Cluster Attention Contrast for Video Anomaly Detection

Ziming Wang∗
wziming@pku.edu.cn
ADSPLAB, School of ECE, Peking University
ShenZhen, China

Yuexian Zou∗†
zouyx@pku.edu.cn
ADSPLAB, School of ECE, Peking University
ShenZhen, China

Zeming Zhang
zhangzm@hit.edu.cn
Harbin Institute of Technology
Harbin, China

ABSTRACT

Anomaly detection in videos is commonly referred to as the discrimination of events that do not conform to expected behaviors. Most existing methods formulate video anomaly detection as an outlier detection task and establish normal concept by minimizing reconstruction loss or prediction loss on training data. However, these methods performances suffer drops when they cannot guarantee either higher reconstruction errors for abnormal events or lower prediction errors for normal events. To avoid these problems, we introduce a novel contrastive representation learning task, Cluster Attention Contrast, to establish subcategories of normality as clusters. Specifically, we employ multi-parallel projection layers to project snippet-level video features into multiple discriminate feature spaces. Each of these feature spaces is corresponding to a cluster which captures distinct subcategory of normality, respectively. To acquire the reliable subcategories, we propose the Cluster Attention Module to draw the cluster attention representation of each snippet, then maximize the agreement of the representations from the same snippet under random data augmentations via momentum contrast. In this manner, we establish a robust normal concept without any prior assumptions on reconstruction errors or prediction errors. Experiments show our approach achieves state-of-the-art performance on benchmark datasets.

CCS CONCEPTS
• Computing methodologies → Computer vision; Scene anomaly detection.

KEYWORDS
video anomaly detection, cluster attention contrast, contrastive learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Requests for permissions to permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION

Anomaly detection in videos is an important task because of its extensive applications in intelligent surveillance, activity recognition, evidence investigation, etc. Since anomalous events are rarely seen in common environments, anomalies are often defined as behavioral or appearance patterns different from usual patterns in previous work [2, 3]. It is extremely challenging because of the ambiguity and complexity of anomaly definition in real applications. Obviously, it is almost infeasible to collect all examples of abnormal events and perform traditional supervised learning methods. Therefore, for a typical setting for anomaly detection tasks, only normal scenarios are given in the training data. Based on this definition, most recently anomaly detection approaches [3, 6, 26, 29, 37, 52, 59] are based on outlier detection and learn a model of normality from training videos. During inference, events are labeled as abnormal if they deviate from the normality model. These methods for modeling normal concept can be categorized into two types, namely the reconstruction-based framework [11, 12, 34, 52] and the prediction-based framework [36, 60]. They both regard events that do not fit into the normality models as abnormal, but their performances
suffer when they cannot guarantee either higher reconstruction errors for abnormal events or lower prediction errors for normal events. However, these methods do not explicitly distinguish between normal conditions that differ greatly. For example, on the sidewalks on campus, there are both normal when no one to pass by or when a large number of pedestrians pass by after school, but the two conditions have completely different motion patterns.

Different from previous methods, we propose a novel contrastive representation learning task, Cluster Attention Contrast, to obtain multiple subcategories of normality. We appreciate the diversity of normal conditions and exploit these gaps to establish a robust normal concept. We propose to employ a contrastive learning paradigm to establish reliable subcategories of normality as clusters. Specifically, we employ multi-parallel projection layers to project snippet-level video features into multiple discriminate feature spaces. Each of these feature spaces is corresponding to a cluster which captures distinct subcategory of normality, respectively. To acquire the reliable subcategories, we propose the Cluster Attention Module and strive to draw the optimal cluster attention representation of each snippet by maximizing the agreement of the representations from the same snippet under random data augmentations. During inference, for each test snippet, the highest similarity between its subcategory-specific representations and corresponding centers is regarded as its regularity score. We detect anomalies based on this intuition: if a sample is far away from the subcategory-specific centers in all corresponding feature spaces, we can determine that an anomaly has occurred.

