applied sciences

MDPI

Article

Ensemble-Based Knowledge Distillation for Video Anomaly Detection

Burçak Asal * and Ahmet Burak Can

Department of Computer Engineering, Hacettepe University, Ankara 06800, Turkey; abc@hacettepe.edu.tr
* Correspondence: basal@cs.hacettepe.edu.tr

Abstract: Video anomaly detection has become a vital task for smart video surveillance systems because of its significant potential to minimize the video data to be analyzed by choosing unusual and critical patterns in the scenes. In this paper, we introduce three novel ensemble and knowledge distillation-based adaptive training methods to handle robust detection of different abnormal patterns in video scenes. Our approach leverages the adaptation process by providing information transfer from multiple teacher models with different network structures and further alleviates the catastrophic forgetting issue. The proposed ensemble knowledge distillation methods are implemented on two state-of-the-art anomaly detection models. We extensively evaluate our methods on two public video anomaly datasets and present a detailed analysis of our results. Finally, we show that not only does our best version model achieve comparable performance with a frame-level AUC of 75.82 to other state-of-the-art models on UCF-Crime as the target dataset, but more importantly our approaches prevent catastrophic forgetting and dramatically improve our model's performance.

Keywords: computer vision; ensemble-based methods; knowledge distillation; video anomaly detection; weak supervision

check for updates

Citation: Asal, B.; Can, A.B. Ensemble-Based Knowledge Distillation for Video Anomaly Detection. *Appl. Sci.* **2024**, *14*, 1032. https://doi.org/10.3390/app14031032

Academic Editor: Zulfiqar Habib

Received: 13 December 2023
Revised: 19 January 2024
Accepted: 23 January 2024
Published: 25 January 2024

1. Introduction

Recently, video surveillance systems have become the center of attention as they are deployed and utilized in nearly every place of human settlements for critical tasks such as public protection, crowd analysis and crime detection. However, these systems produce a substantial rate of video data that cannot be analyzed solely by the human factor. That is where the video anomaly detection task becomes prominent as it inherently helps to alleviate the burden of processing many critical surveillance tasks by focusing on abnormal scenes, consequently reducing the data to be analyzed by human operators.

Similar to the general anomaly detection task, the purpose of the video anomaly detection is detecting anomaly patterns in the videos [1]. However, the implementation of this detection process can be very problematic because there is no exact or fixed detection pattern for generalizing all abnormal scenes in videos [1–4]. Different network structures have been proposed for this task [2,3,5,6]. Newer deep learning-based approaches have much more performance potential compared to their classical machine learning or statistical counterpart approaches because of their superior structure for capturing more complex low and high level details such as lighting, texture, background scenes, behavioral patterns [7] or general appearance and motion-based features [8,9] in videos. But ultimately there is no all-in-one model capable of detecting all different anomaly patterns [10,11].

Additionally, as the new video data are received, the models need to adapt to different complex scenes and capture new anomaly patterns. Since the video anomaly concept cannot be defined with specific patterns, the detection mechanism should be based on a data-driven approach; it should adapt itself to new anomaly patterns in the scenes in time by also remembering old patterns learned previously.

A possible solution can be constantly retraining the models with all video data (including new video data), but this solution is not feasible as it brings a heavy computational burden on systems. Another possible solution could be continuously fine-tuning the models with new video data, but this solution is prone to the potential problem of forgetting old learned patterns in time, which is also called catastrophic forgetting. To alleviate this problem, a knowledge distillation method has been proposed, in which a student model can adapt itself to new video data by using a teacher model's feedback and ground truth annotations. In this way, the student model learns from new data while reducing the catastrophic forgetting problem. Moreover, the student model might learn from multiple networks such as in [12] and can combine different kinds of feedback from teacher models having different network structures or trained on different datasets.

Consequently, we think that a separate adaptation mechanism is crucial for a model in a specific scenario, where the model has to prepare itself to constantly learn the new patterns (anomalous or normal) belonging to incoming video samples while the model must also not forget its previously learned patterns on a continuously expanding dataset in time. In this study, we build our motivation and novelty on developing such an adaptation mechanism on weakly supervised video anomaly detection tasks. We also set our objectives and our experiments for evaluating how feasible our proposed approaches are to integrate such an adaptation mechanism.

Proposed Work and Contributions: With this motivation, we extend the knowledge distillation concept into a promising adaption mechanism utilizing information transfer from multiple teacher models having different network structures or trained on different datasets, while alleviating catastrophic forgetting problems. In this paper, we propose three ensemble-based knowledge distillation approaches for video anomaly detection. As a baseline model, we deploy two state-of-the-art model structures, which are the AR-Net model [13] and GCN model [14]. Both of these models are weak supervision-based models utilizing only video level ground-truth information and they can make use of intermediate features of pre-trained I3D action recognition model [15] to leverage the high-level action/behavioral patterns of crowds/actors while classifying anomaly patterns. To gradually adapt the model while forcing it to prevent catastrophic forgetting, we deploy a knowledge distillation approach [16] as a backbone adaptation mechanism. By this approach, we create an ensemble-based learning method based on two AR-Net and GCN-based teacher models, where the AR-Net-based student model tries to optimize itself with respect to ground-truth label information and respective teachers model's own outputs by using a combined loss function.

Our main contributions in the scope of this paper can be listed as follows:

- We introduce three novel ensemble-based knowledge distillation mechanisms for video anomaly detection, where the student model focuses on adapting itself to new incoming data while also focusing on not forgetting old patterns it learned before completely. The student model is also provided with an information transfer including different perspectives from multiple teacher models, during this adaptation process for new information. Consequently, instead of training/fine-tuning a model on a comprehensive-sized dataset growing continuously, the model can be updated with our ensemble knowledge distillation methods more efficiently while preventing the catastrophic forgetting problem. Although the focus of this paper is video anomaly detection, the proposed knowledge distillation methods can also be used in other computer vision tasks such as image classification, scene classification, etc.
- We adapt AR-Net and GCN models into our ensemble-based knowledge distillation approach but other state-of-the-art methods can be easily adapted to our approach. We extensively evaluate AR-Net and GCN models on two comprehensive datasets, which are UCF-Crime [17] and RWF-2000 [18] to validate our proposed methods with respect to quantitative results. Within experiments, firstly we use the UCF-Crime dataset as the source dataset, which represents the baseline dataset including previous patterns the model first learned and RWF-2000 represents the target dataset including

new patterns to be learned by the model. In later stages, we also switch dataset roles, meaning that we use RWF-2000 as the source dataset and UCF-Crime as the target dataset. Consequently, we try to ensure that the model shows stable performance and behavior in case the dataset roles are reversed.
- We present extensive experiments and analyze different parameters with respect to the knowledge distillation mechanism and three different ensemble-based formulations. Results show that our proposed approaches reduce the catastrophic forgetting problem of the model on the source dataset while generally improving the performance of the model on the target dataset. We also discuss limitations and possible future work plans for our proposed approaches.
- We share our code for future research studies on this link: https://github.com/BurcakAsal/AnomalyEnsembleKD (accessed on 19 January 2024).

Our study continues in Section 2; related studies with respect to our study are provided in Section 3 and our proposed ensemble-based knowledge distillation approaches and different formulations are introduced in detail; in Section 4 we provide and discuss the quantitative results we obtained. Finally, in Section 5, the conclusion of our study and discussion about possible improvements for future work are provided.

2. Related Works

Within the domain of our study, we focus on related literature about supervision factors, weak supervision and knowledge distillation.

2.1. Supervision Factor

With respect to the supervision factor, generally, we can classify the anomaly detection task into three main branches, which are unsupervised, weakly supervised, and supervised-based models [19] :

1. **Unsupervised Models:** These models do not require any pre-annotated label or ground-truth information. They focus on intrinsic feature patterns of video data samples and the classification process is considered as outlier detection. An outlier pattern is assumed as its intrinsic features are not similar enough to other normal intrinsic features.
2. **Supervised Models:** In supervised models, frame level annotations or bounding box level annotations are used in the training process as ground-truth information besides video level annotations.
3. **Weakly Supervised Model:** These types of models use a weak supervision method, where both normal and abnormal labels are provided as video-level annotations in the training process. Frame-level labels or bounding box-level labels for localization are not provided to these models in the training process.

