FedVAD: Enhancing Federated Video Anomaly Detection with GPT-Driven Semantic Distillation

Fan Qi[1], Ruijie Pan[1], Huaiwen Zhang[2], and Changsheng Xu[3⋆]

[1] Tianjin University of Technology, Tianjin, China
[2] College of Computer Science, Inner Mongolia University, China
[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
fanqi@email.tjut.edu.cn; panruijie@stud.tjut.edu.cn;
huaiwen.zhang@imu.edu.cn; csxu@nlpr.ia.ac.cn;

Abstract. The imperative for smart surveillance systems to robustly detect anomalies poses a unique challenge given the sensitivity of visual data and privacy concerns. We propose a novel Federated Learning framework for Video Anomaly Detection that operates under the constraints of data heterogeneity and privacy preservation. We utilize Federated Visual Consistency Clustering to group clients on the server side. Further innovation is realized with an Adaptive Semantic-Enhanced Distillation strategy that infuses public video knowledge into our framework. During this process, Large Language Models are utilized for semantic generation and calibration of public videos. These video-text pairs are then used to fine-tune a multimodal network, which serves as a teacher in updating the global model. This approach not only refines video representations but also increases sensitivity to anomalous events. Our extensive evaluations showcase FedVAD's proficiency in boosting unsupervised and weakly supervised anomaly detection, rivaling centralized training paradigms while preserving privacy. The code will be made available publicly at https://github.com/Eurekaer/FedVAD.

Keywords: Video Anomaly Detection · Federated Learning · Large Language Model

1 Introduction

Advances in Deep Learning for computer vision have significantly enhanced smart surveillance systems, which are widely used now for detecting abnormal activities in various environments. At the heart of this advancement is Video Anomaly Detection (VAD), which aims to identify events in video data that are out of the ordinary, helping these systems quickly spot potential issues.

However, previous data-driven VAD [1,23,28,33,49] raise serious privacy concerns that impede public acceptance and real-world deployment due to the inherent sensitivity of the visual information such as human faces, objects, identities,

⋆ Corresponding author

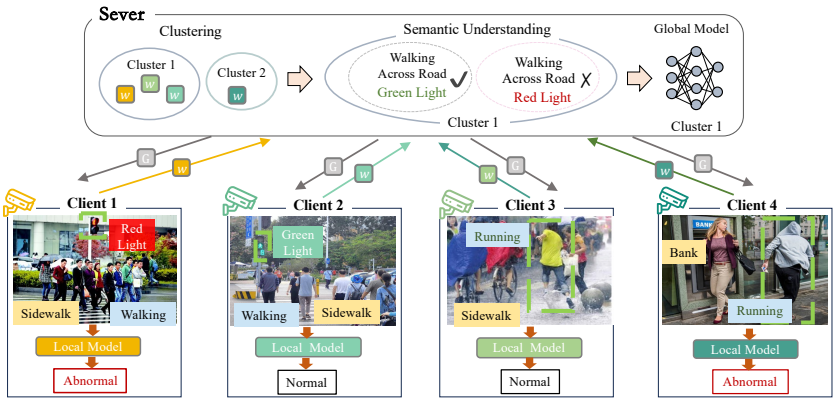


Fig. 1: The pipelines for Federated Video Action recognition and Anomaly Detection.

and activities. Developing privacy-preserving video anomaly detection frameworks is thus imperative for enabling ethical and socially acceptable applications. Federated Learning (FL), as introduced by McMahan et al. [24], enables the collaborative training of global models across multiple clients without the need to share raw data, offering a degree of user privacy preservation. This approach, increasingly applied in visual tasks recently [4, 15, 17, 25], navigates the challenges posed by diverse data distributions across clients. Since each client's dataset may not reflect the overall data distribution, local model updates can inadvertently degrade the global model's performance. A primary hurdle in this domain is *data heterogeneity* [20], which significantly impacts model effectiveness.

