Privacy-Preserving Video Understanding via Transformer-based Federated Learning

1st Keval Doshi
Department of Electrical Engineering
University of South Florida
Tampa, Florida, USA
kevaldoshi@usf.edu

2nd Yasin Yilmaz
Department of Electrical Engineering
University of South Florida
Tampa, Florida, USA
yasiny@usf.edu

2023 IEEE Conference on Dependable and Secure Computing (DSC) | 979-8-3503-8211-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/DSC61021.2023.10354099

***Abstract*—Video understanding has been an active area of research over the past several years, which is dominated by deep neural networks like image understanding and other computer vision tasks. In real-world implementations, even after pre-training on large datasets, state-of-the-art deep neural networks for video understanding tasks, such as anomaly detection and action recognition, can greatly benefit from diverse training data from multiple sources to adapt to the specific application. However, sharing videos collected by multiple sources with a central unit may not be feasible in practice due to privacy and communication constraints. Federated Learning (FL),which allows data parties to collaborate on machine learning models while preserving data privacy and reducing communication requirements, can be used to overcome these challenges. Despite significant progress on various computer vision tasks, FL for video understanding tasks has been largely unexplored. To this end, we propose a novel transformer-based approach for video anomaly detection and action recognition, and extensively benchmark the model performance in FL setting. Our results indicate that the proposed approach outperforms all existing state-of-the-art approaches under the centralized (non-FL) setting and performs competitively under the FL setting.**

***Index Terms*—privacy-preserving machine learning, federated learning, video anomaly detection, action recognition, video transformer**

I. Introduction

Given the availability of increasing number of edge devices like CCTV cameras, further training data from multiple sources can greatly facilitate developing artificial intelligence (AI) for video understanding that can approach human-level performance. However, in such a scenario, centralized training is impractical due to massive communication and storage overheads. Moreover, centralizing data may also lead to security and privacy concerns, and violate regulations such as the General Data Protection Regulation [1]. Federated Learning (FL) is offered as a distributed model training method that does not communicate raw data, therefore maintaining data privacy and saving communication bandwidth [2].

Specifically, in an FL system, multiple parties train a machine learning model cooperatively without exchanging raw data [3]. The system generates a common machine learning model for the parties such that the model learnt via FL is superior to a model learned via local training with the same model architecture. Despite the rapid growth of FL research for computer vision, the vast majority of existing works concentrate on image understanding problems [4]. FL in a variety of video understanding problems has received little attention to date.

Even if a model is trained on a large-scale video dataset, it is typically not representative of all nominal or anomalous patterns. Hence, a sustainable model should be continually updated with new training data. Cooperation among multiple data sources can greatly facilitate training a powerful video understanding model with a rich dataset, representative of various data patterns. As illustrated in Fig. 1, it is usually not easy to locally collect and train on a dataset sufficiently representative of all relevant classes/scenarios. On the other hand, gathering a diverse dataset from multiple sources and training a model in a centralized manner is also not feasible in general due to privacy and communication overhead constraints. An FL setup can enable continually updating a video understanding model with diverse data from multiple sources by sharing only some processed information instead of raw data.

A majority of existing video understanding works primarily employ 3D convolutional models such as C3D [5] or I3D [6] for extracting spatiotemporal feature representations from videos. However, convolution based approaches suffer from several shortcomings. Particularly, the learning capacity of convolutional models on huge datasets is severely limited by inductive biases such as local connection, translation invariance, and a locally constricted receptive field. Furthermore, convolutional kernels are incapable of capturing spatiotemporal correlations that span a large number of time instances. Finally, despite gains in hardware acceleration, training and evaluating deep CNNs on large video datasets continues to be a computationally intensive endeavor. On the other hand, self-attention based transformer models are able to overcome these limitations thanks to a relatively larger receptive field [7]. Also, due to being computationally efficient as compared to the convolutional approaches, self-attention based models can process longer video sequences, thus capturing long-range dependencies more effectively. In Fig. 2, we show the t-SNE representation of the extracted visual features using the proposed transformer based approach as compared to the