Our proposed method uses a weak supervision model in the training process. Therefore, the studies using weak supervision are focused on within the scope of this study.

2.2. Weak Supervision for Video Anomaly Detection

As a critical study for this domain, Ref. [17] represents a video anomaly detection task as a regression problem and utilizes deep multiple instance learning (MIL) based formulation. For providing ground-truth information in the training process, the study only uses video-level labels; the labels represent whether a video sample contains an abnormal scene or not. Another study, Ref. [14] assumes the weak supervision concept as a one-sided label noise problem and also deploys Graph Convolutional Network (GCN) structures by proposing feature similarity and temporal consistency concepts in videos. Ref. [20] proposes that normal scenes in videos contain predictable patterns, while anomaly scenes do not include such patterns. With this assumption, the study uses a margin learning approach with convolutional LSTM and encoder network structures to determine the boundaries between abnormal and normal patterns more precisely. Ref. [21] utilizes a

unique inner bag loss function taking into account the highest and lowest anomaly score values in each bag within the MIL problem. By this formulation, the study's approach tries to make sure that a positive bag has a large difference between the lowest and highest score and a negative bag has a small difference. The approach also makes use of the temporal convolutional network (TCN) using pre-trained C3D features [22]. Ref. [23] deploys a temporal MIL approach by using an attention mechanism to determine which segment of the video is more crucial than another. The study also makes use of a temporal augmented network to leverage motion patterns for the classification process. Ref. [24] develops a Siamese neural network-based decision structure to fit a distance function between pairwise video sub-sequences by a data-driven concept. Consequently, if a video sub-sequence is not similar enough compared with another video sub-sequence by the decision network, then the video sub-scene is supposed to be abnormal. Ref. [25] relies on a multimodal approach that combines the visual and audio features of the video file. Then, these combined features are input to three-sectioned sub-neural networks working concurrently with each other and representing different aspects of relations between video sub-sequences. Ref. [26] advances the network structure by a distance metric-based clustering loss function to minimize the distance interval between clusters of normal videos and maximize the distance interval between clusters of abnormal videos. Also, the study uses the normalcy suppression approach to reduce the anomaly scores belonging to normal sections of the video file. Finally, it also deploys a batch-based training process depending on random selection to decrease inter-batch relations. Ref. [27] develops a multiple instance pseudo label generator structure, which utilizes a sampling approach to generate proper clip-level artificial annotations. The study also uses a self-attention-based encoder structure to focus on abnormal regions of videos and a self-training process. Ref. [28] thinks of the MIL concept as a Robust Temporal Feature Magnitude Learning (RTFM) concept, in which a feature magnitude learning function is used for focusing especially on rare abnormal patterns and preventing the model from having a large bias to normal events in videos. The RTFM approach also deploys dilated convolutions and self-attention mechanisms to handle variable-size temporal interval relations of patterns in videos. Ref. [29] proposes a weak supervision MIL-based spatio-temporal classification mechanism to localize spatio-temporal regions in consecutive frames. Within this mechanism, the study develops a dual-branched structure model, in which each branched structure makes use of a relation module to handle possible object and behavior relationships for anomaly detection by a self-attention mechanism. The study also deploys a separate approach to pass learned representations from one branch structure to another. Ref. [30] proposes a context encoder structure to take into account temporal differentiations and high-level semantic representations for weakly supervised anomaly detection tasks and utilizes a noise stimulation mechanism to minimize false positive anomaly samples. Ref. [31] introduces an approach including four model structures to obtain temporal relations between frames of videos and suitable feature discrimination. Ref. [32] develops a transformer-based approach by also applying a Multi Sequence Learning (MSL) based ranking loss function in the scope of weakly supervised anomaly detection tasks. Lastly, Ref. [33] suggests a model composed of a specific amplification mechanism to improve feature discrimination by the proposed model's two glance and focus modules. These modules also include video clip level transformer structures. Finally, the study also integrates a unique loss function to maximize the distinction between normal features and anomalous features.

2.3. Knowledge Distillation

Before introducing the Knowledge Distillation concept, we also introduce important key rules about Incremental Learning and Continuous Learning concepts within the scope of our study. Firstly according to Incremental Learning these three constraints should be satisfied [34]:

1. While the model is updating itself with respect to the stream data, it should not update itself by using all of the previously collected data; therefore, it should focus on using the new stream information/data.
2. While the model is updating itself with respect to the new data, it should also remember important data patterns, it should not forget important and fundamental patterns within the total big data. Therefore, the proposed model should not be affected by the Catastrophic Forgetting while adapting the new data.
3. In constant stream data, fixed-size data should be used for the updating process.

Also, in the process of adapting the model to new data, three main scenarios are generally encountered [35]

1. **Fine-Tuning Case for New Tasks:** On two datasets, which do not intersect with respect to the samples and classes they include. One dataset of these two is the dataset for the new task and a pre-trained model trained on the latter dataset (used for the old task) beforehand, is re-trained on the dataset for a new task. While in the retraining process, parameters from the specific last layers or the parameters of all layers of the pre-trained model can be updated/optimized for the new task.
2. **Continuous Learning for Known Classes:** In this situation, additional training data are constantly (or in certain intervals) added from a streaming source to the baseline data used before. The task has not changed and the new samples' classes are exactly the same as the samples' classes from the old data. This concept can be also thought as a standard online learning concept.
3. **Continuous Learning of Known and New Classes:** In addition to the second scenario, now in this case, classes from new samples can also be different from old ones.

Our study is actually within the scope of the third scenario with an additional case. Even though data samples from the new dataset compared to the baseline dataset we experiment on have included new patterns and classes, our task especially focuses on the binary (two classes) video anomaly detection task, which means our proposed approach classifies the video data into two possible outputs in which videos include an anomaly scene or not. While our anomaly detection task is binary, the new data samples still include new class information and new pattern information intrinsically so our model has to adapt itself. To satisfy the requirements of baseline rules and constraints mentioned above, we use the Knowledge Distillation (KD) approach within the scope of our study.

Knowledge Distillation is an approach to transfer distilled learned pattern information from a larger network to a smaller network. In this approach, there are two important module structures, which are teacher and student models. During the training process, the teacher model's task is to transfer its intrinsic information from data patterns learned previously from the student model. The student model is trained by a weighted combined loss of two functions, the first loss function is calculated with respect to the difference between the student model's end output logit values and ground-truth label values belonging to new data. The second loss function is generally specified as a "distillation" function and calculated between the student model's end output logit values and the teacher model's end output logit values. Teacher model logit values are obtained by giving the same input vector in the new data that are given to the student model as input. Thus, the teacher model reinterprets the input vector with its own perspective and teaches this perspective to the student model with its own end logit values. That way, while the student model adapts itself to the patterns in the new data, this approach also forces the student model to remember/keep previous data patterns distilled by the teacher model in its neural structure to a certain extent [16].

Within the scope of [36–38] studies, the main focused branches with respect to knowledge strategy are stated below:

1. **Response-Based Knowledge:** This knowledge strategy is the basic approach within the scope of the KD approach. Distillation loss is obtained by calculating the difference between the teacher model's end output logit values and the student model's end

output logit values. Softmax function is generally utilized to obtain output logit values for both of the models. An additional temperature constant is also utilized within the formula to prevent focusing too much on one of the softmax outputs while passing information, thus distributing information transfer uniformly with respect to each softmax output.

2. **Feature-Based Knowledge:** Different from Response-Based Knowledge, the Feature-Based Knowledge approach proposes to use intermediate-level layers of teacher and student models for the knowledge distillation method.
3. **Relation-Based Knowledge:** Relation/correlation matrix representation of outputs for both teacher and student models are focused on this type of knowledge strategy. In the later stage, the distillation-based loss function is minimized by calculating the difference between teacher and student model relation matrices. Additionally, the relation matrices can be calculated by utilizing end-output logits or intermediate-level output logits. Lastly, an inner product or a specialized distance/similarity-based function can be utilized to create a relation matrix for the teacher/student models.

