# Self-Supervised Sparse Representation for Video Anomaly Detection

Jhih-Ciang Wu*[1,2], He-Yen Hsieh*[1], Ding-Jie Chen[1], Chiou-Shann Fuh[2], and Tyng-Luh Liu**[1]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] National Taiwan University, Taiwan

**Abstract.** Video anomaly detection (VAD) aims at localizing *unexpected* actions or activities in a video sequence. Existing mainstream VAD techniques are based on either the one-class formulation, which assumes all training data are *normal*, or weakly-supervised, which requires only video-level normal/anomaly labels. To establish a unified approach to solving the two VAD settings, we introduce a *self-supervised sparse representation* (S3R) framework that models the concept of anomaly at feature level by exploring the synergy between dictionary-based representation and self-supervised learning. With the learned dictionary, S3R facilitates two coupled modules, *en-Normal* and *de-Normal*, to reconstruct snippet-level features and filter out normal-event features. The self-supervised techniques also enable generating samples of pseudo normal/anomaly to train the anomaly detector. We demonstrate with extensive experiments that S3R achieves new state-of-the-art performances on popular benchmark datasets for both one-class and weakly-supervised VAD tasks. Our code is publicly available at `https://github.com/louisYen/S3R`.

**Keywords:** sparse representation, video anomaly detection

## 1 Introduction

These days surveillance/security cameras are ubiquitously deployed in various public places, such as factories, offices, shopping malls, and intersections. To strengthen public safety, it is constructive to automatically detect abnormal events such as accidents, illegal activities, or crimes. In practice, abnormal events are rare and diverse in nature; manually identifying abnormal events is laborious and time-consuming, especially for long-duration video sequences. To facilitate recognizing the varied anomalies, developing intelligent computer vision algorithms, i.e., video anomaly detection (VAD) systems, is a pressing need.

Recent efforts to tackle the VAD task can be categorized into unsupervised and weakly-supervised techniques, depending on the annotations or assumptions about the training video sequences. The unsupervised VAD scenario, which we

* Both authors contributed equally to this work.
** Corresponding author. E-mail:`liutyng@iis.sinica.edu.tw`

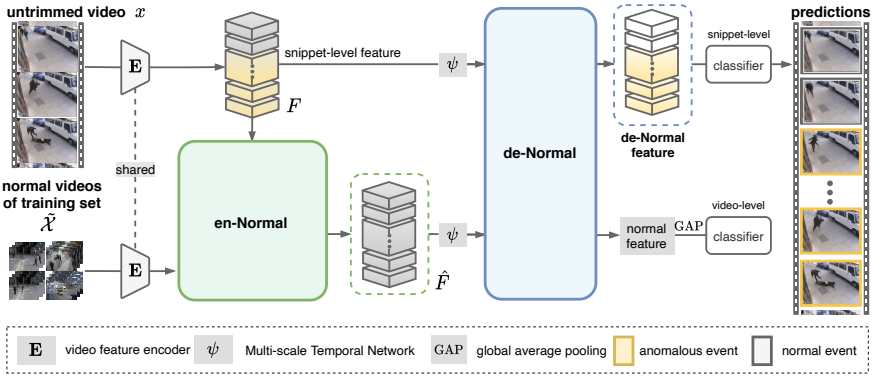


Fig. 1: The proposed S3R framework couples dictionary learning with self-supervised techniques to model the concept of feature-level anomaly. First, a feature extractor **E** represents each untrimmed video $x$ as the snippet-level feature $F$, and all the normal training videos $\tilde{\mathcal{X}}$ are collected to build the task-specific dictionary. Next, the en-Normal module employs $F$ and the dictionary to reconstruct the feature $\hat{F}$. Then, the de-Normal module explores $F$ and $\hat{F}$ differences to filter out the normal-event patterns. Finally, the filtered features are ready to discriminate the normal and anomalous events of the snippet-level and video-level features.

instead refer to as one-class VAD, assumes that only anomaly-free videos are available for training. The widely adopted approaches to discriminating normal and anomalous patterns are embedding-space learning or data reconstructing. The weakly-supervised VAD assumes that video-level normal/anomaly labels are given for training. Compared to the unsupervised VAD, obtaining such video-level labels requires more human effort but could achieve significant performance gains. A popular strategy to tackle weakly-supervised VAD is the inclusion of multiple instance learning (MIL). Specifically, an MIL-based weakly-supervised VAD algorithm treats each video and snippet as the bag and instance respectively, and the annotation for each bag, indicating whether a bag contains at least one anomalous instance, is known in the training stage.

In dealing with a VAD task, exhaustively modeling all possible scenarios of abnormal events is infeasible. Our method casts VAD as an out-of-distribution problem. A video clip that cannot be well reconstructed or explained by the normal-event dictionary is supposed to involve abnormal events. To realize such an idea, we develop the self-supervised sparse representation (S3R) framework to model the concept of feature-level anomaly by generalizing a dictionary-based representation with self-supervised techniques. We further infuse the MIL strategy into the proposed S3R to form a unified reconstruction-based method for effectively solving both unsupervised VAD and weakly-supervised VAD tasks.

In sum, S3R learns a normal-event dictionary for generating two opposite network modules, i.e., en-Normal and de-Normal, to reconstruct snippet-level

features and filter out the normal-event features. These two modules complement each other and enable the processed features to be better discriminated by our snippet-level and video-level anomaly classifiers. With the aid of self-supervised techniques, we can generate more pseudo anomaly data concerning a specific dictionary to optimize the anomaly detector training. Since all samples in inference are unseen from the training stage, S3R indeed can adequately distinguish between unseen normal and unseen anomalous snippets.

