Dynamic Distinction Learning: Adaptive Pseudo Anomalies for Video Anomaly Detection

Supplementary Material

7.1. Dynamics of Anomaly Weight $\sigma(\ell)$ in Model Training (Methodology Supplementary)

To delve deeper into the dynamics of our model's training, it is imperative to scrutinize the behavior of the anomaly weight $\sigma(\ell)$, which serves as a crucial variable in the generation of pseudo-anomalous frames. This parameter, $\sigma(\ell)$, is not merely a static coefficient but a dynamic element that evolves during the training process, reflecting the model's progressing ability to discern between normal and anomalous instances.

As training progresses, the trend in the anomaly weight $\sigma(\ell)$ reflects the model's adaptation and learning curve. Initially, $\sigma(\ell)$ may experience fluctuations, including a rapid increase as the model begins to differentiate between normal and anomalous patterns - as it has in Figure 5 when training the C3DSU model on the Ped2 Dataset. This early rise in $\sigma(\ell)$ is crucial as it indicates the model's initial phase of learning to handle more pronounced anomalies. As the model's ability to reconstruct and identify anomalies improves, we observe a subsequent decline and eventual stabilization of $\sigma(\ell)$, suggesting that the model is settling on an optimal threshold for anomaly detection without being overly influenced by the injected noise levels.

The crux of training with $\sigma(\ell)$ lies in pinpointing the minimal anomaly magnitude that still allows our model f to reliably differentiate anomalous from normal. This threshold is the 'sweet spot' where $\sigma(\ell)$ is neither too insubstantial to be deemed noise nor too dominant to mask the underlying structure of the data. It's at this juncture that our model f is optimally trained to flag deviations from normalcy, while remaining anchored enough to not be swayed by random perturbations or outliers.

In essence, the evolution of $\sigma(\ell)$ during the training is a barometer of the model's growing intelligence in anomaly detection. By carefully calibrating this parameter, we empower our model f to identify the minimum anomaly required to discern abnormalities, a testament to its analytical prowess and the culmination of a successful training regimen.

7.1.1 Visualizing the Effects of Training

To further illustrate the success of our training process and the dynamic adjustment of the anomaly weight $\sigma(\ell)$, Figure 4 presents a visual comparison. These images encapsulate various aspects of our model's performance, including the original middle frame X^t, the reconstruction from the normal input $f(X)$, the pseudo-anomalous middle frame X^t_A, the reconstruction from the pseudo-anomalous input $f(X_A)$, and the respective reconstruction errors $\|X^t - f(X)\|$, $\|X^t_A - f(X_A)\|$, and $\|X^t - f(X_A)\|$.

This visual representation demonstrates the impact of our training methodology. Particularly, the comparison between $\|X^t_A - f(X_A)\|$ and $\|X^t - f(X_A)\|$ reveals a critical insight: the discrepancy between the reconstruction of the pseudo-anomalous frame and the normal frame is significantly less than that between the reconstruction of the pseudo-anomalous frame and its corresponding anomalous input. This outcome underscores the model's capability to more closely align the reconstructed output with the normal frame rather than perpetuating the anomalies present in the pseudo-anomalous input.

7.2. Rationale Behind Weighted Noise for Simulating Anomalies (Methodology Supplementary)

The efficacy of employing weighted noise to simulate a range of human-defined anomalies – such as skipped frames, duplicate frames, random patches, and the insertion of foreign shapes or objects – is rooted in the fundamental operating principles of convolutional neural networks (CNNs) employed in reconstruction-based anomaly detection methods. These networks are adept at learning and reconstructing patterns observed in the training data, which predominantly consists of normal behavioral patterns within video sequences.

When confronted with anomalies, the convolutional kernels of a CNN do not perceive these as distinct types of irregularities per se, but rather as inputs lacking the regular patterns or structures they have been trained to recognize and reconstruct. From the perspective of these kernels, anomalies disrupt the spatial and temporal consistency of the input data, rendering them as pattern-less examples – essentially, noise. This perception is crucial for understanding why weighted noise can serve as a universal proxy for various anomalies in training anomaly detection models.

