Dynamic Distinction Learning: Adaptive Pseudo Anomalies for Video Anomaly Detection

Demetris Lappas
k1838447@kingston.ac.uk

Vasileios Argyriou
vasileios.argyriou@kingston.ac.uk

Dimitrios Makris
d.makris@kingston.ac.uk

Kingston University, London, UK,
School of Computer Science and Mathematics

Abstract

We introduce Dynamic Distinction Learning (DDL) for Video Anomaly Detection, a novel video anomaly detection methodology that combines pseudo-anomalies, dynamic anomaly weighting, and a distinction loss function to improve detection accuracy. By training on pseudo-anomalies, our approach adapts to the variability of normal and anomalous behaviors without fixed anomaly thresholds. Our model showcases superior performance on the Ped2, Avenue and ShanghaiTech datasets, where individual models are tailored for each scene. These achievements highlight DDL's effectiveness in advancing anomaly detection, offering a scalable and adaptable solution for video surveillance challenges. Our work can be found on: https://github.com/demetrislappas/DDL.git

1. Introduction

Anomaly detection is pivotal in the field of video surveillance, where algorithms scan through endless hours of footage to identify activities or events that deviate from the norm—be it unauthorized intrusions, unusual behavior, or safety breaches. Its application in video analysis is indispensable across a multitude of sectors, underpinning security protocols, ensuring public safety, and enhancing operational efficiency. The capacity of video anomaly detection systems to flag deviations in real-time or in hindsight allows organizations to take quick, informed action to mitigate risks.

Nevertheless, the task of distinguishing the ordinary from the extraordinary in videos is exceptionally challenging. Video anomaly detection typically lives within the domain of unsupervised learning due to the inherent scarcity of labeled anomalies and the impracticality of cataloging the large array of possible anomalous events. The unpredictable nature of anomalies further adds to the complexity, making it difficult for models trained on 'normal' behavior to generalize and identify outliers effectively. This difficulty is magnified by the context-sensitive definition of what constitutes an anomaly within video data, as it can vary significantly from one setting to another. In the absence of sufficient examples of anomalous behavior during training, systems often struggle to accurately discern anomalies when they do occur, resulting in a high number of false positives or missed detections.

Traditional approaches to this challenge have relied on neural network architectures like AutoEncoders and UNets [7–10, 12, 16, 17, 23, 28, 30, 31, 33, 35, 37, 38, 40]. These models are trained to recreate 'normality' by learning to compress and then reconstruct input data with minimal loss. The underlying premise is that, by becoming adept at reconstructing normality, these networks would inherently struggle when faced with anomalies, thus allowing for their detection. However, there lies a catch—these systems do not necessarily learn an explicit distinction between normal and anomalous samples, it is only *hoped* that anomalies will pose a greater challenge for the reconstruction process.

To address this predicament, various methodologies have introduced pseudo-anomalies during the training phase, offering models a taste of the 'abnormal' to foster learning [1–4, 24, 41]. These strategies, however, often overlook a critical aspect: the quantification of the 'right level' of pseudo-anomaly. That is, how anomalous should pseudo-anomalies be to represent real anomalies? Too small, and the pseudo-anomalies bare too close a resemblance to normal data; too high, and the model may fail to recognize genuine, more subtle, anomalies.

In our work, the innovation lies not just in the incorporation of pseudo-anomalies, but in the strategic introduction of a dynamic anomaly weight $\sigma(\ell)$. This adapt-

ability is crucial, allowing our model the flexibility to discover the optimal threshold of anomaly intensity for effective learning. Rather than being constrained to a predetermined, static level of pseudo-anomaly — which might risk the model's overfitting to artificial quirks — the dynamic nature of $\sigma(\ell)$ entrusts the model with the autonomy to fine-tune this threshold. By doing so, the model is trained to differentiate between normal and anomalous patterns without being anchored to any specific level of anomaly defined by the user.

Our work also introduces the Distinction Loss, which works in tandem with $\sigma(\ell)$, and is crafted to refine the model's discrimination capabilities. The Distinction Loss encourages the model to rebuild pseudo-anomalous frames to more closely resemble the normal state rather than the inputted anomalous one.

In the forthcoming chapters, we delve into the core of our research on Dynamic Distinction Learning (DDL) for video anomaly detection. We first begin by providing a brief overview of related work in Chapter 2. Chapter 3 outlines the methodology, detailing the DDL framework and its components. Chapter 4 describes the datasets considered for evaluation, leading to Chapter 5, which presents our findings through quantitative results. In our final chapter, Chapter 6, we provide ablation studies, highlighting the improvements offered by our work.

2. Related Work

The challenge of anomaly detection in video data is exacerbated by the predominance of normal behavior within datasets, leading to an inherent bias towards non-anomalous examples. Unsupervised learning, particularly through the use of AutoEncoders (AEs), has emerged as a preferred solution. AEs leverage the discrepancy between input and reconstructed output to identify anomalies, operating under the principle that unfamiliar anomalous inputs will result in significant reconstruction errors [7–10, 12, 28, 30, 33]. However, the challenge of accurately reconstructing normal samples to distinguish them from anomalies remains, with UNets and their skip connections offering a partial solution by improving reconstruction fidelity, albeit complicating the reliance on the latent space for anomaly detection [16, 17, 23, 31, 35, 37, 38, 40].

