Consistency-constrained unsupervised video anomaly detection framework based on Co-teaching

Wenhao shao[a,b,*], Praboda Rajapaksha[b], Noel Crespi[b], Xuechen Zhao[a],
Mengzhu Wang[a], Nan Yin[c], Xinwang Liu[a], Zhigang Luo[a]

[a]*National University of Defense Technology, College of Computer, 410073, Changsha, China*
[b]*Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France*
[c]*Mohamed bin Zayed University of AI - UAE*

Abstract

Intelligent video surveillance continues to be a vibrant research domain within the field of computer vision. However, existing representation learning frameworks primarily focus on static information extraction frame by frame such as appearance features, they often overlook the valuable dynamic information like optical flow feature inherent in the video data, which is most essential characteristics of sequence data. To mining dynamic features and bridge this gap, our paper introduces a novel anomaly detection framework that balance dynamic information with static information and construct a relationship between appearance features and corresponding optical flow features, where we sets strong consistency constraints, which reduce the loss between dynamic information and corresponding static information, and we leverages collaborative teaching network to ensure a consistent representation of both static and dynamic information for predict. The proposed framework consists of two sets of encoder-decoder pairs complemented by a memory storage mod-

---

[*]Corresponding author:Wenhao Shao, E-mail address:shaowenhao007@gmail.com

ule. Operating in parallel with the dual encoder network is a Co-teaching network, with the shared memory module serving as the cornerstone for collaborative training. The Consistency constrained condition guarantees the strong consistency of dynamic and static information in the learned representations. In our experimental phase, we present compelling results that showcase the superior performance of our algorithm across three publicly available datasets.

1. Introduction

Abnormal behavior, seen as a possible threat to public safety, has consistently captivated the attention of security experts. Nevertheless, the industry's ability to acquire an ample quantity of diverse abnormal data remains unrealistic, primarily due to the indistinct demarcation between abnormal and normal events in surveillance video data. Additionally, academia faces challenges in precisely defining all abnormal models within videos. As a result, video anomaly detection has remained an exceptionally formidable task. .

Before the emergence of deep learning, traditional video analysis technologies primarily consisted of methods such as the frame difference method [1], color histogram[2], and HOG feature[3]. These video analysis techniques transform original video data into interpretable feature signals, aiding researchers in more effectively analyzing video data. With the advent of deep learning, video anomaly detection technology based on neural network

learning can be categorized into two main groups: unsupervised learning of anomaly detection and weakly-supervised learning of anomaly detection[4] [5, 6].

The core of the unsupervised framework lies in representation learning or self-supervised learning[7]. It employs video frame reconstruction/prediction as the objective function to establish a fundamental model for identifying abnormal data. On the other hand, the weakly supervised framework relies on multi-instance learning and comparative hierarchical loss[8], leveraging multi-instance learning to construct a ranking loss function and develop an anomaly recognition model. Weakly supervised algorithms offer advantages such as robustness, high detection accuracy, and effective utilization of time features[9, 10]. However, they necessitate anomalous datasets, exhibit limited detectable types, and demonstrate poor transferability. In contrast, unsupervised algorithms exhibit strong generalization, do not require labeled data, possess a simple structure, and offer high portability and scalability. Nonetheless, they face challenges related to poor robustness, underutilization of time series features, and low detection accuracy.

Today, explainable artificial intelligence is gaining increasing attention, and a growing number of researchers advocate moving away from purely data-driven models. The unsupervised learning video anomaly model, being definition-driven and not reliant on a large amount of labeled data, holds broader development prospects.

Academics have been working hard to combine the potential advantages of weakly supervised video anomaly detection with the generalization advantages of unsupervised algorithms. For instance, Wang [11] introduced
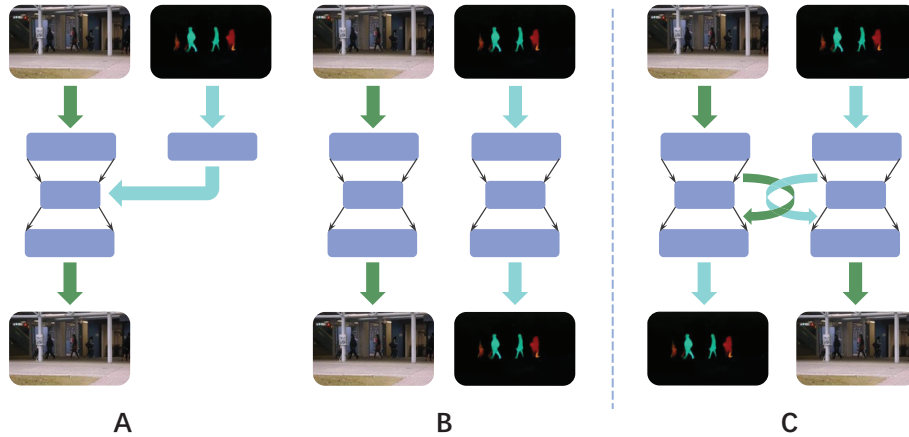
3

Figure 1: Comparison of methods: **A** uses optical flow features as a supplement to video frame appearance features to improve prediction accuracy; **B** uses parallel prediction of appearance features and optical flow features to build a joint prediction loss error; **C** is the proposed strong consistency collaborative training framework.

a novel and robust unsupervised video anomaly detection method that incorporates a frame prediction scheme tailored for surveillance videos. Their approach employs a multipath ConvGRU-based frame prediction network, which adeptly handles semantically rich objects and regions at various scales while capturing spatiotemporal dependencies in normal videos. This algorithm enhances the representation of spatiotemporal features in unsupervised algorithms, thereby enhancing their robustness.

Similarly, Huang et al,[12],introduced the appearance-motion semantic consistency framework, which exploits the difference in appearance and motion semantic representation between normal data and abnormal data. They first designed a two-stream structure to encode the appearance and motion information representation of normal samples, and then proposed a novel

4

consistency loss algorithm to enhance the consistency of feature semantics, enabling the identification of low-consistency anomalies. This algorithm further enhances the consistent representation of dynamic and static features in unsupervised algorithms.

The most advanced semantically consistent model of appearance-motion features is the dual-channel framework proposed in 2022 [13], which proposes a spatiotemporal memory-enhanced dual-stream autoencoder framework and designs two identical and independent proxy tasks to train the dual-stream autoencoder. The structure extracts appearance and motion features separately and decodes them separately. Finally, the optical flow loss and appearance feature loss are calculated to explore the correlation between appearance and motion semantics. In this model, the only consistency constraint is the loss function, but two separate encoding-decoding processes cannot really constrain the consistency of motion features and appearance features[14, 15].

Considering the above-mentioned works, this paper proposes a novel unsupervised learning video anomaly framework CCC-T (Consistency-constrained Framework Based on Co-teaching) as shown in Figure 1-C, which emphasizes the consistent representation of dynamic information and static information by utilizing carefully designed Strong consistency constraints. In this framework, dynamic information (optical flow features) and static information (appearance features) are regarded as equally important input data. The framework designed in this paper mainly contains three parts: two sets of encoding and decoding network structures and memory storage modules. There are two encoding and decoding structures. One is responsible for encoding the appearance features of the video frame as input, and then updating the input

5

features in the memory module, while its decoder outputs the optical flow features corresponding to the video frame. The other encoder is responsible for encoding the optical flow of the video frame, which is used as the input feature; that input feature is updated, and finally the decoder outputs the appearance feature corresponding to the video frame. The memory storage module stores the normal pattern and updates the passing characteristics. To ensure the accuracy of optical flow features in predicting appearance features, the missing background and color information is compensated. The framework utilizes skip connections to connect the encoding layer (appearance features predict optical flow features) and the decoding layer (optical flow features predict appearance features) and reads map features from each layer as a complement. The three modules in the framework are connected through a collaborative teaching network to promote collaborative learning.