Our contribution is summarized as follows:

- We introduce a novel contrastive representation learning task, Cluster Attention Contrast, to establish a robust normal concept for video anomaly detection task. As far as we know, this is the first work in the related field.
- We propose an effective contrastive learning framework for video anomaly detection. Our method achieves state-of-the-art results on two benchmark datasets: CUHK Avenue, ShanghaiTech Campus, where it reaches 87.0% and 79.3% frame-level AUC score, respectively.

We organize the paper as follows. We present related work in Section 2 and describe our approach in Section 3. We conduct the video anomaly detection experiments and elaborate on the implementation details in Section 4. We draw our conclusions in Section 5.

2 RELATED WORK

Video Anomaly Detection. As one of the most challenging problems, anomaly detection in videos has been extensively studied for many years. Video Anomaly Detection is commonly formalized as an outlier detection task, most the research addresses the problem under the assumption that anomalies are rare or unseen, and behaviors deviating from normal patterns are supposed to be anomalous. Researchers attempt to encode regular patterns via a variety of statistic models, e.g., the social force model [37], the mixture of dynamic models on texture [29], Hidden Markov Models on video volumes [15, 27], the Markov Random Field upon spatial-temporal domain [26], Gaussian process modeling [6, 28], and identify anomalies as outliers. With the great success of deep learning, the recent video anomaly detection work can be classified into deep reconstruction methods and deep prediction methods. Hasan et al. [12] employed two autoencoders, one is learned on conventional handcrafted features, while another is learned in an end-to-end fashion with a fully convolutional network. On the other hand, Liu et al. [31] proposed a future frame prediction network for anomaly detection. What more closely related to our work are methods [6, 7, 10, 33, 33, 42] that learn a dictionary of atoms representing normal events during training, they utilize sparse representation to construct a dictionary for normal behavior and detect anomalies as the ones with high reconstruction error. The main difference between these methods and our work is that we do not sparsely encode features to evaluate reconstruction loss but directly judge the occurrence of anomalies by the similarities between its subcategory-specific representations and corresponding centers.

Deep Clustering. Over the years, many methods have been developed for deep clustering. Some earlier deep clustering methods divided representation learning and clustering into two stages, draw on SAE [4, 14] or its variants [35, 38, 48, 57] to extract intermediate features, and then employ k-means [29, 35] or spectral clustering [29]. In contrast to them, some studies [8, 51, 56, 61] integrate the two stages into a unified framework. At the same time, some methods [16, 18, 53] directly use the convolutional neural network (CNN) or multi-layer perceptron (MLP) for representation learning by designing specific loss functions instead of the reconstruction loss. Furthermore, Semantic deep clustering methods have recently gained more attention for clustering, IIC [24] directly trains a classification network by maximizing the mutual information between original data and their transformations. However, the computation of mutual information requires an immense batch size in the training process, which is challenging to apply on large images. The most similar method to our approach is DMC [17], which performs clustering by aligning the cluster centers of two modal features. DMC [17] uses pretrained image features to guide the learning and clustering of speech features in a teacher-student way. Besides, it needs to be clarified that our approach is distinguished from the concept of deep subspace clustering methods [23, 58]. The main difference is our method directly projects the original space into a lower-dimensional feature space instead of using a subset of the original space dimensions.

Contrastive Representation Learning. Contrastive representation learning makes use of internal patterns of data and devises predictive tasks to train a model. Many pretext tasks have been proposed: counting the objects [39], predicting the context [9], filling in missing parts of an image [41], or recovering colors from grayscale images [57]. As for videos, self-supervision strategies involve: predicting future [46], leveraging temporal continuity via tracking [49, 50], or preserving the equivariance of egomotion [22, 40, 62]. CMC [45] presented a contrastive learning framework that enables the learning of unsupervised representations from multiple views of a dataset. The principle of maximization of mutual information enables the learning of powerful representations. These methods can be thought of as building dynamic dictionaries. SimCLR [5] verified the impact of different types of data augmentation and project heads on contrastive learning and pointed out contrastive learning benefits from larger mini-batch sizes and longer training compared to its supervised counterpart. At the same time, contrastive learning
benefits from deeper and wider networks like supervised learning. MoCo[13] present Momentum Contrast as a way of building large and consistent dictionaries for unsupervised learning with a contrastive loss. Considering that video feature extraction requires a considerable amount of computing resources, we use the momentum contrast to train our network to obtain better performance with a small mini-batch size.