Recent advancements have explored the temporal dimension of video anomaly detection, employing AEs and UNets to reconstruct sequences or predict subsequent frames, under the hypothesis that anomalies will disrupt the model's ability to accurately predict future frames based on a sequence of normal frames [8, 12, 16, 17, 22, 23, 31, 33, 35, 37, 38, 40]. The integration of Transformers and attention mechanisms aims to capture the temporal characteristics of video data more effectively, enabling AutoEncoders and UNets to identify anomalies by focusing on the relationships between frames [12, 18, 36, 40]. Optical Flow has been utilized to enhance motion-related anomaly detection, providing a compact yet informative representation of temporal changes by capturing pixel motion between consecutive frames [5, 8, 33, 37, 39].

To improve the performance of AEs and UNets, some studies have incorporated supervised learning techniques, Generative Adversarial Networks (GANs), and Object Detection to refine the distinction between normal and anomalous samples. GANs, in particular, create a generative-discriminative adversarial relationship that enhances the model's ability to reconstruct outputs indistinguishable from the original input [12, 19, 31, 33, 37, 40]. Object Detection focuses the anomaly detection process on significant frame objects, albeit limited by the detection model's scope and accuracy [7, 11, 12, 19, 31, 31, 33, 34, 37, 40]. Memory modules have also been proposed to prevent anomaly reconstruction by referencing normal samples, suggesting enhanced model complexity as a pathway to more effective anomaly detection [7, 14, 23, 28, 30, 34, 35, 39].

The imbalance between normal and anomalous samples in datasets has necessitated the development of innovative approaches that introduce pseudo-anomalies. These methods are designed to enhance the capability of reconstruction-based models to distinguish between normal and anomalous samples with greater precision. Techniques for generating pseudo-anomalies vary widely, some strategies involve the use of external datasets to inject anomalies into a dataset of normal samples. This can involve leveraging attention mechanisms to identify and transfer key features from third-party datasets to normal samples, thus creating pseudo-anomalies [1], or introducing noise into the latent space of models using external data [24]. Other methods consider a more creative approach, which utilize the previous state of the model to generate lower quality reconstructions which would be represented as anomalous samples [41]. More traditional approaches attempt to invoke abnormality during training by directly providing the model with human defined anomalous behavior, such as reversing the sequence of input frames [2]. More recent pseudo anomalous methods attempt to attain superior results by injecting a suite of human defined anomalies, including the manipulation of video frames by reversing their sequence, skipping frames, adding noise, fusing frames, or incorporating random patches [3, 4]. Anomalies, regardless of their specific nature (skipping frames, repeating frames, introducing extraneous shapes, etc), are perceived by convolutional layer kernels as unusual collections of vector representations—noise. Despite their efficacy, these methods rely on manual intervention to simulate anomalies, requiring a subjective determination of the degree of anomaly introduced—raising the question, "What constitutes the appropriate level of noise to be considered anomalous?"

Against this backdrop, our research introduces a sophisticated approach that not only incorporates the concept of dynamic anomaly weighting but also presents a novel distinction loss function. This methodology aims to advance the anomaly detection domain by providing a more refined mechanism for distinguishing between normal and anomalous events, thereby segueing into the detailed explanation of our proposed methodology outlined in Section 3.

3. Methodology

The Dynamic Distinction Learning (DDL) architecture is outlined in Figure 1. Consider a sequence of normal video frames represented as a tensor $X \in \mathbb{R}^{c \times T \times H \times W}$, where c is the number of channels, T is the number of frames (which must be an odd number, as we will be reconstructing the middle frame), and H and W are the height and width of the frames respectively. To simulate anomalies over the sequence, we pass the model through an Object Detection and Tracking model, followed by Random Object Masking - which selects a random tracked object across all frames and returns a sequence of binary masks $M \in \{0,1\}^{c \times T \times H \times W}$ delineating the regions of the frames where the pseudo-anomaly will be present. Alongside, we also introduce a noise tensor $A \in \mathbb{R}^{c \times T \times H \times W}$, which is uniformly random generated.

We also define a trainable parameter $\ell \in \mathbb{R}$, which is passed through a sigmoid function, $\sigma(\ell) \in (0,1)$, to represent the anomaly weight. We chose a sigmoid function to bound the trainable parameter between the values of 0 and 1, so to portray a percentage of anomaly inflicted. The sequence of normal frames X, the masks M, the noise tensor A, and the anomaly weight $\sigma(\ell)$ are passed into the Pseudo Anomaly Creator to fabricate pseudo anomalies X_A.

Both the sequence of normal frames X, and the pseudo anomalies X_A, are passed through a reconstruction model and calibrated using a linear combination of the Reconstruction Loss and Distinction Loss. The anomaly weight is heavily calibrated by the Distinction Loss, a loss function designed to converge the anomaly weight to represent the minimum anomaly capable of being detected. The adaptability of the anomaly weight $\sigma(\ell)$ allows the model to dynamically calibrate the degree of anomaly present in the training data, ensuring an effective balance between the recognition of normal patterns and the detection of deviations. This is critical for preventing the model from either becoming desensitized to subtle anomalies or overreacting to minor irregularities, thus maintaining a nuanced representation of what constitutes an anomaly throughout the training process.

3.1. Pseudo Anomaly Creator

Our approach to fabricating pseudo-anomalies within video sequences begins with the application of object detection and tracking at each frame, then randomly selecting an object from the set of tracked objects for masking. We employ object tracking to consistently mask the same object across all frames within a temporal window, T. These masks, denoted as M, are crucial in defining the regions for anomaly simulation, ensuring the anomalies are contextually integrated around objects.