We validate the usefulness of S3R by conducting experiments on both one-class and weakly-supervised VAD tasks, which include model evaluations on three popular datasets, i.e., ShanghaiTech, UCF-Crime, and XD-Violence. We also ablate each module within S3R to evaluate their effectiveness. To our knowledge, S3R is the first unified framework that can be applied to both one-class and weakly-supervised VAD task. We highlight the main contributions as follows.

- We introduce a novel self-supervised sparse representation (S3R) framework for modeling and generating the feature-level anomalies through the (offline) learned dictionary and self-supervised learning. Our experimental results support the advantage of such a strategy in addressing the VAD task.
- We propose two coupled modules, en-Normal and de-Normal, leading to a unified framework for tackling both one-class and weakly-supervised tasks.
- Our method achieves significant performance gains over other state-of-the-art on one-class and weakly-supervised video anomaly detection tasks.

## 2 Related Work

### 2.1 Anomaly Detection

Anomaly detection aims to discover the irregular pattern with subtle or significant differences to the normal data. With the remarkable progression for deep neural networks, several types of research on anomaly detection are prospering. Ruff *et al.* [29] used the simulated image-based dataset and tackled it in the one-class framework in the early periods due to the absence of the corresponding data. The one-class anomaly detection intends to determine whether the test image belongs to the said class or not. Following the development of one-class anomaly detection, the industrial dataset named MVTec AD [2], which with pixel-level annotation for the manufacturing inspection, is proposed. The purpose of anomaly detection using MVTec AD focuses on image-level anomaly classification or pixel-level anomaly localization. Various works handle MVTec using different manners such as knowledge distillation [3], self-supervised learning [18], and meta-learning [45].

Another more challenging anomaly detection leverages temporal information, known as video anomaly detection, searching for unexpected actions or illegal activities in a video clip. Specifically, it is demanding to estimate whole anomalous patterns for all types of anomaly detection in real-world applications. Therefore, the approaches for numerous types of anomaly detection are usually completed

in an unsupervised manner, assuming that only access normal data during training while it has been unsuitably termed as the unsupervised VAD. Several works engage in one-class VAD from different perspectives. For instance, Liu *et al.* [20] took VAD as a video prediction framework and measured the anomaly score based on the gap between the ground-truth the predicted future frame. Another work [21] proposed a multi-level memory-augmented autoencoder with skip connection conditioned on reconstructed optical flow. Moreover, several approaches [11,15] embedded a pre-trained object detector into the model and used motion cues to deal with VAD. Recently, the VAD with weak supervision [34] that contains video-level labels in the training stage shows noticeable progress. It is considerable and trades the human annotations off against performance. Several approaches have evident improvement compared to unsupervised VAD. For example, RTFM [36] uses feature magnitude with the multi-scale temporal scenario from the video to select the top-k snippets and determine whether it belongs to an abnormal video or not. MSL [19] proposes multi-sequence learning and designs the exclusive ranking loss to select the most anomalous sequence. In contrast to previous works, we introduce an architecture with flexibility to deal with one-class and weakly-supervised VAD together.

### 2.2 Video Feature Extractors

Recently, neural network-based models have achieved a substantial performance boost for tackling the action recognition task and serve as powerful video feature extractors to obtain robust representations in downstream tasks. These popular models fall into two major categories of two-stream networks [9,32,41] and 3D networks [5,28,37,47]. The two-stream network exploits RGB images and stacked optical flow clues separately to generate appearance and motion features. The 3D networks directly employ raw video volumes to learn spatio-temporal representations. In this paper, we follow current efforts on VAD and employ the latter style as the video feature extractor, *i.e.* I3D, to encode untrimmed video and acquire snippet-level representation $F$.

### 2.3 Self-Supervised Sparse Dictionary Learning

The goal of dictionary learning is to find a linear combination using the elements in a dictionary and keep the sparsity of the weights as possible at the same time. With the optimization for dictionary learning, the redundant atoms are filtered out, and pivot ones are preserved [1]. Cong *et al.* [8] proposed sparse reconstruction cost over a dictionary to estimate the anomaly score for local and global abnormal events in the testing stage. Lu *et al.* [22] adopted this strategy to encode normal event patterns in the surveillance video and boost the running time speed by constraint the sparsity coefficient. Luo *et al.* [23] proposed temporally-coherent sparse coding accompanying a stacked recurrent neural network to speed up the time of the testing phase. In contrast to the most of former works focusing on acceleration, we explore the capability of sparse representation learning and optimize all features obtained from the video feature

extractor to formulate the universal dictionary $D_U$ and task-specific dictionary $D_T$. Notably, since the most standard video feature extractors such as C3D or I3D are pre-trained on Kinetics-400 [16], we formulate the $D_U$ and $D_T$ using Kinetics-400 and the target dataset, respectively. The resulting dictionaries are then employed to feature reconstruction and pseudo label generation.

Self-supervised learning aims to increase the labels without manual annotation. [10] Some recent works deal with VAD adopt this strategy by generating pseudo labels. For example, Pang *et al*. [26] proposed formulating the ordinal regression as a pretext task. The model initially learns anomaly scores for pseudo normal and anomaly-free frames and applies the pseudo label to an end-to-end detector. Feng *et al*. [10] proposed to train a generator via MIL and predict the pseudo label for the segments of anomalous videos and address VAD in the self-training scheme. In this paper, we introduce pseudo label generation to deal with VAD. By comparison, we generate video-level pseudo labels in latent representation space with non-parametric sparse dictionary learning.