3 PROPOSED METHOD
3.1 Overview
Our method proposes to establish a robust normal concept without any prior assumptions on reconstruction errors or prediction errors. Unlike previous works, we introduce an instance discrimination pretext task, Cluster Attention Contrast, to obtain reliable subcategories of normality, which is the core contribution of this work. As illustrated in Figure 2, our model consists of two fully symmetric encoders. Each encoder mainly consists of three components: a pretrained feature extractor, a parallel projection layers group, and a Cluster Attention Module. We employ multi-parallel projection layers to project snippet-level video features into multiple discriminative feature spaces for capturing diverse normal patterns. We introduce the Cluster Attention Module to connect the Cluster Attention Contrast task to the Video Anomaly Detection task. We first construct the input of encoders by a temporal cuboid using a sliding window technique. Specifically, we stack $T = 8$ frames together as a snippet. It should be noted, all training and inference phases are performed on such snippets. During Training, We draw two random data augmentations on the snippet and send them to the two encoders, respectively. In each encoder, we first use a pretrained action classifier to extract snippet-level features and perform a layer normalization to improve regularization. Then we employ $K$ parallel projection layers to project snippet-level video features into $K$ discriminate feature spaces. In this work, $K$ is a manually pre-fixed parameter. Each of these $K$ feature spaces is corresponding to a cluster which captures distinct subcategory of normality. In Cluster Attention Module, we redraw the original snippet-level feature through the linear combination of its $K$ subcategory-specific representations. The linear combination weight is obtained by the similarity between the feature’s subcategory-specific representations and the corresponding cluster centers. After this redrawn snippet-level feature passes through a project head $H$, we can obtain the output of our encoder, i.e., the cluster attention representation.

The cluster attention representation is the link between the Video Anomaly Detection task and our contrastive representation learning task. On the one hand, for our instance discrimination pretext task, a high-quality cluster attention representation should satisfy the requirement to agree on the same sample under random data augmentation and to distinguish from different samples. On the other hand, due to this representation is obtained by subcategory-specific representations through our cluster attention mechanism, we need reliable projected feature spaces to support it. We utilize those $K$ feature spaces and corresponding cluster centers to construct a robust normal concept for Video Anomaly Detection task.

To estimate the cluster centers while training, we modify an EM-like optimization mechanism: alternately updating the cluster center and the model parameters, which makes a soft assignment based on the posterior probabilities and converges to a local optimum. It is worth noting that our method only calculates the centroid of features in determining feature space as the cluster center instead of introducing an additional clustering task.

Figure 2: Overview of the Cluster Attention Contrast task. We perform two random data augmentation transformations on the input snippet, and send them to the two encoders, respectively. We regard the output of encoder $q$ and the output of encoder $k$ as a positive pair if they originate from the same snippet, and otherwise as a negative sample pair. Finally, we employ a contrast loss to maximize agreement the cluster attention representation of the positive pairs.
Algorithm 1 Cluster Center Estimating

Input:
- temperature \( \tau_c \), number of clusters \( K \), feature \{\( u_i \)\} \( i=1 \) to \( M \)
- structure of project layers \( \{ p^q_1, p^q_2, \ldots, p^q_K \} \), \( \{ p^k_1, p^k_2, \ldots, p^k_K \} \)

Initialize the \( s^q_{i,j} = s^k_{i,j} = 0 \).