979-8-3503-8211-2/23/$31.00 ©2023 IEEE

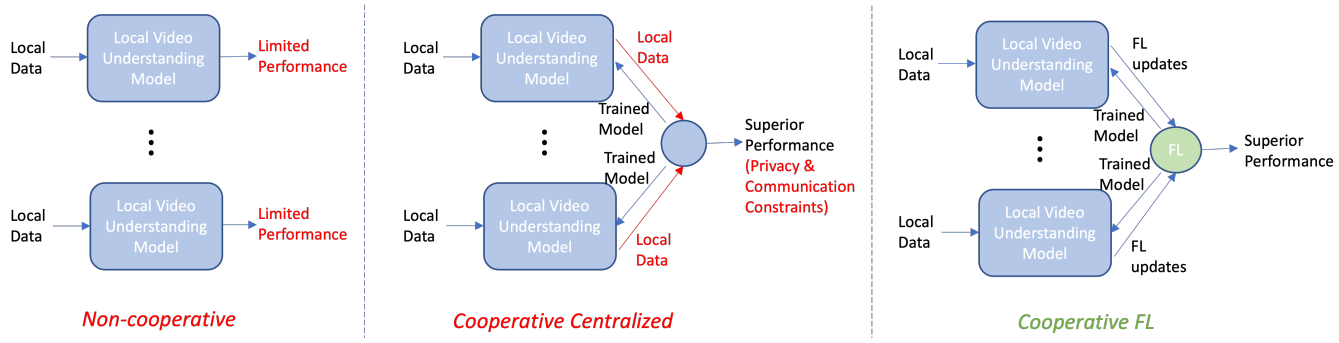

Fig. 1. Left: Non-cooperative training suffers from limited performance due to the limited representation power of local data. Middle: Cooperative centralized training suffers from privacy and communication limitations. Right: FL setup enables training on a diverse dataset while satisfying privacy and communication constraints.

I3D model. We see that the transformer model learns more semantically seperable features as compared to I3D.

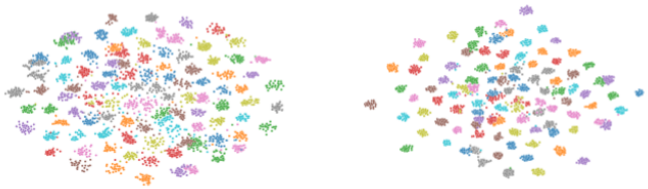

Fig. 2. t-SNE visualization of the I3D (left) and the proposed transformer (right) features extracted from the UCF-101 dataset. Each point represents a video and various classes are represented with different colors. We see that the features learned by the proposed transformer model (right) are semantically more separable than the I3D features (left). Best viewed in color.

Motivated by these observations, we propose an FL setup for video understanding consisting of a transformer architecture and extensively evaluate the proposed approach on several video understanding tasks. Our contributions can be summarized as follows:

- *A federated learning framework for privacy-preserving video understanding.* We show that the proposed approach is able to perform competitively on various video understanding tasks without sharing raw data, thus greatly reducing privacy and communication concerns.
- *A transformer based approach for extracting spatiotemporal features.* There is no existing work that uses a transformer architecture for FL-based weakly-supervised anomaly detection or zero-shot/low-shot action recognition.

II. Related Works

[8] is the first work that applies FL to a real-world image dataset, Google Landmark. FedVision [9] is an FL framework for object detection. It supports object detection models such as FastRCNN and YOLOv3. However, FedVision does not consider other video understanding tasks, such as action recognition and anomaly detection. For FL in other application domains, we refer the reader to the survey in [10].

Over the last few years, there has been a lot of research about video action recognition [11]. Zero-shot learning (ZSL) for video action recognition, on the other hand, has only lately begun to gain traction. In general, ZSL can be divided into two types: inductive [12], where the test data is unknown during training, and transductive [13], where the test data is provided without class labels during training. We study the inductive setting in this paper. On the other hand, in few-shot learning, the meta-learning paradigm which trains the model using few-shot tasks constructed from training data, has been widely used. Methods in this paradigm can be broadly classified as initialization-based or metric-based. Specifically, the approach which tackles the few-shot classification problem by comparing samples from the query set and the support set have become popular [12].

Video anomaly detection is one of the most challenging and long standing problems in computer vision [14]. Several recent works propose using a GAN for detecting anomalies in videos. For example, [15] proposes a future frame prediction network, which attempts to predict the future frame based on a sequence of input frames, and computes the prediction error in terms of the peak signal to noise ratio. Several other works follow reconstruction based approaches [14], which try to classify frames based on the reconstruction error. All these works study a fixed training setting, which is not sustainable since a fixed training set cannot possibly represent all nominal and anomalous scenarios. While there are a few works proposing continual learning approaches for video anomaly detection [16], they do not consider cooperation among multiple sources for a diverse stream of training data.

III. Proposed Approach

In this section, we present our approach for applying FL to video understanding tasks. We begin by carefully defining the problem formulation and describing various video understanding tasks. Next, we introduce the proposed FL setup, which leverages a spatiotemporal transformer to learn visual features (Fig. 3).