Following the identification and masking of objects, we proceed to the creation of pseudo-anomalies, via the Pseudo Anomaly Creator, a two-step noise integration process shown in Figure 2. Initially, we generate noise-infused frames, $X_{\bar{A}}$, by blending the original input frames, X, with a noise tensor, A, using the dynamically learned anomaly weight, $\sigma(\ell)$. This blend is achieved through a linear combination, ensuring the proportionate integration of noise and original content as per the following equation:

$$X_{\bar{A}} = (1 - \sigma(\ell)) \cdot X + \sigma(\ell) \cdot A \tag{1}$$

Here, the element-wise multiplication $(\cdot)$ facilitates the precise control over the extent of noise addition, allowing for variable distortion levels that are directly influenced by the anomaly weight, which evolves during the training phase.

The subsequent phase involves the formulation of the pseudo-anomalous frames, X_A. These frames emerge from overlaying the noise-infused frames, $X_{\bar{A}}$, onto the original input frames, X, strictly within the boundaries defined by the object masks M. The mathematical representation of this process is captured by:

$$X_A = (1 - M) \cdot X + M \cdot X_{\bar{A}} \tag{2}$$

Through this method, we ensure that the noise, symbolizing potential anomalies, is selectively applied to the areas of interest - those being the detected objects within the frame. This approach not only maintains the contextual relevance of the introduced anomalies but also simulates a variety of anomalous patterns by leveraging the variability in noise composition; we elaborate on this in Section 7.2 within the Supplementary Material. By focusing on object regions, our method aims to create realistic and pertinent anomalies, enhancing the model's ability to detect and learn from these fabricated irregularities, which are designed to mimic a diverse spectrum of anomalous behaviors and appearances, including unseen shapes and uncommon motion blurs.

3.2. Reconstruction Model Definition

We define a reconstruction model $f = \mathcal{E} \circ \mathcal{D}$, where $\mathcal{E}$ and $\mathcal{D}$ represent some encoder and decoder parts of a deep learning architecture, respectively. The choice of architecture is flexible and can include, but is not limited to, AutoEncoders,

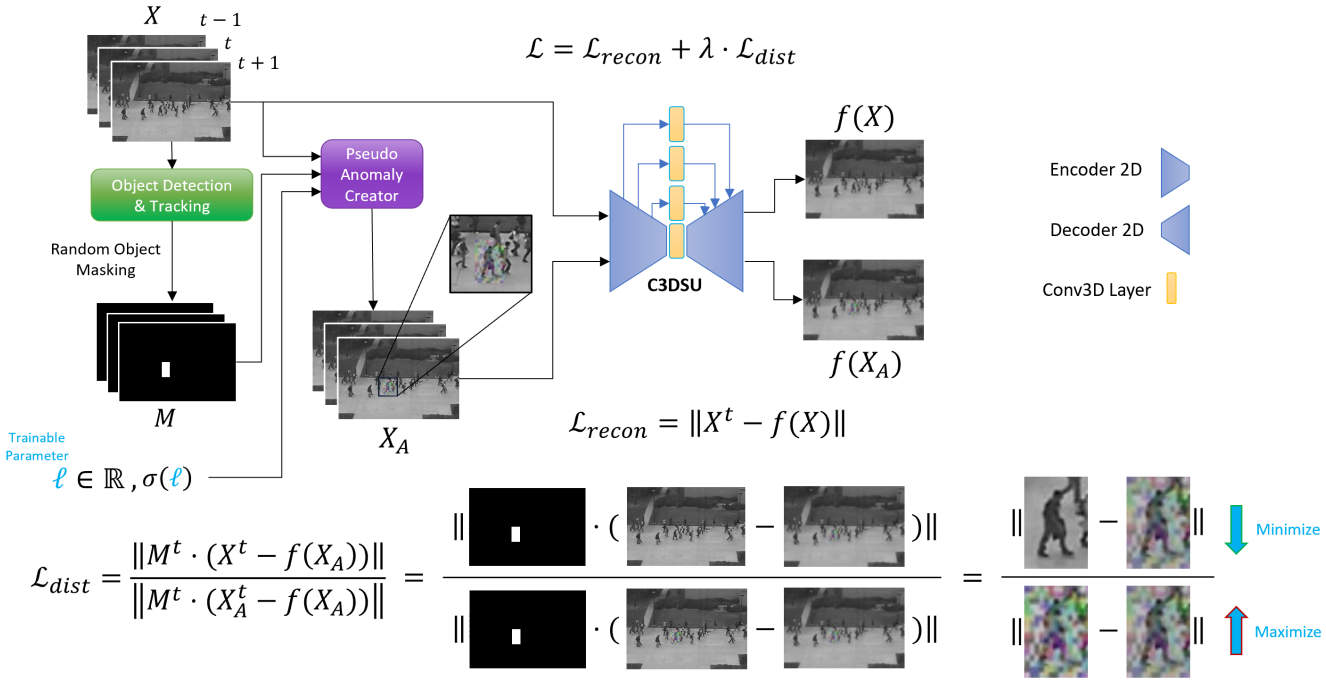

Figure 1. The Dynamic Distinction Learning (DDL) Architecture: This diagram illustrates the DDL model's workflow, including object detection and tracking, random object masking, pseudo anomaly creation, our C3DSU model and the distinction loss calculation. The architecture depicts how the pseudo anomalies are created, then passed through the model along with their normal counter parts. The diagram also provides a visual depiction of the distinction loss calculation, showing how the model learns to minimize the numerator and maximize the denominator.