for all \( q \in \{ q, k \} \) do
  for all \( i = 1 \) to \( M \), \( j = 1 \) to \( K \) do
    \( h^q_{i,j} = p^q_j(u_i) \)
    \( w^q_{i,j} = \text{softmax}(s^q_{i,j}) \)
    \( c^q = \sum_{i=1}^{M} w^q_{i,j} h^q_{i,j} \) # update centers
    \( s^q_{i,j} = \frac{h^q_{i,j}^T c^q}{\| h^q_{i,j} \| \cdot \| c^q \|} \) # update similarity
  end for
end for

return cluster centers for each encoder,
\( \{ c^q_1, c^q_2, \ldots, c^q_K \}_{q \in \{ q, k \}} \in \mathbb{R}^{|q|} \), \( \{ c^k_1, c^k_2, \ldots, c^K_k \}_{k \in \{ q, k \}} \in \mathbb{R}^{|k|} \)

We elaborate in section 3.2 on estimating the center over the features through a differentiable approach, and introduce the Cluster Attention Module and Cluster Attention Contrast task in section 3.3.

3.2 Clustering

Establish Subcategories of Normality as Clusters. For simplicity, consider the task of obtaining the center of the subcategory \( j \) specific feature space of the encoder \( q \), which can be expressed as \( c^q_j \). The simplest idea is selecting the centroid data point of all features in current subcategory specific feature space as its center. Consider all the features in subcategory \( j \) specific feature space:
\( \{ h^q_{1,j}, h^q_{2,j}, \ldots, h^q_{M,j} \in \mathbb{R}^{|q|} \} \), and an alternative center \( \hat{c}^q_j \), hence, the objective function can be formulated as,

\[
\min \mathcal{F}(\hat{c}^q_j) = \sum_{i=1}^{M} d(h^q_{i,j}, \hat{c}^q_j)
\] (1)

where \( d(h^q_{i,j}, \hat{c}^q_j) \) means the distance between current point and the alternative center. However, simply introducing Eq.(1) into the deep networks will make it difficult to optimize by gradient descent, as the minimization function in Eq.(1) is a hard assignment of whether a data point is the center or not, which is not differentiable. Inspired by DMC[17] and DMC[17], we transform the hard assignment in Eq.(1) to a soft assignment problem and be a differentiable one. We regard the center of each cluster as a linear combination of the features in current subcategory specific feature space, and the weight of each feature in the center can be obtained by the following formula,

\[
w^q_{i,j} = \frac{e^{-d^q_{i,j}/\tau_c}}{\sum_{i=1}^{M} e^{-d^q_{i,j}/\tau_c}} = \text{softmax}(s^q_{i,j})
\] (2)

where \( s^q_{i,j} = -d(h^q_{i,j}, \hat{c}^q_j)/\tau_c \) means the similarity between current point and the alternative center.

 Unlike DMC[17], we propose to use \( s^q_{i,j} = h^q_{i,j}^T c_j/\tau_c \| h^q_{i,j} \| \| c_j \| \) as the similarity measure, which can flexibly control the softness of assignment by adjusting clustering temperature hyperparameter \( \tau_c \).

\[c^q_j = \sum_{i=1}^{M} w^q_{i,j} h^q_{i,j}
\] (3)

However, there remains another problem that the computation of \( s^q_i \) depends on the current center \( c \) which makes it difficult to get the direct update rules for the centers. We choose the same way as DMC[17] to alternatively update the coefficient \( s^q_{i,j} \) and center \( c^q_j \), Eq.(3) is modified into,

\[
c^q_j^{(r+1)} = \sum_{i=1}^{M} w^q_{i,j}^{(r)} h^q_{i,j}
\] (4)

Consistent Centers Use a Queue. Considering that mini-batch gradient descent is usually used for optimization during training, which cannot obtain a coherent center representation from the limited mini-batch size. To address this problem, we propose to use a feature queue to maintain the features of several previous batches. The introduction of the feature queue decouples the number of subcategory-specific representations involved in clustering from the mini-batch size. The new centers can be updated by performing clustering on more projected features that benefit from the feature queue. Our queue size can be much larger than a typical mini-batch size and be flexibly and independently set as a hyperparameter. We show the complete cluster center estimating process in Algorithm 1.