To summarize, this section makes the following three contributions

- Proposes a novel unsupervised video anomaly detection framework built using co-teaching networks;

- Achieves the first collaborative training of optical flow and representational features in unsupervised video anomaly detection; and

- After testing on three datasets, the proposed model further improves the accuracy of unsupervised video anomaly detection algorithms.

The organization structure of this article is as follows: Section 2 is an introduction to the relevant work of this article; Section 3 is the algorithm proposed in this article; Section 4 is the experiment used to verify the model proposed in this article; Section 5 is the conclusion.

2. Related Work

Video anomaly detection algorithms within an unsupervised learning framework always focus on a single goal: improving prediction or reconstruction accuracy by extracting more precise video features. Progress in this field can be traced to the seminal work in [16], which introduced a comprehensive unsupervised framework based on appearance feature representation learning. On this basis, the work in [17] and subsequent research further optimized the unsupervised learning framework by optimizing the relevant loss function and improving the update mechanism of the memory storage module. Based on this framework, other solutions have further improved the accuracy of feature extraction through multi-task learning [18]. The core concept is to use object detection to improve the accuracy of feature extraction in an unsupervised framework, which include appearance features and optical flow features.

Many results have appeared in the area of multi-task unsupervised representation learning, including those in [18, 19, 20, 21, 22]. Among these, the multi-task unsupervised framework proposed in [18] contains four different agent tasks. The first is to determine the order; the second is to determine whether the current actions are continuous. The third task predicts intermediate frames, and the fourth task requires training a sub-network (3D convolution). Multiple tasks work together to improve the accuracy of feature pattern extraction. A different approach is offered in [20], which features a novel bidirectional architecture with three consistency constraints to comprehensively regulate the prediction task from the pixel level, cross-modality and time series levels. Prediction consistency is proposed as a priority, to

7

consider the symmetry of motion and appearance in forward and backward time, which ensures a highly realistic appearance and motion prediction at the pixel level. At the same time, the consistency of temporal features and spatial features is also trying to emphasized in multi-task models. For example, in the literature[22], this paper proposes to set up two agent tasks to predict appearance features from frames sequence in forward and reverse order and calculate the bidirectional optical flow feature of the real frame and the predicted frame as the loss, which still belongs to the prediction task. Even in the literature[23], optical flow features are still used to build additional tasks and then serve as supplementary features to the appearance features to achieve the prediction task. As shown in Figure2 (A).

The above-mentioned unsupervised methods are all dedicated to utilizing sub-tasks, including identifying the order or reverse order of the sequence to extract features, thereby enhancing the extraction of dynamic features and static features. However, for video data, multi-tasking only guarantees the accuracy of extracting dynamic features and static features, it cannot constrain the consistency of dynamic features and static features.

The dual-channel unsupervised model [13, 11, 12, 24] is a new attempt to address these issues. Differing from the framework described above, the dual-channel model attempts to directly extract dynamic features as a supplement to static features, and builds a dynamic feature-static feature constraint framework to enhance the integrity of the input features to improve the accuracy of prediction/reconstruction. However, the existing dual-channel model, as shown in Figure 2(A,B), only uses dynamic features as a supplement to static features, which enhances the accuracy of input features, but does not

set consistency constraints. Framework C, on the other hand, designs a completely parallel encoding-decoding structure and relies on interactive loss functions to constrain consistency. This constraint cannot affect the features extracted by the encoder, and the channels are relatively independent, that is, the processing of dynamic features and the processing of static features are independent and cannot act as a real consistency constraint on the extracted features. In addition, while mainstream methods use dynamic features as supplementary elements to enhance the representation capabilities of static features, they cannot achieve simultaneous learning of spatio-temporal features.

To solve this problem, this paper introduces a new dual- channel video anomaly framework to enhance the detection capabilities of unsupervised learning algorithms. This framework treats dynamic information and static information as inputs of equal importance and carefully designs strong consistency constraints between dynamic information and static information to ensure consistent representation of optical flow features and appearance features, and it builds a collaborative learning and memory storage module based on co-teaching. The core of this study is collaborative learning, memory storage modules, and skip connections and other technical means, which strictly follow the consistency constraints of dynamic features and static features.

3. Methodology

This section provides a detailed explanation of our proposed unsupervised learning framework and the models utilized in our experiments. This includes

explaining how the co-teaching architecture works in the training process of two encoder-decoder networks.

In our proposed framework, the Flownet2 network [25] is responsible for extracting optical flow features from video frames. Subsequently, these features from video frames and optical flow are used as the input into two encoder networks. These features are then compressed, followed by their entry into the memory storage module to update the corresponding elements of video frame features and optical flow features.

The mechanism entails retrieving the features of the nearest counterpart and amalgamating them into novel features. Finally, the amalgamated new features feed into the two decoder networks to predict the features of the opposing entity. For example, the optical flow features serve as the input to the encoder-decoder, resulting in the output of video frame features. Conversely, when the input is the video frame feature, the output manifests as the optical flow feature. To address the potential information gap in video frame feature prediction by optical flow features, this study integrates skip connections[26, 27] that bridge the encoding map of video frames to the optical flow decoder (predictive video frames).

The loss function is comprised of the prediction loss in- herent in the video frame features and the optical flow features' prediction, as well as the similarity loss in memory modules. The proposed model greatly ensures the consistent description of optical flow features and appearance features through shared memory entries. .
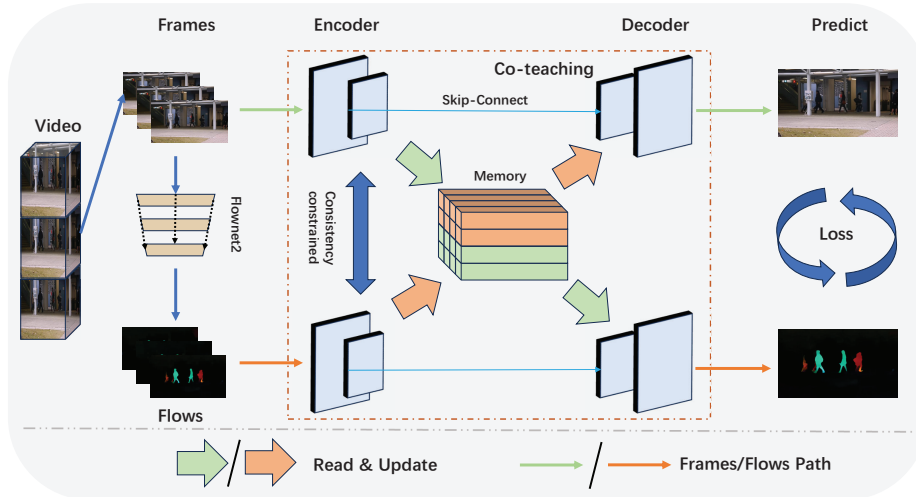
Figure 2: A detailed framework of CCC-T, The first step uses Flownet2 to obtain the optical flow information of the video sequence. The second step inputs the segmented video frames and optical flow information into their respective encoding networks. The third step is to cross-read and collaborate the output of the encoding network with the memory module. Update, the fourth step, the updated input features are input to the decoder network for cross prediction.

