architectural modifications or reliance on sequence models. This approach bridges both static and temporal domains, from fine-grained visual recognition enhanced with synthetic dynamics to naturally evolving video streams, all while maintaining simplicity and broad applicability.

Fine-grained image recognition. In static recognition tasks, we simulate temporal progression by applying smoothly varying augmentations to individual images, such as gradual pose shifts, lighting changes, or appearance perturbations. These transformations create synthetic sequences that expose the classifier to the kinds of structured variations that occur in real-world observations. Through this process, the model learns to maintain consistent and confident predictions over these evolving inputs, effectively acquiring robustness to structured perturbations and intra-class variability, key challenges in fine-grained recognition. Importantly, our method captures not only static class identity but also characteristic prediction trajectories over these pseudo-temporal sequences, serving as an implicit form of temporal supervision even in datasets lacking real temporal signals.

Video anomaly detection. For temporal anomaly detection, we use naturally evolving video sequences to model the typical temporal dynamics of normal behavior. Support sets comprising normal sequences are used to construct prototype trajectories, which capture class-consistent temporal evolution in feature space. The model then performs fine-grained, frame-level anomaly detection by identifying deviations from these learned prototypes. This setup naturally supports early anomaly detection: anomalies can be flagged promptly as soon as prediction trajectories deviate from the expected normal patterns. Such responsiveness is critical in applications requiring timely monitoring and intervention, such as industrial inspection, medical monitoring, or security surveillance.

C. Relation to Existing Frameworks

Our SEQ learning adapts and extends concepts from contrastive learning, few-shot learning, and metric learning to provide a new framework for fine-grained visual recognition and detecting anomalous patterns in sequential data.

Contrastive learning framework. Our SEQ learning framework draws inspiration from contrastive learning, which learns robust representations by encouraging similarity between positive pairs and dissimilarity between negatives (Chen et al. 2020; Kuang et al. 2021; Oord, Li, and Vinyals 2018). However, unlike conventional instance-level contrastive learning, SEQ operates at the sequence level, using temporal alignment rather than pointwise distance. It introduces an implicit temporal inductive bias by aligning prediction trajectories, promoting consistency across samples that share underlying class semantics and temporal evolution.

Remark 1. *In SEQ learning, positive pairs are sequences from the same class, such as normal videos in anomaly detection or temporally-augmented samples from the same fine-grained category. The model minimizes alignment cost between these sequences, learning class-consistent temporal dynamics without requiring negative sampling.*

Crucially, SEQ avoids explicit negative pairs and instead focuses on capturing intra-class temporal consistency. In fine-grained recognition, it models the subtle trajectory shifts introduced by smooth temporal augmentations, enabling static classifiers to learn temporal structure. In video anomaly detection, it captures the diverse but coherent normal behaviors, improving robustness and enabling early anomaly detection. This class-conditional sequence-level supervision offers a lightweight yet powerful alternative to conventional contrastive or recurrent models, especially in domains where temporal labels are scarce.

Few-shot learning framework. Our training strategy adopts an episodic structure reminiscent of few-shot learning, where each episode includes a query sequence and a support set drawn from the same class. This design encourages the model to generalize from limited supervision by learning class-specific patterns from few examples.

Remark 2. *Unlike standard few-shot learning, SEQ learning dynamically samples diverse support instances per episode, forming class prototypes that capture intra-class variation. In fine-grained image recognition, support sets are sampled from temporally augmented views of the same class, fostering rich, class-conditional dynamics. For video anomaly detection, the model aligns anomalous query sequences with normal support sets to detect deviations, using the exemplar as a learned prototype of normal behavior.*

This episodic formulation allows the model to build robust temporal prototypes from few but diverse samples, crucial in scenarios like anomaly detection, where abnormal events are rare and hard to enumerate (Zhu et al. 2024). By training to generalize from limited examples and learning deviations from normality, our approach naturally fits the few-shot paradigm while using temporal alignment for stronger inductive bias.