Conventional FL depend on the upload of either local model weights [24, 40, 48] or prototypes [14, 38, 47] to facilitate knowledge clustering [10, 35] and sharing among participating clients. In the real-world VAD, mobile surveillance devices are ubiquitous throughout every corner of cities, and background differences are the primary cause of data heterogeneity. Simple weighted averaging for model aggregation, as utilized in action recognition tasks [7, 54], fails to address the nuanced scene-specific contexts that are critical for accurate abnormal event detection across various clients. Moreover, centralizing local prototypes fails to rectify these contextual inadequacies. For example, as shown in Fig.1, the captured footage shows jaywalking scenes (Client 1 vs Client 2), where pedestrians crossing at a red light are considered anomalous events; Similarly, two people suddenly running may appear the same, but could represent completely different activities - one person snatching a purse (Client 3) versus hurriedly taking shelter from the rain (Client 4). *The cooperative analysis of both foreground and background elements is essential for accurately interpreting anomalies within a visual scene [23, 37, 53].*

In practical scenarios, the prevalence of normal events in contrast to the rarity of anomalous ones results in substantial labeling costs. This imbalance severely limits the efficacy of unsupervised anomaly detection techniques. Furthermore,

the dearth of anomaly samples significantly impedes the development of robust inference mechanisms in anomaly detection models, presenting a critical challenge in this domain. *Leveraging high-level semantic information, as suggested by the previous works [12, 30, 34], can yield substantial benefits in vision-related tasks, particularly when server infrastructure supports the utilization of large pre-trained models.*

To address these challenges, we propose FedVAD, which empowers surveillance devices with enhanced capabilities and security. For foreground-background cooperative analysis, we design a **Federated Visual Consistency Clustering** by examining similarities in motion model weightings and background prototypes, inherently upholding privacy. For anomaly detection capabilities enrichment, our server-side architecture incorporates a Large Language Models (LLMs) to enhance **Adaptive Semantic-Enhanced Distillation** paradigm. This innovative approach augments public datasets—originally comprising only video content and associated labels—with expansive descriptive narratives generated via Blip-2 [16]. To ensure these narratives precisely reflect the semantic content of the videos, they are refined by the linguistic prowess of GPT-4 [27], allowing for a nuanced differentiation between normalcy and anomaly. Building upon this enriched semantic foundation, we fine-tune a multimodal network equipped with pre-trained visual and textual models. This model inputs the newly generated public video-text pairs, setting the stage for a more nuanced and client-specific anomaly detection capability. This step tailors the anomaly detection to the distinct characteristics of specific client clusters. Employing knowledge distillation, the multimodal network plays the role of a **teacher**, imparting its learned subtleties to the cluster-specific global model. This pedagogical exchange not only consolidates the model's understanding but also sharpens the precision of the anomaly detection mechanism.

Our FedVAD is specifically designed to handle the complexities inherent in varying surveillance scenarios. This paper makes the following significant contributions:

- We propose the first federated learning approach for Video Anomaly Detection (VAD), tackling the heterogeneity challenge across different surveillance views and adapting to the unique data distributions of each client.
- We introduce a Federated Visual Consistency Clustering strategy tailored for distributed video anomaly detection. By harnessing the inferential power of large language models, we enhance the visual representations in our global clustering model, boosting the generalization capabilities of local models.
- We conduct a comprehensive benchmark analysis of our federated approach across multiple VAD datasets, including ShanghaiTech [22], UBnormal [2], UCF-Crime [36], and XD-Violence [31], under both unsupervised and weakly-supervised learning paradigms. To be noted, we also compare our method with **GPT-4V [45]**. The experimental results indicate that our framework represents the current state-of-the-art (SOTA) best suited for federated VAD tasks.

2 Related Work

2.1 Video Anomaly Detection

Video Anomaly Detection is a machine learning paradigm to autonomously detect and analyze anomalous behavior in surveillance footage by scrutinizing frame-level pixel deviations from learned activity patterns. Given the rarity of anomalous instances in real-world datasets, video anomaly detection is predominantly characterized by two settings: *Unsupervised video anomaly detection* constructs a normative model from surveillance footage, detecting irregularities by assessing divergences from established behavioral patterns [1, 11, 28, 46]. Unsupervised VAD typically looks for statistical outliers or relies on clustering, and might fail to detect anomalies that require more nuanced, context-aware analysis. Anomalies are often context-dependent and may require understanding of complex patterns and relationships within the data. The *Weakly-supervised paradigm* utilizes training samples with video-level annotations to effectively differentiate normal from abnormal events, facilitating refined anomaly detection in surveillance data with limited labeling [23, 33, 49, 50, 55]. Weakly supervised VAD is hindered by its reliance on domain-specific knowledge, which can restrict the model's interpretability and adaptability when weak labels are ambiguous. Additionally, these models often struggle to generalize to new scenarios due to their training on a limited scope of anomaly patterns, leading to decreased performance in dynamic or unfamiliar contexts. Sato et al. [34] recently develop a framework that employs textual embeddings guided by user input to detect anomalies not represented in the training distribution. Drawing inspiration from their approach, we integrate the sophisticated inferential capabilities of state-of-the-art language models to refine visual feature extraction with contextual semantics, thereby enhancing our system's ability to identify and characterize unusual behaviors in video data with increased accuracy.