3.1. Preliminary

The fundamental algorithms highlighted in this chapter contain Flownet2, the encoder-decoder structure, the memory module, and the co-teaching framework. Notably, The encoder-decoder structure and memory module already well described in previous paper[17, 16]. Consequently, the ensuing content will provide a succinct overview of Flownet2, outlining its objectives and structural attributes, followed by an outline of the co-teaching architecture.

FLOWNET2: FlowNet2 represents an evolution of the original FlowNet architecture, both of which were developed by researchers at NVIDIA[28]. The primary objective of these architectures is to precisely predict displacement vectors that explain the movement of pixels between frames. They find applications across various research fields, including computer vision, video analysis, motion tracking, and visual effects.

The principal characteristics and components of FlowNet2 are; 1. Siamese Network: FlowNet2 comprises two identical sub-networks that share weights. Each subnetwork processes an image from the input pair, Co-process two different input vectors to compute a comparable output vector. 2. Feature Extraction: This process employs a sequence of convolutional layers to extract hierarchical features from the input images. 3. Pyramid Processing: FlowNet2 leverages pyramid processing to capture information across various scales. Pyramid processing is a model of multi-scale signal representation. 4. Correlation Layer: The correlation layer is instrumental in determining the similarity between blocks within two input images. FlowNet2 presents significant enhancements over the original FlowNet architecture, enhancing accuracy and robustness in optical flow estimation. It achieves state-of-the-art performance on benchmark datasets designed for optical flow estimation tasks.

Co-Teaching[29]: A collaborative teaching network is a framework in which multiple neural network models collab- orate to solve specific problems or achieve a common goal. For example, multiple actors merge their predictions through techniques such as voting, averaging, or weighted averaging. Classic co-teaching networks are one of the following four types: 1. Knowl-
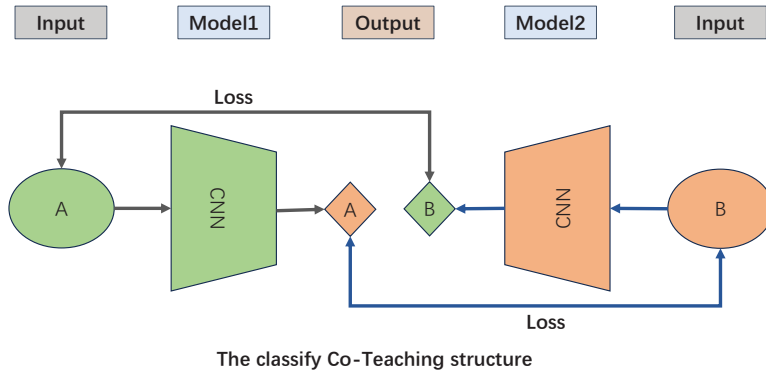
Figure 3: The Co-Teaching structure

edge distillation[30]: A broader or more com- plex model (teacher model) is trained together with a smaller or simpler model (student model). The student model learns to imitate the behavior of the teacher model, reduce parameters, and/or to build multi-task models; 2. Collaborative training [31]: Multiple models are trained simultaneously and exchange training data or gradients during the optimization process; 3. An Adversarial Network [32]: Multiple models with complementary effects, such as a generator network and a discriminator network in a generative adversarial network (GAN) work together to achieve a specific result; and 4. Federated learning [33]: Many models are trained on different data subsets and then merged or averaged to generate a global model. This approach can improve privacy and data distribution issues. In this paper, we adopt two encoder-decoder structures to share the memory module, cross-read the video frame feature pool and the optical flow feature pool, and to promote the collaborative training of the model. These two encoding structures are similar to two teacher networks,

13

learning from each other.

3.2. Consistency-constrained Framework Based on Co-teaching

This section introduces the operation and interaction of each module of the framework(CCC-T: Consistency-constrained Framework Based on Co-teaching) proposed by this paper in detail. This CCC-T employs two interconnected encoder-decoder structures facilitated by the co-teaching network 2. These structures are designed to encode optical flow and video frame features separately while predicting the corresponding features of the opposite type (i.e., optical flow to video frame and vice versa). The predicted loss resulting from these predictions is then utilized to update the model. The following section outlines the detailed steps involved in the comprehensive formalization.

Formalization: There is an existing video denoted as V, which is divided into a sequence of continuous video frames: $V = v_1, v_2, v_3, ..., v_N$, where N represents the total number of frames in the video. The optical flow features of these video frames are extracted using Flownet2, denoted as $F_{flows} = Flownet2(V)$, with individual flow features represented as $f_{flow} \in f_{flows_1}, f_{flows_2}, f_{flows_3}, ..., f_{flows_N}$. The read library of OpenCV2 is employed to directly extract frame features from the video frames, yielding $F_{frames} = Ir(V)$, with frame features represented as $f_{frames} \in f_{frames_1}, f_{frames_2}, f_{frames_3}, ..., f_{frames_N}$.

As stated earlier, this paper presents a model that encompasses two encoder-decoder structures, as illustrated in Figure 1. where ψ represent the Encoder function and ϕ the Decoder, The upper structure is the video frame feature encoderψ_{frames}, while the lower one is the optical flow feature

14

encoder, referred to as ψ_{flows}. The decoder positions are the opposite: the upper one is ϕ_{flows}, and the lower one is ϕ_{frames}.

During the training phase, the extracted video frame features F_{frames} are input into the ψ_{frames} to focus and refine the quality of the appearance feature representation. Subsequently, these features are passed through a memory module. The error is calculated with the nearest video frame feature entry, leading to an update of the video frame feature storage module. Simultaneously, the module queries the optical flow entry that is closest to the input feature and then reads and updates the input feature. The updated input feature is then fed into the ϕ_{frames} to predict the optical flow feature. This process can be expressed in an equation as:

$$\begin{aligned}
F_{frames}^{E} &= \psi_{frames}(f_{frames}) \\
&= \psi_{frames}(Ir(V)), \\
V &= v_1, v_2, v_3, ..., v_N
\end{aligned} \tag{1}$$

$$\hat{F}_{flows} = \phi_{frames}(\theta(F_{frames}^{E}, M)) \tag{2}$$

where, M signifies the memory storage module, and θ embodies the interaction between input data and the memory storage module, encompassing functions such as reading, updating, and the integration of novel features. Comprehensive insights into the memory storage module are explained in 3.3 . F_{frames}^{E} denotes the features emanating from the encoder, while $\hat{F}_{flows}$ encapsulates the optical flow features prognosticated by the decoder.

Conversely, the optical flow features F_{flows}, obtained from Flownet2, are input into the ψ_{flows}. This step help to refine the high-quality optical flow

feature representation. These features are then processed through a memory module. Similar to the video frame features, the error is computed with the nearest optical flow feature entry, resulting in an update of the optical flow feature storage module. Furthermore, the module queries the appearance feature entry closest to the input feature, reading and updating the input feature. The updated input feature is directed into the ϕ_{flows} to predict the appearance feature.

$$
\begin{aligned}
F_{encoder}^{flows} &= \psi_{flows}(f_{flows}) \\
&= \psi_{flows}(Fownets2(V)), \\
V &= v_1, v_2, v_3, ..., v_N
\end{aligned}
\tag{3}
$$

$$
\hat{F}_{frames} = \phi_{flows}(\theta(F_{frames}^E, M))
\tag{4}
$$

The loss function contains three components: optical flow prediction loss L_{flows}, appearance feature prediction loss L_{flows}, and memory storage module loss L_M. Similarly, during the test phase, the anomaly score is composed of three parts.

$$
loss = \begin{cases}
\left\| \hat{f}_{flows} - f_{flows} \right\| \\
\left\| \hat{f}_{frames} - f_{frames} \right\| \\
L_M
\end{cases}
\tag{5}
$$

The role of this module is to align input data with their corresponding entries in the memory module, thereby capturing and recording trained normal patterns. The memory module loss L_M is described in detail next.