Metric learning framework. At the heart of our approach lies a differentiable metric, γ-soft-DTW (Cuturi and Blondel 2017), which enables alignment-based comparison of sequences with varying lengths.

Remark 3. *Our SEQ framework uses metric learning to minimize alignment cost between positive sequence pairs. In fine-grained recognition, queries are pulled closer to their class prototypes. In anomaly detection, sequences with normal behavior are aligned to form consistent, low-distance pairs.*

Unlike traditional DTW, γ-soft-DTW is fully differentiable, enabling end-to-end optimization. It not only guides training via smooth gradients but also yields interpretable alignment matrices, revealing well-matched (or misaligned) temporal segments, useful for both classification and anomaly localization.

By embedding sequence alignment within a metric learning framework, SEQ effectively captures temporal similarity, offering a general and lightweight solution for tasks with limited labeled data.

D. Proof of Smoothness Preservation

We formally prove that temporal smoothness in the feature space is preserved when the features are passed through a single fully connected (FC) layer followed by a Softmax or Sigmoid activation. This ensures that temporally adjacent features yield prediction scores that do not fluctuate abruptly, which is crucial for tasks such as video anomaly detection and temporally consistent fine-grained recognition.

Let $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_\tau] \in \mathbb{R}^{\tau\times d}$ denote a sequence of d-dimensional feature vectors ordered in time, where τ is the temporal length of the sequence. We assume that the feature sequence is temporally smooth in the ℓ_2 norm:

$$\|\boldsymbol{z}_{i+1} - \boldsymbol{z}_i\|_2 \le \epsilon, \quad \forall i = 1, \ldots, \tau-1, \tag{16}$$

for a small constant $\epsilon > 0$.

Let the classifier be defined as:

$$\boldsymbol{\phi}_i = \sigma(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}), \tag{17}$$

where $\boldsymbol{W} \in \mathbb{R}^{C\times d}$ is the weight matrix, $\boldsymbol{b} \in \mathbb{R}^C$ is the bias vector, $\boldsymbol{\phi}_i \in \mathbb{R}^C$ is the output vector of class scores (probabilities), and $\sigma : \mathbb{R}^C \to \mathbb{R}^C$ is either the Softmax or Sigmoid activation.

We aim to show that the prediction score sequence $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\tau]$ inherits the temporal smoothness of the feature sequence $\boldsymbol{Z}$. That is, we want to prove that:

$$\|\boldsymbol{\phi}_{i+1} - \boldsymbol{\phi}_i\|_2 \le \delta, \quad \forall i = 1, \ldots, \tau-1, \tag{18}$$

for some small $\delta > 0$ that depends on ϵ and the classifier parameters.

To establish this, we proceed with the following steps.

Step 1: linearity of the fully connected layer.

The linear transformation $g(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$ is Lipschitz continuous with Lipschitz constant $\|\boldsymbol{W}\|_2$, where $\|\boldsymbol{W}\|_2$ denotes the spectral norm of $\boldsymbol{W}$:

$$\|g(\boldsymbol{z}_{i+1}) - g(\boldsymbol{z}_i)\|_2 = \|\boldsymbol{W}(\boldsymbol{z}_{i+1} - \boldsymbol{z}_i)\|_2 \le \|\boldsymbol{W}\|_2 \cdot \|\boldsymbol{z}_{i+1} - \boldsymbol{z}_i\|_2. \tag{19}$$

Since $\|\boldsymbol{z}_{i+1} - \boldsymbol{z}_i\|_2 \le \epsilon$, it follows that:

$$\|g(\boldsymbol{z}_{i+1}) - g(\boldsymbol{z}_i)\|_2 \le \|\boldsymbol{W}\|_2 \cdot \epsilon. \tag{20}$$

Let us define $\boldsymbol{u}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b} \in \mathbb{R}^C$, so $\boldsymbol{\phi}_i = \sigma(\boldsymbol{u}_i)$. We now analyze the effect of applying the activation σ to the vector $\boldsymbol{u}_i$.

Step 2: smoothness of the activation function.