2.2 Federated Learning for Video Surveillance

Federated learning has emerged as a pivotal solution in computer vision for addressing privacy concerns, enabling on-device model training, and minimizing the transmission of sensitive data [32, 44]. This paradigm aligns with the trend of balancing computational workloads between cloud services and edge devices, especially in video surveillance applications [42]. Zhao et al. [54] introduced a semi-supervised FL framework that combines unsupervised representation learning on edge devices with supervised learning on the cloud server, enhancing activity recognition. Doshi [7] further demonstrated FL's adaptability and privacy-centric advantages in driver action recognition using a decentralized FedAvg-based model [24]. Recently, Gan et al. [9] introduced a cloud-device collaborative adaptation mechanism for continuous object detection, employing a ResNet-101 model on the cloud for knowledge distillation into a lightweight on-device ResNet-18 model, thus maintaining efficiency across environmental changes.

Building on these advancements, we incorporate a public video dataset augmented with detailed captions, thereby significantly enhancing the videos' semantic context. Furthermore, we exploit the superior reasoning capabilities of Large Language Models for the refinement and optimization of these captions. This method effectively harnesses server computational resources, marking a significant step forward in Federated Learning applications for video anomaly detection.

3 Method

3.1 Overall framework

In federated video anomaly detection, our primary objective is to enhance the client models' performance. This enhancement is achieved by adapting parameters received from the server, while also ensuring data privacy. The framework supports multiple clients ($i \in \{1, 2, \ldots, K\}$), where each client maintains its own private training dataset D_i^{train}, testing dataset D_i^{test}, a local visual model ω_i and a background object detector. Each client uploads their visual model parameters W_i and the object distribution vector O_i to the server.

The server's role is to cluster by assessing similarities between client models and to aggregate these into M model groups. The overarching goal is to refine the performance of these aggregated models through the optimization of a global objective function F, defined as:

$$\underset{\{\Omega_m\}}{\text{minimize}} \frac{1}{m} \sum_{k=1}^{K} \sum_{i=1}^{m} r_{i,k} \mathcal{L}(\Omega_m, D_i) \tag{1}$$

Here, Ω_m represents the centroid of cluster m.

The process involves three key steps: Firstly, we update the cluster assignment $r_{i,k}$ with a fixed W_i and O_i. Secondly, we update the cluster center $\tilde{W}^{(m)}$, $\tilde{O}^{(m)}$ based on W_i, O_i and $r_{i,k}$. (The specifics of these two steps are further elucidated in section 3.2.) Thirdly, we conceptualize $\tilde{W}^{(m)}$ as a student network S_m and develop an adaptive semantic-enhanced distillation strategy for its update. This innovative strategy aims to refine $\tilde{W}^{(m)}$ by incorporating additional public visual and semantic information, significantly improving the learning efficiency and effectiveness of S_m. Finally, the updated student network, now represented as $\hat{W}^{(m)}$, is designated as the new global model for cluster m. This global model is then distributed to each client within the cluster. Upon receiving $\hat{W}^{(m)}$, each client updates its local model. This update is followed by a fine-tuning phase, where the local model's parameters W_i are adjusted based on the client's private training data. This approach ensures that each local model is not only aligned with the global model but also tailored to the specific data and characteristics of each client, thereby optimizing performance while maintaining privacy. (The intricate details of these last two steps are illustrated in section 3.3.) The algorithmic process is described in the supplementary material.