3.3. Co-teaching within Memory Module

The co-teaching structure in training is designed in the memory storage module.

Following the blueprint of the conventional memory storage module[34], this unit serves two primary functions. The first involves reading—wherein the module identifies and retrieves the entry most closely aligned with the input feature, subsequently updating the input feature. The second function entails updating, which transpires as an ongoing process throughout the training. The memory matrix continuously evolves based on the proximity between feature maps, effectively consolidating data from the training set that corresponds to the set entries.

In our framework, the memory storage module is bifurcated into two distinct components, as illustrated in Figure 5.3. The green segment denotes the video appearance feature memory mode, while the orange segment signifies the optical flow feature memory mode. The act of reading and updating each input datum transpires in disparate sections of the memory module, so that the update operation takes place within the respective memory mode, and the reading operation unfolds in the complementary memory mode. These modes are illustrated in Figure 4.

3.3.1. Reading and Updating Mechanisms of the Memory Module

Reading Mechanism:. Reading Mechanism for the Memory Module as shown in Figure 4: The input to the memory module is the optical flow feature. This involves calculating the cosine similarity between the query feature and the entries within the video frame appearance feature memory module. The aim here is to identify the entry or multiple entries with the closest proximity to

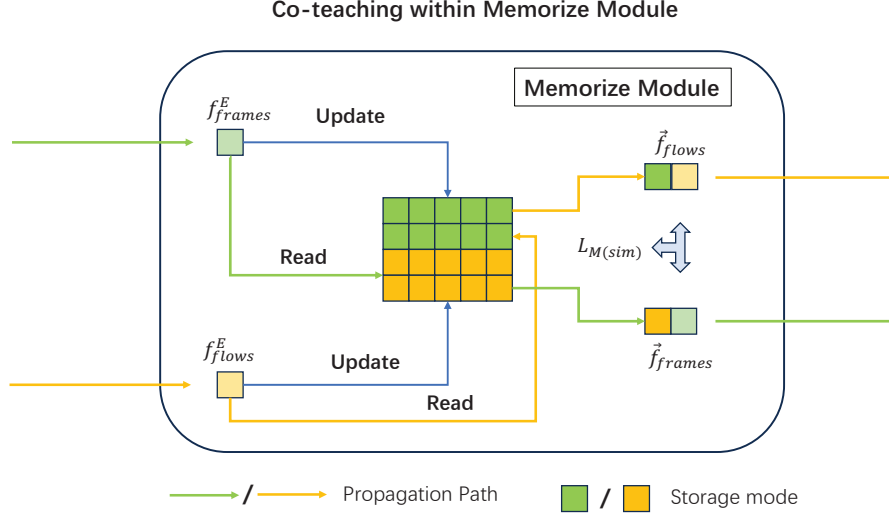**Co-teaching within Memorize Module**

Figure 4: Co-teaching within a memory module: Green indicates the transfer of static features in the memory module, and orange represents the transfer of dynamic features; correspondingly, the memory module is composed of multiple static feature category entries and multiple dynamic feature category entries

the query feature, thus determining their respective distances. The softmax function is applied to establish an average probability match. Subsequently, the probability value is utilized to compute the inner product with the appearance feature entry from the memory module. This process leads to the feature update. Finally, the updated features are merged with the original query features to predict the corresponding video appearance features.

First, the cross-cosine similarity between each entry q_{flows}^{k}, q_{frames}^{k} ($q_{flows}^{k} \in f_{flows}^{E}$, $q_{frames}^{k} \in f_{frames}^{E}$) and memory items p_{m}^{frames}, p_{m}^{flows} is calculated, where q_{flows}^{k} and q_{frames}^{k} are from the corresponding two query encoding features f_{frames}^{E}, f_{flows}^{E}; p_{m}^{frames}, p_{m}^{flows} is set during initialization, and two

two-dimensional correlation maps of size MK are obtained. The softmax function along the vertical direction and obtain the matching probabilities $w_{frames}^{k,m}, w_{flows}^{k,m}$ show in 6:

$$w_{k,m}^{frames} = \frac{exp((p_m^{flows})^T)q_k^{frames}}{\sum_{m'}^M exp((p_{m'}^{flows}))q_k^{frames}} \tag{6}$$

$$w_{k,m}^{flows} = \frac{exp((p_m^{frames})^T)q_k^{flows}}{\sum_{m'}^M exp((p_{m'}^{frames}))q_k^{flows}} \tag{7}$$

For the query items of optical flow features q_{flows}^k and appearance features q_{frames}^k, the opposite memory module is read through the calculated weight $(q_{flows}^k \rightarrow p_{m'}^{frames}, q_{frames}^k \rightarrow p_{m'}^{flows})$, which obtains the desired cross prediction information. The reading process shows in 8 :

$$\hat{p}_k^{frames} = \sum_{m'}^M w_{k,m'}^{frames} p_{m'}^{flows} \tag{8}$$

$$\hat{p}_k^{flows} = \sum_{m'}^M w_{k,m'}^{flows} p_{m'}^{frames} \tag{9}$$

After reading the memory module, the closest cross feature map $\hat{p}_k^{flows}, \hat{p}_k^{frames}$ is obtained, We concatenate $\hat{p}_k^{flows}, \hat{p}_k^{frames}$ with the query map q_{flows}^k , q_{frames}^k along the channel dimension, and send $\vec{f}_{frames}, \vec{f}_{flows}$ into the corresponding decoder.

$$\begin{cases} \vec{f}_{frames} = \sum(\hat{p}_k^{flows}, q_{frames}^k) \\ \vec{f}_{flows} = \sum(\hat{p}_k^{frames}, q_{flows}^k) \end{cases} \tag{10}$$

Updating mechanism:. Update mechanism of memory modules: In this case, the cosine similarity of appearance encoding features f_{frames}^E and optical flow encoding features f_{flows}^E with the corresponding memory modules is calculated. Next the probability value is caculated through the softmax function. Then read the compared memory entry by calculating the probability value, The next steps involve using the query features to increase the inner product of the obtained probability values. This sum is added to the corresponding memory entry of the original appearance feature. As a result of these operations, the memory module is effectively updated. The function of this step is to find the memory feature that is most similar to the query feature, and through its similarity loss, continuously improve the adhesion of the memory entry to the real normal pattern.

The first step is to calculate the cosine similarity between the optical flow query encoding features f_{frames}^E and the optical flow memory entries f_{flows}^E, and the appearance query encoding features and the appearance memory entries. This process is the opposite of the reading mechanism.

$$u_{k,m}^{frames} = \frac{exp((p_m^{frames})^T q_{frames}^k)}{\sum_{k'}^{K} exp((p_m^{frames})^T q_{frames}^{k'})} \tag{11}$$

$$u_{k,m}^{flows} = \frac{exp((p_m^{flows})^T q_{flows}^k)}{\sum_{k'}^{K} exp((p_m^{flows})^T q_{flows}^{k'})} \tag{12}$$

After obtaining the cosine similarity between the memory entry and the query point $u_{k,m}^{frames}$, $u_{k,m}^{flows}$, use the probability value cosine similarity to read the query entry, accumulate it with the original memory entry in the

same channel, and update the memory entry. The updated storage module $< \check{p}_{flows}^m, \check{p}_{frames}^m >$ is shown in the figure below:

$$\check{p}_{flows}^m = \sum(p_{flows}^m + \sum_{k=1}^{K} u_{k,m}^{flows} q_{flows}^k) \tag{13}$$

$$\check{p}_{frames}^m = \sum(p_{frames}^m + \sum_{k=1}^{K} u_{k,m}^{frames} q_{frames}^k) \tag{14}$$

Different from cross-reading, the memory update corresponds from optical flow to optical flow and appearance to appearance. By calculating the similarity matrix, it is accumulated to the memory module.