A. Problem Definition

The weakly-supervised video anomaly detection task can be defined as a *binary* classification problem where given a set of nominal $X^{\mathcal{N}}$ and anomalous $X^{\mathcal{A}}$ videos, we aim to classify a set of videos $X^T = \{x_1^t, \ldots, x_M^t\}$ such that $x_i^t \in \{\mathcal{N} \cup \mathcal{A}\}$,

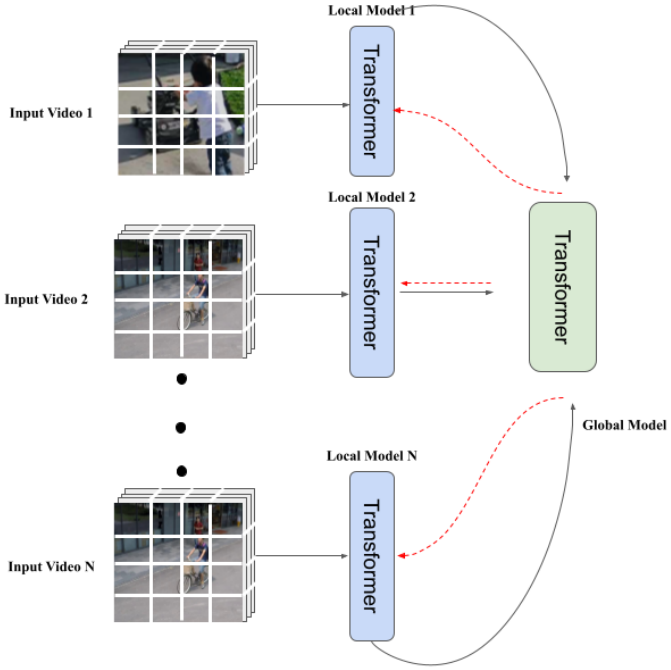

Fig. 3. Proposed FL architecture.

where M is the number of videos in the test set. On the other hand, action recognition is a *multi-class* classification problem where given a training set of videos X^s and labels S from known classes $\{(x_1^s, s_1), \ldots, (x_N^s, s_N)\}$, we aim to accurately classify a set of videos $X^u = \{x_1^u, \ldots, x_M^u\}$ from previously known/unknown classes $U = \{u_1, \ldots, u_M\}$, where N is the number of training videos. Depending on the task, several factors like the number of classes, type of labels, number of training instances per class, etc., might vary drastically. We next briefly summarize the video understanding tasks we address under an FL setup in this work.

Video Anomaly Detection: We consider the weakly-supervised formulation, which assumes the availability of video-level labels for the anomalous activities in training, i.e., a video is either nominal or contains an anomaly somewhere in the video, but no frame-level annotation is available. In the existing literature, the training data consists of videos from several cameras, which is more akin to a cooperative centralized setup. However, in a practical scenario, sharing of streaming video data constitutes a massive communication overhead and also raises serious privacy and security concerns. Hence, FL is crucial for continual cooperative training in a video anomaly detection framework.

We also consider cross-domain adaptability in video anomaly detection. Given videos from different scenes but a similar environment, it is fair to assume that the type of nominal activities remains consistent. Then, a model trained on one scene should be able to adapt to other scenes without needing any additional training. For example, in the benchmark video surveillance datasets discussed in Section IV, the same type of nominal activities are shared.

Video Action Recognition: In the ZSL paradigm, we assume that the training set consists of samples from a set of known classes S, and we aim to classify videos from previously unseen classes U, i.e., $U \not\subset S$. The challenge compared to the regular action recognition task lies in the fact that no direct mapping from the input videos to the output unseen class labels can be learned during training. Typically, semantic embeddings are used to bridge the input videos and the output unseen class labels, which consist of words. The idea behind this mainstream ZSL approach is to learn a semantic embedding model $f(x)$ for the input videos and choose the class that is semantically most similar. Recently, Brattoli et al. [17] proposed a novel training protocol which involves removing classes from the training set if

$$\min_{s_i \in S, u_j \in U} D_{cos}(\phi(s_i), \phi(u_j)) < \tau, \tag{1}$$

where D_{cos} is the cosine distance, and τ is set as 0.05. Since there is a significant overlap between Kinetics-600, commonly used training dataset, and the test classes from other datasets, we use Eq. (1) to remove the overlapping classes and have a fair setup for zero-shot learning.

We also consider the few-shot learning (FSL) setup, which is similar to the ZSL paradigm except a limited number of samples belonging to the test classes are available during training. Since it is unfeasible to collect a sufficient number of training samples for each new class, FSL serves as a potential solution, and hence is crucial for the advancement of action recognition research. In this work, we consider the standard few-shot video action classification problem definition [18]. Given a test dataset, we sample a N-way K-shot classification task to test a learned FSL action model.

B. FL Setup

In [3], FL is suggested as an alternative to centralized learning. In an FL system, a server coordinates with local nodes (clients) and sends them a global deep neural network model. The clients utilize their own data to train the model locally, then communicate it back to the server to be aggregated into a new global model. The server repeats this approach until the global model's performance on a task converges. Thus, the data at a local node is never shared with third parties, providing privacy. To optimize training loss across clients, FL algorithms try to obtain a global model. We primarily focus on the Federated Averaging (FedAvg) algorithm proposed in [3] due to its effectiveness and versatility. Despite being the first FL algorithm and having a simple procedure, FedAvg exhibits versatile convergence in various settings compared to its more complex successors [19]. Specifically, FedAvg optimizes the local training loss using Stochastic Gradient Descent (SGD). The objective for FedAvg can be expressed as $\min_w F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w)$, where $F_k(w)$ is the local loss of client k, n_k is the number of training samples on client k, with a total of n training samples partitioned across all K clients.