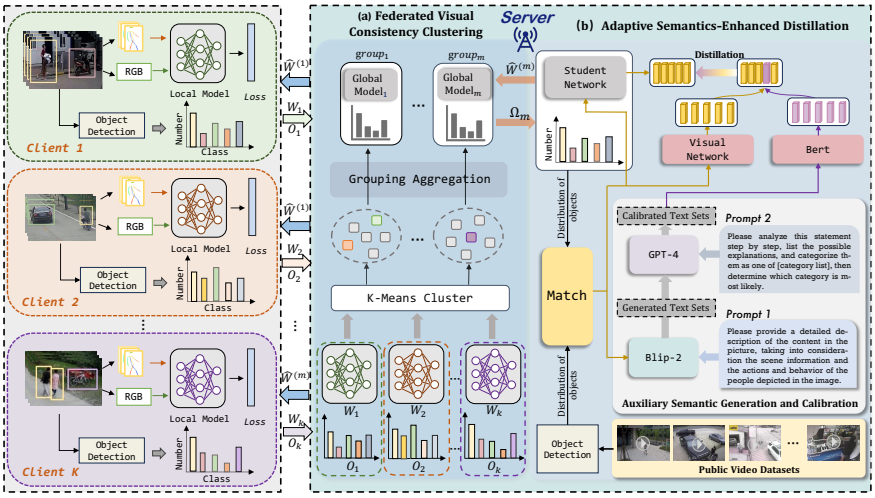


Fig. 2: The whole framework. **On the Client-side**: Local Training and Testing, we select the model parameters W_i and object target distribution O_i to send to the server. **On the Server-side**: **a) Federated Visual Consistency Clustering**, we update the cluster center $\tilde{W}^{(m)}$, $\tilde{O}^{(m)}$ based on W_i, O_i generate M groups Ω_m. **b) Adaptive Semantic-Enhanced Distillation**, responsible for updating $\hat{W}^{(m)}$ in each group. The videos with more similar object distributions in the public dataset D_p are matched by $\tilde{O}^{(m)}$, then these videos are semantically enhanced by Auxiliary Semantic Generation and Calibration module, and finally distill the knowledge to the student network S_m in each group by a multimodal network with the pre-trained visual model Φ and text model Ψ.

3.2 Federated Visual Consistency Clustering

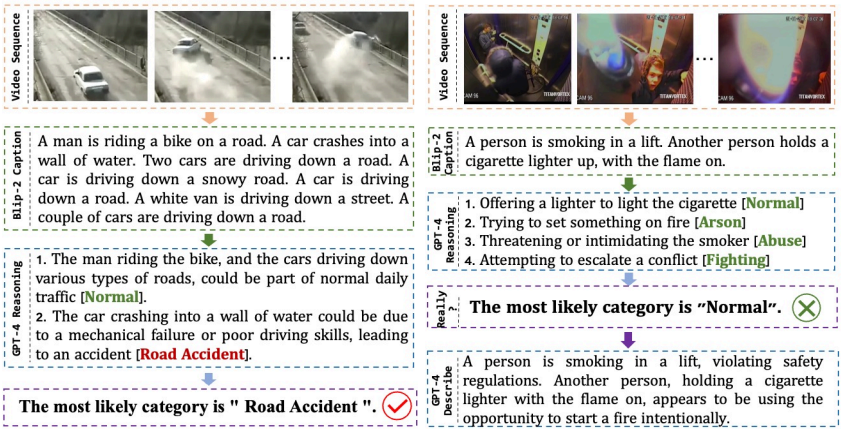
We assume $\{W_i\}_{i=1}^{K}$ to be the collection of model weights gathered from clients, and $\{O_i\}_{i=1}^{K}$ to be the corresponding set of object distributions identified from target detection. The method unfolds in two critical steps:

1. Cluster Assignment Update: Each client's data is assigned to a cluster based on the proximity of its model weights and object distributions to the existing cluster centers. This assignment is mathematically articulated as follows:

$$r_{i,k} = \begin{cases} 1, & \text{if } k = \arg\min_j \big(\alpha \cdot \mathrm{Dist}(W_i, \tilde{W}^{(j)}) \\ & +\beta \cdot \mathrm{Dist}(O_i, \tilde{O}^{(j)})\big) \\ 0, & otherwise. \end{cases} \tag{2}$$

Here, α and β are weighting factors that balance the influence of model weights and object distributions in determining cluster membership.