## 3 Our Method

Learning to carry out video anomaly detection is often cast in two different settings. The first is a one-class formulation that the provided training data include only the video samples describing the underlying normal activities. Despite that the one-class scenario has explicitly assumed the training data are all from the normal category, it has been unsuitably termed as the *unsupervised* VAD task in most previous works [12,35,43]. Departing from the anomaly-free assumption, the other popular setting is called the *weakly-supervised* VAD task. In this case, video samples in the training set are categorized by their video-level label into normal (label 0) and anomaly (label 1); however, the frame-level labels are not available to precisely locate exact segments of abnormal activities. For the ease of presentation, we hereafter refer to the two settings of video anomaly detection as oVAD ("o" for one-class) and wVAD ("w" for weakly-supervised).

We aim at developing a unified reconstruction-based method that can be effectively applied to solve both oVAD and wVAD tasks. To this end, we consider establishing a dictionary learning approach, coupling with self-supervised techniques, to model the concept of *anomaly* at the feature level, no matter which of the two VAD settings we are exploring. In the following sections, we will first elaborate our method for tackling the oVAD task as the problem is more challenging due to the lack of anomaly samples in the training data, and then explain how our method is also applicable to solving the wVAD task.

### 3.1 Sparse Representation for oVAD

One-class VAD assumes that only anomaly-free videos are accessible in the training set $\mathcal{X} = \{\mathbf{x}_i\}$. Now given an untrimmed frame-level video $\mathbf{x} \in \mathcal{X}$, we decompose $\mathbf{x}$ into the snippet-level video sequence $\mathcal{V} = \{\mathbf{v}_t\}_{t=1}^{T}$ of $T$ snippets, where each snippet $\mathbf{v}_t$ comprises 16 consecutive frames. We follow previous work

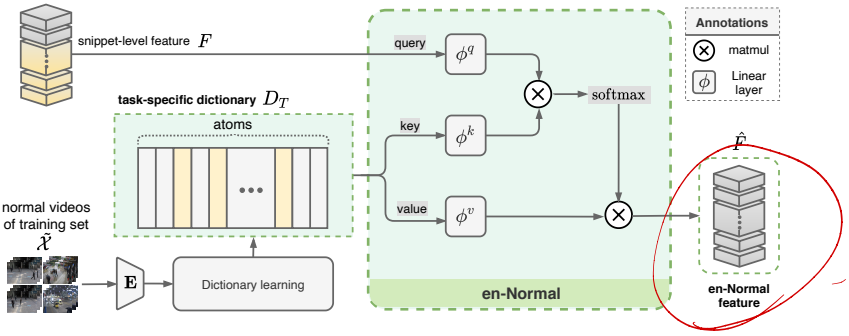


Fig. 2: The pipeline for en-Normal. This module takes the snippet-level feature $F$ and task-specific dictionary $D_T$ to reconstruct feature $\hat{F}$ via an attention mechanism.

[6, 10, 19, 36] to adopt a pre-trained I3D network [5] as the default feature extractor $\mathbf{E}$ to each snippet-level video $\mathcal{V}$, resulting in snippet-level representations $F = \{\mathbf{f}_t\}_{t=1}^{T}$, where $\mathbf{f}_t \in \mathbb{R}^C$ stands for each encoded snippet feature.

The dictionary learning [17, 25] presumes an overcomplete basis, and prefers a sparse representation to succinctly explain a given sample. With the training set $\mathcal{X}$, whose video samples are anomaly-free, we are motivated to learn its corresponding dictionary $D$ of $N$ atoms. More specifically, we apply dictionary learning technique to each representation $F = \mathbf{E}(x) \in \mathbb{R}^{T\times C}$ and optimize as

$$\underset{D,\{\mathbf{w}_t\}}{\operatorname{argmin}} \sum\nolimits_{\mathbf{x}\in\mathcal{X}} \sum\nolimits_{t=1}^{T} (\|\mathbf{f}_t - D\mathbf{w}_t\|^2 + \lambda\|\mathbf{w}_t\|_0)\,, \tag{1}$$

where $D \in \mathbb{R}^{C\times N}$ is the resulting VAD dictionary, and $\mathbf{w}_t \in \mathbb{R}^N$ is the coefficient vector constrained by the sparsity prior. Since the derivation of $D$ is specific to the training dataset $\mathcal{X}$, we will use the notation $D_T$ to emphasize that the underlying dictionary from (1) is *task-specific*.

### 3.2 A Dictionary with Two Modules

With the learned task-specific dictionary $D_T$ from (1), we can design two opposite network components: the *en-Normal* and *de-Normal* modules. Given a snippet-level feature $F$, the former is used to obtain its reconstructed normal-event feature, while, on the contrary, the latter is applied to filter out the normal-event feature. The two modules complement each other and are central to our approach to anomaly video detection.

*en-Normal Module.* With the learned task-specific dictionary $D_T$, we design a dictionary-based attention module to better correlate the snippet-level feature $F$ and the resulting $D_T$, leading to the corresponding normal-event feature $\hat{F}$. That is, since $D_T$ is assumed to span the feature space of all normal-event patterns,

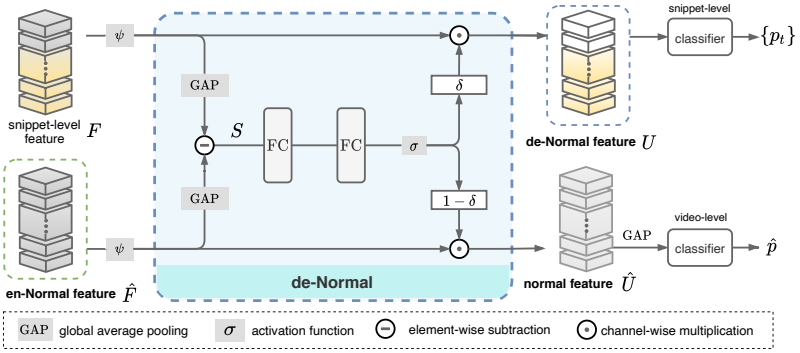


Fig. 3: The illustration of the de-Normal module. This module takes the channel-wise difference between $F$ and $\hat{F}$ to form the cross-video semantics $S$. Then, the channel scale $\delta$ is derived to depress $S$ for describing normal events.