There are also ensemble-based approaches in which a multi-teacher structure is deployed instead of a single-teacher module.

Ref. [39] introduces a multiple teacher-based distillation method by using the mean operation of end outputs of teacher models and by using a voting approach on how dissimilar an intermediate output of a teacher model is with another teacher model. Ref. [40] proposes a sequence-based multiple-teacher distillation approach in which student and teacher model structures are identical and for the specific size of iterations, each teacher model is trained within the supervision of the previous generation teacher model. Finally, an ensemble structured model is obtained by applying average operation on outputs of all previous generation versions of teacher models. Ref. [41] develops an intermediate feature-level-based distillation method on multiple teachers concurrently by utilizing additional non-linear layers. The study also deploys a sequential version of its proposed method by recursively distilling the previous teacher model from the student model and setting the student model as the new teacher model for the next iteration. Ref. [42] follows a distillation strategy in which both end soft output logits and intermediate outputs are used. The study calculates a weighted combination of soft outputs of multiple teachers with respect to data samples while it also utilizes intermediate feature outputs of multiple teachers with its specific early-layer and random-based selection strategy. Ref. [43] makes use of a weighted ensemble method by assignment with respect to the gradients of teacher models. Lastly, Ref. [44] deploys a specific reinforcement learning method to dynamically weighting teacher models to pass a weighted combination of different perspective information to student models.

3. Proposed Approaches

In this paper, we propose three ensemble-based knowledge distillation methods and use state-of-the-art AR-Net [13] and GCN [14] networks by combining these network structures with an I3D action recognition network [15] as a feature extractor to realize our approach (Figure 1).

AR-Net is a weakly supervised model that combines two specific loss functions within its training process. The first loss function ensures that anomalous and normal class samples are distant from each other in the representation space and the second loss function handles the proximity level between normal class samples. Consequently, these two loss functions provide the AR-Net model with an accurate representation of training samples to properly accomplish the detection task.

GCN is another weakly supervised model utilizing two unique graph modules in its inner structure. The first graph module provides the similarity information within different sub-clips of video samples, while the second graph module adds additional temporal information feedback by comparing temporal positions of the sub-clips with each

other. Combined information from these two graph modules provides the GCN model to effectively learn a proper representation of abnormal and normal samples in a dataset.

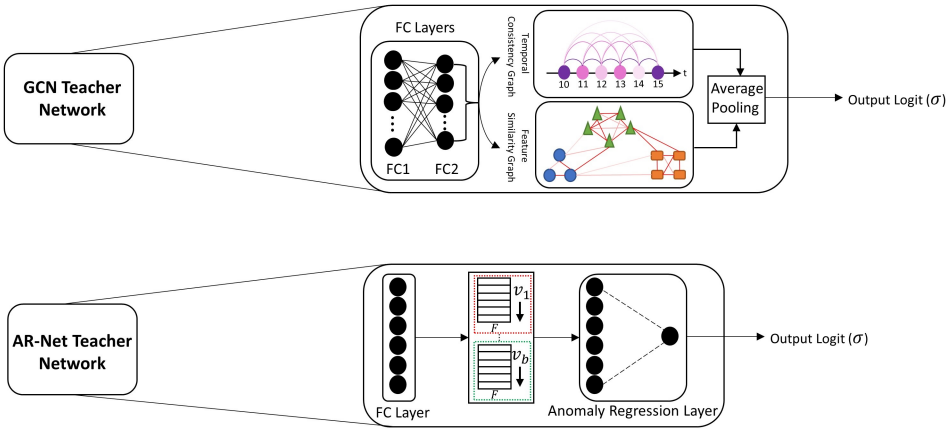

Figure 1. Visual overview of individual GCN and AR-Net teacher modules.

Both of these weakly supervised models require a pretrained network as a module for the feature extraction process. For this purpose, we decided to use the I3D network as a backbone mechanism for our baseline AR-Net and GCN models. I3D is a specific action recognition network utilizing a double stream of three-dimensional inflated convolutional modules. With these modules, the I3D network can produce a feature representation for video samples, which includes a rich summary of a video sample with respect to its behavioral patterns of actors, different level scene details and background details. Consequently, the I3D network helps the AR-Net and GCN models to properly translate a video sample to an accurate feature representation for further processing in their inner network structure.

AR-Net and GCN models use weak supervision and these models are relatively feasible for integrating our ensemble-based knowledge distillation approaches; we decided to use these models as a baseline and I3D as a backbone feature extractor network within our proposed approaches.

Firstly, for all three formulations, video samples are represented as clips, where t_i is the i^{th} clip for a specific video sample. Also, each video includes N video clips and padding operation is applied in case of a video is shorter than N clips. Then, a feature vector (size of F) is obtained for each video clip by utilizing a video action network, we used an I3D pretrained video action network but it is also suitable to deploy other video feature extraction models in this process. After obtaining a mini-batch of clip features from b video samples, we give clip-level features in each video sample of the batch as input to both GCN and AR-Net networks. Within the scope of the three approaches, GCN and AR-Net teacher models are trained on the same source dataset in which different information perspectives from different network structures are planned to pass to the student model.

All of our proposed approaches adopt the knowledge distillation concept to handle the catastrophic forgetting problem. Within our approaches, the student model concurrently obtains feedback signals from both the ground-truth labels of the target dataset and the end output logits of the teacher models which are pre-trained on the source dataset during the training process. Therefore, we try to ensure that while the student model effectively learns the new patterns on the target dataset, it also keeps the previously learned patterns from the source dataset provided indirectly by the teacher models' supervision. Moreover, our proposed approaches accomplish this adaptation process by combining (selecting or merging depending on the approach) the supervision signals from two different teacher networks (AR-Net and GCN) during the information transfer on the AR-Net student model.

In the following three sections, we show how we integrate these three ensemble-based distillation approaches on AR-Net and GCN models.

3.1. Equally Weighted Combination (EWC) Approach

In the EWC approach, the distillation process is formulated as below to provide the AR-Net student model with an equal-weighted distillation signal from both AR-Net and GCN teacher models (Figure 2):

$$L_{Total} = (1 - \alpha) * L_{GT} + L_{EWC} \tag{1}$$

$$\begin{aligned} L_{EWC} = \frac{\alpha}{2} * H(\sigma(Z_{t_1}; T = \tau), \sigma(Z_s; T = \tau)) + \\ \frac{\alpha}{2} * H(\sigma(Z_{t_2}; T = \tau), \sigma(Z_s; T = \tau)) \end{aligned} \tag{2}$$

where Z_{t_1} and Z_{t_2} represent the end output logits of GCN and AR-Net teacher models, respectively, and Z_s represents the end output logit of the AR-Net student model. H represents the cross entropy function and σ represents the sigmoid function. L_{GT} represents the specific cross-entropy-based ground-truth loss function used within the context of the AR-Net student model. Finally, α is the balancing hyperparameter between L_{GT} and L_{EWC} and T is the temperature constant.