Avoiding Trivial Solutions. Furthermore, trivial solutions may appear when the cluster centers are not constrained. For example, in an extreme case, all cluster centers are highly similar, which is not
guaranteed to obtain distinguished subcategories. This will result in all subcategory-specific feature spaces focus on a few patterns and lose the remaining essential information for the normal concept. To overcome this problem, we recommend attaching constraints to align the centers obtained by the two encoders. In practice, we regard the cluster centers in the encoder q as the query, while the cluster centers in the encoder k as the key. The encoder q and encoder k are designed to have the same structure, and the cluster centers obtained from the corresponding project layers in the two encoders are considered as positive pairs. Then we employ the InfoNCE [47] loss to force the similarity between the subcategory-specific representations and corresponding centers.

Another typical trivial solution may appear in other clustering methods is that the vast majority of samples are assigned to a few clusters, but did not appear in our experiments. This is possible because this type of trivial solution does not contribute to minimizing the loss of the contrastive representation learning task we proposed in the next section.

Once we have obtained the current clustering center of each subcategory, we can perform Cluster Attention Contrast according to the similarity between the subcategory-specific representations and their corresponding centers.

### 3.3 Cluster Attention Contrast

#### Cluster Attention Module

In section 3.2, we have obtained the cluster centers in subcategory-specific feature spaces. The next problem to be solved is the approach to redraw the original feature using the K projected features and their corresponding centers. A naive idea is to directly formalize the K similarity scores between the subcategory-specific representations and corresponding centers as the final representation of each original snippet. However, this scheme is not ideal in our early experiments. We speculate that this manner may result in losing the abundant information in features and confuse our instance discrimination pretext task.

In this work, we propose to represent the original snippet-level features in a similar way as estimating the soft-assignment center in the K feature spaces. We redraw the snippet-level video features through a linear combination of its K subcategory-specific representations. The linear combination weight is evaluated by the similarity between the subcategory-specific representations and their corresponding centers. The higher the similarity score measures the larger the proportion occupies in the final representation. The above process can be formulated as,

\[
    w_{ij} = \frac{e^{-d_{ij}/\tau_c}}{\sum_k e^{-d_{ij}/\tau_c}} = \text{softmax} \left( s_{ij} \right) \tag{6}
\]

where \( s_{ij} = h_q^Tz_i / (\tau_c \| h_q \| \| z_i \|) \) means the similarity between current point and the corresponding center. Then the linear combination of the subcategory-specific representations is regarded as

\[
    z_i = \sum_{j=1}^K w_{ij} z_j
\]

#### Algorithm 2 Cluster Attention Contrast

<table>
<thead>
<tr>
<th>Input</th>
<th>batch size ( N ), structure of ( T_r ), structure of encoder ( q ), encoder ( k ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialize the queues and cluster centers.</td>
<td></td>
</tr>
<tr>
<td>for sampled minibatch ( { x_i }_{i=1}^N ) do</td>
<td></td>
</tr>
<tr>
<td>for all ( i \in {1, \ldots, N} ) do</td>
<td></td>
</tr>
<tr>
<td>draw two augmentation functions ( t \sim T, t' \sim T )</td>
<td></td>
</tr>
<tr>
<td>( x_i^t = t(x_i), x_i^{t'} = t'(x_i) )</td>
<td></td>
</tr>
<tr>
<td>( h_i^q = f_q \left( x_i^t \right), h_i^k = f_k \left( x_i^{t'} \right) )</td>
<td></td>
</tr>
<tr>
<td>( z_i^q = g_q \left( h_i^q, c_i \right), z_i^k = g_k \left( h_i^k, c_i \right) )</td>
<td></td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
<tr>
<td>update encoder network ( q ) to minimize ( L'(\mathcal{W}) ) in Eq.(9).</td>
<td></td>
</tr>
<tr>
<td>momentum update encoder network ( k ) parameters.</td>
<td></td>
</tr>
<tr>
<td>update key queue and feature queues.</td>
<td></td>
</tr>
<tr>
<td>update cluster centers ( {q_j^K}<em>{j=1}^K, {k_j^K}</em>{j=1}^K ) use algorithm 1.</td>
<td></td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
<tr>
<td>return encoder network ( q )</td>
<td></td>
</tr>
</tbody>
</table>

Finally, previous work [5] have empirically shown that using an additional projection before the contrastive loss can improve performance by suppressing the loss of information. Our early experiment concluded consistent with them, so we use an additional projection head \( z_i = H(r_i) \) in our later experimental settings. After the redraw features pass through the project head, we can obtain the output of the Cluster Attention Module, i.e., the cluster attention representation. We show the complete Cluster Attention Module design in Figure 3.