we use the attention mechanism [38,42] to reweigh snippet-level input feature $F$ with respect to $D_T$ to obtain its reconstructed normal-event feature $\hat{F}$. In particular, we employ linear embeddings $\phi$ to project $F$ and $D_T$. The attention is computed in the embedding space and defined as

$$\hat{F} = \text{softmax}\left(\phi^q(F)\phi^k(D_T)^\intercal\right)\phi^v(D_T)\,, \tag{2}$$

where $\phi^q$, $\phi^k$, and $\phi^v$ separately represent linear functions to derive *query*, *key*, and *value* embeddings as in [38,42]. Thus, we adaptively involves normal-event patterns from the dictionary $D_T$ based on $F$ to reconstruct normal feature $\hat{F}$. Fig. 2 depicts how the normal-event feature $\hat{F}$ is obtained from an input feature $F$ and $D_T$.

*de-Normal Module.* Opposite to the previous design, the de-Normal module aims to depress the normal-event patterns within the input video feature. Thus, patterns related to normal events are expected to be filtered out, and the remaining can be used to infer whether the input video includes anomalous events or not. In practice, given the snippet-level feature $F \in \mathbb{R}^{T\times C}$ and the reconstructed normal feature $\hat{F} \in \mathbb{R}^{T\times C}$, we first explore the temporal dependency via the multi-scale temporal network (MTN) [36] and retrieve the enhanced features as $\psi(F) \in \mathbb{R}^{T\times C}$ and $\psi(\hat{F}) \in \mathbb{R}^{T\times C}$, where $\psi$ denotes the MTN operation. Next, we employ the *global average pooling* (denoted as $g(\cdot)$) temporally to collect global events, where each channel includes normal or anomalous semantics. We express the retrieved cross-video semantics of $g(\psi(F)), g(\psi(\hat{F})) \in \mathbb{R}^C$ as $S \in \mathbb{R}^C$ which is formulated as their channel-wise difference,

$$S = g(\psi(F)) - g(\psi(\hat{F}))\,, \tag{3}$$

Notice that the cross-video semantics $S$ from (3) remove the normal-event semantic channels. Hence, the cross-video semantics $S$ is able to depress semantic

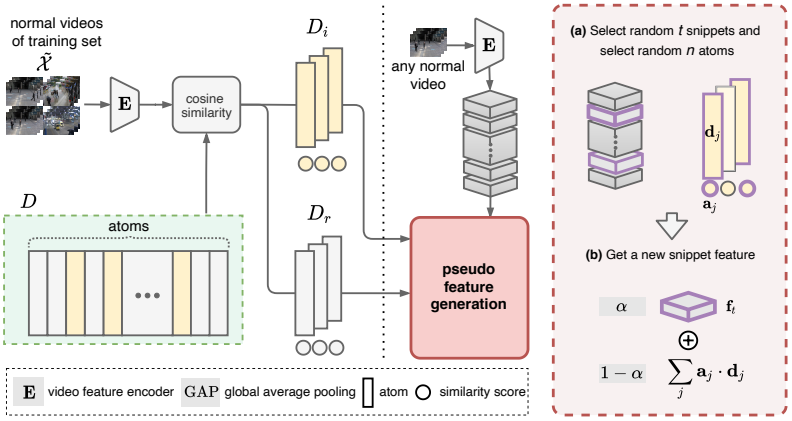


Fig. 4: Pseudo feature generation. A learned dictionary $D$ is first divided into two equal-size dictionaries involving irrelevant atoms $D_i$ and relevant atoms $D_r$. Then, randomly selected atoms are fused to generate the new snippet features.

features for describing those normal events. To further keep the anomalous-event channels of cross-video semantics and simultaneous depress normal-event channels, we employ SENet-style [14] operations to explore the channel-wise relationship and derive the corresponding channel scales for depressing normal event within the input video representation as

$$\delta = \sigma(\mathrm{MLP}(S))\,, \tag{4}$$

$$U = \delta \odot F, \quad \hat{U} = (1-\delta) \odot \hat{F}\,. \tag{5}$$

where MLP comprises two fully-connected layers to probe the channel-wise relationship, $\sigma$ denotes the *sigmoid* activation, and $\odot$ means the channel-wise multiplication. The scale vector $\delta \in \mathbb{R}^c$ indicates the channel-level weights to keep anomalous events, while $1-\delta$ denotes channel weights for focusing on normal events. Finally, we use multiple fully-connected layers to predict snippet-level $P = \{p_t\}$ and video-level $\hat{p}$ probability using $U$ and $g(\hat{U})$, respectively. (See Fig. 3.)

### 3.3 Dictionary-based Self-Supervised Learning

We have described how to learn a task-specific dictionary $D_T$ from anomaly-free training data, and use it to establish two useful modules for achieving video anomaly detection. However, as illustrated in Fig. 1, the overall training of the proposed model has implicitly assumed the availability of training data comprising anomaly events (*i.e.*, of label 1). Whereas the oVAD setting does not provide training data other than anomaly-free, we propose effective self-supervised techniques to generate *pseudo anomaly* data with respect to a given dictionary $D$.

Assume now we are given a training set $\tilde{\mathcal{X}}$, the steps to generate a pseudo anomaly video based on $D$ are listed below.