Let us consider both Softmax and Sigmoid separately.

Case 1: Sigmoid. The Sigmoid function applied element-wise is known to be $\frac{1}{4}$-Lipschitz. For any two inputs $\boldsymbol{u}_i, \boldsymbol{u}_{i+1} \in \mathbb{R}^C$:

$$\|\sigma(\boldsymbol{u}_{i+1}) - \sigma(\boldsymbol{u}_i)\|_2 \le \frac{1}{4} \cdot \|\boldsymbol{u}_{i+1} - \boldsymbol{u}_i\|_2. \tag{21}$$

Combining with the result of Step 1:

$$\|\boldsymbol{\phi}_{i+1} - \boldsymbol{\phi}_i\|_2 \le \frac{1}{4} \cdot \|\boldsymbol{W}\|_2 \cdot \epsilon. \tag{22}$$

Case 2: Softmax. The Softmax function is also Lipschitz continuous under ℓ_2 norm. It is known that:

$$\|\text{Softmax}(\boldsymbol{u}_{i+1}) - \text{Softmax}(\boldsymbol{u}_i)\|_2 \le L_{\text{softmax}} \cdot \|\boldsymbol{u}_{i+1} - \boldsymbol{u}_i\|_2, \tag{23}$$

where $L_{\text{softmax}} \le 1$ and depends on the norm and input scale. Therefore:

$$\|\boldsymbol{\phi}_{i+1} - \boldsymbol{\phi}_i\|_2 \le L_{\text{softmax}} \cdot \|\boldsymbol{W}\|_2 \cdot \epsilon. \tag{24}$$

Step 3: conclusion.

In both cases, the final inequality takes the form:

$$\|\boldsymbol{\phi}_{i+1} - \boldsymbol{\phi}_i\|_2 \le L \cdot \epsilon, \tag{25}$$

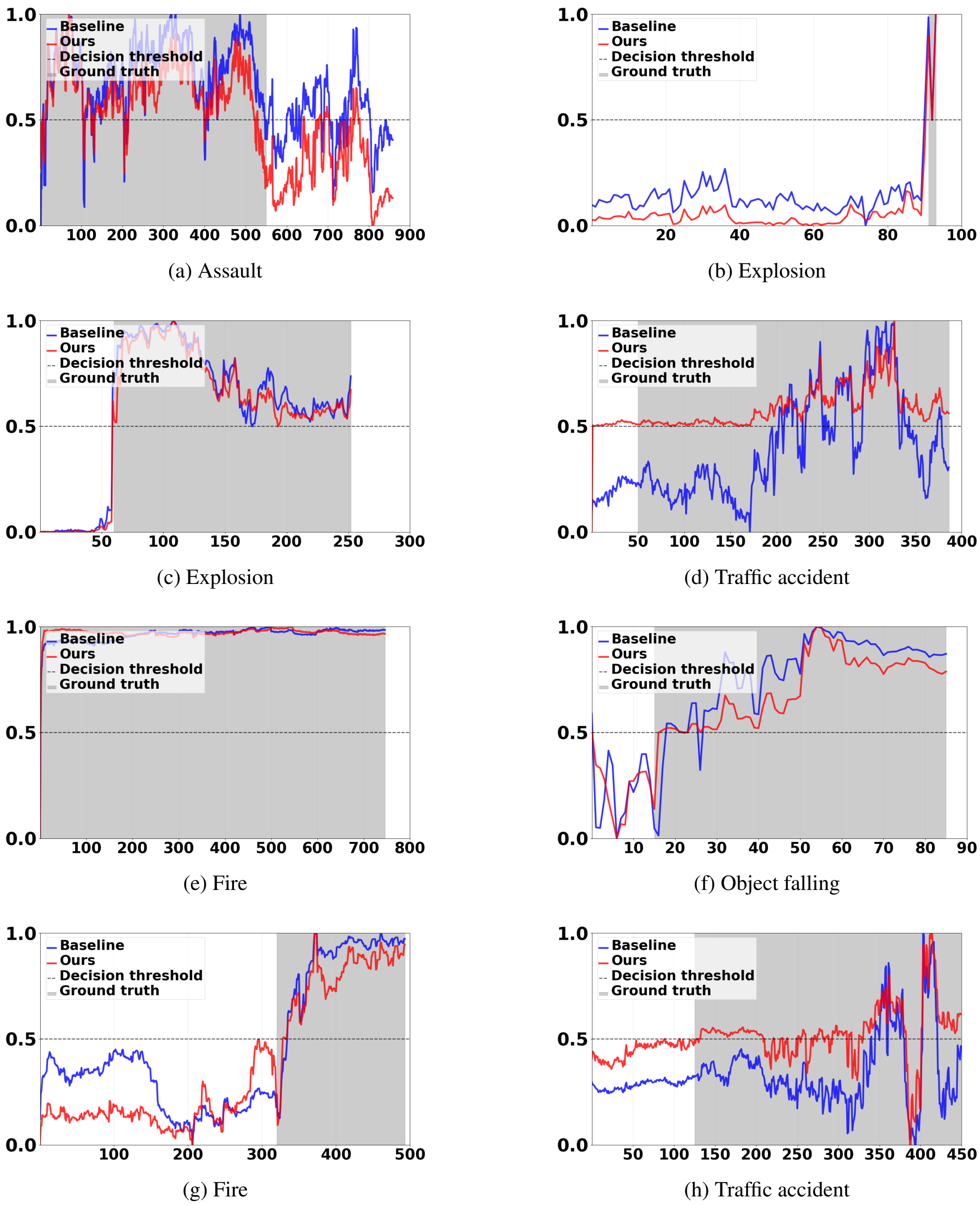
where $L = \frac{1}{4}\|\boldsymbol{W}\|_2$ for Sigmoid activation, and $L = L_{\text{softmax}} \cdot \|\boldsymbol{W}\|_2$ for Softmax activation.