UNet structures, or other suitable convolutional neural networks designed for video reconstruction.

In practice we employ an adaptation of a 2D UNet model, tailored for the analysis of temporal data through the integration of 3D convolutional layers between skip connections. We call this architecture a *Conv3DSkipUNet* (C3DSU or f for the context of this work); more detail of our architecture can be found in Section 7.3 within the Supplementary Material. The model receives an input such as X and returns a reconstructed output $f(X) \in \mathbb{R}^{c \times t \times H \times W}$, where t represents the middle frame in T; that is, the model receives an odd sequence of frames as an input and returns the reconstructed middle frame.

3.3. Loss Function

We define a loss function, $\mathcal{L}$, which integrates the standard reconstruction loss with our novel distinction loss to fine-tune the model's sensitivity to anomalies.

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda \cdot \mathcal{L}_{dist} \tag{3}$$

where λ is a hyperparameter that modulates the impact of the distinction loss relative to the reconstruction loss. This adjustment is crucial for ensuring that the model effectively balances learning to reconstruct normal frames while also distinguishing them from pseudo-anomalous frames.

3.3.1 Reconstruction Loss

The first function is the standard reconstruction loss:

$$\mathcal{L}_{recon} = \|X^t - f(X)\| \tag{4}$$

where X^t is the middle frame of X. This loss function encourages the model to accurately reconstruct the normal input frame, thus learning the distribution of normal frames.

3.3.2 Distinction Loss

The distinction loss is the second loss function in our model, designed to fine-tune the distinction between normal frames and their pseudo-anomalous counterparts. This differentiation is crucial for the model to recognize and identify anomalies effectively. The distinction loss function is articulated through the following mathematical formulations:

$$P = \|M^t \cdot (X^t - f(X_A))\| \tag{5}$$

$$N = \|M^t \cdot (X_A^t - f(X_A))\| \tag{6}$$

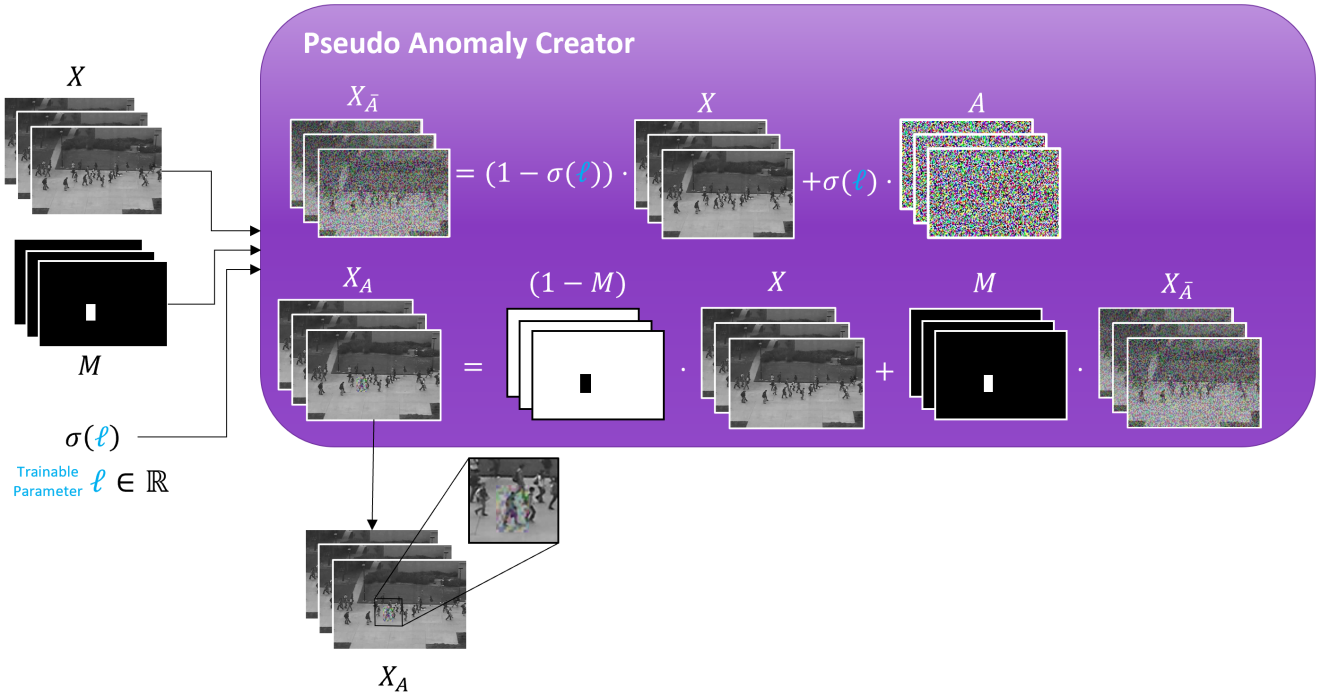

Figure 2. Pseudo-Anomaly Creation Process: This figure demonstrates the step-by-step procedure for generating pseudo-anomalies within video frames. It begins by receiving the normal input frames, the masked frames, and a dynamically learned anomaly weight followed by the application of a noise tensor modulated by the anomaly weight.

$$\mathcal{L}_{dist} = \frac{P + \epsilon}{N + \epsilon} \tag{7}$$

Here, P serves to penalize the differences between the original normal frame X^t and the model's reconstruction of the pseudo-anomalous frame X_A^t within the masked anomalous regions. The term N captures the reconstruction error when the model tries to reconstruct the pseudo-anomalous frame within these same regions. The parameter ϵ is a small constant to prevent division by zero, thus ensuring numerical stability.