1. Collect all normal videos in $\tilde{\mathcal{X}}$ to form the training set $\mathcal{X}$. (This step is for the consideration of wVAD; otherwise, we already have $\mathcal{X} = \tilde{\mathcal{X}}$.)
2. For each atom in $D$, compute its averaged cosine similarity to $\mathcal{X}$, and obtain a ranking list according to their similarity scores in ascending order.
3. Divide $D$ into two equal-size dictionaries, $D = D_i \cup D_r$, where $D_i$ includes those *irrelevant* atoms from the first half of the ranking list, and $D_r$ comprises the remaining *relevant* atoms. The self-supervised scheme uses $D_i$ to generate pseudo anomaly features and $D_r$ for pseudo normal features.
4. By sampling from $\mathcal{X}$, a normal video with representations $F = \{\mathbf{f}_t\}_{t=1}^{T}$, we create an anomalous video by replacing its $2 \times t$ snippets as follows:
   (a) We randomly select $t$ snippets from $F$. For each snippet, we randomly select $n$ atoms $\{\mathbf{d}_j\}_{j=1}^{n}$ from $D_i$ as pseudo anomalous candidates.
   (b) We further apply the weighted fusion to get a new snippet feature $\hat{\mathbf{f}}_t = \alpha \mathbf{f}_t + (1-\alpha)\sum_j \mathbf{a}_j \cdot \mathbf{d}_j$, where $\mathbf{a}_j$ denotes the weight vector. (We set $\alpha = 0.01$ for pseudo anomaly, and 0.5 for pseudo normal.)
   (c) Repeat 4-(a) and 4-(b) steps for replacing the other $t$ snippets with pseudo normal features from relevant atoms of $D_r$.
5. The process of creating a pseudo anomaly video is completed.

In our study, we have considered two reasonable choices of $D$. The first is simply the task-specific dictionary $D_T$, and the second is a task-independent *universal dictionary* $D_U$, which is optimized via (1) over the Kinetics-400 [16] dataset. Notice that in learning $D_U$, we do not need any label information, and instead consider $D_U$ as a general dictionary to account for a rich variety of activities. The steps of pseudo anomaly generation are illustrated in Fig. 4.

### 3.4 S3R: A Unified VAD Framework

To show the proposed self-supervised sparse representation (S3R) is indeed a unified framework for solving both the oVAD and wVAD problems, we are left to justify that our method works equally well for the weakly-supervised scenario. Assume that we are given the training dataset $\tilde{\mathcal{X}}$ for solving the wVAD task. We can readily collect those anomaly-free videos (with label 0) to form the dataset $\mathcal{X}$ and obtain the corresponding dictionary $D_T$ from (1). Then all other procedures remain the same as before except that we now have the choice to decide whether the technique to generate pseudo anomaly data is employed or not.

Our network is end-to-end trained and built upon RTFM [36] with respect to the following multi-task objective/loss:

$$\mathcal{L} = \mathcal{L}_{sep} + \gamma \mathcal{L}_{cls}\,, \tag{6}$$

where $\mathcal{L}_{sep}$ measures the separability of normal and anomalous videos, and $\mathcal{L}_{cls}$ optimizes the snippet-level and video-level classifiers. The weight $\gamma$ balances the

two loss terms and is set to 0.001. In addition, we have

$$\mathcal{L}_{sep} = \sum_{k} \left( \left| m - \|U^{+} = \{u_t^{+}\}_k\|_2 \right| + \|U^{-} = \{u_t^{-}\}_k\|_2 \right), \tag{7}$$

where $m$ is the adopted margin ($m = 100$ in our experiments) and $\|\cdot\|_2$ represents the $\ell_2$-norm operation. We denote $U^{+} = \{u_t^{+}\}_k$ and $U^{-} = \{u_t^{-}\}_k$ as the top-$k$ feature magnitude of $U$ when $y = 0$ and $y = 1$, respectively. Finally, $\mathcal{L}_{cls}$ is the binary logistic regression loss defined by

$$\mathcal{L}_{cls} = BCE(P = \{p_t\}_k, y) + BCE(\hat{p}, y), \tag{8}$$

where $P = \{p_t\}_k$ denotes the top-$k$ snippet probabilities based on the feature magnitude of $U$ as in RTFM ($k$ is set to 3). Following [34], we also adopt temporal smoothness and sparsity regularization in our implementation. Please refer to [36] for the details of training the MIL model.

## 4 Experiments

### 4.1 Dataset and Metric

We evaluate our S3R against SOTA methods on three datasets: ShanghaiTech [20], UCF-Crime [34], and XD-Violence [46]. Notably, ShanghaiTech is used for one-class and the others are for weakly-supervised VAD primitively. To facilitate the evaluation for both settings, we choose the existing variants or follow the previous procedure to create corresponding types of supervision for VAD. Specifically, Zhong *et al.* [49] transferred ShanghaiTech to weak supervision VAD by reorganizing the dataset. Sun *et al.* [35] collects all normal training videos as the training set and remains the same in the testing set in UCF-Crime to perform oVAD. We use the same criteria as the former one to obtain an one-class version XD-Violence. We briefly state the composition of each dataset in the following and report details in the form of a table in supplementary material.

**ShanghaiTech** The ShanghaiTech contains 437 videos from 13 scenes of campus surveillance. The original one-class version comprises 330 regular videos and 107 irregular videos, carrying 130 abnormal events for training and testing, respectively. After reorganization, it retains 238/199 videos that cover all 13 scenes for training/testing in the weakly-supervised setting.

**UCF-Crime** The UCF-Crime has 1900 surveillance videos covering 13 real-world anomalous classes such as robbery, explosion, and road accident. Compared to ShanghaiTech, which nearly includes pedestrian activities in the university, the scenes in this dataset are more diverse and more complex. The number of videos for training/testing is 1610/290 in incipient weakly-supervised requirement and reduce the number of training to 800 by discarding anomalous videos for the unsupervised assumption.