3.3.2. Strong consistency constraints

The strong consistency constraints in this article are mainly implemented through the reading and updating of the memory module. The reading and updating rules have been described in detail above. This article proposes that the memory module of the model is divided into two parts, one is optical flow feature storage, and the other is appearance feature storage; and the reading and updating of each branch are implemented for different parts of the memory module. That is, the prediction in this article is cross prediction, inputting optical flow features, then constructing a similarity matrix, and reading the most similar appearance feature storage entries to construct joint features to predict appearance features. When updating, only the optical flow memory entries corresponding to the optical flow features are updated. Optical flow features and appearance features are cross-read and updated. Through the loss function, appearance features and optical flow features are

21

strengthened, and the consistent representation of dynamic information and static information is enhanced. And according to the consistent description of optical flow features and appearance features, the similarity between appearance and optical flow of videos of the same category is the highest. Based on this, this article sets up the consistent description probability for dynamic information and static information as $Cst_{(S,D)}$:

$$Cst_{(S,D)} = \sum_{i=1}^{k} \sum_{j=1}^{k} < f_{frames}^{i}, f_{flows}^{j} > \tag{15}$$

where f_{frames}^{i} and f_{flows}^{j} are the encoded appearance features and optical flow features, and k is the number of storage entries designed by the memory module. Only when $i = j$, the consistency probability can reach the maximum value.

Therefore, the **strong consistency constraints** set as 16:

$$STC = \sum_{i=j}^{k} |f_{frames}^{i}, f_{flows}^{j}|$$
$$= \sum_{Max(Cst)}^{k} (|f_{frames}^{i}, M_{flows}^{j}| \oplus |f_{flows}^{i}, M_{frames}^{j}|) \tag{16}$$

During the memory module reading process, this article sets up to read the cross feature entries that are closest to the query entries and predict the corresponding cross features $Max(Cst)$. We calculate the similarity by covariance and retrieve the intersection entries with the highest similarity. Final predicted cross information features

3.3.3. Loss function Memorize Module

The loss function of the training process mainly consists of three parts, namely $L_{flows}, L_{frames}, L_M$, Among them, L_{flows} represents the error between the predicted optical flow and the real optical flow, and L_{frames} represents the error between the appearance characteristics of the predicted video frame and the real video frame. These designations are employed for partitioning the distances between distinct entries within the memory module. Here L_M loss is divided into two parts, namely Strong consistency constraint loss $L_{M(Sim)}$ and segmentation loss $L_{M(Seg)}$. $L_{M(Sim)}$ is achieved by enhancing the similarity between the optical flow features in the query entry and the most approximate flows features in the memory entry, and at the same time enhancing the appearance features in the query. The similarity between the feature and the closest optical flow feature in memory is used to ensure the consistency of optical flow features and appearance features, while the $L_{M(Seg)}$ is used to enlarge the distance between the query point and the next closest memory entry, ensuring the orthogonality of the memory module, that is, the entries are kept far enough apart.

The loss function of the memory module L_M is expressed as

$$L_M = L_{M(Sim)} + L_{M(Seg)} + L_{M(STC)} =$$

$$\begin{cases} < \left\| f_{frames}^E - M_{frames}(P_{nearest}, f_{frames}^E) \right\| \\ \left\| f_{flows}^E - M_{flows}(P_{nearest}, f_{flows}^E) \right\| > + \\ < - \left\| f_{frames}^E - M_{frames}(P_{sec-nearest}, f_{frames}^E) \right\| \\ - \left\| f_{flows}^E - M_{flows}(P_{sec-nearest}, f_{flows}^E) \right\| > + \\ < \left\| f_{frames}^E - M_{flows}(P_{nearest}, f_{frames}^E) \right\| \\ \left\| f_{flows}^E - M_{frames}(P_{nearest}, f_{flows}^E) \right\| > \end{cases} \quad (17)$$

In the equation, M represents the memory block. M_{frames} signifies the appearance pattern within the memory module, while M_{flows} represents the optical flow pattern within the same module. The variable p denotes an entry in the memory module, where $M_{frames}(P_{nearest}, f_{frames}^E)$ designates the memory entry that is nearest to the query feature, and $M_{frames}(P_{sec-nearest}, f_{frames}^E)$ denotes the second closest memory entry to the query feature.

3.4. Anomaly detection stage

The primary procedure of the anomaly detection stage maintains consistency with the training process.

The initial step involves preprocessing the dataset, which entails segmenting the test video into video frames and extracting optical flow features. Subsequently, the second step utilizes two distinct encoder structures to compress both the appearance and optical flow features of the video independently. In the third step, the compressed final features are directed into the memory module, where they are combined to generate novel query features. Moving on to the fourth step, these newly generated features are input into the de-

24

coder network to anticipate the corresponding optical flow and appearance representations.

The computation of the anomaly score predominantly encompasses two components: the prediction loss and the similarity loss originating from the memory module. The specific pseudo code is shown in 1:

Algorithm 1 Anomaly Detection Phase

1: Initialization:

Flownet2, Random $M \in R^{K \times 2M}$, $V = v_1, v_2, v_3, ..., v_N$;

2:
$$\begin{cases} F_{frames} = Ir(V) \\ F_{flows} = Flownet2(V) \end{cases} ;$$

3:
$$\begin{cases} F^E_{frames} = \psi(F_{frames}) \\ F^E_{flows} = \psi(F_{flows}) \end{cases} ;$$

4: ; $\vec{f}_{frames}, \vec{f}_{flows} = Co - teach(M, F^E_{frames}, F^E_{frames})$

5:
$$\begin{cases} F^D_{flows} = \phi(\vec{F}_{frames}) \\ F^D_{frames} = \phi(\vec{F}_{flows}) \end{cases} ;$$

Output: Calculate anomaly scores.

Score=$\{\alpha\|F^D_{flows} - F_{flows}\|, \beta\|F^D_{frames} - F_{frames}\|\}$

---

where The core part of the anomaly score is the prediction error, which includes optical flow feature prediction error and appearance feature prediction error. In the testing phase, after a large number of verification experiments, this article sets two prediction losses combined with hyperparameters $\alpha = 0.3$ and $\beta = 0.7$.

Table 1: The result for Avenue dataset

Name	Technology	Journal	AUC
Unmasking[35]	VGG-f	ICCV2017	80.6
StackRNN[36]	Temporally-coherent	ICCV2017	81.7
MemAE[16]	Memory module	ICCV2019	81.0
MNAD[17]	Learning Memory module	CVPR2020	80.6
TAC-Net[37]	Temporal-aware contrastive	IEEE TII	87.3
ITAE[38]	Two-path Generative	PR 2022	88.0
Two-P[13]	Two-path AE	ICME 2022	89.8
DEDDnet[22]	Doub-AE, Fusion	TCSVT 2022	89.0
VABD[39]	Wasserstein GAN	TIP 2022	86.6
Amp-Net[24]	Two-encoder,one decoder	TII 2023	**92.2**
SSAGAN[40]	GAN	TNNLS 2023	88.8
CCC-T	Consistency Co-teaching		89.2

4. Experiments

The experimental settings outlined in this paper are primarily categorized into three groups. According to the experimental settings, this paper evaluates the performance of the framework proposed in this paper from three aspects: advancement comparison, ablation experiment, and effect display.