Feature Extractor: While any feature extraction approach, such as 3D CNNs and video transformers, will also work, we here propose using the TimeSformer architecture, which

comprises of A parallel self-attention heads in each L sequential encoding block. To classify a video, we first sample a clip y of F frames and size $H \times W$ from the input video x. We proceed by converting the entire video clip y to a sequence of 2D patches of size $P \times P$, given by $e_t^p \in \mathbb{R}^{3\times P^2}$. Here $p = 1, \ldots, N$ represents the spatial position of each patch (i.e., patch index), $t = 1, \ldots, F$ denotes the temporal location of each frame and 3 is the number of color channels. Finally, we flatten each patch e_t^p into $v_t^p \in \mathbb{R}^{3P^2}$ and linearly map it into a score vector using a trainable linear projection $E \in \mathbb{R}^{q\times 3P^2}$: $z_t^p(0) = Ev_t^p + \mu_t^p$, where $\mu_t^p \in \mathbb{R}^q$ is a latent vector learned to encode the spatiotemporal position of each (p, t) pair. Following the NLP transformer BERT [20], we add a latent vector $z_0^0(0) \in \mathbb{R}^q$ for an additional fictional patch that will be trained to represent the video's score vector through time and spatial self-attention. The input to the model is $\{z_0^0(0), z_t^p(0)\}_{p,t}$. Each block l receives $\{z_0^0(l-1), z_t^p(l-1)\}$ and produces $\{z_0^0(l), z_t^p(l)\}$. Finally, the output of the last step $z_0^0(L)$ is used as the video's visual feature representation.

Learning Model: After extracting the visual feature representations from the transformer architecture, we propose using a task specific fully connected deep neural network (DNN) $f(\cdot)$ to classify the videos. The fully connected DNN consists of three hidden layers with 600 neurons in the output layer for action recognition and a single neuron for anomaly detection.

Video Anomaly Detection: Based on the existing anomaly detection literature, we consider two popular setups, specifically weakly-supervised and cross-domain adaptability. In both setups, we are provided with a set of anomalous videos $X^{\mathcal{A}}$ from various edge devices consisting of an anomalous event but without any temporal frame-level annotation. Since supervised anomaly detection approaches require exact frame-level annotations, we leverage the Multiple Instance Learning (MIL) approach which only requires annotations at a video-level. In MIL, the learning model is given a collection of labeled bags, each of which contains a large number of instances. A bag is called negative B_n if all of the instances contained within it belong to the nominal class. On the other hand, a bag is called positive B_a if at least one of the instances contains an anomalous event. We follow the standard multiple instance ranking objective function discussed in [21]:

$$l(B_a, B_n) = \max(0, 1 - \max_{i\in B_a} f(X_i^{\mathcal{A}}) + \max_{i\in B_n} f(X_i^{\mathcal{N}})). \quad (2)$$

Since an anomalous event is generally a rare event, we enforce sparsity constraints which is given by

$$Sp(B_a) = \lambda_2 \sum_i^n f(X_i^{\mathcal{A}}). \quad (3)$$

Furthermore, due to the sequential nature of anomalies, we assume temporal smoothness in the predicted anomaly statistic. Hence, we also include a smoothness constraint which is given by

$$Sm(B_a) = \lambda_1 \sum_i^{(n-1)} (f(X_i^{\mathcal{A}}) - f(X_{i+1}^{\mathcal{A}}))^2. \quad (4)$$

Finally, the objective function is given by

$$\mathcal{L}(w) = l(B_a, B_n) + Sm(B_a) + Sp(B_a) \quad (5)$$

Following the popular approach in the literature, we divide each video into 32 overlapping segments. Each segment is then passed through the transformer model to extract visual features. Finally, the weights of the fully connected DNN are optimized to minimize Eq. (5).

Video Action Recognition: The traditional video action recognition setup is a supervised classification problem in which we train and test on a set of videos belonging to the same classes. However, in this work, we focus on the more challenging zero-shot and few-shot action recognition problems. In the zero-shot action recognition (ZSAR) setup, we are given training videos X^s belonging to a set of seen classes S, and we aim to classify videos X^u belonging to a set of unseen classes U. The challenging part in ZSAR as compared to traditional action recognition is to learn a mapping between the input videos and unseen class labels even though no overlap exists between S and U. The mainstream idea behind existing ZSAR approaches is to learn a mapping between input videos and their semantic label embeddings, and then choose the semantically closest label from the set of unseen classes using nearest neighbor approaches. Hence, after extracting visual feature representations, we train the fully connected DNN by minimizing the loss function

$$C = \|f(x_i^s) - \phi(s_i)\|^2, \quad (6)$$

where $\phi(s_i)$ is the semantic embedding of the class description s_i for the training video x_i^s from a Sent2Vec model [22]. We then find the semantically closest label by computing