Table 1: Comparison of frame-level AUC performance for VAD on ShanghaiTech. We present the current SOTA with the corresponding feature and published year. S3R$^\star$ and S3R$^\dagger$ indicate using $D_U$ to generate pseudo labels and reconstruct features, respectively. TSN$_{flow}$ and I3D$_{flow}$ represent only access flow, and I3D$_{f2}$ means access to both frame and flow.

| oVAD | | | | wVAD | | | |
|---|---|---|---|---|---|---|---|
| Method | Feature | Year | AUC (%) | Method | Feature | Year | AUC (%) |
| Conv-AE [13] | - | 2016 | 60.85 | Sultani *et al.* [34] | I3D | 2018 | 85.33 |
| Stacked-RNN [23] | - | 2017 | 68.00 | GCN-Anomaly [49] | C3D | 2019 | 76.44 |
| Frame-Pred [20] | - | 2018 | 73.40 | GCN-Anomaly [49] | TSN$_{flow}$ | 2019 | 84.13 |
| Mem-AE [12] | - | 2019 | 71.20 | GCN-Anomaly [49] | TSN | 2019 | 84.44 |
| MNAD [27] | - | 2020 | 70.50 | AR-Net [39] | I3D$_{flow}$ | 2020 | 82.34 |
| VEC [48] | - | 2020 | 74.80 | AR-Net [39] | I3D | 2020 | 85.38 |
| STC Graph [35] | - | 2020 | 74.70 | AR-Net [39] | I3D$_{f2}$ | 2020 | 91.24 |
| CAC [44] | - | 2020 | 79.30 | MIST [10] | C3D | 2021 | 93.13 |
| AMMC [4] | - | 2020 | 73.70 | MIST [10] | I3D | 2021 | 94.83 |
| HF2-VAD [21] | - | 2021 | 76.20 | RTFM [36] | C3D | 2021 | 91.51 |
| ROADMAP [43] | - | 2021 | 76.60 | RTFM [36] | I3D | 2021 | 97.21 |
| SVD-GAN [30] | - | 2021 | 78.42 | MSL [19] | C3D | 2022 | 94.81 |
| BDPN [7] | - | 2022 | 78.10 | MSL [19] | I3D | 2022 | 96.08 |
| S3R | I3D | 2022 | 79.89 | S3R | I3D | 2022 | **97.48** |
| S3R$^\star$ | I3D | 2022 | **80.47** | S3R$^\dagger$ | I3D | 2022 | 97.47 |

**XD-Violence** The XD-Violence is the latest and the most large-scale dataset, which involves 4754 untrimmed videos together with audio signals. The sources of scenery are various, including surveillance, movies, dashcam, games, etc. The videos number for one-class and weakly-supervised scenarios are 3954/800 and 2049/800, respectively. To measure the effectiveness of our model fairly, we use the same features as previous works that access video only.

**Metric** For evaluating the model performance in VAD, we calculate the Area Under Curve (AUC), a conventional threshold-independent metric used for earlier works. We follow [46] for evaluating the XD-Violence experiment and use the same Average Precision (AP) metric to compare the performance.

### 4.2 Implementation Details

For a fair comparison, we adopt the I3D network [5] pre-trained on Kinetics-400 [16] as [6, 10, 19, 36] for the video feature extraction. During training, we train our S3R through Adam optimizer with a batch size of 64 for 50 epochs on all dataset, and sample each video with 32 snippets via the linear interpolation, *i.e.* $T = 32$. Furthermore, we randomly sample 32 normal and 32 anomalous videos to form a mini-batch under wVAD and oVAD settings. Notably, we establish anomalous

Table 2: Comparison of frame-level AUC performance for VAD on UCF-Crime. S3R$^\star$ and S3R$^\dagger$ indicate using $D_U$ to generate pseudo labels and reconstruct features, respectively.

| oVAD | | | | wVAD | | | |
|---|---|---|---|---|---|---|---|
| Method | Feature | Year | AUC (%) | Method | Feature | Year | AUC (%) |
| SVM Baseline | - | | 50.00 | Sultani *et al.* [34] | I3D | 2018 | 77.92 |
| Conv-AE [13] | - | 2016 | 50.60 | GCN-Anomaly [49] | TSN | 2019 | 82.12 |
| S-SVDD [33] | - | 2018 | 58.50 | MIST [10] | I3D | 2021 | 82.30 |
| Lu *et al.* [22] | C3D | 2013 | 65.51 | Wu *et al.* [46] | I3D | 2020 | 82.44 |
| BODS [40] | I3D | 2019 | 68.26 | RTFM [36] | I3D | 2021 | 84.30 |
| GODS [40] | I3D | 2019 | 70.46 | Chang *et al.* [6] | I3D | 2021 | 84.62 |
| STC Graph [35] | RPN | 2020 | 72.70 | MSL [19] | I3D | 2022 | 85.30 |
| S3R | I3D | 2022 | 77.15 | S3R | I3D | 2022 | **85.99** |
| S3R$^\star$ | I3D | 2022 | **79.58** | S3R$^\dagger$ | I3D | 2022 | 85.00 |

videos when training through a dictionary for the oVAD setting , *e.g.* $D_T$ or $D_U$, as mentioned in Sec. 3.3. Following the previous work [36], our S3R uses the learning rate of 0.001 for ShanghaiTech and UCF-Crime, and 0.0001 for XD-Violence.

### 4.3 Results of oVAD

Previous methods [4, 7, 12, 13, 20, 21, 22, 23, 27, 30, 31, 33, 35, 40, 43, 44, 48] deal with VAD in the one-class setup. The left part in Table 1, 2 and 3 show the comparison results of the oVAD on the corresponding dataset. We provide a variant S3R$^\star$ that adopts the universal dictionary $D_U$ for pseudo label generation. As seen in Table 1, 2 and 3, our model outperforms the other state-of-the-art models for all benchmarks. Our model achieves new art on ShanghaiTech, UCF-Crime, XD-Violence, improving around 1.2%, 6.9% and 2.7%, respectively.