4.1. Experiment 1

4.1.1. Comparative Experiment

The first group pertains to a comparison of prediction accuracy with mainstream video anomaly detection algorithms. In this set of experiments, this paper compares the detection accuracy of the model proposed in this paper and the mainstream unsupervised model. We conducted independent comparative analyzes on three public data sets: Ped2[41] Avenue[42],

Shanghaitech[43]. The results are shown in the table below:

Table 2: The result for UCSD(Ped1,Ped2) dataset

Name	Technology	Journal	AUC	
			Ped1	Ped2
Unmasking[35]	VGG-f	ICCV2017	68.4	82.2
StackRNN[36]	Temporally-coherent	ICCV2017	N/A	92.2
MemAE[16]	Memory module	ICCV2019	N/A	91.7
STFF [44]	Fast sparse coding	PR 2020	82.4	92.8
MNAD[17]	Learning Memory module	CVPR2020	N/A	97.0
DPU[45]	Dynamic Prototype	CVPR2021	85.1	96.9
TAC-Net[37]	Temporal-aware contrastive	IEEE TII	N/A	98.1
ITAE[38]	Two-path Generative	PR 2022	N/A	98.7
Two-P[13]	Two-path AE	ICME 2022	N/A	98.1
DEDDnet[22]	Doub-AE, Fusion	TCSVT 2022	94.2	98.1
VABD[39]	Wasserstein GAN	TIP 2022	81.1	97.1
SSAGAN[40]	GAN	TNNLS 2023	84.2	97.6
CCC-T	Consistency Co-teaching		85.0	99.1

Tables I, II, and Table III shows accuracy comparisons between the framework CCC-T proposed in this paper and mainstream algorithms across three datasets (Avenue, UCSD(ped1,ped2), Shanghaitech). Because the Shanghaitech dataset is too large, some of the baseline models only tested with the Avenu and Ped2, and some models use the Ped1 dataset [41]. Therefore, in this paper we used three different tables (Table I, II and III) to illustrates the results for each dataset with different baseline models. The result show that the prediction accuracy AUC achieved by the CCC-T algorithm proposed in this paper has shown superior performances, thereby substantiating the effectiveness of the proposed algorithm. Specifically, while considering con-

sistency, the way in which optical flow and appearance features are combined (either complementary or equal) becomes the primary aspect of differentiation between video data features and image data. When analyzing video data, special attention should be paid to the processing of dynamic features. From Tables I, II, and Table III, it can be concluded that the CCC-T model proposed in this article has more advanced performance. The second core store is the consistency constraint of optical flow features and appearance features. Simply making optical flow and appearance completely independent and predicting them separately does not conform to the essential characteristics of video data. Forcing the consistency of optical flow and appearance through loss functions is the key to video representation learning.

Table 3: The result for ShanghaiTech dataset

Name	Technology	Journal	AUC
StackRNN[36]	Temporally-coherent	ICCV2017	68.0
MemAE[16]	Memory module	ICCV2019	69.7
BMAN[46]	Appearance-motion joint	TIP 2019	76.2
Few-Shot[47]	Few-shot scene-adaptive	ECCV2020	**77.9**
MNAD[17]	Learning Memory module	CVPR2020	70.5
DPU[45]	Dynamic Prototype	CVPR2021	73.8
TAC-Net[37]	Temporal-aware contrastive	IEEE TII	77.2
DissociateAE[48]	Dissociate spatio-temporal	PR 2022	73.7
ITAE[38]	Two-path Generative	PR 2022	76.3
Two-P[13]	Two-path AE	ICME 2022	73.8
VABD[39]	Wasserstein GAN	TIP 2022	78.2
SSAGAN[40]	GAN	TNNLS 2023	74.3
CCC-T	Consistency Co-teaching		77.1

4.1.2. Equal Error Rate (EER) Calculation and Analysis

EER stands for equal error rate, which refers to the error rate when the false positive rate (FAR) is equal to the false negative rate (FRR) in a binary classification task. The false positive rate is the probability that a negative class (non-target speaker) is mistakenly classified as a positive class; the false negative rate is the probability that a positive class is mistakenly classified as a negative class. In the equal error rate, we hope that both FAR and FRR are as close as possible.
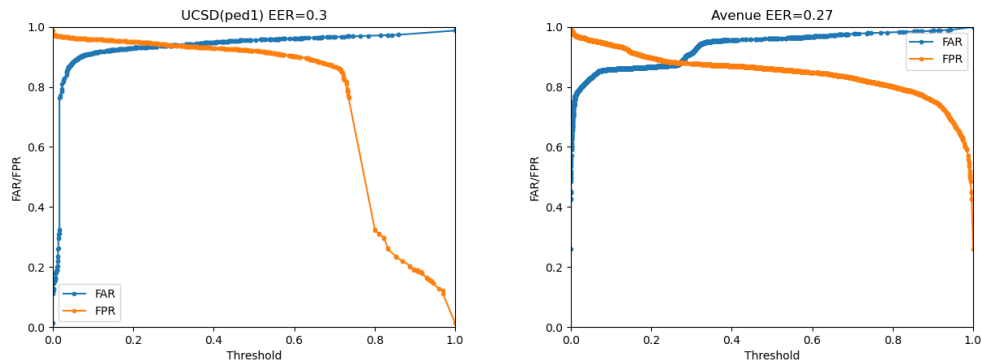
(a) UCSD (Ped1)	(b) Avenue

Figure 5: EER calculation results for datasets UCSD(Ped1) and Avenue.

Figure 5 shows the EER of UCSD(Ped1) and Avenue, which show that the value of EER for Ped1 is 0.30, and the value for Avenue is 0.27.

Relationship between FAR and FPR: As the threshold increases, FAR gradually increases, while FPR gradually decreases. The two show a clear intersection in the figure, showing the performance of the system at different thresholds.

For the Avenue datasets, the EER in the figure is 0.27, which means that

in this system, when FAR equals FPR, the error rate of the system is 27%. EER is usually used as an indicator to measure the performance of a binary classification system. The lower the EER, the better the performance of the system. For the Ped1 datasets, because its data involves distortion, the accuracy is low and this dataset is ignored in most anomaly video detection algorithms. The different thresholds in the figure show the trade-off of the system increasing another error type (such as FPR) while reducing one error type (such as FAR). When selecting a threshold, it is necessary to find a reasonable balance between the two error rates based on the actual application scenario.

4.2. Experiment 2

The experiments are focused on ablation studies. This experiment involves the separation of various modules such as skip-connecting and Consistency Co-Teaching within the framework for distinct training tests, followed by an assessment of accuracy in the current dual-channel training approach. This article set up three groups of ablation experiments to study the comparison between single channel and dual channel, the performance comparison of different components of the model, and the intrinsic relationship between the dual channel loss hyperparameters. In the diagram in the table IV, V, blue represents the propagation path of appearance features, and yellow represents the propagation path of optical flow features.

4.2.1. The Performance comparison of single-channel and various dual-channel models

In this experiment, we set up four groups of models to compare the performance of single-channel and dual-channel and their different variants: 1) prediction from frame appearance to frame appearance; 2) prediction from appearance features and optical flow features to appearance features; 3) prediction from appearance to appearance and optical flow to optical flow; and 4) As well as the CCC-T framework proposed in this article, appearance predicts optical flow, and optical flow predicts appearance. The experimental results are shown in Table IV.