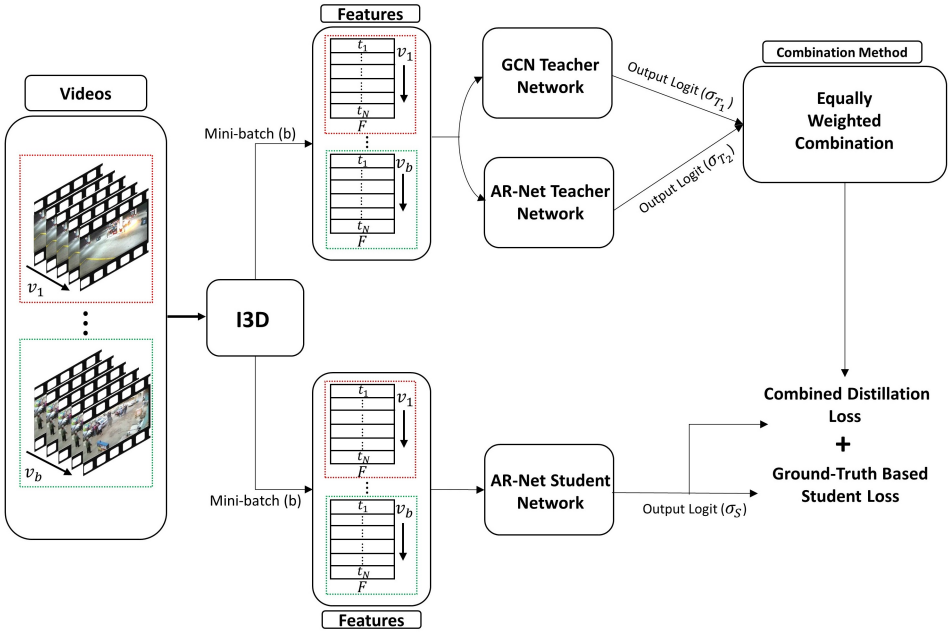

Figure 2. Visual overview of Equally Weighted Combination approach.

The EWC approach uses equal weight values ($\frac{\alpha}{2}$) for merging the two feedback signals from AR-Net and GCN teacher models in the training process. In this way, the student model can use multiple information feedback from different teacher networks and utilize different interpretations of a sample with respect to each teacher model's respective output. Within our first approach, we decided to especially observe how naively distributing the α parameter value equally affects the student model's performance in our task.

Compared to the standard response-based knowledge distillation strategy, this formulation allows equally weighted different perspective-based information transfer from multiple different model teacher models to student models by utilizing a combined distillation loss function.

3.2. Confidence Based Maximum Selection (CBMS) Approach

In this approach, a selection mechanism has been integrated within the loss function compared to the EWC approach as shown in Figure 3:

$$L_{Total} = (1 - \alpha) * L_{GT} + L_{CBMS} \tag{3}$$

$$L_{CBMS} = \alpha * \underset{C_{T_1}, C_{T_2}}{argmax}(H(\sigma(Z_{t_1}; T = \tau), \sigma(Z_s; T = \tau)), H(\sigma(Z_{t_2}; T = \tau), \sigma(Z_s; T = \tau))) \quad (4)$$

$$C_{T_i} = \begin{cases} \frac{0.5 - \sigma_{T_i}}{0.5} & \text{if } \sigma_{T_i} < 0.5 \\ \frac{\sigma_{T_i} - 0.5}{0.5} & \text{if } \sigma_{T_i} \geq 0.5 \end{cases} \quad (5)$$

where C_{T_i} represents the confidence scores for the GCN and AR-Net teacher models calculated from sigmoid outputs (σ_{T_i}) of the GCN and AR-Net teacher models. Because directly using sigmoid outputs is inappropriate, this transformation method is designed.

Unlike our previous approach, CMBS deploys a selection process, in which the approach forces the student model to choose a single feedback signal from one of the teacher models (GCN or AR-Net) for a specific input sample. We designed this selection mechanism by comparing confidence scores obtained from teacher models' sigmoid outputs and choosing the teacher model with a bigger confidence score for information transfer to the student model. Since directly using the sigmoid output of a teacher model is not appropriate for determining confidence level, we developed a special function (Equation (5)) that transforms the sigmoid output of each teacher model to a confidence score, which represents how much each teacher trusts itself for determining the class (anomalous or normal) of a specific sample. Please notice that the function naturally reaches its minimum value on the "0.5" sigmoid function value, which represents the lowest level of confidence and reaches its maximum values on "0" and "1" sigmoid function values, which represent the highest level of confidence.

Compared to the EWC approach, this method provides a student model to select one of the teacher models (GCN or AR-Net) with respect to their current confidence scores calculated on their end outputs.

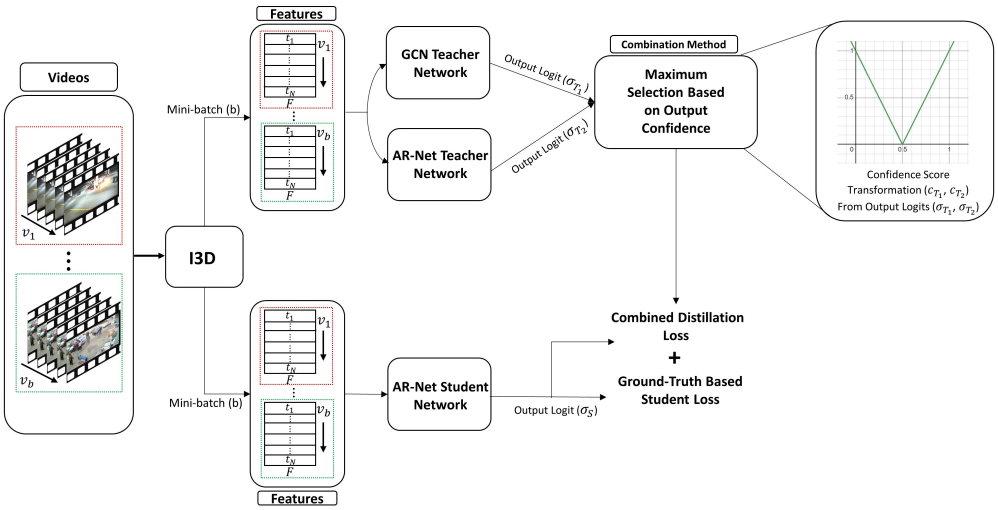

Figure 3. Visual overview of Confidence Based Maximum Selection approach.

3.3. Confidence Based Weighted Combination (CBWC) Approach

In this approach, a weighted combination mechanism based on confidence scores has been integrated within the total loss function (Figure 4):

$$L_{Total} = (1 - \alpha) * L_{GT} + L_{CBWC} \quad (6)$$

$$L_{CBWC} = \alpha * (\beta_{T_1} * H(\sigma(Z_{t_1}; T = \tau), \sigma(Z_s; T = \tau)) + \beta_{T_2} * H(\sigma(Z_{t_2}; T = \tau), \sigma(Z_s; T = \tau))) \quad (7)$$

$$\beta_{T_1} = \frac{C_{T_1}}{C_{T_1} + C_{T_2}} \quad \beta_{T_2} = \frac{C_{T_2}}{C_{T_1} + C_{T_2}} \quad (8)$$

where β_{T_1} and β_{T_2} are dynamic weight values recalculated from respective confidence scores C_{T_1} and C_{T_2}.

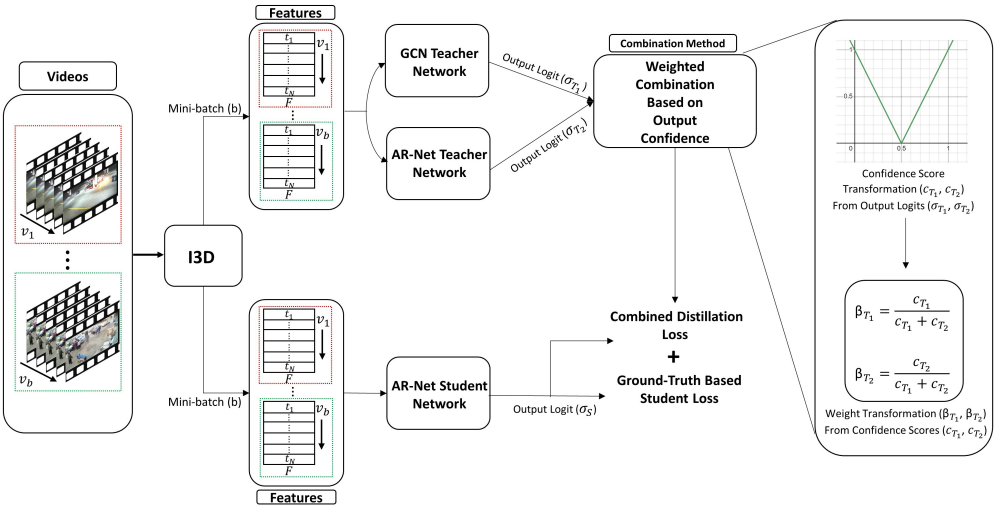

Figure 4. Visual overview of Confidence Based Weighted Combination approach.