### 4.4 Results of wVAD

We consider VAD approaches under weakly-supervised fashions in recent year, including [6, 10, 19, 34, 36, 39, 46, 49]. The right part in Table 1, 2 and 3 show the comparison results of the weakly-supervised VAD on the corresponding dataset. The feature column without emphasis shows that the extractor accesses the frame solely. In particular, we report the AP scores that utilize video but discard audio for proper comparison on XD-Violence.We provide a variant S3R$^\dagger$ that adopts the universal dictionary $D_U$ for the en-Normal. As seen in Table 1, 2 and 3, our model outperforms the other state-of-the-art models for all datasets. Our model achieves new art on all benchmarks, improving around 0.3%, 0.7% and 2%, respectively.

Table 3: Comparison of AP performance on XD-Violence. S3R$^\star$ and S3R$^\dagger$ indicate using $D_U$ to generate pseudo labels and reconstruct features, respectively.

| oVAD | | | | wVAD | | | |
|---|---|---|---|---|---|---|---|
| Method | Feature | Year | AUC (%) | Method | Feature | Year | AP (%) |
| - | - | - | - | Sultani *et al.* [34] | I3D | 2018 | 75.68 |
| SVM Baseline | - | - | 50.78 | Wu *et al.* [46] | I3D | 2020 | 75.41 |
| OCSVM [31] | - | 1999 | 27.25 | RTFM [36] | I3D | 2021 | 77.81 |
| Conv-AE [13] | - | 2016 | 30.77 | MSL [19] | I3D | 2022 | 78.28 |
| S3R | I3D | 2022 | 51.64 | S3R | I3D | 2022 | **80.26** |
| S3R$^\star$ | I3D | 2022 | **53.52** | S3R$^\dagger$ | I3D | 2022 | 79.54 |

### 4.5 Ablation Study

To verify the effectiveness of each module in the S3R, we consider four configurations for the model deal with wVAD, *i.e.*, baseline without de-Normal and dictionary, S3R with de-Normal using $\tilde{\mathcal{X}}_{avg}$, $D_U$ or $D_T$, respectively. Table 4 shows the ablation study on these configurations. All the models are end-to-end trained and under the same remaining configuration. Precisely, the baseline model is similar to RTFM since we adopt MTN and also build an MIL-based model. Consequently, the AUC of baseline does not perform much of a difference to RTFM. The second configuration employs de-Normal without any dictionary but uses the averaged feature of all normal training videos. The configuration shows the benefit and effectiveness of the proposed de-Normal module, which significantly improves AUC on ShanghaiTech and UCF-Crime. The last two configurations ablate our full model by utilizing the different dictionaries. Particularly, S3R using the task-specific dictionary obtains the best score with a broad margin on UCF-Crime.

Another ablation exploits the composition of the pseudo label. As shown in Table 5, we generate pseudo normal and pseudo anomaly by referring to several ratios. Notably, the snippets and atoms are selected according to their mutual similarity rather than the hand-crafted annotations [24]. The ratio for anomaly and normal is 25%, *i.e.*, $T/4$ snippets are replaced, which obtains the best score in our framework.

Table 6 ablates the channel reduction rate for en-Normal and de-Normal modules on the ShanghaiTech dataset under the oVAD setting, respectively. The first row shows different rates of the embedding layers, *i.e.* $\phi^q$ and $\phi^k$. With the 25% reduction rate, we obtain the best performance of 80.47 in AUC. As the reduction rate increases or decreases, the performance drops at least 2.59% in AUC. The second row ablates the channel reduction of MLP in (4) for the de-Normal module. With the 25% rate, we get the worst performance of 66.14% in AUC. Using the rate of 6.25%, we improve the performance by 14.33% AUC.

Table 4: Ablation study on *S3R's modules* tackling wVAD task with AUC metric and AUC's improvement against the baseline on ShanghaiTech and UCF-Crime.

| Configuration | | ShanghaiTech | | UCF-Crime | |
|---|---|---|---|---|---|
| de-Normal | en-Normal | AUC (%) | improvement | AUC (%) | improvement |
| - | - | 96.97 | - | 83.42 | - |
| ✓ | $\tilde{\mathcal{X}}_{avg}$ | 97.28 | ↑ 0.77 | 84.19 | ↑ 0.77 |
| ✓ | $D_U$ | 97.47 | ↑ 0.19 | 85.00 | ↑ 0.81 |
| ✓ | $D_T$ | **97.48** | ↑ 0.20 | **85.99** | ↑ 1.80 |

Table 5: Ablation study on *snippet ratio* tackling oVAD task with AUC metric on ShanghaiTech. A and N represent the ratio of anomaly and normal, respectively.

| ratio (A / N) | 25% / 25% | 25% / 0% | 25% / 50% | 25% / 12.5% | 50% / 25% | 12.5% / 25% |
|---|---|---|---|---|---|---|
| AUC (%) | **80.47** | 79.59 | 79.18 | 76.46 | 78.02 | 60.43 |

Table 6: Ablation study on *channel reduction rate* in en-Normal module (2) and de-Normal module (4) tackling oVAD task with AUC metric on ShanghaiTech.

| channel reduction rate | 50% | 25% | 12.5% | 6.25% | 3.125% |
|---|---|---|---|---|---|
| en-Normal (embedding layers $\phi^q$, $\phi^k$ in (2)) | 70.38 | **80.47** | 77.88 | 76.01 | 73.36 |
| de-Normal (MLP in (4)) | 72.73 | 66.14 | 70.31 | **80.47** | 68.77 |

