Bi-directional Frame Interpolation for Unsupervised Video Anomaly Detection

Hanqiu Deng, Zhaoxiang Zhang, Shihao Zou, Xingyu Li
University of Alberta
{hanqiu1, zhaoxia2, szou2, xingyu}@ualberta.ca

Abstract

Anomaly detection in video surveillance aims to detect anomalous frames whose properties significantly differ from normal patterns. Anomalies in videos can occur in both spatial appearance and temporal motion, making unsupervised video anomaly detection challenging. To tackle this problem, we investigate forward and backward motion continuity between adjacent frames and propose a new video anomaly detection paradigm based on bi-directional frame interpolation. The proposed framework consists of an optical flow estimation network and an interpolation network jointly optimized end-to-end to synthesize a middle frame from its nearest two frames. We further introduce a novel dynamic memory mechanism to balance memory sparsity and normality representation diversity, which attenuates abnormal features in frame interpolation without affecting normal prototypes. In inference, interpolation error and dynamic memory error are fused as anomaly scores. The proposed bi-directional interpolation design improves normal frame synthesis, lowering the false alarm rate of anomaly appearance; meanwhile, the implicit "regular" motion constraint in our optical flow estimation and the novel dynamic memory mechanism play blocking roles in interpolating abnormal frames, increasing the system's sensitivity to anomalies. Extensive experiments on public benchmarks demonstrates the superiority of the proposed framework over prior arts.

1. Introduction

Unsupervised anomaly detection is a challenging task with a wide range of real-world applications, such as industrial defect detection [2], medical diagnosis [40], and video surveillance [27, 21, 17]. In particular, unsupervised anomaly detection plays an increasingly important role in intelligent video surveillance systems. Video is high-dimensional spatiotemporal data. Detection of abnormal patterns from such huge data volume is challenging.

In literature, unsupervised video anomaly detection primarily follows the paradigm of either frame reconstruction or future frame prediction. Spatial appearance and temporal motion patterns are usually exploited as mutually complementary cues to tackle this problem [17, 3]. Since both paradigms rely on frame synthesis, generative models such as auto-encoders [11, 14] naturally serve as the backbone architecture. Specifically, reconstruction-based approaches treats video frames independently and targets to synthesize the input picture [8]. It hypothesizes that an anomalous frame incurs large reconstruction errors by an anomaly-free trained model. However, this hypothesis is not always true. Instead, abnormal appearance in video frames may be reconstructed partially, or even completely [9, 32]. In addition, frame reconstruction-based methods usually don't consider temporal continuity among video frames. Consequently, it is weak to detect abnormal motion patterns. To address these problems, future frame prediction-based methods are proposed [17]. With a hypothesis that anomalous events are unpredictable from previous sequence [29], these methods target the generation of future frames from the previous sequence and take the prediction error as an indicator of anomaly. Since the targeted future frame is not fed into the generative model, the problem of abnormal

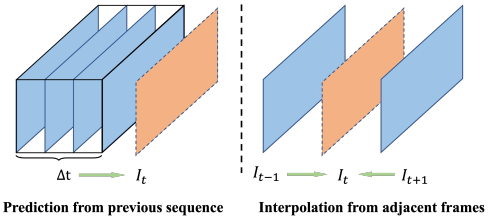

Figure 1. Left: Conceptual demonstration of conventional frame prediction-based methods where the current frame (in orange) is predicted from previous continuous sequence (in blue). Right: the proposed bi-directional interpolation method that only use the nearest two frames (in blue) to interpolate the middle frame (in orange). Such a design exploits spatial similarity and temporal continuity between adjacent frames and exhibits two benefits. First, forward and backward information facilitates normal frame interpolation, which lowers the false detection rate of anomaly appearance. Second, the minimized data volume (i.e. two frames) required in our frame interpolation remedies the anomalous motion leakage problem in conventional prediction-based methods.

appearance residue in frame reconstruction-based methods is alleviated. To further improve performance, SIGnet cooperates two independent U-Nets to predict a frame forward and backward with a bi-directional consistency term in training [8]. Despite the specific design, conventional frame prediction-based approaches take a short sequence as input, which may leak anomalous motion patterns existing in those sequences to the synthesized future frame [17, 10, 4, 26, 32, 8], hurting the detection of abnormal motion patterns.

To facilitate detection of abnormal appearance and motion patterns in videos, we explore the spatial similarity and temporal continuity among adjacent frames and propose the use of bi-directional frame interpolation as a novel paradigm for unsupervised video anomaly detection. We conceptually demonstrate the proposed frame interpolation-based method and its difference from conventional frame prediction-based methods in Fig.1. On one hand, instead of taking video sequences as input, our interpolation method minimizes the input data volume (i.e. two frames only), which remedies the anomalous motion pattern leakage problem in conventional prediction-based methods. On the other, it is noteworthy that our method incorporates forward and backward knowledge from adjacent frames for better interpolating normal frames, lowering the false alarm rate in anomaly detection.