Different from EWC and CMBS, the CBWC approach follows a merging operation of different feedback signals from teacher models by dynamically weighting each feedback signal of a teacher model with respect to its confidence level. The confidence level of each teacher model again is calculated by the same transformation function (Equation (5)) with respect to the specific sample. Subsequently, β_{T_1} and β_{T_2} weight values are calculated for each teacher model by applying a normalization operation (Equation (8)) on the confidence scores C_{T_1} and C_{T_2}. Consequently, the CBWC approach provides the student model with a mixture of information feedback comprised of different interpretations of samples by AR-Net and GCN teacher models, but additionally with a dynamic weighting factor.

In summary, compared to the EWC and CBMS approach, this method provides a student model to combine a weighted combination of teacher models (GCN or AR-Net) with respect to their current confidence scores calculated on their end outputs.

4. Experimental Evaluation

In this section, we perform extensive experiments on UCF-Crime [17] and RWF-2000 datasets [18] for both AR-Net and GCN-based teacher and student models and show quantitative results. In our knowledge distillation-based training and testing procedure, we always train the teacher model on the source dataset's training set, while training the student model on the target dataset's training set. A self-distillation mechanism is deployed in the experiments, which means that the AR-Net teacher and AR-Net student model networks are structurally identical and the AR-Net student model initially starts with the same weight parameters as the AR-Net teacher model for all experiments.

We evaluate the performance of our models without knowledge distillation, with the standard response-based knowledge distillation and with our ensemble-based knowledge distillation methods. Also, within experiments, we reversed the source and target datasets (by first using UCF-Crime as the source and RWF-2000 as the target, and then vice versa) to ensure that our approaches show stable performance even if the dataset roles are reversed.

Video-level AUC (Area Under Curve) metric (V-AUC) is utilized to calculate the quantitative performance of the models. All the quantitative results are extracted by testing the student model on the source dataset's test set, the target dataset's test set, and ultimately the source and target dataset's combined test sets (UCF-Crime + RWF-2000). We also show the student model's performance on test sets with respect to variable Alpha (α) parameters.

4.1. Datasets

In this study, we deploy UCF-Crime and RWF-2000 datasets for training and again evaluate the performance of the proposed approaches on these datasets. The UCF-Crime dataset includes 1900 videos from CCTV cameras, in which, 950 videos have anomaly (Abnormal) patterns in 13 different anomaly classes and the remaining 950 videos consist of only normal patterns. The training set of UCF-Crime includes 800 normal and 810 anomalous videos while the test set of UCF-Crime includes 150 normal and 140 anomalous videos. The RWF-2000 dataset includes 2000 videos. Each video of the RWF-2000 dataset is annotated with a class label of "Violent" or "Non-Violent"; 1000 videos have violent behavior patterns, while the other 1000 videos consist of solely non-violent behavior patterns in which both violent and non-violent video sets consist of 800 videos for training and 200 videos for testing.

With respect to the scope of our study, the main task is the binary classification of video anomaly detection. Hence, we represent any video including an anomaly pattern belonging to one of 13 different anomaly classes as the "Anomaly" class video and videos belonging to the normal class as the "Normal" class video for UCF-Crime. Similarly again, we reconsider violent behavior patterns as a subset of anomaly class patterns; consequently, we set "Violent" class videos as videos containing anomaly patterns and "Non-Violent" class videos as videos including purely normal patterns for RWF-2000.

4.2. Results Obtained without Knowledge Distillation

In order to demonstrate the performance contribution of our knowledge distillation strategies, as a beginning, AR-Net and GCN models are trained on UCF-Crime and RWF-2000 datasets without knowledge distillation. Firstly, we train the AR-Net teacher model with the training set of the UCF-Crime dataset and the student model with the training set of the RWF-2000 dataset. Then, we test these models on the test sets of the UCF-Crime and RWF-2000 datasets. Table 1 shows that the teacher and student models have better performance on their training datasets than the other datasets. Then, we repeat the same experiment with the GCN teacher and student models, and we obtain similar results (Table 2). We can also observe that while the teacher model shows poor performance on the test of the RWF-2000 dataset and again the student model shows poor performance on the test set of the UCF-Crime dataset, the teacher model shows relatively much higher performance on the test of UCF-Crime and the student model shows again much higher performance on the test set of the RWF-2000 dataset. This is a consistent result we expected when we do not deploy the knowledge distillation process. This means also that both teacher and student models cannot exhibit high performance on their counterpart datasets.

Table 1. V-AUC values of AR-Net-based student and teacher models without knowledge distillation.

	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
Teacher	82.1	59.1	70.7
Student	60.8	77.3	68.9

Table 2. V-AUC values of GCN-based student and teacher models without knowledge distillation.

	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
Teacher	81.3	57.6	69.3
Student	59.2	75.2	67.2

4.3. Results Obtained with Response-Based Knowledge Distillation

For obtaining a baseline experiment for the performance comparison with our ensemble-based knowledge distillation methods, additional experiments are carried out on UCF-Crime and RWF-2000 datasets with standard response-based knowledge distillation method [16].

Firstly, we train the AR-Net-based teacher and student models by using response-based knowledge distillation on training sets of UCF-Crime and RWF-2000 datasets, respectively. The results of these models on the test sets are demonstrated in Table 3 with respect to the alpha parameter. As the alpha parameter increases, the student model's V-AUC values generally increase on the test set of the source dataset and decrease on the test set of the target dataset. Thus, we can conclude that the teacher model has a gradual influence on the student model with increasing alpha parameter values. The same experiment is replicated with GCN-based teacher and student models and also concluded the similar pattern as shown in Table 4.

If the results of Table 1 are compared with Table 3 and the results of Table 2 with Table 4, we can observe that both AR-Net and GCN-based student models trained with the knowledge distillation approach have better V-AUC performance on all test sets compared to the student model without the distillation approach. Thus, it can be concluded that the knowledge distillation process leverages the student model to remember and not to forget completely the patterns of the source dataset, and hence catastrophic forgetting is alleviated to a certain extent.

Table 3. V-AUC values of AR-Net-based student model with standard response-based knowledge distillation. (**Source: UCF-Crime, Target: RWF-2000**).

Alpha (α)	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
0.1	68.4	77.5	73.5
0.2	68.9	77.1	73.2
0.3	70.3	74.8	72.4
0.4	72.1	74.0	73.0
0.5	73.0	70.2	71.6

Table 4. V-AUC values of GCN-based student model with standard response-based knowledge distillation. (**Source: UCF-Crime, Target: RWF-2000**).

Alpha (α)	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
0.1	69.4	76.8	73.1
0.2	69.2	74.9	72.2
0.3	71.2	74.1	72.6
0.4	72.3	71.2	71.8
0.5	73.7	69.0	70.9

Within the experiments of Tables 3 and 4, the student model is initialized with the teacher model's weights before the training process. Also, as an alternative approach, the training process of the student model is started with random weights instead of starting with teacher model weights. With respect to this alternative approach, we observed that initializing with the teacher model's weights provides better performance of the student model compared to initializing with random weights. In the following sections, we also obtained the same observation for the proposed methods. Consequently, we decided to not present the results with random weights in this section and the following sections.

4.4. Results of EWC Approach

In Tables 5 and 6, the V-AUC results are obtained by training the AR-Net student model with the EWC Approach. The experiments are also replicated by changing the datasets setting RWF-2000 as the source and UCF-Crime as the target dataset. Compared to response-based knowledge distillation approaches in Tables 3 and 4, it can be observed that the results are slightly improved. We also observed a similar pattern with respect to

increasing alpha parameters, i.e., increased alpha values raise the teacher's model influence on the student model and lead up to gradually increased V-AUC values on the test set of the source dataset and gradually decreased V-AUC values on the test set of the target dataset. These results also show that, by the EWC approach, the AR-Net student model leverages the equally combined feedback information from both GCN and AR-Net teacher models to a certain extent.