#### Momentum Contrast Training

Since Cluster Attention Module is the final component of our encoders, the cluster attention representation is the final output of our model. Next, follow the classic contrastive learning paradigm [5], we consider the output of encoder \( q \) and the output of encoder \( k \) as a positive pair if they originate from the same snippet, and otherwise as a negative sample pair. Then we employ the InfoNCE [47] loss to force the positive pairs are similar and the negative pairs are dissimilar,

\[
    L_C = \frac{1}{N} \sum_{i=1}^N \log \frac{\exp \left( z_i^q \cdot z_i^k / \tau_c \right)}{\sum_p \exp \left( z_i^q \cdot z_i^k / \tau_c \right)} \tag{8}
\]

By incorporating the constraint on the clusters, the total loss function is:

\[
    L'(\mathcal{W}) = \alpha_1 L_C + \alpha_2 L_D + \|\mathcal{W}\|_F \tag{9}
\]

where \( \mathcal{W} \) represents model weights, \( \alpha_i (i = 1, 2) \) are hyperparameters to weight the losses. However, the contrastive learning paradigm commonly consumes lots of resources particularly with respect to video feature extraction. It brings great challenges to model training. Inspired by MoCo [13], we recommend performing momentum contrast training to address our proposed Cluster Attention
DATA AUGMENTATION FOR CONTRASTIVE REPRESENTATION LEARNING

In this section, we evaluate our proposed method as well as the functionalities of different components on two publicly available anomaly detection datasets, including the CUHK Avenue dataset[33], and the ShanghaiTech Campus dataset[34].

4.1 Datasets and Evaluation Metric

Here we briefly introduce the datasets used in our experiments.

CUHK Avenue dataset. It contains 16 training videos and 21 testing ones with a total of 47 abnormal events, including throwing objects, loitering, and running. The size of people may change because of the camera position and angle.

ShanghaiTech Campus dataset. ShanghaiTech Campus dataset is a challenging anomaly detection dataset. It contains 330 training videos and 107 testing ones, which is taken in 13 different scenes with various camera angles and illumination. Abnormal activities are diverse and realistic, which include appearance anomalies like bicycles, skateboards, motorbikes, cars, strollers, and motion anomalies such as fighting, chasing, pushing, jumping, etc.

Evaluation Metric. In most previous works[12, 31, 33, 34, 60], a popular evaluation metric is to calculate the Receiver Operating Characteristic (ROC) by gradually changing the threshold of regular scores. Then the Area Under Curve (AUC) is accumulated to a scalar useful for performance evaluation. A higher value indicates better anomaly detection performance. In this paper, following the work[12, 33, 34], we leverage frame-level AUC for performance evaluation.

4.2 Implementation Details

Data Augmentation for Contrastive Representation Learning. For data augmentation, we use: (1) random multiscale crop and resize, (2) random horizontal flip, (3) random color distortions, (4) random rotation, and (5) random grayscale for contrastive representation learning. For each training snippet which stacking 8 frames, we first uniformly crop the frames with size 256 × 256 at a random location and then randomly perform the data augmentations on the 8-frame-group. Some examples are shown in Figure 4.

Pretrained Feature Extractor. Considering the contrastive learning paradigm commonly requires a larger batch size and more training steps than supervised learning, we chose a light and fast action classifier, PAN[55], as our feature extractor. PAN[55] designed a concise motion cue called Persistence of Appearance (PA), which focuses on modeling the small displacements of motion boundaries by calculating the pixel-wise differences between two adjacent frames in the feature space. We choose BN-Inception[19] pretrained on Kinetics-400[25] as the backbone, and extract features from its global_pool layer. As mentioned above, all training and inference phases are performed on the snippets which are temporal cuboids stacking 8 frames, while a typical video segment length in PAN[55] is 4. We strongly recommend performing layer normalization on the extracted features to improve regularization.