The essence of this loss function is to compel the model to prefer transforming pseudo-anomalous frames back into their normal state. In simpler terms, when the model encounters an anomalous frame, the goal is for its reconstructed output to bear a closer resemblance to a normal frame rather than retaining the anomalous characteristics. Though this is a hopeful outcome for any standard reconstruction model, the distinction loss explicitly trains the model to target this outcome, evidence of this is shown in Supplmentary Material, Section 7.4 within Figures 7 and 8.

The underlying intuition of the distinction loss $\mathcal{L}_{dist}$ is to foster a reconstruction process that pulls the pseudo-anomalous frame towards the normal frame more than it does towards itself. This is achieved by aiming to reduce

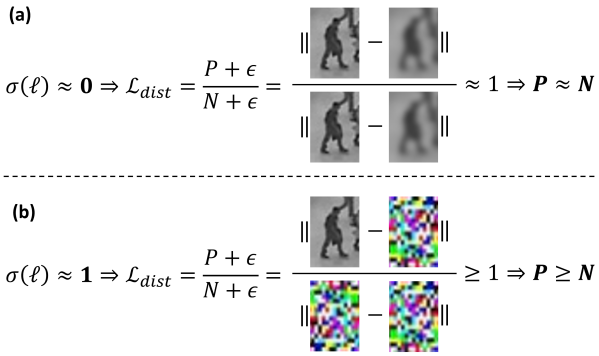

Figure 3. Panel (a) depicts a scenario where $\sigma(\ell)$ approaches zero, leading to minimal deviation from the original frame and challenging the model's ability to distinguish between normal and anomalous regions due to the lack of significant noise. Panel (b) illustrates the opposite extreme, where $\sigma(\ell)$ is near one, resulting in an overly distorted anomalous region dominated by noise, which challenges the model's reconstruction capabilities and undermines the distinction loss's effectiveness.

P—the difference between the normal frame and its reconstruction from a pseudo-anomalous input—and to increase N—the difference between the pseudo-anomalous

frame and its reconstruction. By doing so, the model is incentivized to differentiate between normal and anomalous frames, thereby enhancing its anomaly detection capabilities. This approach contrasts with methodologies employed by our competitors [2, 3], who focus on maximizing the discrepancy between pseudo anomalous inputs and their reconstructions. Such a strategy often results in the emergence of unusual patches within the reconstructed images, a side effect not observed with our model. In contrast, the distinction loss aims to transform pseudo anomalies to resemble normalcy. A visual representation of the distinction loss can be seen in Figure 1.

For the model's reconstruction function f, the ideal scenario is to replicate the normal regions with high fidelity while transforming the anomalous regions towards normalcy. This ability is reflected in the dynamics of P and N:

- A lower P indicates the model's proficiency in reconstructing the normal aspects of a frame, even when presented with a pseudo-anomalous input.
- A higher N indicates that the model is not simply replicating the anomalous features present in the pseudo-anomalous frames, but rather is challenged to reconstruct those features, reflecting a discrepancy between the input and the output.

The impact of $\sigma(\ell)$, the anomaly weighting factor, on the distinction loss is pivotal:

- With $\sigma(\ell)$ approaching zero, the noise's influence on X_A is minimized, leading to a scenario where X_A is almost identical to X. This presents a challenge in distinguishing between normal and anomalous frames, as P and N become similar, pushing $\mathcal{L}_{dist} \approx 1$. This can be visualized in Figure 3 (a).
- On the other hand, as $\sigma(\ell)$ tends towards one, the anomalous region is replaced with something which almost entirely resembles noise. The model then faces the nearly impossible challenge of reconstructing the anomalous regions, thus rendering the distinction loss redundant, as shown in Figure 3 (b).

However, striking the right balance for $\sigma(\ell)$ is essential: it should be low enough so that the model, f, is able to reconstruct normality from a pseudo anomalous frame, but not so low where the pseudo-anomalous frame is too similar to the normal frame; causing $\mathcal{L}_{dist} \approx 1$. The adjustment of $\sigma(\ell)$ is carried out through backpropagation during training, allowing the model to iteratively find the optimal balance to maximize its proficiency in anomaly detection, aiming to pinpoint the smallest discernible anomaly from normalcy.

3.4. Inference

During the inference phase, the components involved in training, specifically the anomaly weight, object detection and tracking, and the Pseudo Anomaly Creator, are not utilized. The inference stage is streamlined to function through a conventional reconstruction approach. This process entails imposing a sliding window across each video, then submitting a sequence of video frames directly into the reconstruction model, which then processes these frames to output a reconstructed version of the middle frame.

4. Datasets

Our investigation utilizes a suite of video datasets to evaluate the adaptability and effectiveness of our proposed pseudo-anomalous loss approach in more complex scenarios. Specifically, we focus on three prominent video datasets: Ped2, CUHK Avenue, and ShanghaiTech. These datasets, with their varied and intricate anomaly instances, offer a robust testing ground to assess the performance of our model under diverse conditions.