Specifically, our framework consists of an optical flow estimation network and an interpolation network jointly trained from scratch on normal video sequences. The former learns to estimate *regular* optical flows corresponding to normal motions only, prone to generate poor optical flows for unseen, abnormal motion patterns. The latter regresses the target frame from the adjacent frames and corresponding "regular" optical flows. In inference, a large interpolation error indicates an anomalous frame. We also design a novel dynamic memory mechanism that stores embedding associated with regular motion and appearance and sparsely address these normal prototypes in frame interpolation, thereby increasing the generative error of anomalous samples. In sum, the implicit "regular" motion constraint in our bi-directional optical flow estimation and the novel dynamic normal prototype addressing strategy in the proposed memory mechanism play blocking roles in interpolating abnormal frames, increasing the system's sensitivity to anomalies. Our contributions are summarized as follows:

• We introduce a simple yet effective bi-directional frame interpolation framework as a novel paradigm for unsupervised video anomaly detection. Spatial similarity and temporal continuity among adjacent video frames are exploited. The novel design greatly reduce the input data volume as well as model complexity.

• We improve the memory module by dynamically selecting the Top-K representative memory items to represent normal features. It well balances memory sparsity and prototype diversity in normal representation.

• Extensive experiments on public video anomaly detection benchmarks demonstrate superiority of our approach over prior arts.

2. Related Work

Unsupervised video anomaly detection. Most efforts to tackle this problem deploy generative models for either frame reconstruction or future frame prediction. Reconstruction-based approaches aim to train a model to retain prototypical normal patterns, from which normal samples can be reconstructed well while anomalies cannot. These models include sparse coding [24, 9], auto-encoding [11, 39], etc. Later, future frame prediction, a task-specific paradigm, is introduced for video anomaly detection [17]. These approaches take prediction errors as anomaly clues and demonstrate promising performance [4, 32, 10]. Aware of the high-dimensional spatiotemporal properties of video sequences, many studies exploit various complexity models such as 3D convolution [39], recurrent neural networks [24] ,and long short term memory [23] to estimate spatio-temporal dependency and appearance-motion association for video anomaly detection [30, 4].

This study proposes a new frame interpolation based method for video anomaly detection. Unlike previous works [30, 38, 4, 19, 10, 8] that either use a pre-trained FlowNet to estimate optical flow or designs two independent generative models for forward and backward prediction, our method train one model from scratch without any performance degradation. Despite the simple architecture, we demonstrate its effective and superior performance in public video anomaly detection benchmark.

Frame interpolation. Video frame interpolation is generally based on time-varying information in the optical flow to produce a continuous motion association [1]. More recently, convolutional neural networks enable optical flow estimation to be trained in an end-to-end fashion, and most video frame interpolation methods incorporate this approach into the frame synthesis process [12, 13, 37]. Alternative, flow-free methods attempt precise frame interpolation without explicit optical flow estimation. These methods include, but not limited to, PixelShuffle [35], PhaseNet [28], and channel attention [6]. While some methods [38, 8, 7, 31] propose to interpolate a frame from a continuous sequence, we follow the frame-based paradigm that bi-directionally predict the middle one from its nearest two frames.

Memory mechanism. Memory mechanisms are widely used to constrain the generalization capabilities of generative models for anomaly detection. In memory-augmented auto-encoders, anomaly-free features are reorganized in the memory bank according to similarity-weighted memory