## 5 Conclusion

We establish a self-supervised sparse representation framework, a unified model for simultaneously tackling both oVAD and wVAD tasks. At the core of S3R is to model the feature-level anomaly through the offline trained dictionary and self-supervised learning. Our design results in two opposite modules. The first module, en-Normal, is in charge of reconstructing normal-event features, while the second one, de-Normal, filters out the normal-event feature. By using the self-supervised techniques, we are able to further generate the pseudo anomaly/normal data concerning the learned dictionary to guide the training of our anomaly detector. The extensive experiments on three public benchmarks show that S3R consistently surpasses state-of-the-art oVAD and wVAD methods, demonstrating that our unified reconstruction-based framework effectively solves both one-class and weakly-supervised video anomaly detection tasks.

**Acknowledgements**. This work was supported in part by the MOST grants 110-2634-F-007-027, 110-2221-E-001-017 and 111-2221-E-001-015 of Taiwan. We are grateful to National Center for High-performance Computing for providing computational resources and facilities.

## References


1. Barlow, H.B.: Single units and sensation: a neuron doctrine for perceptual psychology? Perception **1**(4), 371–394 (1972)
2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In: CVPR. pp. 9592–9600 (2019)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: CVPR. pp. 4183–4192 (2020)
4. Cai, R., Zhang, H., Liu, W., Gao, S., Hao, Z.: Appearance-motion memory consistency network for video anomaly detection. In: AAAI. pp. 938–946 (2021)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR. pp. 4724–4733 (2017)
6. Chang, S., Li, Y., Shen, J.S., Feng, J., Zhou, Z.: Contrastive attention for video anomaly detection. IEEE Transactions on Multimedia (2021)
7. Chen, C., Xie, Y., Lin, S., Yao, A., Jiang, G., Zhang, W., Qu, Y., Qiao, R., Ren, B., Ma, L.: Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In: AAAI (2022)
8. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: CVPR. pp. 3449–3456 (2011)
9. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR. pp. 1933–1941 (2016)
10. Feng, J.C., Hong, F.T., Zheng, W.S.: Mist: Multiple instance self-training framework for video anomaly detection. In: CVPR. pp. 14009–14018 (2021)
11. Georgescu, M.I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: Anomaly detection in video via self-supervised and multi-task learning. In: CVPR. pp. 12742–12752 (2021)
12. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV. pp. 1705–1714 (2019)
13. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR. pp. 733–742 (2016)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
15. Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L.: Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: CVPR. pp. 7842–7851 (2019)
16. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
17. Kreutz-Delgado, K., Murray, J.F., Rao, B.D., Engan, K., Lee, T.W., Sejnowski, T.J.: Dictionary learning algorithms for sparse representation. Neural computation **15**(2), 349–396 (2003)
18. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: CVPR. pp. 9664–9674 (2021)
19. Li, S., Liu, F., Jiao, L.: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In: AAAI (2022)
20. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: CVPR. pp. 6536–6545 (2018)

21. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: ICCV. pp. 13588–13597 (2021)
22. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: ICCV. pp. 2720–2727 (2013)
23. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnn framework. In: ICCV. pp. 341–349 (2017)
24. Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J.: Localizing anomalies from weakly-labeled videos. IEEE transactions on image processing **30**, 4505–4515 (2021)
25. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML. pp. 689–696 (2009)
26. Pang, G., Yan, C., Shen, C., Hengel, A.v.d., Bai, X.: Self-trained deep ordinal regression for end-to-end video anomaly detection. In: CVPR. pp. 12173–12182 (2020)
27. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: CVPR. pp. 14372–14381 (2020)
28. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: ICCV. pp. 5534–5542 (2017)
29. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: ICML. pp. 4393–4402 (2018)
30. Samuel, D.J., Cuzzolin, F.: Svd-gan for real-time unsupervised video anomaly detection. In: BMVC (2021)
31. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: NIPS. pp. 582–588 (1999)
32. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. pp. 568–576 (2014)
33. Sohrab, F., Raitoharju, J., Gabbouj, M., Iosifidis, A.: Subspace support vector data description. In: ICPR. pp. 722–727 (2018)
34. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488 (2018)
35. Sun, C., Jia, Y., Hu, Y., Wu, Y.: Scene-aware context reasoning for unsupervised abnormal event detection in videos. In: ACMMM. pp. 184–192 (2020)
36. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV. pp. 4975–4986 (2021)
37. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
39. Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: ICME. pp. 1–6 (2020)
40. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: ICCV. pp. 8201–8211 (2019)
41. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36 (2016)
42. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)

43. Wang, X., Che, Z., Jiang, B., Xiao, N., Yang, K., Tang, J., Ye, J., Wang, J., Qi, Q.: Robust unsupervised video anomaly detection by multipath frame prediction. IEEE Transactions on Neural Networks and Learning Systems (2021)
44. Wang, Z., Zou, Y., Zhang, Z.: Cluster attention contrast for video anomaly detection. In: ACMMM. pp. 2463–2471 (2020)
45. Wu, J.C., Chen, D.J., Fuh, C.S., Liu, T.L.: Learning unsupervised metaformer for anomaly detection. In: ICCV. pp. 4369–4378 (2021)
46. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV. pp. 322–339 (2020)
47. Xu, H., Das, A., Saenko, K.: R-C3D: region convolutional 3d network for temporal activity detection. In: ICCV. pp. 5794–5803 (2017)
48. Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., Kloft, M.: Cloze test helps: Effective video anomaly detection via learning to complete video events. In: ACMMM. pp. 583–591 (2020)
49. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: CVPR. pp. 1237–1246 (2019)