Table 4: The Ablation Study : The Performance comparison of single-channel and various dual-channel models, blue represents the propagation path of appearance features, and yellow is optical flow features

	Number	Input	Output	Model	AUC
UCSD(Ped2)	a	frames	frames	→	97.0
	b	frames,flows	frames	→	98.9
	c	frames,flows	frames,flows	→	98.7
	d	frames,flows	flows,frames	→	75.3
Avenue	a	frames	frames	→	70.5
	b	frames,flows	frames	→	88.0
	c	frames,flows	frames,flows	→	89.8
	d	frames,flows	flows,frames	→	73.6

Table IV is the performance comparison between the classic single-channel model and the multi-channel dual-channel model, 1) the initial frame appearance to the prediction/reconstruction of frame appearance; 2) the optical flow as a supplementary feature, and then 3) the separate prediction of optical

flow and appearance features and reconstruction; 4) the basic model proposed but without constraint. The final prediction accuracy of the model shows an upward trend. Without the assistance of the Co-Teaching module and skip connections, the performance of the model Init (Un-Constraint) proposed in this chapter is far inferior to the classic model. After analysis, it was found that this is because optical flow features lack more appearance information (such as background, color, etc.), and appearance features cannot be predicted directly from optical flow. This ablation study results are presented in Table V..

4.2.2. The impact of skip connections and co-teaching on the performance of dual-channel models

In this set of ablation experiments, this paper set up five sets of models: 1) the baseline model of cross prediction; 2) the skip connection model that only includes appearance to optical flow; 3) the skip connection model that only includes optical flow features to appearance features. ;4) Double-skip connection model; 5) The final framework CCC-T including consistency co-teaching and double-skip connection; and compare their performances with Ped2 and Avenue datasets. The experimental results are shown in Table V.

The results in Table V show the performances of different modules in the proposed CCC-T framework. From the Table V, it can be concluded that the main reason for the low cross-prediction performance is that the optical flow feature has less appearance feature information and cannot be completely restored. Therefore, in the channel where optical flow features predict appearance features, whether there are skip connections that provide appearance feature input has a greater impact on the performance of

Table 5: The Ablation Study: The impact of skip connections and co-teaching on the performance of dual-channel models

Method	Model	UCSD(ped2)	Avenue
Init (Un-constraint)		75.3	73.6
Skip(Flows-Frames)		95.4	87.2
Skip(Frames-Flows)		76.2	76.7
Fully-Skip		95.6	86.9
CCC-T		**99.1**	**89.2**

the Skip(Flows-Frames) model. Table V shows that the performance of the Skip(Flows-Frames) model containing only this core skip connection basically reaches the performance of the double-hop connection. Secondly, whether to set up a strong co-teaching network structure also has a great impact on performance. Therefore, each component of the CCC-T model performance proposed in this article is essential.

4.3. Experiment 3

Experiment 3 is a visual evaluation experiment and the fluctuation of abnormal scores between normal frames and abnormal frames.

The experimental findings from the test phase have been visually presented in Figure 6 7. From Figure 6, we can get that a recognizable shift in the abnormal score is observed when confronted with irregular video frames, exhibiting a significant increase. It shows that this phenomenon helps us effectively pinpoint anomalies in video data streams. Exceptions in the graphical representation include various situations, particularly the use of bicycles, skateboards, and other unconventional vehicles on sidewalks. From Figure 7 which shows that the current unsupervised algorithm has insufficient perfor-
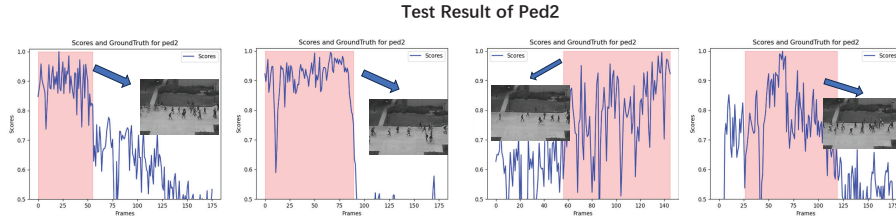
Figure 6: Test results on the dataset UCSD(ped2). The pink background is the area where the real anomaly occurs; The blue curve represents the change of the anomaly score with the time series.

mance indicators in the Shanghaitech dataset and is difficult to distinguish not obvious abnormal events. Combining the results displayed by the two effects, we can infer that the video anomaly is not for the detection of a certain frame, but for the analysis of a segment. Since the abnormality score in the picture fluctuates violently, it is difficult to locate abnormal from several other frames, but considering overall situation of video or the entire segment, abnormal events can be clearly located. This once again proves that abnormalities are continuous and indivisible.

5. Conclusion and Future Work

We introduces an innovative approach to unsupervised video anomaly detection framework CCC-T which is leveraging the inherent consistency between optical flow features and appearance features. The framework capitalizes on the correlation properties of these two types of features, marking the first instance of their fusion within an unsupervised algorithm.

In this framework, We set strong consistency constraints to achieve con-

A. Test Result of Avenue
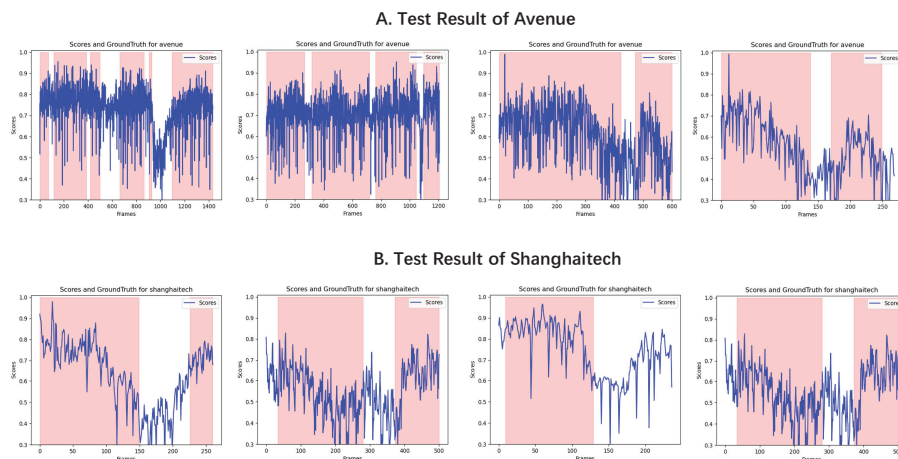

B. Test Result of Shanghaitech

Figure 7: Test results on the dataset Avenue and Shanghaitech; A represents the result of Avenue and B is the result of Shanghaitech. The pink background is the area where the real anomaly occurs; The blue curve represents the prediction anomaly score with the time series.

sistent alignment of appearance features and motion features, and introduce a novel prediction mechanism. This mechanism is bidirectional predicting both optical flow from appearance and appearance from optical flow. This ingenious strategy effectively mitigates the robustness challenges that typically afflict unsupervised learning, thereby generates enhancements in algorithmic performance. Furthermore, the framework employs a co-teaching network, which fosters coordination between the two channels. This approach skillfully averts distortions that can arise from the neural network's potent representation capacity. The empirical findings underscore the superior and more resilient performance of the algorithm proposed in this paper, as compared to conventional methods for predicting video frames. However, the model

35

proposed has some limitations. For example, extracting both dynamic and static features from videos at the same time puts a lot of strain on servers and slows down the processing speed. Additionally, because video scenes vary so much. This means the most similar dynamic and static features might not come from the same video or a normal situation. Most importantly, the model still struggles to clearly explain video anomalies.

In future, our team is committed to delving deeper into the placement of optical flow features within unsupervised anomaly detection algorithms. We aim to explore the potential synergies between a broader range of unsupervised and weakly supervised algorithms, with the goal of pushing the boundaries of anomaly detection even further.