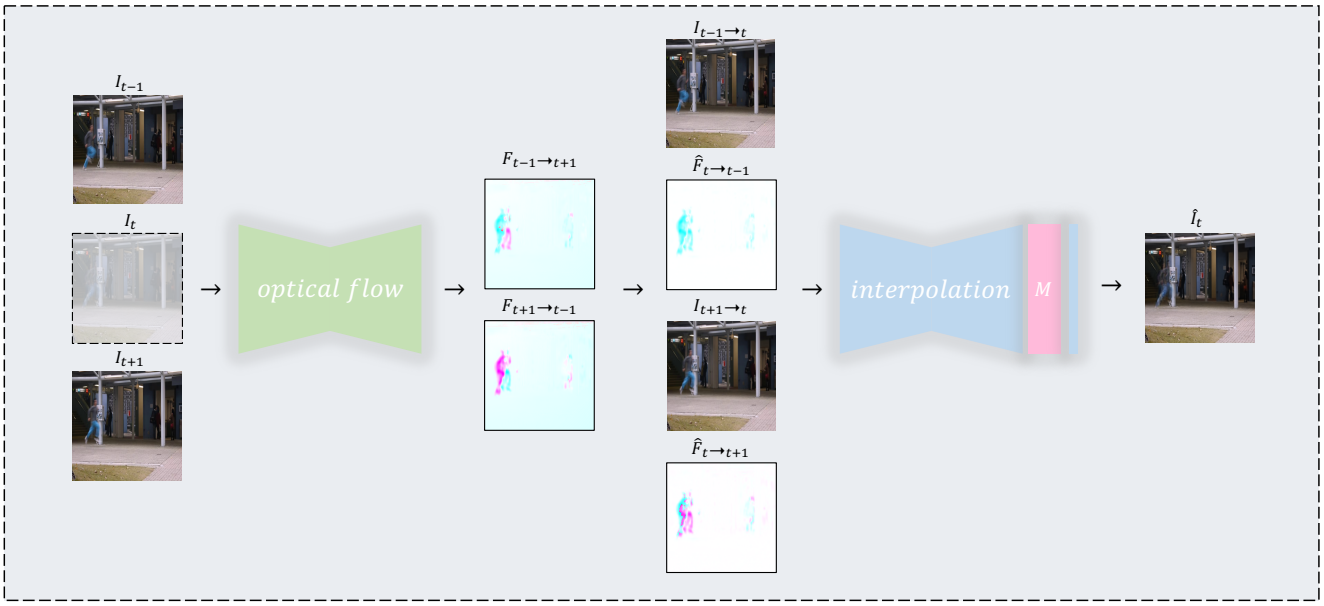

Figure 2. Systematic diagram of the proposed method. Our end-to-end trained model consists of optical flow estimation and frame refinement augmented with a dynamic memory module. Assuming the current frame I_t being the interpolation target, we feed I_{t-1} and I_{t+1} into the optical flow estimation network. Then, we calculate the forward/backward optical flows and synthesize interpolation candidates $I_{t-1\to t}$ and $I_{t+1\to t}$ from I_{t-1}/I_{t+1}, respectively. The refinement network is deployed to regresses the target frame I_t from those coarse interpolations and generates the final result $\hat{I}_t$. In inference, the interpolation error between I_t and $\hat{I}_t$ indicate anomalous degree.

items for reconstruction [9]. Park et al. [32] propose the compact loss and separation loss to reduce the distance of the nearest feature and enhance the diversity of memory patterns, respectively. Moreover, the memory module remains effective and outperforms the reconstructed results on the predictive model [32]. Cai et al. [4] introduce the motion information from a pre-trained flow estimation network signals and utilize the appearance-motion memory consistency to increase the gap between the abnormal and regular events. To cope with unseen test scenarios, a dynamic attention mechanism is proposed to encode normal patterns [26]. Liu et al. [19] proposed a multi-stage memory module combined with skip connections to guarantee reconstruction quality. In this study, we incorporate a novel dynamic memory mechanism to balance memory sparsity and prototype diversity for normal representation, which helps to interpolate normal frames well while interpolating abnormal frames poorly, thus promoting discrimination of anomalies.

3. Methodology

In the proposed method, we full exploit the spatial similarity and temporal continuity among video frames for frame interpolation and anomaly detection. As shown in Fig. 2, for the anomaly detection of frame I_t, our proposed method takes its previous frame I_{t-1} and future frame I_{t+1} as input and infers the forward and backward optical flows between these two frames. Then the optical flows, together with the two frames, are concatenated and fed into the interpolation network to synthesize the "normal" version of the current frame, $\hat{I}_t$. The interpolation error between the target I_t and its "normal" version $\hat{I}_t$ is a strong indicator of anomaly. In addition, we introduce a dynamic memory mechanism in the interpolation network, which strengthens the normality representation but impairs the representation of anomalies.

3.1. Frame interpolation

Given a video sequence $\{I_0, ..., I_{t-1}, I_t, I_{t+1}, ..., I_N\}$ of $N+1$ frames, our goal is to predict a sequence $\{\hat{I}_1, ..., \hat{I}_{t-1}, \hat{I}_t, \hat{I}_{t+1}, ..., \hat{I}_{N-1}\}$ by interpolation of the neighboring two frames for each time step t. More specifically, at time t, we use frame I_{t-1} and I_{t+1} as the input to infer the current frame $\hat{I}_t$, whose difference with the given frame I_t is considered as the criterion of anomaly detection.

Following previous works [4, 19, 38], the model for optical flow prediction is trained via unsupervised learning by warping the input two frames with the estimated flows. Since the optical flow interprets the motion between frames, it can be employed to interpolate the intermediate frames. Let $F_{t-1\to t+1}$ and $F_{t+1\to t-1}$ denote the optical flow from I_{t-1} to I_{t+1} and I_{t+1} to I_{t-1}, respectively. Given the flows between frame $t-1$ and $t+1$, the intermediate frame can be obtained by warping with the interpolated optical flow. Since frame at t located at the center of two frames at $t-1$

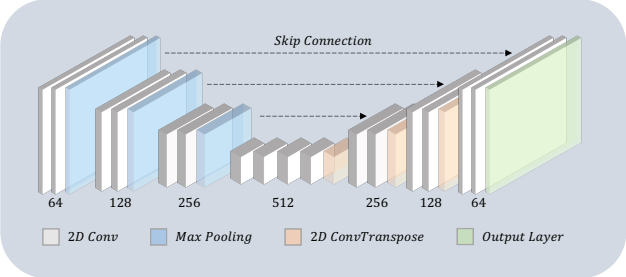

Figure 3. The U-Net architecture used in our framework. The 2D convolution layer consists of 3×3 convolutional kernel with stride of 1, batch normalization and ReLU activation. We use max pooling layer for down-sampling and transposed convolution for up-sampling. The output layer is also a 2D convolution but uses Tanh activation.

and $t+1$, we have the optical flow from I_t to I_{t-1} as

$$F_{t\to t-1} = \frac{1}{2}F_{t+1\to t-1} = -\frac{1}{2}F_{t-1\to t+1}, \tag{1}$$

and the optical flow from I_t to I_{t+1} as

$$F_{t\to t+1} = \frac{1}{2}F_{t-1\to t+1} = -\frac{1}{2}F_{t+1\to t-1}. \tag{2}$$

Then the optical flows from I_t to I_{t-1} and from I_t to I_{t+1} are interpolated linearly as follows,

$$\hat{F}_{t\to t-1} = -\frac{1}{4}F_{t-1\to t+1} + \frac{1}{4}F_{t+1\to t-1}, \tag{3}$$

$$\hat{F}_{t\to t+1} = \frac{1}{4}F_{t-1\to t+1} - \frac{1}{4}F_{t+1\to t-1}. \tag{4}$$

When anomalous motion enters the network, the bi-directional optical flow estimates produce more cumulative motion bias that contributes to detecting anomalies. In this study, we use a U-Net[34] to predict the bidirectional optical flows $F_{t-1\to t+1}$ and $F_{t+1\to t-1}$. The detailed architecture of U-Net [34] used in our work is shown in Fig. 3.