$$\arg\min_j D_{cos}(f(x_i^u), \phi(u_j)), \quad (7)$$

where $\phi(u_j)$ is the Sent2Vec semantic embedding of the class label u_j

C. Implementation Details

The TimeSformer architecture is used in the proposed framework to extract spatiotemporal features [23]. We pre-train the transformer model on the Kinetics-600 datatset. The shorter side of the input video is scaled to 256 pixels and then randomly cropped to generate a 224×224 ($H \times W$) video sample for training the visual feature extractor model. The patch size is set to 16×16, resulting in a frame with $N = 196$ patches. The number of self-attention heads is $A = 12$, and the size of the score vectors ($z_t^p(l)$) at each encoding block for each patch is $q = 768$. For video action recognition, the learned semantic embedding $f(x)$ is 600 in size, as is the Sent2Vec embedding for class descriptions. The length of the input sequence to the model in all of our trials is 8 frames. The models are trained on four NVIDIA A40 GPUs with a batch size of four. We set the learning rate as 0.002 and the categorical loss function is minimized via synchronized SGD. For video anomaly detection, following the existing literature, we segment each video into 32 parts of 8 frame each. To

form the positive and negative bags, we randomly select 20 positive and negative segments to form a mini-batch. The fully connected DNN is trained to minimize the MIL loss given in Eq. (5). The model is trained using the Adagrad optimizer with a learning rate of 0.001. For the IID setup, we randomly select videos from the entire training set and assign them to each node, whereas in the non-IID setup each node is assigned a random set of classes and only observes videos belonging to those classes.

D. Computational Efficiency

The transformer model is more computationally efficient than the popular I3D model with 8 input frames, which requires 10.8 TFLOPS for inference. Whereas, the proposed transformer model only requires 0.79 TFLOPS. This shows that the proposed approach can be implemented on edge devices where computational efficiency is critical. Thanks to the scalability of the transformer model, the number of input frames per segment can be increased beyond 8 to obtain higher performance.

IV. Experiments

A. Datasets

Video Anomaly Detection: For weakly-supervised video anomaly detection, we consider the popular UCF-Crime dataset [21], which consists of a total of 1900 videos with 800 normal and 810 anomalous training videos. The testing set comprises 150 nominal and 140 anomalous videos from 13 real-world anomaly classes. The total dataset length is approximately 128 hours, with a variable number of frames per video. While the ShanghaiTech dataset is originally an unsupervised video anomaly detection dataset, it has been modified to work in a weakly-supervised setting. We use the split suggested in [24], which consists of 238 training videos and 199 testing videos. We show a few nominal and anomalous events from both datasets in Fig. 4.

Video Action Recognition: The majority of recent studies use three publicly available benchmark datasets: the UCF-101, HMDB-51, and Olympics datasets. The UCF-101 dataset contains 13,320 videos from 101 classes, with the majority of the videos concentrating on five different types of behaviors. The HMDB-51 dataset contains 6767 videos divided into 51 categories based on everyday human actions. The Olympics dataset is divided into 16 categories, each of which is tied to an Olympic sport. Brattoli et al. [17] recently tested their approach on the ActivityNet dataset by extracting tagged frames from every video. ActivityNet is significantly more thorough than the other benchmark datasets, with 27,801 videos from 200 classes relating to daily activities. With over 500K videos in 700 categories obtained from YouTube, the Kinetics-700 dataset is the largest dataset available for video action detection. We used the Kinetics-600 dataset in our experimental setup because numerous classes from Kinetics-700 were not available or had files that were corrupted. We do not use the Olympics dataset in our evaluations due to its limited size and high overlap with Kinetics. For the few-shot action recognition setup, we use the splits proposed by [25]. We evaluate the performance for action recognition tasks using the accuracy metric averaged over 10 trials.

TABLE I
Video anomaly detection comparison with the state-of-the-art methods on the UCF-Crime dataset. We show that both our federated and centralized models outperform the existing approaches.

Method	Supervision	Feature Extractor	AUC
Hasan et al. [26]	Unsup.	AE^{RGB}	50.6
Lu et al. [27]	Unsup.	Dictionary	65.51
SVM	Weak	$C3D^{RGB}$	50
Sultani et al. [21]	Weak	$C3D^{RGB}$	75.4
Zhang et al. [28]	Weak	$C3D^{RGB}$	78.7
Zhu et al. [29]	Weak	AE^{Flow}	79.0
Zhong et al. [24]	Weak	$C3D^{RGB}$	81.08
Liu et al. [30]	Full	$C3D^{RGB}$	70.1
MIST [31]	Weak	$I3D^{RGB}$	82.30
(Ours)-Federated	Cross-Domain	TimeSformer	80.2
(Ours)-Federated	Weak	TimeSformer	82.9
(Ours)-Centralized	Weak	TimeSformer	**86.3**

TABLE II
Video anomaly detection comparison with the state-of-the-art methods on the ShanghaiTech dataset. We show that both our federated and centralized models perform competitively with respect to the existing approaches.