This proves that the transition from temporally smooth features to prediction scores via a fully connected classifier followed by a smooth activation preserves temporal smoothness. That is, the classifier does not introduce abrupt changes in the predicted scores when the input features change gradually over time. This property is critical for tasks involving temporal modeling, where consistency across frames or clips must be maintained.

Figure 7: Examples from Flowers-102, SoyAging, Stanford Dogs, and Cars show how augmentations create temporal variations from one image. The first column shows originals (green); others apply augmentations by color: flip (red), zoom (blue), rotation (purple), color jitter (orange), shear (brown), translation (pink), blur (gray), and cutout (cyan), enriching the feature space with varied appearances.

Figure 7: (continued) Examples from Flowers-102, SoyAging, Stanford Dogs, and Cars show how augmentations create temporal variations from one image. The first column shows originals (green); others apply augmentations by color: flip (red), zoom (blue), rotation (purple), color jitter (orange), shear (brown), translation (pink), blur (gray), and cutout (cyan), enriching the feature space with varied appearances.

Figure 7: (continued) Examples from Flowers-102, SoyAging, Stanford Dogs, and Cars show how augmentations create temporal variations from one image. The first column shows originals (green); others apply augmentations by color: flip (red), zoom (blue), rotation (purple), color jitter (orange), shear (brown), translation (pink), blur (gray), and cutout (cyan), enriching the feature space with varied appearances.