The Ped2 dataset [25], sourced from pedestrian area surveillance footage, is notable for its range of anomalous events such as biking, skating, or irregular movement patterns. This dataset provides video clips with a frame resolution of 360×240, enabling a diverse sampling environment for anomaly detection research.

The CUHK Avenue dataset [32], originating from surveillance systems at the Chinese University of Hong Kong's Avenue, documents typical anomalies like running, loitering, and object throwing. These activities are unusual for the setting, making it an ideal dataset for testing anomaly detection models. Videos in this dataset are presented at a resolution of 640×360, offering a detailed view for analysis.

Comprising surveillance footage from a variety of indoor and outdoor scenes, the ShanghaiTech dataset [21] introduces a wide range of anomalies, including burglary, climbing, and fighting. The dataset's videos feature a resolution of 856×480, with variable frame numbers across clips. This diversity makes the ShanghaiTech dataset a comprehensive platform for challenging and evaluating the capabilities of anomaly detection systems.

5. Results

Our experimental setup and performance evaluation, aligning with established benchmarks in anomaly detection, leverages FastRCNN [13] for object detection and OC-Sort [6] for object tracking during the training phase. Notably, our Conv3DSkipUNet (C3DSU) model processes sequences of 3 frames. We benchmark our model against leading competitors identified in the comprehensive review by Astrid et al. [3], implementing a median window filtering approach with a window size of 17, as effectively demonstrated by Liu et al. [22]. It is crucial to note that, while our innovative approach leverages object detection to

generate pseudo-anomalies during the training phase, the core functionality of our model during inference strictly adheres to the principles of reconstruction-based anomaly detection. Though object detection methods have shown superior performance on the datasets described in Section 4, they are limited in practical application due to their incapability of detecting non-object related anomalies, such as explosions or debris falling off of buildings. Therefore, we strictly compare our methodology to other reconstruction-based methods. This strategic choice differentiates our work from methods reliant on object detection or frame prediction techniques for anomaly identification.

To quantitatively assess the performance of our model in video anomaly detection tasks, we employ a detailed anomaly scoring mechanism. Each frame's anomaly score is derived by computing the Euclidean distance at the pixel level between the frame and its reconstructed counterpart. To refine this evaluation, the calculated distances are segmented into patches sized 16×16, with the frame score determined by the highest mean value among these patch scores.

Method	Year	Ped2	Avenue	SHT
AE-Conv2D [15]	2016	90.00	70.20	60.85
AE-Conv3D [42]	2017	91.20	71.10	-
AE-ConvLSTM [27]	2017	88.10	77.00	-
TSC [26]	2017	91.03	80.56	67.94
StackRNN [26]	2017	92.21	81.71	68.00
MemAE [14]	2019	94.10	83.30	71.20
MNAD [30]	2020	90.20	82.80	69.80
PseudoBound [3]	2023	98.44	87.10	73.66
MAMC [29]	2024	96.70	87.60	71.50
C^2Net [20]	2024	98.00	87.50	-
C3DSU with DDL	Ours	**98.46**	**90.35**	**74.25**

Table 1. Comparative AUC Scores across Ped2, Avenue, and ShanghaiTech datasets. The table presents the AUC performance of our DDL model against a range of competing methodologies, highlighting the best performing results in bold.

The performance of our proposed methodology, as quantified through the Area Under the Curve (AUC) scores across three benchmark datasets, demonstrates its superior capability in detecting anomalies within video sequences. Table 1 showcases a comparative analysis of our model, denoted as C3DSU with DDL, against a variety of established methods in the field.

On the Ped2 dataset [25], our approach achieves an AUC score of **98.46**%, surpassing the previous state-of-the-art, PseudoBound [3], by a slight margin. This indicates an improvement in the model's ability to detect anomalies, illustrating the effectiveness of the dynamic anomaly weighting and distinction loss mechanism implemented in our methodology.

In the context of the Avenue dataset [32], our DDL model demonstrates a notable leap in performance, registering an AUC score of **90.35**%. This represents not only an improvement over the PseudoBound method [3] but also a substantial advancement compared to other reconstruction-based approaches such as MemAE [14] and MNAD-Reconstruction [30]. The results underscore our method's adeptness at handling the dataset's complex anomaly scenarios, further establishing the efficacy of incorporating pseudo-anomalies in training to enhance anomaly detection accuracy.

For the ShanghaiTech (SHT) dataset [21], our model achieves an AUC of **74.25%**, representing the best performing model amongst those compared. It is important to note that, in addressing the SHT dataset's diverse and dynamic anomaly instances, we trained a unique model for each scene, acknowledging that each scene warrants a different anomaly weight, $\sigma(\ell)$. This scene-specific approach allows for a more tailored anomaly detection mechanism, catering to the unique characteristics and challenges of each scene. The median score of all scenes is then taken to represent the overall performance on the SHT dataset. This methodological nuance underscores the adaptability of our approach, demonstrating its robustness across varied surveillance contexts despite the inherent challenges of the SHT dataset.

6. Ablation Studies

To elucidate the impact of Dynamic Distinction Learning (DDL) on video anomaly detection, we conducted ablation studies comparing the performance of two models, UNet and Conv3DSkipUNet (C3DSU), on the Ped2 and Avenue datasets, both with and without the implementation of DDL. The UNet model serves as a baseline, employing a traditional architecture without the convolutional 3D (Conv3D) layers between skip connections, and processes single frames independently. In contrast, the C3DSU model, designed for temporal data analysis, incorporates Conv3D layers between skip connections to capture temporal dynamics between frames.