With these interpolated flows, we can get the frame $I_{t-1\to t}$ and $I_{t+1\to t}$ at time t by backward warping operation respectively, which are represented as

$$I_{t-1\to t} = \phi(I_{t-1}, \hat{F}_{t\to t-1}), \tag{5}$$

$$I_{t+1\to t} = \phi(I_{t+1}, \hat{F}_{t\to t+1}), \tag{6}$$

where $\phi(I, F)$ is the backward warping operation. Since $I_{t-1\to t}$ and $I_{t+1\to t}$ are roughly estimated by temporal continuity, another interpolation model, $\mathcal{R}$, is applied for the refinement of frame at t, whose output is the final interpolated result $\hat{I}_t$:

$$\hat{I}_t = \mathcal{R}(\hat{F}_{t\to t-1}; I_{t-1\to t}; \hat{F}_{t\to t+1}; I_{t+1\to t}). \tag{7}$$

Note that previous video interpolation methods [13] use a residual connection to the output image in the refinement model, which is shown to produce high quality images. However, it causes anomalous information leakage into the restored image, weakening the interpolation discrepancy for anomalies. Therefore, we use direct generation path without the residual connection on the output image.

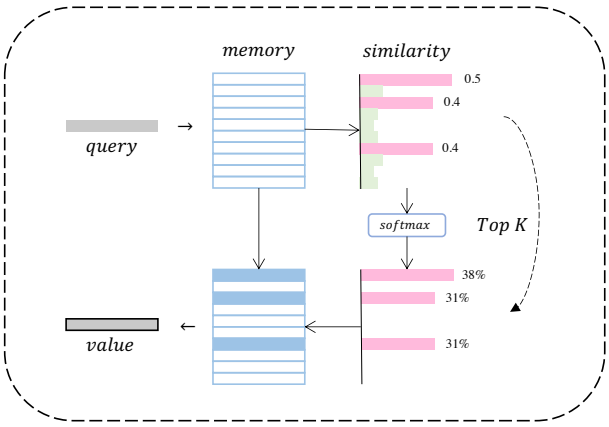

Figure 4. Overview of the dynamic memory mechanism. Given a feature map as a query, we calculate the similarity between it and the memory items. For each query, we select the top K items in terms of the similarity. Finally, we calculate a weighted average value in terms of the matching probability of the selected image as the addressing output.

3.2. Dynamic memory mechanism

The generative U-Net model itself is not able to learn different representation between normal and anomaly inputs. Thus we introduce a memory module in the refinement model as is shown in Fig. 4. Intuitively, the memory module in the unsupervised anomaly detection is able to remember normal patterns which have distinct representations with anomalies. This module is plugged into the refinement model located before the output layer as shown in Fig. 2, which avoids anomalous information leakage into the restored image. In this paper, we propose a *dynamic* memory mechanism, where top K memory keys are addressed according to the rank of cosine similarity. This is different from previous works [9, 32, 26, 4, 19] that either uses a hard threshold to select memory keys, select only the most similar memory key or linearly combine all the keys. Our dynamic strategy balances the memory capacity and normal prototype diversity, thus boosting performance on anomaly detection (see Sec. 4.3).

Specifically, the memory pool $M \in \mathbb{R}^{N\times C}$ is defined as a matrix consisting of N keys with each key of dimension C. The memory pool is learnable and is expected to learn typical patterns of anomaly-free features while training. Given the feature map $Q \in \mathbb{R}^{H\times W\times C}$ extracted from a frame, we linearly query its mapped features on the normality space from the memory pool. As we discussed above,

the memory module is located before the output layer, so the height H and width W of the feature map are consistent with the frame. For each input feature vector $q_{ij} \in \mathbb{R}^C$ in the feature map Q, we use cosine similarity to represent the matching extent,

$$\omega(k_m, q_{ij}) = \frac{k_m^\top q_{ij}}{||k_m||\,||q_{ij}||}, \tag{8}$$

where k_m is the m-th key in the memory pool.

The goal of memory addressing is to find the prototypical normal patterns in the memory and perform reconstruction that is able to present discrepancy between the anomaly and normal queries. For a given feature vector q_{ij}, we dynamically select top K of N memory keys as the addressing target according to the similarity ranking, as these keys are the most representative for normal features. Then the matching probabilities between q_{ij} and the top K selected keys are given by

$$\hat{p}_m = \frac{\exp(\omega(k_m, q_{ij}))}{\sum_{n=1}^{K} \exp(\omega(k_n, q_{ij}))}. \tag{9}$$

Using the matching probabilities as the addressing weights, the addressing result v_{ij} is obtained by the weighted average of all selected keys,

$$v_{ij} = \sum_{n=1}^{K} \hat{p}_n \cdot k_n. \tag{10}$$

Then, we add the value v_{ij} as a residual term to the query q_{ij} as the output of the memory module. Finally, we have the output as the reconstructed frame $\hat{I}_t$ at time t via frame interpolation.