2. Updating Cluster Centers: We update $\tilde{W}^{(m)}$ and $\tilde{O}^{(m)}$ in each clustering center based on W_i, O_i and $r_{i,k}$ to ensure that only the data points that are the most relevant to the current clustering center will have an impact on the



(a) Correct inference **(b)** Incorrectly inferred

Fig. 3: Auxiliary Semantic Generation and Calibration

center update.

$$\tilde{W}^{(m)} = \frac{\sum_{i=1}^{M} r_{i,k} W_i}{\sum_{i=1}^{M} r_{i,k}}, \quad \tilde{O}^{(m)} = \frac{\sum_{i=1}^{M} r_{i,k} O_i}{\sum_{i=1}^{M} r_{i,k}}. \tag{3}$$

These steps are pivotal in ensuring that the clusters accurately reflect the underlying patterns in the data, taking into account both foreground vision and background object, thereby enhancing the clustering accuracy in a privacy-preserving manner.

3.3 Adaptive Semantic-Enhanced Distillation

We aim to enhance the model parameters $\tilde{W}^{(m)}$ within each group Ω_m with the objective of strengthening their generalization capabilities upon redistribution to the clients. Here, we introduce an Adaptive Semantic-Enhanced Distillation process, wherein the core lies in leveraging video data from a public dataset D_p, initially containing only videos. Our process begins by selecting a subset of videos from D_p that exhibit object distributions similar to $\tilde{O}^{(m)}$. This step is crucial for ensuring that the semantic information closely aligns with the specific characteristics and distribution of the target object, thereby facilitating a more targeted and effective optimization of the model parameters in subsequent stages. Next, we utilize LLMs for Auxiliary Semantic Generation and Calibration to generate corresponding textual descriptions for these videos. This integration of video-text pairs is crucial for the meticulous optimization of $\tilde{W}^{(m)}$ through knowledge distillation. Such a strategy is instrumental in significantly elevating the model's adaptability and performance, particularly when it is deployed across the diverse data environments of various clients.

Auxiliary Semantic Generation and Calibration. The robust development of LLMs inspired us to explore the importance of text for the task of video anomaly detection. This involves two key issues: (1) caption generation and (2) semantic calibration. To aptly represent video sequences and capture the continuity of anomalous events, we employ a strategy of extracting frames at regular intervals. Each frame is processed using Blip-2 with a specific prompt, enabling the generation of textual descriptions for individual frames. These descriptions are then aggregated to form a comprehensive narrative of the entire video. We utilize GPT-4 to calibrate these textual descriptions. GPT-4 classifies them into various anomaly categories, as detailed in Fig.3. Additionally, for descriptions with inference errors, we re-engage GPT-4 using category-specific prompts. This iterative process, involving the re-description of these segments without semantic alteration, refines and calibrates the text for inclusion in the public dataset.

Adaptive Distillation Unit. Once we have the video-text pairs, for model parameters $\tilde{W}^{(m)}$ in each group Ω_m, we conceptualize $\tilde{W}^{(m)}$ as a student network S_m. The corresponding video index i in the public dataset is subsequently found, and the video text pairs $\langle V_{i^*}, T_{j^*} \rangle$ adapted to each group are composed. Finally, the features are extracted and fused using the pre-trained visual models Φ and text models Ψ. Then the features extracted from the student network S_m with V_i as input are adaptively optimized by feature knowledge distillation.

$$Z_m = \mathrm{MHSA}(\mathrm{LN}([\Phi(V_i); \Psi(T_j)])), \tag{4}$$

$$L_{\text{distill}} = \frac{1}{N} \sum_{i=1}^{N} \|S_m - Z_m\|^2, \tag{5}$$

where $\mathrm{MHSA}(\cdot)$ denotes the multi-head self-attention layer, and $\mathrm{LN}(\cdot)$ denotes the layer normalization. $[;]$ denotes the concatenation.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method using four datasets, i.e., ShanghaiTech [22], UBnormal [2], UCF-Crime [36], and XD-Violence [31]. For convenience, we refer to these datasets as 'ST', 'UB', 'UCF', and 'XD' respectively. Our evaluation encompasses two different settings: unsupervised and weakly supervised video anomaly detection. Following Liu et al. [21], in unsupervised video anomaly detection, training data includes only normal videos, while test data comprises both normal and anomalous videos. In the weakly supervised setting, both training and test data contain a mix of normal and anomalous videos.