References

[1] D. A. Migliore, M. Matteucci, M. Naccari, A revaluation of frame difference in fast and robust motion detection, in: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, 2006, pp. 215–218.

[2] J.-H. Han, S. Yang, B.-U. Lee, A novel 3-d color histogram equalization method with uniform 1-d gray scale histogram, IEEE Transactions on Image Processing 20 (2) (2010) 506–512.

[3] A. Omid-Zohoor, C. Young, D. Ta, B. Murmann, Toward always-on mobile object detection: Energy versus performance tradeoffs for embedded hog feature extraction, IEEE Transactions on Circuits and Systems for Video Technology 28 (5) (2017) 1102–1115.

[4] W. Shao, R. Xiao, P. Rajapaksha, M. Wang, N. Crespi, Z. Luo, R. Minerva, Video anomaly detection with ntcn-ml: A novel tcn for multiinstance learning, Pattern Recognition (2023) 109765.

[5] Y. Zhao, W. Wu, Y. He, Y. Li, X. Tan, S. Chen, Good practices and a strong baseline for traffic anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3993–4001.

[6] B. Ramachandra, M. Jones, R. R. Vatsavai, A survey of single-scene video anomaly detection, IEEE transactions on pattern analysis and machine intelligence (2020).

[7] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection–a new baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6536–6545.

[8] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.

[9] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, G. Carneiro, Weakly-supervised video anomaly detection with robust temporal feature magnitude learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4975–4986.

[10] H. Mu, R. Sun, M. Wang, Z. Chen, Spatio-temporal graph-based cnns for anomaly detection in weakly-labeled videos, Information Processing & Management 59 (4) (2022) 102983.

[11] X. Wang, Z. Che, B. Jiang, N. Xiao, K. Yang, J. Tang, J. Ye, J. Wang, Q. Qi, Robust unsupervised video anomaly detection by multipath frame prediction, IEEE transactions on neural networks and learning systems 33 (6) (2021) 2301–2312.

[12] X. Huang, C. Zhao, Z. Wu, A video anomaly detection framework based on appearance-motion semantics representation consistency, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[13] Y. Liu, J. Liu, M. Zhao, D. Yang, X. Zhu, L. Song, Learning appearance-motion normality for video anomaly detection, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 1–6.

[14] S. Sun, X. Gong, Hierarchical semantic contrast for scene-aware video anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22846–22856.

[15] T.-N. Nguyen, J. Meunier, Anomaly detection in video sequence with appearance-motion correspondence, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1273–1283.

[16] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, A. v. d. 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, 2019, pp. 1705–1714.

[17] H. Park, J. Noh, B. Ham, Learning memory-guided normality for anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14372–14381.

[18] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, M. Shah, Anomaly detection in video via self-supervised and multi-task learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12742–12752.

[19] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, Y. Wang, Abnormal event detection using deep contrastive learning for intelligent video surveillance system, IEEE Transactions on Industrial Informatics 18 (8) (2021) 5171–5179.

[20] C. Chen, Y. Xie, S. Lin, A. Yao, G. Jiang, W. Zhang, Y. Qu, R. Qiao, B. Ren, L. Ma, Comprehensive regularization in a bi-directional predictive network for video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 230–238.

[21] H. Deng, Z. Zhang, S. Zou, X. Li, Bi-directional frame interpolation for unsupervised video anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2634–2643.

[22] S. Zhang, M. Gong, Y. Xie, A. K. Qin, H. Li, Y. Gao, Y.-S. Ong, Influence-aware attention networks for anomaly detection in surveillance videos, IEEE Transactions on Circuits and Systems for Video Technology 32 (8) (2022) 5427–5437.

[23] C. Guo, H. Wang, Y. Xia, G. Feng, Learning appearance-motion synergy via memory-guided event prediction for video anomaly detection, IEEE Transactions on Circuits and Systems for Video Technology (2023).

[24] Y. Liu, J. Liu, K. Yang, B. Ju, S. Liu, Y. Wang, D. Yang, P. Sun, L. Song, Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system, IEEE Transactions on Industrial Informatics (2023).

[25] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470.

[26] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 4799–4807.

[27] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, B. Kainz, Video summarization through reinforcement learning with a 3d spatio-temporal u-net, IEEE Transactions on Image Processing 31 (2022) 1573–1586.

[28] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2758–2766.

[29] W. Ren, L. Wang, Y. Piao, M. Zhang, H. Lu, T. Liu, Adaptive co-teaching for unsupervised monocular depth estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 89–105.

[30] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022, pp. 11953–11962.

[31] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, S. Kwong, Global-and-local collaborative learning for co-salient object detection, IEEE transactions on cybernetics 53 (3) (2022) 1920–1931.

[32] X. Zhang, Z. Zheng, D. Gao, B. Zhang, P. Pan, Y. Yang, Multi-view consistent generative adversarial networks for 3d-aware image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18450–18459.

[33] L. Zhang, G. Gao, H. Zhang, Spatial-temporal federated learning for lifelong person re-identification on distributed edges, IEEE Transactions on Circuits and Systems for Video Technology (2023).

[34] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, Meta-learning with memory-augmented neural networks, in: International conference on machine learning, PMLR, 2016, pp. 1842–1850.

[35] R. Tudor Ionescu, S. Smeureanu, B. Alexe, M. Popescu, Unmasking the abnormal events in video, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2895–2903.

[36] W. Luo, W. Liu, S. Gao, A revisit of sparse coding based anomaly detection in stacked rnn framework, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 341–349.

[37] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, Y. Wang, Abnormal event detection using deep contrastive learning for intelligent video surveillance system, IEEE Transactions on Industrial Informatics 18 (8) (2021) 5171–5179.

[38] M. Cho, T. Kim, W. J. Kim, S. Cho, S. Lee, Unsupervised video anomaly detection via normalizing flows with implicit latent features, Pattern Recognition 129 (2022) 108703.

[39] J. Li, Q. Huang, Y. Du, X. Zhen, S. Chen, L. Shao, Variational abnormal behavior detection with motion consistency, IEEE Transactions on Image Processing 31 (2022) 275–286. doi:10.1109/TIP.2021.3130545.

[40] C. Huang, J. Wen, Y. Xu, Q. Jiang, J. Yang, Y. Wang, D. Zhang, Self-supervised attentive generative adversarial networks for video anomaly detection, IEEE Transactions on Neural Networks and Learning Systems 34 (11) (2023) 9389–9403. doi:10.1109/TNNLS.2022.3159538.

[41] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE transactions on pattern analysis and machine intelligence 36 (1) (2013) 18–32.

[42] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 fps in matlab, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 2720–2727.

[43] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 589–597.

[44] P. Wu, J. Liu, M. Li, Y. Sun, F. Shen, Fast sparse coding networks for anomaly detection in videos, Pattern Recognition 107 (2020) 107515.

[45] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, J. Yang, Learning normal dynamics in videos with meta prototype network, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15425–15434.

[46] S. Lee, H. G. Kim, Y. M. Ro, Bman: Bidirectional multi-scale aggregation networks for abnormal event detection, IEEE Transactions on Image Processing 29 (2019) 2395–2408.

[47] Y. Lu, F. Yu, M. K. K. Reddy, Y. Wang, Few-shot scene-adaptive anomaly detection, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 125–141.

[48] Y. Chang, Z. Tu, W. Xie, B. Luo, S. Zhang, H. Sui, J. Yuan, Video anomaly detection with spatio-temporal dissociation, Pattern Recognition 122 (2022) 108213.