3.3. Joint training

We train the model end-to-end according to the loss defined below,

$$\mathcal{L} = \mathcal{L}_{\text{warp}} + \mathcal{L}_{\text{frame}} + \lambda_1 \mathcal{L}_{\text{SSIM}} + \lambda_2 \mathcal{L}_{\text{con}} + \lambda_3 \mathcal{L}_{\text{div}}, \tag{11}$$

where $\mathcal{L}_{\text{warp}}$ is the bidirectional warping loss between frame I_{t-1} and I_{t+1}, defined as

$$\begin{aligned}\mathcal{L}_{\text{warp}} = &\|I_{t-1} - \phi(I_{t+1}, F_{t-1\to t+1})\|^2 \\ &+ \|I_{t+1} - \phi(I_{t-1}, F_{t+1\to t-1})\|^2.\end{aligned} \tag{12}$$

It is used as a regularization term to jointly train the flow estimation networks. $\mathcal{L}_{\text{frame}}$ denotes the loss between the predicted frame $\hat{I}_t$ and ground-truth frame I_t,

$$\mathcal{L}_{\text{frame}} = \|I_t - \hat{I}_t\|_2, \tag{13}$$

and $\mathcal{L}_{\text{SSIM}}$ is the structure similarity (SSIM) loss [36] to measure the perceptual difference between $\hat{I}_t$ and I_t,

$$\mathcal{L}_{\text{SSIM}} = \text{SSIM}(I_t, \hat{I}_t). \tag{14}$$

In addition, we use a feature constraint loss to minimize the discrepancy between the query and addressed keys. So the normal queries are sufficiently reconstructed with only the top K memory keys, while for unknown anomalous queries, these K keys are not sufficient to represent it and give larger reconstruction error. The constraint loss is defined as:

$$\mathcal{L}_{\text{con}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \|v_{ij} - q_{ij}\|_2. \tag{15}$$

Dynamic addressing involves only some of the memory keys for each query, and the feature constraint loss enforces the query and the aggregated memory key close to each other. To prevent all memory keys from being close to each other, we impose a diversity loss [26] as the memory regularization, which increases the distinction between memory features, especially between normal and abnormal. The object function is maximize the mean square error between memory items,

$$\mathcal{L}_{\text{div}} = -\frac{1}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} ||k_n - k_{n'}||_2. \tag{16}$$

Note that λ_1, λ_2 and λ_3 are hyper-parameters to balance each loss during training.

3.4. Anomaly Score

The frame interpolation model is trained on video sequences with normal frames, while for unknown anomalous samples it yields higher interpolation errors. Thus, the anomaly score is represented by the interpolation error as below,

$$\mathcal{S}_{\text{int}} = \mathcal{L}_{\text{frame}} = ||I_t - \hat{I}_t||_2. \tag{17}$$

Since abnormal queries incur errors when addressing in the memory pool, we also refer to $\mathcal{L}_{\text{con}}$ as the anomaly score:

$$\mathcal{S}_{\text{con}} = \mathcal{L}_{\text{con}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} ||v_{ij} - q_{ij}||_2. \tag{18}$$

The overall anomaly score is defined as

$$\mathcal{S} = \alpha \phi(\mathcal{S}_{\text{int}}) + (1-\alpha)\phi(\mathcal{S}_{\text{con}}), \tag{19}$$

where $\phi(*)$ denotes the min-max normalization and α is a hyper-parameter to balance $\mathcal{S}_{\text{int}}$ and $\mathcal{S}_{\text{con}}$.

4. Experiments

4.1. Setup

Datasets. Three benchmark datasets are used to valuate our proposed method.

1) UCSD Ped2 [15] dataset is composed of 16 training videos and 12 testing video depicting pedestrian moving

parallel to the camera plane, the samples are in 240 × 360 pixels resolution. The crowd density varies from sparse to crowded following natural undulation. Anomalous pedestrian motions and non pedestrian entities including bikers, skaters, and people walking across a walkway, are considered as abnormal cases.

2) The CUHK Avenue[22] dataset contains 16 training videos and 21 testing videos from camera overlooking a busy sidewalk. The It includes 47 unusual events such as throwing objects, loitering, and running, the spatial resolution of each frame is 600×360.

3) ShanghaiTech Campus [18] dataset consists of 274k training and 42k testing frames with 130 irregular events, covering 13 different scenes of size 856×480. Compared with other datasets, ShanghaiTech is more challenging because of the diversity of scenes, multiple view angles, complicated light conditions, and the introduction of sudden motion like chasing and brawling.

Implementation Details. We use PyTorch [33] to implement the proposed method. The frames are resized to 256×256 and normalize the pixel values to the range of $[-1, 1]$ for all three datasets. The memory size is same as MNAD [32] and MPN [26] at 10. Various value of dynamic selection number K in constraint loss are also explored from 1 to 9. The balance weights in the objective function are empirically set as $\lambda_1 = 0.0001, \lambda_2 = 1, \lambda_3 = 0.0001$. The model is optimized by Adam optimizer. The learning rate is initialized to be 0.0002, and decayed to 0 in the last epoch monitored by Cosine Annealing [20] scheduler. Training epochs are set to $100, 50, 20$ for Ped2, Avenue and ShanghaiTech respectively. We set the batch size as 8 for all datasets. The experiments are conducted with dual Nvidia RTX-3090 GPUs in the form of parallel training, it takes about $8, 12, 40$ hours for training phase on Ped2, Avenue and ShanghaiTech respectively. For inference, we choose $\alpha = 0.3$ to balance the anomaly score between interpolation error and feature constraint error. Our code will be released once the paper is accepted.