Project Layers. For each subcategory-specific feature space, we employ two linear layers to project the PAN[55] features to a 128-dimensional representation. One layer directly projects the PAN[55] features to 128-dimensional by a fully-connected layer, while the other one firstly performs a timescale max pooling, then projects the feature to 128-dimensional through another fully-connected layer. Finally, we weight sum the two 128-dimensional vector through a manually set parameter as the projected representation.

Training. We train our models using an SGD[43] optimizer with the default minibatch size of 128 snippets on 4 GPUs (i.e., 32 snippets per GPU). The SGD weight decay is $1e^{-4}$ and the SGD momentum is 0.9. We employ Shuffling Batch Normalization reported in [13] to avoid the information leakage caused by Batch Normalization[19]. We train models for 100 epochs with the initial learning rate of 0.015. Hyperparameters $a_1$, $a_2$ were empirically set to 1, $1e^{-3}$, respectively. The default clustering temperature $\tau_c$ and cluster attention temperature $\tau_a$ are both empirically set to 0.5. The best empirical setting for $\tau$ in Eq.10 is 0.999. The temperature $\tau$ of InfoNCE loss is 0.2. Our method introduces two types of queues to help accomplish the contrastive representation learning task. The default size of the feature queue is 16384 and those of the key queue is 8192. In the training phase, we firstly random crop all the input snippets to 256 × 256-pixel before employ the data augmentation transformations.

4 EXPERIMENTS

In this section, we evaluate our proposed method as well as the functionalities of different components on two publicly available anomaly detection datasets, including the CUHK Avenue dataset[33], and the ShanghaiTech Campus dataset[34].

Contrast task. In practice, we maintain a queue during training, store the keys of the latest several optimization steps, and continuously update the queue as the training process. In addition, to maintain the key representations’ consistency in the key queue, we follow MoCo[13] to employ momentum update on the parameters of encoder k during training, which is an exponential moving average of the encoder q parameters$^3$.

Formally, denoting the parameters of encoder q as $\theta_q$ and parameters of encoder k as $\theta_k$, we update $\theta_k$ as:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$

(10)

where $m \in [0, 1)$ is a momentum coefficient.

We show the complete Cluster Attention Contrast algorithm in Algorithm 2.
Table 1: AUC(%) of various methods on the CUHK Avenue and ShanghaiTech Campus datasets. All methods are listed by the published year. Missing results due to the lack of reports.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CUHK Avenue</th>
<th>ShanghaiTech Campus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lu et al.[33]</td>
<td>80.9</td>
<td>-</td>
</tr>
<tr>
<td>ConvAE [12]</td>
<td>80.0</td>
<td>60.9</td>
</tr>
<tr>
<td>s-RNN [34]</td>
<td>81.7</td>
<td>68.0</td>
</tr>
<tr>
<td>STAE [28]</td>
<td>80.9</td>
<td>-</td>
</tr>
<tr>
<td>Liu et al.[31]</td>
<td>85.1</td>
<td>72.8</td>
</tr>
<tr>
<td>Liu et al.[32]</td>
<td>84.4</td>
<td>-</td>
</tr>
<tr>
<td>Ionescu et al. [21]</td>
<td>88.9</td>
<td>-</td>
</tr>
<tr>
<td>Abati et al.[1]</td>
<td>-</td>
<td>72.5</td>
</tr>
<tr>
<td>AnoPCN [54]</td>
<td>86.2</td>
<td>73.6</td>
</tr>
<tr>
<td>our method</td>
<td>87.0</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Table 2: Ablation Studies on the CUHK Avenue Dataset.