Table 5. V-AUC values of AR-Net-based student model with EWC approach. (**Source: UCF-Crime, Target: RWF-2000**).

Alpha (α)	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
0.1	70.8	78.1	74.1
0.2	69.7	76.5	73.0
0.3	69.9	75.3	72.2
0.4	71.6	73.9	72.7
0.5	72.8	69.6	71.4

Table 6. V-AUC values of AR-Net-based student model with EWC approach. (**Source: RWF-2000, Target: UCF-Crime**).

Alpha (α)	RWF-2000	UCF-Crime	UCF-Crime + RWF-2000
0.1	69.1	75.2	72.2
0.2	71.4	75.7	73.7
0.3	71.0	73.3	72.1
0.4	73.8	72.6	73.4
0.5	74.3	70.4	72.3

4.5. Results of CBMS Approach

The experiments with the CBMS approach produce the results in Tables 7 and 8. These results again show that the AR-Net student model follows a similar pattern compared to the EWC formulation with respect to increasing alpha parameters. More importantly, all the V-AUC results are again slightly improved, especially with specific parameters compared to results of the EWC method in Tables 5 and 6. This shows that the CBMS approach generally provides better information transfer by helping the student model to select which individual teacher network (GCN or AR-Net) is better to pass information by comparing and selecting the teacher model which has a bigger confidence score. Also, when the source and target datasets are switched, we can observe that the approach shows similar and consistent behavior compared to previous distillation approaches.

Table 7. V-AUC values of AR-Net-based student model CBMS approach. (**Source: UCF-Crime, Target: RWF-2000**).

Alpha (α)	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
0.1	71.3	78.4	74.7
0.2	71.8	77.2	74.4
0.3	72.5	75.1	73.5
0.4	73.1	74.9	73.6
0.5	73.9	70.3	72.1

Table 8. V-AUC values of AR-Net-based student model with CBMS approach. (**Source: RWF-2000, Target: UCF-Crime**)

Alpha (α)	RWF-2000	UCF-Crime	UCF-Crime + RWF-2000
0.1	72.9	76.7	74.9
0.2	73.8	75.5	74.7
0.3	75.4	76.3	76.0
0.4	74.5	72.9	73.7
0.5	74.2	73.3	73.6

4.6. Results of CBWC Approach

The results of experiments with the CBWC approach are presented in the Tables 9 and 10. The normal results and results with reversed source and target datasets show that the student model again exhibits a similar behavioral pattern compared to EWC and CBMS approaches with respect to increasing alpha parameters. Additionally, the general performance of the student model is again slightly improved on specific alpha parameters compared to Tables 5 and 6 with respect to the EWC method, but no drastic improvement is observed compared to the CBMS method (Tables 7 and 8). We explain a possible reason for this in Section 4.7. Compared to the CBMS, instead of selecting a feedback distillation signal from one of the GCN or AR-Net teacher models in CBMS, the CBWC approach provides the AR-Net student model with the weighted combined signal feedback synthesized from both the GCN and AR-Net model teacher models and helps to alleviate catastrophic forgetting by improving the performance on the test set of the source dataset.

Table 9. V-AUC values of AR-Net-based student model with CBWC approach. (**Source: UCF-Crime, Target: RWF-2000**).

Alpha (α)	UCF-Crime	RWF-2000	UCF-Crime + RWF-2000
0.1	71.0	77.6	74.2
0.2	71.9	77.4	74.8
0.3	73.7	76.8	75.1
0.4	73.6	74.9	74.0
0.5	72.5	71.1	71.9

Table 10. V-AUC values of AR-Net-based student model with CBWC approach. (**Source: RWF-2000, Target: UCF-Crime**).

Alpha (α)	RWF-2000	UCF-Crime	UCF-Crime + RWF-2000
0.1	73.8	77.0	75.3
0.2	73.4	76.9	75.2
0.3	74.4	76.1	75.4
0.4	75.0	76.2	75.7
0.5	75.6	73.2	74.2

4.7. An Additional Experiment

Besides the experiments mentioned above, an additional experiment is also carried out on individual GCN and AR-Net-based teacher models and AR-Net-based student model with EWC method by applying a specific sigmoid output threshold value **"0.5"** from Tables 11–17.

Table 11. Confusion matrix of GCN teacher model with 0.5 threshold value on test set of Source Dataset. (UCF-Crime).

GCN Teacher (UCF-Crime)	Actual Positive	Actual Negative
Predicted Positive	109	27
Predicted Negative	31	123

Table 12. Confusion matrix of GCN teacher model with 0.5 threshold value on test set of Target Dataset. (RWF-2000).

GCN Teacher (RWF-2000)	Actual Positive	Actual Negative
Predicted Positive	117	86
Predicted Negative	83	114

In Tables 11 and 12, two confusion matrices for individual GCN teacher models are shown. These matrices are obtained by following a similar procedure as in Section 4.2. An individual GCN Teacher model is firstly trained on a training set of the Source Dataset (UCF-Crime) and then directly tested on a test set of the Source Dataset and Target Dataset (RWF-2000). Tables 11 and 12 results show consistent behavior with respective Table 2 results in which the teacher model's performance is better on the test set of Source Dataset compared to the test set of the Target Dataset.

Table 13. Confusion matrix of AR-Net teacher model with 0.5 threshold value on test set of Source Dataset. (UCF-Crime).

AR-Net Teacher (UCF-Crime)	Actual Positive	Actual Negative
Predicted Positive	114	22
Predicted Negative	26	128

Table 14. Confusion matrix of AR-Net teacher model with 0.5 threshold value on test set of Target Dataset. (RWF-2000).

AR-Net Teacher (RWF-2000)	Actual Positive	Actual Negative
Predicted Positive	120	84
Predicted Negative	80	116

In Tables 13 and 14, again, two confusion matrices for individual AR-Net teacher models are presented. These matrices are extracted again by following a similar procedure as in Section 4.2. This time, the individual AR-Net Teacher model is again firstly trained on a training set of Source Dataset (UCF-Crime) and then directly tested on a test set of the Source Dataset and Target Dataset (RWF-2000). Tables 13 and 14 results also show consistent behavior with respect to Table 1 results, in which the teacher model has drastically better performance on the test set of Source Dataset compared to the test set of the Target Dataset.

Table 15. Confusion matrix of AR-Net student model with 0.5 threshold value on test set of Source Dataset (UCF-Crime) with EWC approach.

AR-Net Student (UCF-Crime)	Actual Positive	Actual Negative
Predicted Positive	106	49
Predicted Negative	34	101

Table 16. Confusion matrix of AR-Net student model with 0.5 threshold value on test set of Target Dataset (RWF-2000) with EWC approach.

AR-Net Student (RWF-2000)	Actual Positive	Actual Negative
Predicted Positive	175	60
Predicted Negative	25	140

In Tables 15 and 16, two confusion matrices are calculated by training the AR-Net student model with the EWC approach (Table 5) with the best overall alpha parameter (0.1) and testing on a test set of UCF-Crime and RWF-2000 Dataset. Please notice, that in Table 16, the student model's performance is drastically improved on RWF-2000 test set (Target Dataset) compared to the Tables 12 and 14 and more importantly, in Table 15, the student model also shows comparable performance on the UCF-Crime test set (Source Dataset) compared to Tables 11 and 13 and shows consistent behavior with respect to Table 5.

Table 17. Misclassified example statistics for both GCN and AR-Net-based teacher models for UCF-Crime (Source Dataset) and RWF-2000 (Target Dataset).