Evaluation. We use the AUC (Area Under Curve) as the measurement for frame-level score. It's obtained by computing the areas under the ROC curve through a varying threshold for the anomaly score of the frame interpolation.

4.2. Quantitative Comparison

As shown in Tab.1, we compare state-of-the-art frame-level unsupervised video anomaly detection methods on UCSD Ped2 [27], CUHK Avenue [21] and ShanghaiTech Campus [17], including both reconstruct-based and predict-based methods. The primary comparison methods are memory-augmented models, which are $MemAE$ [9], $MNAD$ [32], $AMMC$ [4] and MPN [26]. In particular, $MemAE$ [9] is a reconstruction-based model and $AMMC$ [4] is a prediction-based model, while $MNAD$ [32] and MPN [26] provide results for both. In addition, we also compare the proposed model with other reconstruction-based and prediction-based methods, including 2D Auto-Encoder ($2DAE$) [11], 3D Auto-Encoder ($3DAE$) [39], Temporally-coherent Sparse Coding (TSC) [24], stacked Recurrent Neural Network ($sRNN$) [25], Spatiotemporal Consistency-enhanced Network ($STCEN$) [10], Appearance-motion Correspondence (AMC) [30], Frame Prediction ($Frame - Pred$) [17], Video Prediction and Compression (VPC) [16] and Spatio-temporal Dissociation (STD) [5]. Note that the prediction-based methods are generally better than the reconstruction-based methods. In particular $MNAD$ [32] and MPN [26] set up for comparison in the same settings also show this trend. Also, we present the comparison with the sequence-based interpolation approach, Siamese generative network ($SIGnet$) [8]. Our proposed frame-based interpolation method achieves better results overall, by avoiding direct input of both appearance and motion information. Additionally, the object-level methods VEC [38] and HF [19] are also shown as references. In contrast to frame-level settings [9], object-level methods use an extra object detection model to extract objects in the video. Although comparisons with frame-level methods [38, 19] are unfair as they use a prior knowledge model, our method still achieves comparable results.

	Method\Dataset		Ped2	Avenue	Campus
Object	Rec.	HF-R [19]	98.8%	86.8%	73.1%
	Pre.	VEC [38]	97.3%	89.6%	74.8%
		HF-P [19]	94.5%	90.2%	76.2%
Frame	Rec.	2DAE [11]	85.0%	80.0%	60.9%
		3DAE [39]	91.2%	77.1%	-
		MNAD-R [32]	90.2%	82.8%	69.8%
		TSC [24]	91.0%	80.6%	67.9%
		sRNN [25]	92.2%	81.7%	68.0%
		MemAE [9]	94.1%	83.3%	71.2%
		STCEN [10]	96.9%	86.6%	73.8%
		AMC [30]	96.2%	86.9%	-
		MPN-R [26]	96.2%	87.1%	71.9%
	Pre.	MPN-P [26]	92.6%	85.2%	71.1%
		VPC [16]	93.6%	85.4%	-
		Frame-Pred [17]	95.4%	84.9%	72.8%
		STD [5]	96.7%	87.1%	73.7%
		MNAD-P [32]	97.0%	88.5%	70.5%
		AMMC [4]	96.6%	86.6%	73.7%
	Int.	SIGnet [8].	96.2%	86.8%	-
		Ours w/o Mem.	98.2%	86.9%	73.4%
		Ours w/ Mem.	**98.9%**	**89.7%**	**75.0%**

Table 1. Comparison with state-of-the-art methods on the UCSD Ped2 [27], CUHK Avenue [21] and ShanghaiTech Campus [17] in terms of AUC score. The results in bold denote the best performance of frame-level anomaly detection. **Rec.**, **Pre.**, and **Int.** indicate reconstruction-based methods, prediction-based methods, and frame interpolation-based method, respectively.

4.3. Ablation Study

Method\Dataset		Ped2	Avenue	Campus
Rec.	MemAE [9]	91.7%	81.0%	69.7%
	MNAD-R [32]	86.4%	80.6%	65.8%
Pre.	MNAD-P [32]	94.3%	84.5%	66.8%
	AMMC [4]	95.1%	84.9%	71.5%
	MPN [26]	95.1%	83.9%	66.7%
Int.	Ours	**98.2%**	**86.9%**	**73.4%**

Table 2. Comparison with the **baselines** with out memory module used in state-of-the-art in terms of AUROC.

Comparison on baselines. We compare with the baselines of memory-augmented methods, such as $MemAE$ [9], $MNAD$ [32], $AMMC$ [4] and MPN [26]. The detailed comparison is shown in Table 2, showing the results of methods without the memory module. Our approach and $AMMC$ [4] achieve promising results on all three datasets because both involve a concern on motion abnormalities. However, for the detection of motion anomalies, $AMMC$ [4] utilizes an extra network for the prediction of the optical flow. As a baseline, our proposed frame interpolation method is superior in unsupervised video anomaly detection.