Data Splitting. *Unsupervised.* ShanghaiTech, sourced from 13 campus cameras, and UBnormal, a synthetic benchmark, are partitioned into 7 and 6 clients to ensure data balance and avoid data scarcity, respectively, for heterogeneous data distribution. This partitioning emphasizes data diversity by focusing on

Table 1: Quantitative comparison of different FL methods in unsupervised and weakly supervised settings. Ub indicates the upper bound of FL algorithms, while ↑ and ↓ represent increases and decreases, respectively, compared with FedAvg. Detailed analyses are offered in Sec. 4.2.

Method	Unsupervised		Weakly Supervised		
	ShanghaiTech	UBnormal	ShanghaiTech	UCF-Crime	XD-Violence
Centralized(Ub)	85.90(4.05) ↑	71.80(3.83) ↑	98.14(12.43) ↑	86.76(3.95) ↑	85.59(21.16) ↑
FedAvg [24]	81.65(0.00)	67.27(0.00)	89.93(0.00)	82.81(0.00)	64.43(0.00)
FedProx [18]	81.46(0.19) ↓	68.08(0.11) ↑	90.91(0.98) ↑	80.39(2.42) ↓	64.47(0.04) ↑
FedBN [19]	81.87(0.22) ↑	68.02(0.75) ↑	94.13(4.20) ↑	82.86(0.05) ↑	63.93(0.50) ↓
FedRep [5]	81.92(0.27) ↑	67.95(0.68) ↑	94.62(4.69) ↑	82.92(0.11) ↑	64.53(0.10) ↑
FedALA [51]	81.98(0.33) ↑	68.04(0.77) ↑	95.40(5.47) ↑	82.97(0.16) ↑	64.66(0.23) ↑
FedPAC [43]	82.12(0.47) ↑	68.24(0.97) ↑	95.58(5.65) ↑	83.34(0.53) ↑	64.98(0.55) ↑
FedAvg SGD [26]	81.95(0.30) ↑	67.98(0.21) ↑	95.18(5.25) ↑	83.02(0.21) ↑	64.65(0.22) ↑
FedAdam SGD [26]	82.13(0.48) ↑	68.17(0.90) ↑	95.62(5.69) ↑	83.35(0.54) ↑	64.72(0.29) ↑
FPS [3]	82.21(0.56) ↑	68.34(1.07) ↑	95.57(5.64) ↑	83.43(0.62) ↑	64.87(0.44) ↑
FedFTG [52]	82.36(0.71) ↑	68.42(1.15) ↑	95.96(6.03) ↑	83.55(0.76) ↑	65.32(0.89) ↑
Ours	**84.07(2.42)** ↑	**69.83(2.56)** ↑	**97.63(7.70)** ↑	**85.13(2.32)** ↑	**66.78(2.35)** ↑
Ours with DP	**83.42(1.77)** ↑	**69.36(2.09)** ↑	**97.07(7.14)** ↑	**84.68(1.87)** ↑	**66.21(1.78)** ↑

unique perspectives or scenarios. *Weakly Supervised.* UCF-Crime and XD-Violence, important benchmarks in weakly-supervised anomaly detection, are partitioned into 13 and 6 subsets, respectively. This partitioning aligns with their anomaly categories and ensures a balanced distribution of normal and anomalous data. The ShanghaiTech dataset is similarly partitioned as in the unsupervised setting.