Method	Supervision	Feature Extractor	AUC
Sultani et al. [21]	Weak	$I3D^{RGB}$	85.33
Zaheer et al. [32]	Weak	$C3D^{RGB}$	89.67
Zhong et al. [24]	Weak	$C3D^{RGB}$	76.44
Wan et al. [33]	Weak	$I3D^{RGB}$	91.24
MIST [31]	Weak	$I3D^{RGB}$	**94.83**
(Ours)-Federated	Weak	TimeSformer	90.3
(Ours)-Centralized	Weak	TimeSformer	94.4

B. Results

Video Anomaly Detection: In Table I, we compare the performance of our proposed approach to state-of-the-art algorithms in terms of the Area under the Curve (AUC) metric on the UCF-Crime (Table I) and ShanghaiTech (Table II) datasets. Here, we consider 10 local nodes and randomly divide the available training data among all the local nodes (IID setup). It is seen in Table I that the proposed approach is able to significantly outperform the rest of the existing approaches on the UCF-Crime dataset. Notably, even our global model trained using the FL setup outperforms the existing centralized approaches by a small margin. On the ShanghaiTech dataset, both the federated and centralized versions of the proposed method attain competitive performance with respect to the state-of-the-art results (Table II). These results support our hypothesis that a federated setup is capable of performing competitively while preserving privacy and minimizing the communication overhead.

Cross Domain Adaptability in Non-IID Setup: We then test the proposed approach's cross-domain scene adaption capabilities and see how well it generalizes to other situations. In this scenario, we assume a non-IID setup in which each local

Fig. 4. Examples of nominal and anomalous frames in the UCF-Crime and ShanghaiTech datasets. Anomalous events are shown with red box.

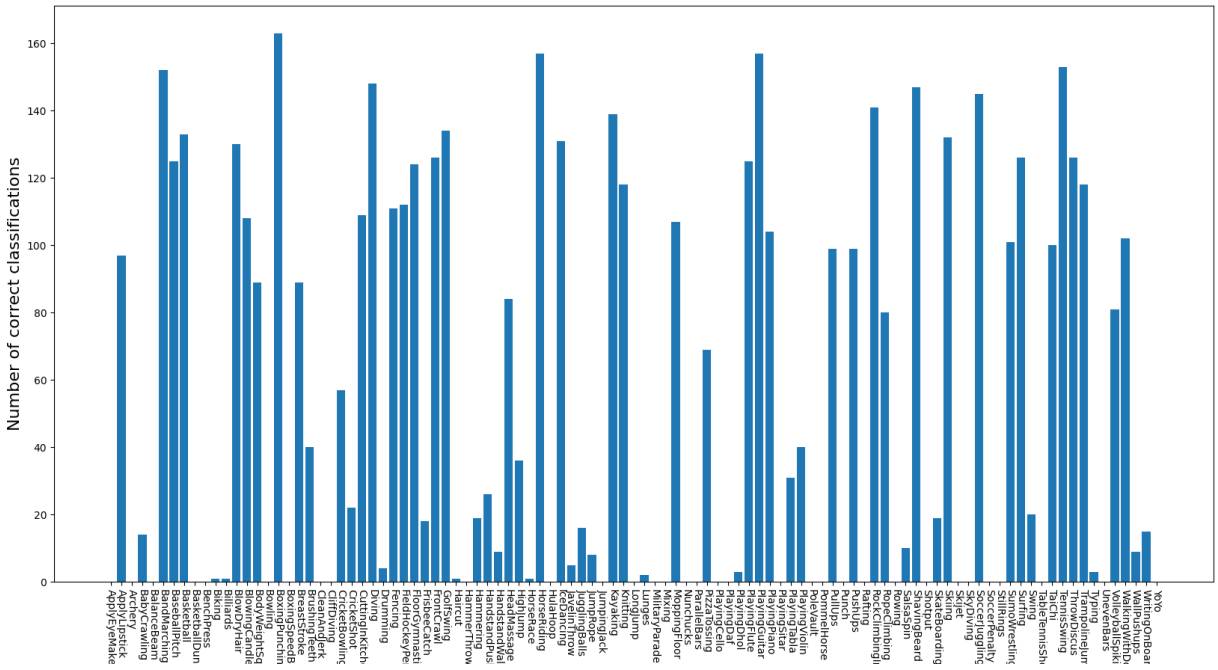

Fig. 5. Performance of the proposed FL approach on all classes of the UCF-101 dataset.

node observes a separate camera scene from the ShanghaiTech dataset for training. Then, we test the FL-trained model on test videos from the UCF-Crime dataset. Cross-domain scene adaptation is mostly unexplored, with only [34] discussing a similar few-shot adaptation approach to our knowledge. However, the proposed strategy outlined in [34] necessitates multiple anomaly-free video frames for their model to adjust to the new scenario, which may not always be possible. The cross-domain result presented in Table I is close to the IID FL result trained on the UCF-Crime dataset, demonstrating the cross-domain adaptation capability of our approach.