Misclassification	GCN Teacher	AR-Net Teacher	Common
UCF-Crime	58	48	45
RWF-2000	169	164	157

Finally in Table 17, different misclassification statistics are shown for both individual GCN and AR-Net teacher models. The first column represents the total misclassified sample count for the GCN teacher model with respect to Tables 11 and 12, while the second column represents the total misclassified sample count for the AR-Net teacher model with respect to Tables 13 and 14. More importantly, the third column (Common) represents the sample count both misclassified by GCN and AR-Net teacher models. With respect to this table, we concluded that very often, when the GCN teacher model mistakenly classifies a sample, the AR-Net teacher model also shows the same behavior. We think that this behavior explains, to a certain extent, the limited performance increase in our EWC and CBWC ensemble-based distillation formulations in this study compared to the CBMS method.

4.8. Comparison with State-of-the-Art Studies

In this experiment, our best-performed model is quantitatively compared with other state-of-the-art (SOTA) weakly supervised approaches on the UCF-Crime dataset (Table 18). While for the previous experiments, the video-level AUC (V-AUC) metric is utilized to measure the performance of our models because the RWF-2000 dataset does not have frame-level annotations; the UCF-Crime dataset has frame-level annotations and other weakly supervised methods previously experimented on this dataset also use frame-level AUC (F-AUC) as a performance metric. Consequently, we repeat the testing process with our best ensemble-based student model trained on the UCF-Crime dataset with respect to frame-level AUC (F-AUC) metric for comparison with other SOTA studies. For this case, we select the student model trained with the CBMS method with $\alpha = 0.3$ which has the best overall performance as shown in Table 8.

In Table 18, it can be concluded that the best ensemble-based student model (CBMS) we proposed exhibits comparable performance compared to the other state-of-the-art approaches. Besides, our CBMS-based student model also shows reassuring performance in the RWF-2000 dataset as shown in Table 8. Another important observation is that compared to the standard response-based method (Tables 3 and 4), our ensemble-based methods (Tables 5–10) have drastic performance increase regardless of the target dataset. This represents the increased generalization capability of our ensemble-based distillation methods.

Table 18. Comparison with State-of-the-Art (SOTA) Studies on UCF-Crime Dataset.

Method	F-AUC
(Sultani et al., 2018) [17]	75.41
(Zhong et al., 2019) [14]	82.12
(Zhang et al., 2019) [21]	78.66
(Zhu and Newsam, 2019) [23]	79.00
(Wu et al., 2020) [25]	82.44
(Zaheer et al., 2020) [26]	83.03
(Feng et al., 2021) [27]	82.30
(Tian et al., 2021) [28]	84.30
(Lv et al., 2021) [30]	85.38
(Wu and Liu, 2021) [31]	84.89
(Li et al., 2022) [32]	85.62
(Chen et al., 2023) [33]	86.98
Ours (CBMS)	**75.82**

In the case of the training case of AR-Net and GCN models on a specific dataset without knowledge distillation, it is observed that results are dramatically worse on the other dataset (RWF-2000, Tables 1 and 2 in Section 4.2). All SOTA models in Table 18 are trained without a knowledge distillation approach on the UCF-Crime dataset. Because they are trained on the UCF-Crime dataset without distillation, we conclude that while they have results better than our models on this specific dataset, their generalization performance will be lower than our models on another dataset, such as the RWF-2000 dataset as in Section 4.2.

4.9. Discussion on Results

Overall, we can conclude that within the scope of our ensemble-based knowledge distillation-based experiments with our three formulation approaches, respectively, each subsequent formulation generally shows a slight performance increase compared to the previous formulation except the CBWC model with potential reasons explained in Section 4.7. Also, each of the three ensemble-based knowledge distillation methods shows a general behavior with respect to preventing catastrophic forgetting issues. More specifically, we also observe that the selection or combination mechanism within our approaches helps the student model to integrate different supervision feedback from AR-Net and GCN teacher models and passes this feedback information to the student model for keeping the old learned patterns in the source dataset. Also, our approaches provide the student model with stable performance on the target dataset and even increase the student model's baseline performance in specific configurations in the experiments. Ultimately we can also conclude that selection operation or combination operation can have a different effect on the student model's performance depending on the structures of the utilized baseline models. As we mentioned in Section 4.7, because of our baseline models' rather high common misclassification rate, selection operation can have more contribution to the student model's performance compared to combination operation in specific cases.

In our experiments, generally, it can be observed that as the alpha parameter value increases in the knowledge distillation loss function, and the student model's performance gradually increases on the test set of the source dataset, it has a gradual performance decrease on the test set of the target dataset. This situation confirms that increased operation of the alpha parameter also intensifies the teacher model's influence on the student model and brings about closer bias to the source dataset and further bias from the target dataset for the student model. This observation is also crucial for handling the catastrophic forgetting issue because we can also observe that as the teacher model's influence increases on the

student model, the student model performs gradually better on the source dataset, which indicates that the student model gradually focuses on keeping previously learned patterns from the source dataset with increasing alpha parameter values.

In order to also test the robustness of our ensemble-based distillation methods, we additionally have reversed the UCF-Crime and RWF-2000 datasets for source and target dataset roles and replicated our experiments in this way. From experiments, it can be concluded that our experimental results show similar patterns in the case where the source and target datasets are reversed. Consequently, it can be also interpreted as our proposed approaches showing robust and independent performance from the datasets used in our experiments.

Finally, we also prepared a table (Table 19) for the qualitative comparison of the other recent three weakly supervised studies with our ensemble-based knowledge distillation approaches with respect to their advantages and disadvantages. Study [30] utilizes an enhanced context encoding method for handling high-level semantic and temporal variations and a unique noise simulation strategy based on weak supervision for handling noises in video samples. Another study [32] deploys a transformer-based convolutional structure to improve the encoding process of high-level features from video samples by using a two-stage multi-sequence-based training method. Finally, the study [33] again makes use of transformer-based glance and focus sub-modules to effectively capture global and local level information details from video samples by using a special amplification mechanism for obtaining an improved feature representation.

The crucial point can be noticed when we examine the disadvantages of these recent studies compared to our approaches. Firstly, the models in these studies are trained and tested in fixed-size datasets. These studies do not include a scenario where there is a dataset continuously growing in time and the model has to constantly adapt itself for learning incoming samples while also preserving the old patterns learned from previous samples. Secondly, depending on this scenario, these studies do not integrate an additional adaptation mechanism designed to prepare the model for this scenario. With respect to this case, even though these studies outperform the baseline model trained with our proposed adaptation mechanism in Table 18, we conclude that their performance will be dramatically reduced in the aforementioned scenario compared to our adaptation mechanism.

Table 19. Qualitative Comparison with State-of-the-Art (SOTA) Studies.

Method	Advantages	Disadvantages
(Lv et al., 2021) [30]	- More Advanced Temporal and Semantic Context Encoding - More Advanced Weak Supervision Strategy	- Trained and Tested on Fixed-Size Datasets - No Separate Adaptation Mechanism
(Li et al., 2022) [32]	- Transformer Based Convolutional Network Structure - Two Stage Self Training Strategy	- Trained and Tested on Fixed-Size Datasets - No Separate Adaptation Mechanism
(Chen et al., 2023) [33]	- Transformer Based Glance and Focus Modules - Feature Amplification Mechanism	- Trained and Tested on Fixed-Size Datasets - No Separate Adaptation Mechanism

Another possible explanation for the performance limitation in Table 18 is that since our proposed mechanisms require baseline teacher and student models for the adaptation process, the structures of baseline networks utilized for our proposed mechanisms (in our case, GCN and AR-Net networks) can inherently limit the performance of the student model, because of the specific baseline network's individual restricted performance related to its designed structure.

5. Conclusions and Future Work

In this study, we have presented three novel ensemble-based knowledge distillation approaches for video anomaly detection. Our approaches provide a model to adapt itself to incoming new video data, while also preventing the model from completely forgetting learned patterns from previous data by alleviating catastrophic forgetting problems. Additionally, the proposed formulations improve the performance of the student model on the source dataset and help the student model to obtain consistent performance on the target dataset by using the ensemble of multiple different network structures as a teacher model.