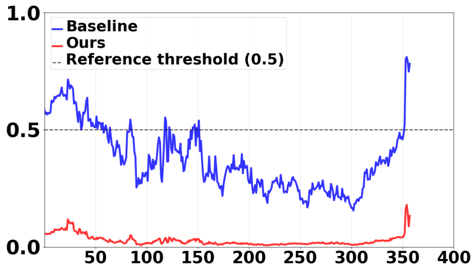
(a) Assault (b) Explosion

(c) Explosion (d) Traffic accident

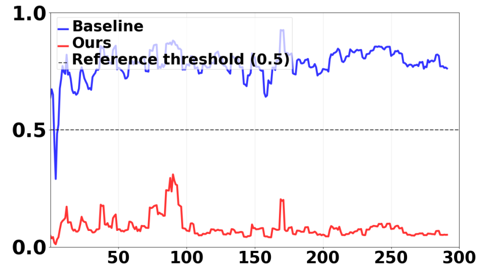
(e) Fire (f) Object falling

(g) Fire (h) Traffic accident

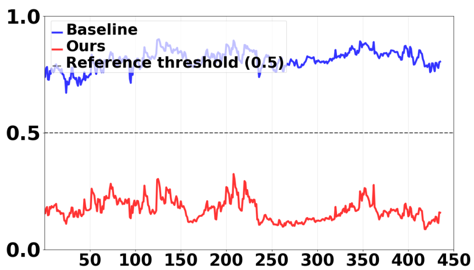
Figure 8: Anomaly prediction comparison. Grey regions indicate ground-truth anomalies. Blue and red curves show the baseline and our method. Our approach detects anomalies more accurately and earlier, with scores crossing the 0.5 threshold in closer alignment with the ground truth.



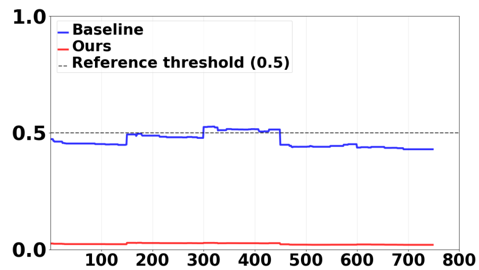
(a) Road



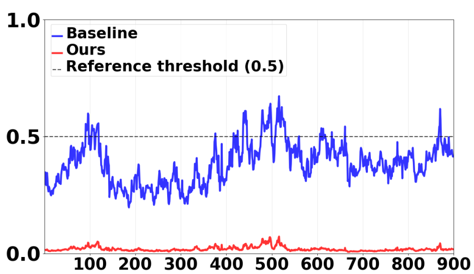
(b) Train



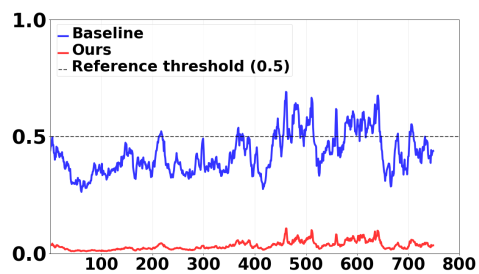
(c) Street highview



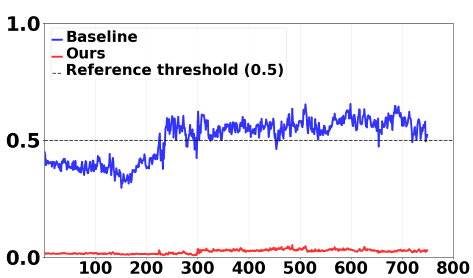
(d) Sidewalk



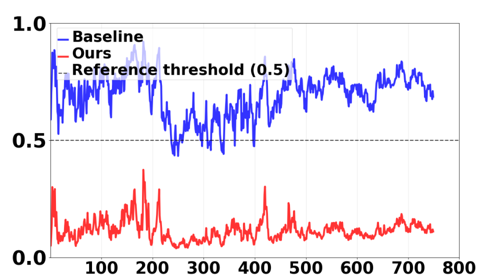
(e) Train



(f) Sidewalk



(g) Road



(h) Sidewalk

Figure 9: Anomaly prediction comparison. In these normal scenarios, our methods correctly avoid false positives, while the baseline method incorrectly flags them as anomalies.

E. Additional Visualizations

E.1. Additional Smooth Temporal Augmentations Below (Fig. 7), we provide additional visualizations showcasing our smooth temporal augmentations applied to fine-grained and ultra-fine-grained image datasets, including Flowers-102, SoyAging, Stanford Dogs, and Stanford Cars. Each example starts with the original image (green, first column), followed by a sequence of augmented variants: horizontal flip (red), zoom (blue), rotation (purple), color jitter (orange), shear (brown), translation (pink), blur (gray), and cutout (cyan). These augmentations introduce realistic temporal variations from a single static image, enriching the feature space with diverse yet semantically consistent appearances, crucial for modeling temporal dynamics in the absence of natural video sequences.