Video Action Recognition: In Table III, we compare the proposed method to state-of-the-art ZSL approaches using the commonly used random split setup. Specifically, we divide the test datasets in half at random using the seed described in [17]. Here, we consider 10 local nodes and split the dataset in such a way that each local node receives distinct classes, i.e, there is no overlap among the classes at each node. Under all conditions, the suggested spatiotemporal transformer-based ZSL technique consistently outperforms all other state-of-the-art approaches. Furthermore, our decentralized federated setup performs competitively and still outperforms all existing approaches. On the UCF-101 dataset, our FL approach shows an improvement of 22% over the next best result, which is notably trained in the centralized setup. In Fig. 5, we show the average performance of the global federated model on the UCF-101 dataset for each class.

In Table IV, we compare the proposed method to state-of-the-art approaches for the few-shot action recognition setup. Particularly, we use the 5-way 5-shot split proposed in [25]. For the FL setup, we use 5 local nodes and randomly distribute the training data among them. We see that in contrast to the anomaly detection and zero-shot setups, the proposed approach performs poorly. This can be attributed to the few-samples being distributed among multiple local nodes, thus diminishing their advantage.

V. Ablation Study

Impact of Number of Local Nodes: In Fig. 6, we show the ZSL performance of the global model trained using the

TABLE III
ZERO-SHOT COMPARISON WITH STATE-OF-THE-ART METHODS ON BENCHMARK DATASETS. METHODS ARE EVALUATED BY RANDOMLY SPLITTING DATASET IN HALF AND AVERAGING RESULTS OVER 10 TRIALS.

Method	UCF	HMDB	ActivityNet
DataAug [35]	18.3	19.7	-
InfDem [36]	17.8	21.3	-
Bidirectional [37]	21.4	18.9	-
TARN [38]	19	19.5	-
Action2Vec [12]	22.1	23.5	-
OD [39]	26.9	30.2	-
GGM [39]	20.3	20.7	-
(Ours)-Federated	48.9	31.3	37.3
(Ours)-Centralized	**51.7**	**33.2**	**41.7**

TABLE IV
5-WAY 5-SHOT COMPARISON WITH STATE-OF-THE-ART METHODS ON BENCHMARK DATASETS. METHODS ARE EVALUATED BY RANDOMLY SPLITTING DATASET IN HALF AND AVERAGING RESULTS OVER 10 TRIALS.

Method	UCF	HMDB
GenApp [40]	78.6	52.5
ProtoGAN [41]	80.2	54.0
ARN [25]	83.1	60.6
TRX [42]	**96.1**	**75.6**
(Ours)-Federated	67.5	52.7
(Ours)-Centralized	89.3	69.2

federated setup on the UCF-101 and HMDB-51 datatsets. We notice an increase in the classification accuracy as we increase the number of local nodes contributing to the model update, which shows the efficacy of our proposed approach.

Impact of the Spatiotemporal Transformer: We investigate how well the suggested transformer model can learn spatiotemporal visual features. For comparison, we consider the state-of-the-art I3D 3D-CNN model, which has been a popular choice for video action recognition and has been employed by numerous previous methods for zero-shot action recognition [39], [43]. To extract I3D features, we use the same method as [39], [43] and average the output from the *Mixed 5c* layer throughout the temporal dimension, then pool it by four in the spatial dimension, and lastly flatten it to a 4096-dimensional vector. We concatenate both RGB and flow features to create an 8192-pixel vector. The t-SNE visualizations for the I3D and transformer features on UCF-101 are shown in Fig. 2, where each point represents a video in the UCF dataset. We also use other statistical criteria to compare them quantitatively in Table V. The average silhouette score indicates how closely all of the points in the cluster are related. The adjusted rand index determines how similar the clusterings are to the ground truth. The homogeneity score determines if a cluster comprises samples from the same class. Finally, we compute the accuracy for traditional video classification using a simple k-NN classification algorithm on the retrieved features. We observe that the transformer features create a better separation in all measurements.

VI. CONCLUSION

We proposed a privacy-preserving FL setup for weakly-supervised anomaly detection, zero-shot and few-shot action recognition. The proposed methods for all considered tasks utilize a spatiotemporal transformer for extracting visual features from videos and a fully connected DNN for decision making. The anomaly detection and action recognition methods differ in their loss functions to train the fully connected DNN. Using the FedAvg algorithm we demonstrated the proposed FL framework for the considered video understanding tasks. The experimental results showed that the proposed transformer based centralized method achieves superior performance compared to several state-of-the-art methods in both anomaly detection and zero-shot action recognition. It achieved competitive performance in the few-shot action recognition problem. The decentralized version trained via FedAvg achieved a close performance (within 4%) with respect to the centralized version in all tasks except few-shot action recognition. The large performance drop in the few-shot task can be attributed to the distribution of few samples among multiple nodes, i.e., instead of 5-shot each of 5 nodes trains on a single shot. The FL results staying close to the centralized results showed that the proposed approach can be effectively trained in a privacy-preserving and communication-efficient way.