Analysis on components. To evaluate the effectiveness of each component of the proposed framework, a detailed ablation comparison is implemented on the CUHK Avenue [21] dataset as shown in Table. 3. In addition to pixel-level interpolation loss $\mathcal{L}_{\text{frame}}$, we also implement structural similarity loss $\mathcal{L}_{\text{SSIM}}$ to improve the perceptual effect and thus obtain a small gain. As skip-connects can cause anomalies to leak and decrease anomaly scores [19], we insert a memory module in the last layer to prevent this. The dynamic memory mechanism is promoted from two aspects. At first, we implement a feature constraint $\mathcal{L}_{\text{con}}$ to make the dynamically addressed memories represent the normal query and add the diversity $\mathcal{L}_{\text{div}}$ loss to enlarge the distance between memory items. By doing so, a compact normal feature space is formed without affecting the memory capacity. Secondly, the feature constrain can be seen as feature-level reconstruction, which can be used as a unique anomaly score $\mathcal{S}_{\text{con}}$. The feature-level anomaly score $\mathcal{S}_{\text{con}}$ achieves the AUROC of 85.5% and improve the interpolation-based anomaly score $\mathcal{S}_{\text{int}}$ slightly. Compared to the baseline and vanilla memory module, the proposed method improves by 2.8% and 2.2% AUROC scores respectively.

Impact of memory quantity. The impact of the dynamic quantity K for the memory addressing is demonstrated in Fig. 5. The anomaly scores $\mathcal{S}_{\text{int}}$, $\mathcal{S}_{\text{con}}$ and $\mathcal{S}$ are denoted as $int.$, $con.$ and $mix.$ respectively. We also explore the evaluation results for different memory capacities N with respect to the ratio of K. The memory effectiveness reaches an optimal distinction between normal and abnormal events at $K = 5, N = 10$, while achieving the best AUROC result.

Network	Loss			Score		AUROC
Memory	$\mathcal{L}_{\text{SSIM}}$	$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{con}}$	$\mathcal{S}_{\text{con}}$	$\mathcal{S}_{\text{int}}$	
✗	✗	-	-	-	✓	86.2%
✗	✓	-	-	-	✓	86.9%
✓	✓	✗	✗	✗	✓	87.5%
✓	✓	✓	✗	✗	✓	87.7%
✓	✓	✓	✓	✗	✓	89.5%
✓	✓	✓	✓	✓	✗	88.5%
✓	✓	✓	✓	✓	✓	89.7%

Table 3. Ablation study results under different settings on CUHK Avenue in terms of AUROC.

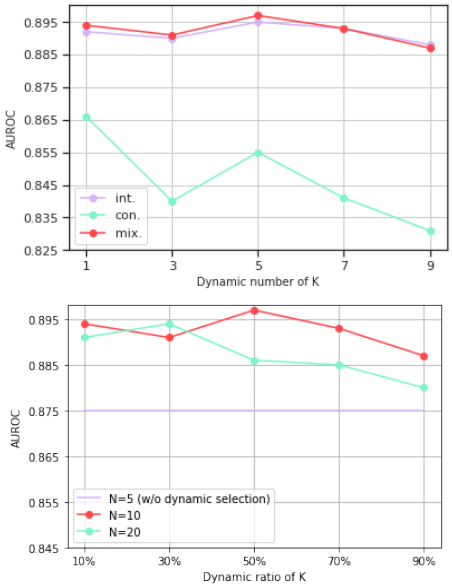

Figure 5. Analysis on the memory capacity N and quantity of dynamic memory items K in terms of AUROC on CUHK Avenue [21] dataset

Visualization. Three examples of anomaly scores over time for all frames in a sequence on UCSD Ped2 [27], CUHK Avenue [21] and ShanghaiTech Campus [17] are shown in Figure 6, respectively. The line charts show the anomaly scores of all frames of a video sequence, by which temporal changes in both normal and abnormal events can be intuitively observed. We release a more clear visualization in Fig. 7 to show the effect of frame interpolation on anomalous events, where the anomaly map represents the pixel-level interpolation error. For better visual expression, we apply the Gaussian smoothing on the anomaly map. The interpolation results on UCSD Ped2 [27] represent an enhancement of bidirectional interpolation for anomalous mo-

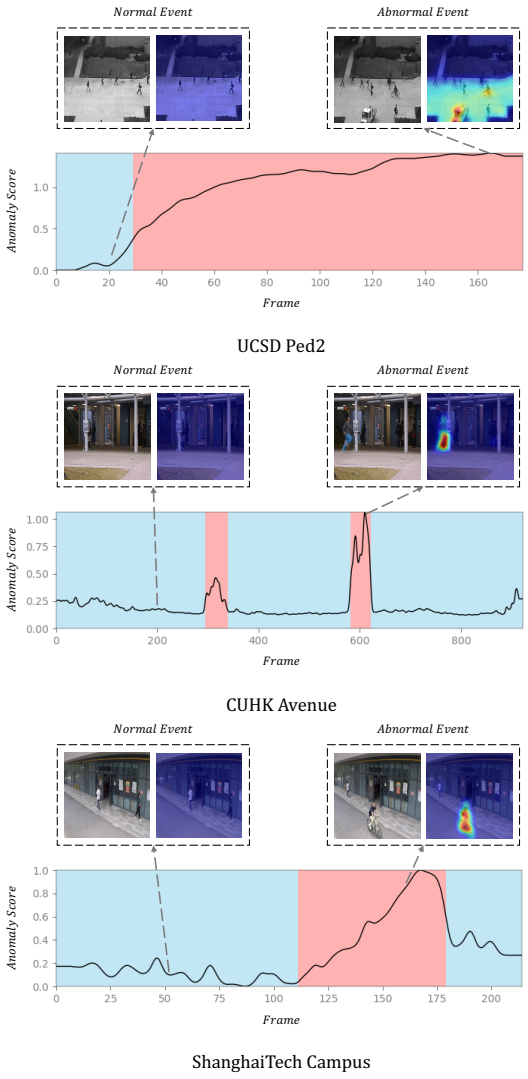

Figure 6. Visualization of variations in anomaly scores for normal and abnormal events. Abnormal scores for normal events are highlighted in blue, while abnormal events are red. The heat map shows the anomaly localization, with blue to red indicating the rise in the anomaly score.