<table>
<thead>
<tr>
<th>Moment contrast</th>
<th>Cluster attention</th>
<th>Feature Queue</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>fail</td>
<td>75.5</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>86.1</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>87.0</td>
</tr>
</tbody>
</table>

Inference. During inference, for each test samples, the highest similarity between its $K$ feature space representations and corresponding centers will be regarded as its regularity score. To maintain comparability with training data, we oversample each video snippet with the Group Five-crop Augmentation in $256 \times 256$-pixel and take the lowest regularity score evaluated on the cropped snippets as the regularity score of the original snippet. The intuitive idea is that once a position in video detects the irregularity, we believe an abnormality has occurred in the current snippet. Finally, we employ a $1-D$ Gaussian filter to temporally smooth the frame-level regularity scores.

4.3 Comparison with Existing Methods

Table 1 compares our method with a series of state-of-the-art methods [1, 12, 21, 28, 31–34, 54] on the CUHK Avenue and the ShanghaiTech Campus Datasets. Table 1 contains all the evaluation results reported in the published papers of these methods on the two datasets. Since ShanghaiTech is published in 2017, some early methods have not been evaluated (or reproduced) on it. Besides, some methods, example as Ionescu et al.[21], did not report the other missing results in their paper.

On the ShanghaiTech Campus Dataset, our method establishes a new state-of-the-art performance at 79.3% frame-level AUC. Comparison with the existing deep reconstruction methods and deep prediction methods, our work provides an absolute gain of +5.7% in terms of frame-level AUC, which demonstrating the improvements in our work is appreciable. Notably, Ionescu et al.[21] applying a pretrained SSD detector[30] on each frame, and implement their method on detected objects. For fairness, our method did not compare with their approach.

On the CUHK Avenue dataset, our method achieves comparable performance with the state-of-the-art methods. Our method has a significant improvement over the existing methods based on establishing normal behavior dictionaries, which indicates the effectiveness of our introduced contrastive learning pretext task. Our method is slightly inferior to the work of Ionescu et al.[21]. They employ a two-stage algorithm based on k-means clustering and one-class Support Vector Machines (SVM) [44] to eliminate outliers. In contrast, our method does not have additional SVM classifiers and achieves comparable performance to their reported results.

4.4 Discussion

Ablation Study. We conduct ablation study of each component of our method on the CUHK Avenue Dataset. Table 2 present the ablation results. These experimental results demonstrate the method we proposed can significantly improve the performance. Firstly,
72
74
76
78
80
82
84
86
88
1 2 5 10 16 20
w/ cluster constraint w/o cluster constraint

while the light red shaded regions represent ground truth temporal segments of the abnormal events. Figure 5 indicates that the regularity scores obtained by our model are strongly correlated with the ground truth on these samples, which demonstrates the effectiveness of our approach. Figure 7 shows some of the failed cases of our method on the ShanghaiTech Campus Dataset. We noticed that a considerable number of our failed cases contain abnormal objects away from the camera, which reveals that our method requires to explore improving the performance of detecting small abnormal objects.

5 CONCLUSION

In this paper, we explore addressing video anomaly detection task by a contrastive representation learning pretext task. We introduce a novel instance discrimination pretext task, Cluster Attention Contrast, to obtain reliable subcategories of normality as clusters. Experiments on benchmark datasets demonstrate our approach is effective and interpretable. The Cluster Attention Contrast task can help to establish reliable subcategory-specific feature spaces for video anomaly detection. By utilizing the similarity between subcategory-specific representations and corresponding centers, the proposed method achieves state-of-the-art performance on benchmark datasets. In this manner, we established a robust normal concept without any prior assumptions on reconstruction errors or prediction errors.

In the future, we would like to explore the following directions to improve this work. Firstly, we will explore to improve the performance of detecting small abnormal objects. Furthermore, we will design a more reliable method for determining the number of \( K \) feature spaces in encoders automatically. Finally, we will explore different distance measures to obtain a more suitable one for Cluster Attention Contrast task.

ACKNOWLEDGMENTS

This paper was partially supported by the IER foundation (No. HT-JD-CXY-201904) and Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing). Special acknowledgments are given to Aoto-PKUSZ Joint Lab for its support.

REFERENCES