The terminology used to describe the training configurations of the models—specifically, 'without DDL', 'with SDL', and 'with DDL'—reflects the incorporation of our Dynamic Distinction Learning (DDL) framework at different levels. The 'without DDL' configuration represents the standard reconstruction training process where the models, UNet and Conv3DSkipUNet (C3DSU), are trained purely on the task of reconstructing normal frames, leveraging only the reconstruction loss and omitting the introduction of pseudo anomalies. In contrast, the 'with SDL' (Static Distinction Learning) setup incorporates both the reconstruction loss and a static version of the distinction loss, where the anomaly weight, $\sigma(\ell)$, is fixed at 0.5 and not subject to training adjustments. This static distinction approach aims

to introduce a consistent level of challenge in distinguishing anomalies but lacks the adaptability of dynamic weighting. Finally, 'with DDL' employs our proposed methodology, integrating the dynamic anomaly weighting mechanism alongside the distinction loss into the training of the models.

Model	without DDL	with SDL	with DDL
UNet	86.90	95.28	**97.76**
C3DSU	95.55	97.12	**98.46**

Table 2. This table illustrates the performance improvement on the Ped2 dataset facilitated by the Dynamic Distinction Learning (DDL) approach across two different architectures: UNet and C3DSU.

As shown in Table 2, the implementation of DDL significantly enhances model performance. For the UNet model, the Area Under the Curve (AUC) score increases from 86.90% without DDL to 95.28% with SDL and further to **97.76%** with DDL, underscoring the effectiveness of DDL in enhancing anomaly detection accuracy. The introduction of a static distinction loss already marks a notable improvement, demonstrating the value of integrating anomaly differentiation into the training process. Similarly, the C3DSU model benefits from the addition of DDL, with its AUC score improving from 95.55% to 97.12% with SDL and then to **98.46%**. These results highlight the pivotal role of DDL in refining the model's ability to differentiate between normal and anomalous frames, particularly when temporal dynamics are considered. The improvement seen with SDL indicates the initial benefits of incorporating distinction mechanisms, which are significantly amplified upon transitioning to dynamic weighting.

The impact of DDL is also evident in the performance on the Avenue dataset, as depicted in Table 3. The UNet model experiences an improvement in AUC score from 84.18% without DDL to 87.06% with SDL and further to **88.96%** with DDL. The introduction of SDL showcases a tangible improvement, setting the stage for the more substantial enhancements afforded by the dynamic approach. The C3DSU model, however, showcases a more pronounced improvement, with the AUC score increasing from 82.54% without DDL to 89.41% with SDL and then to **90.35%** with DDL. These findings demonstrate the utility of DDL across different architectural frameworks and datasets, especially in scenarios involving complex anomaly patterns. The stepwise enhancements from static to dynamic distinction learning illustrate the methodological progression and its impact on the models' anomaly detection capabilities, highlighting the critical role of adaptively learning the anomaly weight for maximizing detection accuracy.

The ablation studies highlight the incremental value offered by each component of our methodology. The introduction of the distinction loss with a static pseudo anomaly weight significantly improves model performance by explicitly training the model to map pseudo anomalies towards normality. Further refinement is achieved with the implementation of a dynamic anomaly weight, which empowers our methodology to adaptively fine-tune and identify the minimum level of anomaly that can be detected.

Model	without DDL	with SDL	with DDL
UNet	84.18	87.06	**88.96**
C3DSU	82.54	89.41	**90.35**

Table 3. This table presents a comparison of model performance on the Avenue dataset, with and without the incorporation of Dynamic Distinction Learning (DDL), across UNet and C3DSU architectures.

7. Conclusion

This paper introduced Dynamic Distinction Learning (DDL), a novel approach designed to enhance the accuracy of video anomaly detection through the integration of pseudo-anomalies, dynamic anomaly weighting, and a unique distinction loss function. Our methodological innovation lies in its ability to adaptively learn the variability of normal and anomalous behaviors without relying on fixed anomaly thresholds, thereby significantly improving detection performance.

Our experiments, conducted on benchmark datasets such as Ped2, CUHK Avenue, and ShanghaiTech, have demonstrated the superior performance of the DDL framework. The model achieved remarkable AUC scores, outperforming existing state-of-the-art methods on the Ped2 and Avenue datasets, and delivering competitive results on the ShanghaiTech dataset. These achievements underscore the effectiveness of DDL in addressing the challenges of video anomaly detection, offering a scalable and adaptable solution that can be tailored to specific scene requirements.

The ablation studies further highlighted the impact of incorporating DDL into different model architectures, including UNet and Conv3DSkipUNet (C3DSU). The significant improvements in anomaly detection accuracy with DDL underscore its role in refining models' ability to distinguish between normal and anomalous events effectively, showcasing its broad applicability across different architectural frameworks and complex anomaly patterns.

In conclusion, Dynamic Distinction Learning represents a significant advancement in the field of video anomaly detection. Its ability to dynamically adapt and learn from pseudo-anomalies, coupled with the distinction loss function, provides a robust framework for accurately identifying anomalous events in video data.