tion detection, with significant interpolation errors seen in the interpolation column where the bicycle appears to be overlapped. The overlap arises because motion delays occur in both relative directions. The example from CUHK Avenue [21] dataset shows the abnormal behavior "$running$" with attention on the "$legs$" of the object. The anomalous event "$fight$" appears on the ShanghaiTech Campus [17] and our model attention is on the "arm".

5. Conclusions

This paper proposed an intermediate frame interpolation framework as a novel paradigm for unsupervised anomaly detection in video surveillance. This framework included

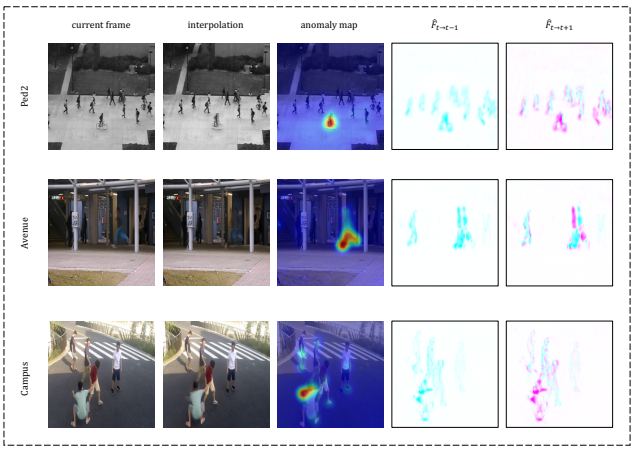

Figure 7. Visualization examples of interpolated frame, anomaly map and optical flow on UCSD Ped2 [27], CUHK Avenue [21] and ShanghaiTech Campus [17]. The first two columns indicate the current frame and the interpolated frame. The abnormal events for the above three datasets are "$bicycle$", "$running$" and "$fight$" respectively. The anomaly region is visualized by heatmap in the third column. The last two columns show the estimated optical flow from the moment t to the moments $t-1$ and $t+1$.

an optical flow estimation network and an interpolation network jointly trained on normal video sequences. In the inference phase, an anomaly was reflected by the interpolation error. Unlike previous approaches, our pipeline blocked the direct input of anomalous appearance and abnormal motion patterns, enlarging the cumulative error on anomalous events. In addition, we introduced the dynamic memory mechanism to enhance the discrepancy between normality and abnormality in the feature space. Our dynamic memory mechanism constrained the representation of abnormal features without compromising the memory capacity for normal features. The excellent results of frame-level video anomaly detection on public benchmarks verified the effectiveness of the proposed framework.

Limitations. The key of the proposed method is to interpolate the normal frames with small, or no, errors, but abnormal frames with large errors. Since only two frames are fed into the model, video content with obstacle views or occlusions poses a challenge to our approach. We demonstrate some failure cases in the supplementary document.

Potential negative social impact. Although video anomaly detection was developed to solve real-life problems, such as traffic control and urban policing. However, there are still factors that make this technology potentially harmful to society. For example, frame interpolation can be used as a means of video forgery to create false evidence and evade surveillance, etc. We call for attention to such issues and encourage researchers to develop more socially friendly technologies in the future.

References

[1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[3] Sovan Biswas and R Venkatesh Babu. Real time anomaly detection in h. 264 compressed videos. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 1–4. IEEE, 2013.

[4] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proc. AAAI*, pages 938–946, 2021.

[5] Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition*, 122:108213, 2022.

[6] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020.

[7] Valentin Durand De Gevigney, Pierre-François Marteau, Arnaud Delhay, and Damien Lolive. Video latent code interpolation for anomalous behavior detection. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3037–3044. IEEE, 2020.

[8] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and Feng Yang. Anomaly detection with bidirectional consistency in videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[9] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] Yi Hao, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognition*, 121:108232, 2022.

[11] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[13] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

[16] Bowen Liu, Yu Chen, Shiyu Liu, and Hun-Seok Kim. Deep learning in latent space for video prediction and compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 701–710, 2021.

[17] 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.

[18] 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.

[19] 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, October 2021.

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[21] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[22] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[23] 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 439–444. IEEE, 2017.

[24] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017.

[25] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1070–1084, 2019.

[26] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta

prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15425–15434, June 2021.

[27] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1975–1981. IEEE, 2010.

[28] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 498–507, 2018.

[29] Rashmiranjan Nayak, Umesh C Pati, and Santos K Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106, 2011.

[30] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019.

[31] Jonathan Pan. Physical integrity attack detection of surveillance camera with deep learning based video frame interpolation. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 79–85. IEEE, 2019.

[32] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch, 2017.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[35] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[37] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

[38] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020.

[39] 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*, pages 1933–1941, 2017.

[40] David Zimmerer, Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, Klaus Maier-Hein, Tobias Roß, Tim Adler, Annika Reinke, and Lena Maier-Hein. Medical out-of-distribution analysis challenge 2022, Mar. 2022.