The concept of weighted noise as a universal anomaly proxy is further justified by the intrinsic adaptability and learning mechanisms of CNNs. These networks, through their deep architecture, are designed to capture and encode complex patterns in the data they process. The introduction of weighted noise challenges these networks in a unique way, compelling them to discern between the 'normal' patterns they've learned to reconstruct and the 'abnormal' patterns represented by the noise. This challenge is akin to

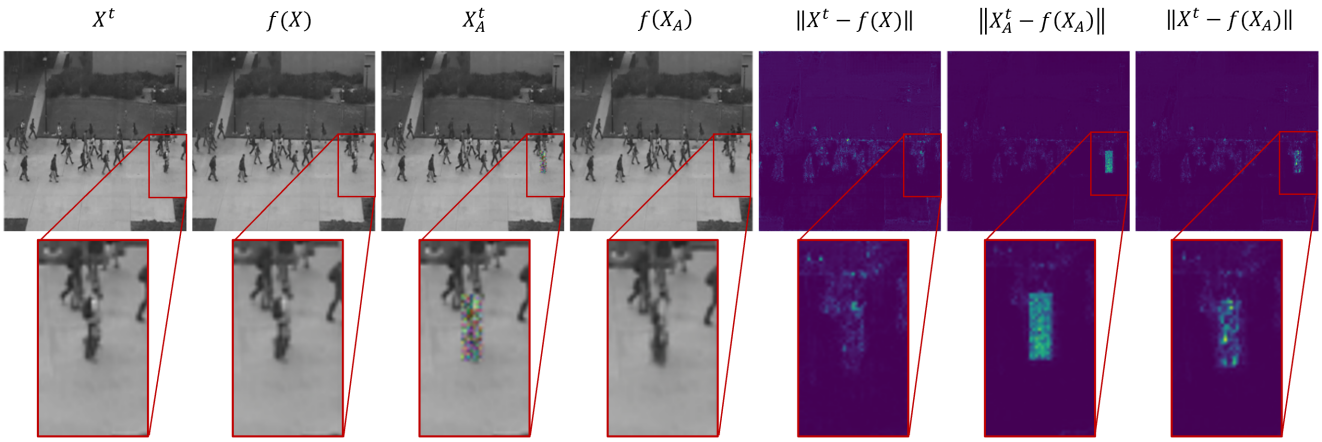

Figure 4. Visual Comparison of Model Training Effects: This figure provides a comprehensive visualization of the model's performance across different frames and stages of reconstruction. It features the original middle frame X^t, the reconstructed frame from normal input $f(Xt$, the pseudo-anomalous middle frame X_A^t, the reconstructed frame from the pseudo-anomalous input $f(X_A)$, and the reconstruction errors $\|X^t - f(X)\|$, $\|X_A^t - f(X_A)\|$, and $\|X^t - f(X_A)\|$.

exposing the network to a wide variety of anomalies without the need for explicit enumeration or replication of each possible anomalous event, which in real-world applications is infeasible due to the vast and unpredictable nature of such events.

7.3. C3DSU Architecture (Results Supplementary)

The design of our Conv3DSkipUNet (C3DSU) model incorporates a Conv2D UNet structure, enhanced with custom ConvBlocks for both the Encoder and Decoder components, and augmented by the integration of Conv3D layers within the skip connections for handling the temporal dimension.

A ConvBlock in this context is engineered to facilitate multi-headed convolution. Specifically, it comprises four Conv2D layers, each employing "same" padding to maintain dimensional consistency and a kernel size of 3 for capturing spatial details. These layers are executed in parallel and their outputs are concatenated along the channel dimension, ensuring a rich feature representation. Following the concatenation, the process involves batch normalisation and activation through the ReLU function, aiming to stabilize learning and introduce non-linearity, respectively. The output is then restructured to reduce the Height and Width dimensions, the output is then passed through an additional Conv2D layer with "same" padding, followed by another round of batch normalisation and ReLU activation. The Encoder ConvBlock's operational flow is depicted in Figure 6

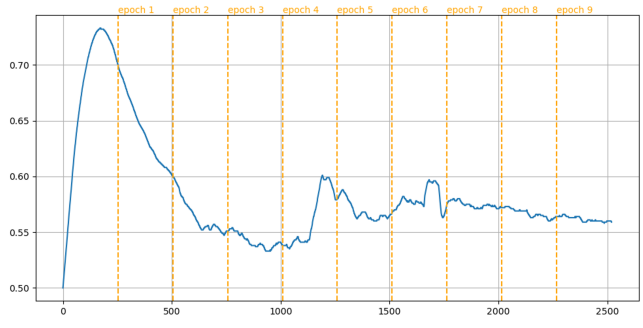

Figure 5. Evolution of the anomaly weight $\sigma(\ell)$ during the training of the C3DSU model on the Ped2 Dataset for 10 epochs.