E.2. Additional Video Anomaly Detection Visualizations Below (Fig. 8 and 9), we present additional visualizations of video anomaly detection results on the MSAD dataset. These examples demonstrate that our method detects anomalies not only with greater accuracy but also at earlier time steps compared to competing approaches, highlighting its effectiveness in capturing subtle temporal deviations.

F. Additional Discussions

This section provides further explanation of several aspects of SEQ, including its generalization behaviour, computational characteristics, temporal augmentation design, and the role of backbone choices and ablations.

Generalizability. SEQ models temporal structure by learning class-level temporal distributions rather than memorizing fixed temporal patterns. During training, each episode samples diverse support sequences whose augmentation parameters evolve smoothly over time. The resulting exemplars act as barycenters of multiple trajectories, capturing characteristic prediction-space evolution for each class. This episodic diversity encourages the model to form robust temporal prototypes that generalize across input variations. Moreover, the FC+Softmax mapping preserves trajectory smoothness, allowing temporal continuity to be maintained throughout the prediction sequence.

Complexity. Although Soft-DTW-based alignment is used during training, it operates on $\tau \times \tau'$ matrices of class probability vectors and thus has moderate computational cost. In the image domains considered in this work, $\tau \leq 5$ and the support sets are small, making alignment efficient and lightweight. Importantly, SEQ introduces no additional cost at inference time: the classifier reduces to a single fully connected layer identical to the static baseline, with no need for alignment or temporal matching. This preserves the architectural simplicity that motivates the framework and differentiates it from recurrent or transformer-based temporal models.

Augmentation. The temporal augmentation strategy is designed to impose a temporal continuity bias rather than simulate physically accurate motion. Augmentation parameters such as rotation, translation, or color intensity evolve linearly across virtual timesteps, ensuring that the induced variation is smooth and coherent. Applying identical augmentation schedules to both support and query sequences prevents artificial discrepancies and ensures that the learning signal is governed by feature-level temporal evolution rather than augmentation artifacts. This provides a controlled setting in which the classifier can learn trajectory consistency while remaining agnostic to the exact nature of image-level transformations. Extensions based on generative temporal augmentations or video test-time adaptation represent promising avenues for future work.

Backbone fairness. All comparisons within each dataset use the same frozen backbone for both the baseline and the temporal variants. This ensures that improvements arise solely from modeling temporal structure rather than from differences in representation quality. Different datasets use different backbones only to match established conventions in prior work and dataset-specific domain characteristics (*e.g.*, CLIP-ViT for Cars, ViT-B/16 for Dogs, CLE-ViT (Swin-B/448, IN-21K) for ultra-fine SoyAging). Experiments using a unified backbone across Cars and Dogs demonstrate that the relative gains of SEQ remain stable, indicating backbone-invariance of the temporal modeling benefits.

Ablations & statistics. Ablation studies emphasize the complementary contributions of the different loss components. Removing any single loss term leads to a decrease in accuracy. For example, on the Flowers-102 dataset, we observe the following results: Baseline 97.5 ± 0.5, without $\mathcal{L}_{\text{align}}$ 97.6 ± 0.4, without $\mathcal{L}_{\text{smooth}}$ 98.0 ± 0.3, without $\mathcal{L}_{\text{CE}}$ 96.8 ± 0.6, and the full SEQ objective 98.4 ± 0.3 ($p < 0.01$ across tasks).

Specifically, removing alignment, smoothness, or semantic supervision consistently degrades performance across datasets. The complete objective, which integrates temporal prototype alignment, semantic consistency, and smooth temporal evolution, produces the most coherent prediction trajectories and achieves the highest recognition accuracy, highlighting the importance of combining all components.