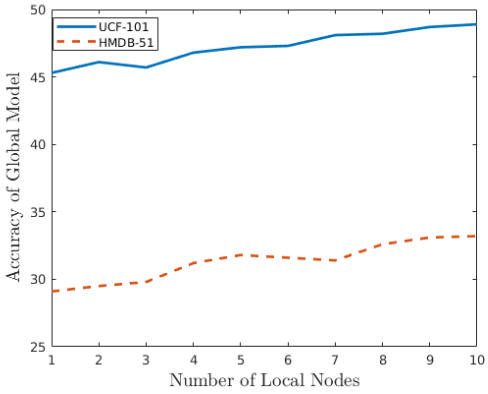

Fig. 6. ZSL performance of the proposed FL approach as a function of the number of local nodes.

TABLE V
COMPARISON BETWEEN TIMESFORMER AND I3D FEATURES IN TERMS OF CLUSTERING AND CLASSIFICATION PERFORMANCE USING DIFFERENT METRICS. CLASSIFICATION ACCURACY IS FOR THE k-NN CLASSIFIER.

Method	I3D	TimeSformer
Average Silhouette	0.119	**0.196**
Adjusted Rand Index	0.80	**0.88**
Homogeneity Score	0.92	**0.96**
Classification	0.93	**0.96**

REFERENCES

[1] Sanchit Alekh, "Eu general data protection regulation: A gentle introduction," *arXiv preprint arXiv:1806.03253*, 2018.

[2] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[4] Chaoyang He et al., "Fedcv: A federated learning framework for diverse computer vision tasks," *arXiv preprint arXiv:2111.11066*, 2021.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[6] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[7] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe, "Vidtr: Video transformer without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13577–13587.

[8] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown, "Federated Visual Classification with Real-World Data Distribution," *arXiv e-prints*, p. arXiv:2003.08082, Mar. 2020.

[9] Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang, "Fedvision: An online visual object detection platform powered by federated learning.," in *AAAI*, 2020, pp. 13172–13179.

[10] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning, Vol 14, Issue 1–2*, 2021.

[11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[12] Meera Hahn, Andrew Silva, and James M Rehg, "Action2vec: A crossmodal embedding approach to action learning," *arXiv preprint arXiv:1901.00484*, 2019.

[13] Xun Xu, Timothy Hospedales, and Shaogang Gong, "Transductive zero-shot action recognition by word-vector embedding," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 309–333, 2017.

[14] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.

[15] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua 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.

[16] Keval Doshi and Yasin Yilmaz, "Rethinking video anomaly detection-a continual learning approach," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3961–3970.

[17] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4613–4623.

[18] Linchao Zhu and Yi Yang, "Compound memory networks for few-shot video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 751–766.

[19] Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang, "On the unreasonable effectiveness of federated averaging with heterogeneous data," *arXiv preprint arXiv:2206.04723*, 2022.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[22] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," in *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

[23] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," *arXiv preprint arXiv:2102.05095*, 2021.

[24] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li, "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*, 2019, pp. 1237–1246.

[25] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz, "Few-shot action recognition with permutation-invariant attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 525–542.

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

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

[28] Jiangong Zhang, Laiyun Qing, and Jun Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4030–4034.

[29] Yi Zhu and Shawn Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.

[30] Kun Liu and Huadong Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1490–1499.

[31] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14009–14018.

[32] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee, "Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 358–376.

[33] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[34] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang, "Few-shot scene-adaptive anomaly detection," *arXiv preprint arXiv:2007.07843*, 2020.

[35] Xun Xu, Timothy M Hospedales, and Shaogang Gong, "Multi-task zero-shot action recognition with prioritised data augmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 343–359.

[36] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen, "Informed democracy: voting-based novelty detection for action recognition," *arXiv preprint arXiv:1810.12819*, 2018.

[37] Qian Wang and Ke Chen, "Zero-shot visual recognition via bidirectional latent embedding," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 356–383, 2017.

[38] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras, "Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition," *arXiv preprint arXiv:1907.09021*, 2019.

[39] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9985–9993.

[40] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal, "A generative approach to zero-shot and few-shot action recognition," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 372–380.

[41] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain, "Protogan: Towards few shot learning for action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[42] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen, "Temporal-relational crosstransformers for few-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 475–484.

[43] Shreyank N Gowda, Laura Sevilla-Lara, Frank Keller, and Marcus Rohrbach, "Claster: Clustering with reinforcement learning for zero-shot action recognition," *arXiv preprint arXiv:2101.07042*, 2021.