As a future work, because our proposed ensemble-based knowledge distillation approaches can be adapted into other baseline networks independently, we plan to integrate our proposed approaches for other newer model structures in the literature, especially for transformer-based networks. We think that this will alleviate the performance limitation issue we mentioned in the previous section.

Another potential limitation is the explainability of our proposed approaches. We also plan to integrate a specific interpretable layer for our adaptation mechanisms for taking different interactions between objects and actors into account in videos such as in study [45]. This will provide additional interpretability for further analyzing the details of the information transfer related to our adaptation mechanisms. We think Large Language Models (LLMs) can have a dramatic contribution to the interpretability issue, because of their inherent ability to capture complex patterns and relationships in data samples and produce human-interpretable text. Furthermore, we think integrating our approaches with Visual Language Models (VLMs) has great potential for adding the explainability concept into our approaches by utilizing Signature Transform-based techniques [46,47].

Finally, we plan to further utilize feature-based knowledge distillation techniques with our ensemble-based approaches for providing richer information transfer for the student model. Additionally, classical ensemble techniques such as Bagging and Boosting can contribute to our study for further analysis.

Author Contributions: B.A. provided methods, conducted relevant experiments and wrote the manuscript; A.B.C. supervised the study, reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant No. 119E098 and Hacettepe University Scientific Research Projects Coordination Department under Grant No. FHD-2022-20044.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used in this study are publicly available on Internet by studies [17,18].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ramachandra, B.; Jones, M.; Vatsavai, R.R. A survey of single-scene video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2293–2312. [CrossRef]
2. Suarez, J.J.P.; Naval, P.C., Jr. A survey on deep learning techniques for video anomaly detection. *arXiv* **2020**, arXiv:2009.14146.
3. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–38. [CrossRef]
4. Pawar, K.; Attar, V. Deep learning approaches for video-based anomalous activity detection. *World Wide Web* **2019**, *22*, 571–601. [CrossRef]
5. Chalapathy, R.; Chawla, S. Deep learning for anomaly detection: A survey. *arXiv* **2019**, arXiv:1901.03407.
6. Mohammadi, B.; Fathy, M.; Sabokrou, M. Image/video deep anomaly detection: A survey. *arXiv* **2021**, arXiv:2103.01739.
7. Şengönül, E.; Samet, R.; Abu Al-Haija, Q.; Alqahtani, A.; Alturki, B.; Alsulami, A.A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. *Appl. Sci.* **2023**, *13*, 4956. [CrossRef]
8. Liu, Y.; Liu, J.; Yang, K.; Ju, B.; Liu, S.; Wang, Y.; Yang, D.; Sun, P.; Song, L. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Trans. Ind. Inform.* **2023**, *20*, 2843–2855. [CrossRef]

9. Wang, L.; Tian, J.; Zhou, S.; Shi, H.; Hua, G. Memory-augmented appearance-motion network for video anomaly detection. *Pattern Recognit.* **2023**, *138*, 109335. [CrossRef]
10. Ren, J.; Xia, F.; Liu, Y.; Lee, I. Deep Video Anomaly Detection: Opportunities and Challenges. In Proceedings of the 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand, 7–10 December 2021; pp. 959–966.
11. Raja, R.; Sharma, P.C.; Mahmood, M.R.; Saini, D.K. Analysis of anomaly detection in surveillance video: Recent trends and future vision. *Multimed. Tools Appl.* **2023**, *82*, 12635–12651. [CrossRef]
12. Panagiotatos, G.; Passalis, N.; Iosifidis, A.; Gabbouj, M.; Tefas, A. Curriculum-based teacher ensemble for robust neural network distillation. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019; pp. 1–5.
13. Wan, B.; Fang, Y.; Xia, X.; Mei, J. Weakly supervised video anomaly detection via center-guided discriminative learning. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
14. Zhong, J.X.; Li, N.; Kong, W.; Liu, S.; Li, T.H.; Li, G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1237–1246.
15. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
16. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
17. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
18. Cheng, M.; Cai, K.; Li, M. RWF-2000: An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4183–4190.
19. Zhu, S.; Chen, C.; Sultani, W. Video Anomaly Detection for Smart Surveillance. In *Computer Vision: A Reference Guide*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–8.
20. Liu, W.; Luo, W.; Li, Z.; Zhao, P.; Gao, S. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 3023–3030.
21. Zhang, J.; Qing, L.; Miao, J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 4030–4034.
22. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
23. Zhu, Y.; Newsam, S. Motion-aware feature for improved video anomaly detection. *arXiv* **2019**, arXiv:1907.10211.
24. Ramachandra, B.; Jones, M.; Vatsavai, R. Learning a distance function with a Siamese network to localize anomalies in videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 2598–2607.
25. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 322–339.
26. Zaheer, M.Z.; Mahmood, A.; Astrid, M.; Lee, S.I. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020, pp. 358–376.
27. Feng, J.C.; Hong, F.T.; Zheng, W.S. Mist: Multiple instance self-training framework for video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14009–14018.
28. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised video anomaly detection with contrastive learning of long and short-range temporal features. In Proceedings of the 2021 18th IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
29. Wu, J.; Zhang, W.; Li, G.; Wu, W.; Tan, X.; Li, Y.; Ding, E.; Lin, L. Weakly-supervised spatio-temporal anomaly detection in surveillance video. *arXiv* **2021**, arXiv:2108.03825.
30. Lv, H.; Zhou, C.; Cui, Z.; Xu, C.; Li, Y.; Yang, J. Localizing anomalies from weakly-labeled videos. *IEEE Trans. Image Process.* **2021**, *30*, 4505–4515. [CrossRef] [PubMed]
31. Wu, P.; Liu, J. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Trans. Image Process.* **2021**, *30*, 3513–3527. [CrossRef] [PubMed]
32. Li, S.; Liu, F.; Jiao, L. Self-training multi-sequence learning with Transformer for weakly supervised video anomaly detection. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), Virtual, 22 February–1 March 2022; Volume 36, pp. 1395–1403. *Proc. AAAI Virtual* **2022**, *36*, 1395–1403.
33. Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; Wu, Y.C. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 387–395.
34. Losing, V.; Hammer, B.; Wersing, H. Incremental on-line learning: A review and comparison of state of the art algorithms. *Neurocomputing* **2018**, *275*, 1261–1274. [CrossRef]

35. Käding, C.; Rodner, E.; Freytag, A.; Denzler, J. Fine-tuning deep neural networks in continuous learning scenarios. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 588–605.
36. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
37. Wang, L.; Yoon, K.J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3048–3068. [CrossRef]
38. Ruffy, F.; Chahal, K. The state of knowledge distillation for classification. *arXiv* **2019**, arXiv:1912.10850.
39. You, S.; Xu, C.; Xu, C.; Tao, D. Learning from multiple teacher networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 1285–1294.
40. Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; Anandkumar, A. Born again neural networks. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1607–1616.
41. Park, S.; Kwak, N. Feed: Feature-level ensemble for knowledge distillation. *arXiv* **2019**, arXiv:1909.10754.
42. Liu, Y.; Zhang, W.; Wang, J. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing* **2020**, *415*, 106–113. [CrossRef]
43. Du, S.; You, S.; Li, X.; Wu, J.; Wang, F.; Qian, C.; Zhang, C. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12345–12355.
44. Yuan, F.; Shou, L.; Pei, J.; Lin, W.; Gong, M.; Fu, Y.; Jiang, D. Reinforced multi-teacher selection for knowledge distillation. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 14284–14291.
45. Doshi, K.; Yilmaz, Y. Towards interpretable video anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2655–2664.
46. De Zarzà, I.; de Curtò, J.; Calafate, C.T. Socratic Video Understanding on Unmanned Aerial Vehicles. *Procedia Comput. Sci.* **2023**, *225*, 144–154. [CrossRef]
47. De Curtò, J.; de Zarzà, I.; Roig, G.; Calafate, C.T. Summarization of Videos with the Signature Transform. *Electronics* **2023**, *12*, 1735. [CrossRef]