References

[1] Abhishek Aich, Kuan-Chuan Peng, and Amit K. Roy-Chowdhury. Cross-domain video anomaly detection without target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2579–2591, 2023. 1, 2

[2] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Limiting reconstruction capability of autoencoders using moving backward pseudo anomalies. pages 248–251. IEEE, 2022. 2, 6

[3] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Pseudobound: Limiting the anomaly reconstruction capability of one-class classifiers using pseudo anomalies. *Neurocomputing*, 534:147–160, 2023. 2, 6, 7

[4] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, 2023. 1, 2

[5] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):938–946, 2021. 2

[6] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696, 2023. 6

[7] S. Chandrakala, P. Shalmiya, V. Srinivas, and K. Deepak. Object-centric and memory-guided network-based normality modeling for video anomaly detection. *Signal, Image and Video Processing*, 2022. 1, 2

[8] Xingya Chang, Yuxin Zhang, Dingyu Xue, and Dongyue Chen. Multi-task learning for video anomaly detection. *Journal of Visual Communication and Image Representation*, 87: 103547, 2022. 2

[9] K. Deepak, S. Chandrakala, and C. Krishna Mohan. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. *Signal, Image and Video Processing*, 15:215–222, 2021.

[10] K. Deepak, G. Srivathsan, S. Roshan, and S. Chandrakala. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders. *Circuits, Systems, and Signal Processing*, 40:1333–1349, 2021. 1, 2

[11] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114, 2021. 2

[12] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. *Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection*. Association for Computing Machinery, 2021. 1, 2

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 6

[14] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. *arXiv*, pages 1705–1714, 2019. 2, 7

[15] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. 2016. Temporal laplacian Eigen maps and dimensionality reduction. 7

[16] Yujun Kim, Jin Yong Yu, Euijong Lee, and Young Gab Kim. Video anomaly detection using cross u-net and cascade sliding window. *Journal of King Saud University - Computer and Information Sciences*, 2022. 1, 2

[17] Viet-Tuan Le and Yong-Guk Kim. Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence*, 2022. 1, 2

[18] Viet Tuan Le and Yong Guk Kim. Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence*, 53:3240–3254, 2023. 2

[19] Gang Li, Ping He, Huibin Li, and Fan Zhang. Adversarial composite prediction of normal video dynamics for anomaly detection. *Computer Vision and Image Understanding*, 232, 2023. 2

[20] Jiafei Liang, Yang Xiao, Joey Tianyi Zhou, Feng Yang, Ting Li, and Zhiwen Fang. C^2net: content dependent and independent cross-attention network for anomaly detection in videos. *Applied Intelligence*, 54(2):1980–1996, 2024. 7

[21] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7

[22] Yang Liu, Zhengliang Guo, Jing Liu, Chengfang Li, and Liang Song. Osin: Object-centric scene inference network for unsupervised video anomaly detection. *IEEE Signal Processing Letters*, 2023. 2, 6

[23] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13588–13597, 2021. 1, 2

[24] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24500–24510, 2023. 1, 2

[25] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. pages 2720–2727, 2013. Sparse Coding. 6, 7

[26] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349, 2017. 7

[27] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection.

In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1238–1243. IEEE, 2017. 7

[28] Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Şeymanur Akti, Hazim Kemal Ekenel, and Giuseppe Amato. Mocca: Multilayer one-class classification for anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2313–2323, 2022. 1, 2

[29] Zhiyuan Ning, Zile Wang, Yang Liu, Jing Liu, and Liang Song. Memory-enhanced appearance-motion consistency framework for video anomaly detection. *Computer Communications*, 216:159–167, 2024. 7

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 14360–14369, 2020. 1, 2, 7

[31] Pankaj Raj Roy, Guillaume-Alexandre Bilodeau, and Lama Seoud. Predicting next local appearance for video anomaly detection. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5, 2021. 1, 2

[32] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra. Moayed, and Reinhard Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172:88–97, 2018. 6, 7

[33] Savath Saypadith and Takao Onoye. Video anomaly detection based on deep generative network. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021. 1, 2

[34] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22846–22856, 2023. 2

[35] Stanislaw Szymanowicz, James Charles, and Roberto Cipolla. Discrete neural representations for explainable anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 148–156, 2022. 1, 2

[36] Waseem Ullah, Tanveer Hussain, Fath U.Min Ullah, Mi Young Lee, and Sung Wook Baik. Transcnn: Hybrid cnn and transformer mechanism for surveillance anomaly detection. *Engineering Applications of Artificial Intelligence*, 123, 2023. 2

[37] Tuan Hung Vu, Jacques Boonaert, Sebastien Ambellouis, and Abdelmalik Taleb-Ahmed. Multi-channel generative framework and supervised learning for anomaly detection in surveillance videos. *Sensors*, 21:1–16, 2021. 1, 2

[38] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021. 1, 2

[39] Yang Wang, Tianying Liu, Jiaogen Zhou, and Jihong Guan. Video anomaly detection based on spatio-temporal relationships among objects. *Neurocomputing*, 532:141–151, 2023. 2

[40] Hongchun Yuan, Zhenyu Cai, Hui Zhou, Yue Wang, and Xiangzhi Chen. Transanomaly: Video anomaly detection using video vision transformer. *IEEE Access*, 9:123977–123986, 2021. 1, 2

[41] Muhammad Zaigham Zaheer, Jin-Ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14171–14181, 2020. 1, 2

[42] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1933–1941, New York, NY, USA, 2017. Association for Computing Machinery. 7