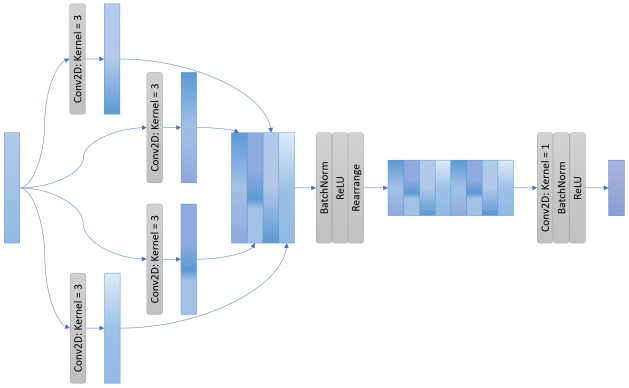

Figure 6. Schematic illustration of an Encoder ConvBlock within the C3DSU architecture, highlighting the multi-headed convolution process. This diagram details the structure and flow through the ConvBlock, including the initial parallel Conv2D layers, concatenation, batch normalisation, ReLU activation, dimension reduction, and final Conv2D layer, followed by another round of batch normalisation and ReLU activation.

for visual clarification.

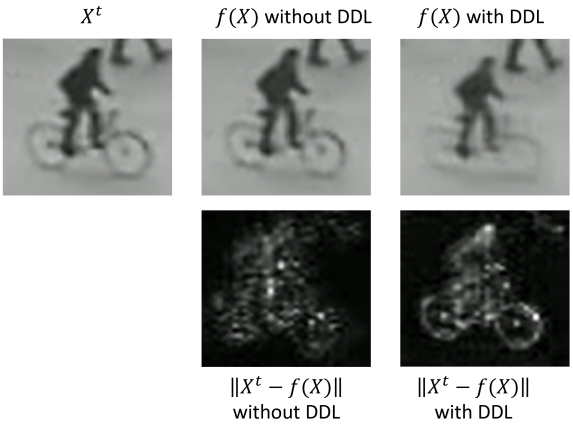

Figure 7. Visual anomaly detection comparison in the Ped2 dataset featuring a cyclist riding a bicycle. From left to right: the original frame, reconstruction without Dynamic Distinction Learning (DDL), reconstruction with DDL, and beneath the residual difference highlighting the anomaly.

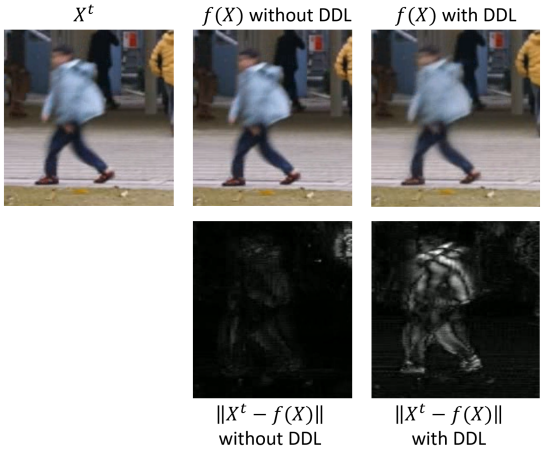

Figure 8. Visual anomaly detection comparison in the Avenue dataset illustrating a boy skipping. From left to right: the original frame, reconstruction without Dynamic Distinction Learning (DDL), reconstruction with DDL, and beneath the residual difference emphasizing the anomaly.

The overarching UNet architecture is assembled with eight such ConvBlocks, evenly split between the Encoder and Decoder. Each ConvBlock in the Encoder is paired with a corresponding ConvBlock in the Decoder via skip connections. These connections are uniquely designed to pass through Conv3D layers, thereby incorporating the temporal aspect into the spatial information flow. This mechanism ensures that while individual frames are initially processed as separate images by the Conv2D layers in the Encoder and Decoder, the Conv3D layers within the skip connections facilitate the integration of temporal information, essential for effective video analysis.

7.4. Qualitative Results (Ablations Supplementary)

To complement the quantitative analysis, qualitative assessments were conducted to visually inspect the model's performance in identifying anomalies within the Ped2 and Avenue datasets. These assessments provide insight into the model's ability to reconstruct scenes and highlight anomalous activities when DDL is applied versus when it is not.

Figure 7 showcases a visual comparison of anomaly detection in a scenario involving a cyclist riding a bicycle, an anomalous event within the Ped2 dataset. The sequence displays the original frame, followed by reconstructions without and with DDL, and finally, the residual differences between the reconstructions and the original frame. Notably, the application of DDL results in a reconstruction where the bicycle is almost entirely erased, signifying the model's training to poorly reconstruct unfamiliar shapes. This illustrates DDL's effectiveness in forcing the model to focus on normal patterns, thereby making anomalies, such as the bicycle in this case, more pronounced.

Similarly, Figure 8 presents a visual analysis involving a boy skipping, an anomalous event in the Avenue dataset. The illustration includes the original frame along with reconstructions without and with DDL, supplemented by the residual differences highlighting the anomaly. The comparison clearly demonstrates that with DDL, the anomaly of the boy skipping is accentuated more effectively than without DDL. This enhancement is evident in the residual images, where DDL's reconstruction struggles more with the skipping motion, thereby amplifying the distinction from normal activity.