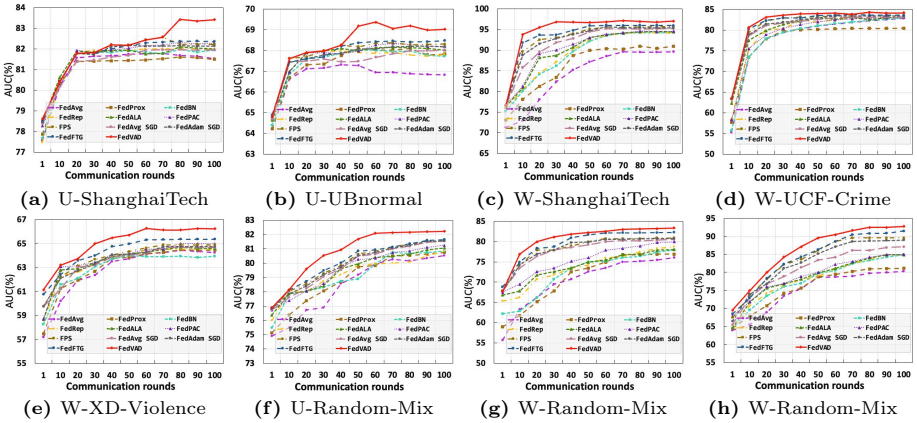
Evaluation Metrics. Similar to past works [23,28,33,50], we use the frame-level area under the curve (AUC) scores as an evaluation metric for ShanghaiTech, UBnormal, and UCF-Crime. For XD-Violence, following [49, 55], we use the AUC of the frame-level precision-recall curve (AP) as the criterion. In order to more objectively grasp the overall effectiveness of the method and its adaptability in different application contexts, we combine the AUC scores or AP scores on each client and calculate the average value. This evaluation strategy not only reveals to us the overall strength and consistency of the method but also confirms its strong generalization ability over multiple client data distributions.

Implementation Details. To ensure that the client is lightweight, we use the lightweight target detector yolov8[4] and pose extractor [8]. Considering the different settings, we use a lightweight model structure on the client side relative to the server side, while on the server side, we use pre-trained visual models [13,29], and the text model Bert [6]. Furthermore, to mitigate the risk of privacy breaches during the communication process of FedVAD, we apply Differential Privacy (DP) [41], thereby preventing potential privacy leakage during communication exchanges. We train for 100 federated rounds to ensure that the model converges steadily, and the local update epoch is set to 1. More implementation details are provided in supplementary material.

[4] https://docs.ultralytics.com/

Table 2: Clients quantitative comparison of different FL methods in unsupervised and weakly supervised settings. C* indicates the number of the client.

Method	ShanghaiTech(Unsupervised)								ShanghaiTech(Weakly Supervised)							
	C1	C2	C3	C4	C5	C6	C7	Avg	C1	C2	C3	C4	C5	C6	C7	Avg
FedAvg	84.33	84.00	82.54	83.61	79.33	76.15	81.56	81.65	91.92	87.76	88.92	91.98	89.93	90.94	88.06	89.93
FedProx	84.63	83.66	83.07	82.48	80.54	77.71	78.06	81.46	91.99	88.34	89.45	92.00	91.81	91.94	90.81	90.91
FedBN	84.84	84.69	81.80	84.71	80.45	77.01	79.64	81.87	95.85	90.15	92.00	97.74	97.96	92.99	92.23	94.13
FedRep	85.72	84.30	83.11	82.33	79.89	78.36	79.68	81.92	96.04	90.17	90.04	97.94	97.89	97.66	92.57	94.62
FedALA	85.87	85.02	83.72	83.52	81.41	76.37	77.91	81.98	96.67	90.00	90.45	98.00	97.98	97.85	96.81	95.40
FedPAC	85.90	85.13	85.14	82.82	80.93	76.19	78.80	82.12	96.48	90.92	93.71	98.00	97.69	97.34	95.27	95.58
FedAvg SGD	83.44	82.81	82.42	81.94	80.32	80.81	81.90	81.95	94.01	86.94	91.87	99.31	98.74	98.12	96.50	95.18
FedAdam SGD	84.71	84.36	84.96	82.36	79.07	77.94	81.52	82.13	94.70	87.67	94.78	98.98	98.67	97.39	97.16	95.62
FPS	85.60	84.07	83.98	82.92	79.01	78.48	81.42	82.21	93.04	86.16	96.67	98.89	98.74	98.96	96.19	95.57
FedFTG	85.62	85.58	84.23	82.70	79.40	77.67	81.27	82.36	95.23	85.58	95.85	99.25	99.72	98.74	97.59	95.96
Ours	**86.90**	85.25	84.93	84.95	**81.83**	81.31	**83.32**	**84.07**	**97.63**	**92.57**	**97.57**	99.08	**99.74**	**99.08**	**98.12**	**97.63**
Ours with DP	86.01	**85.98**	**85.66**	**85.97**	**81.48**	76.44	**82.50**	**83.42**	97.02	90.88	95.70	**99.48**	99.73	99.05	97.64	**97.07**



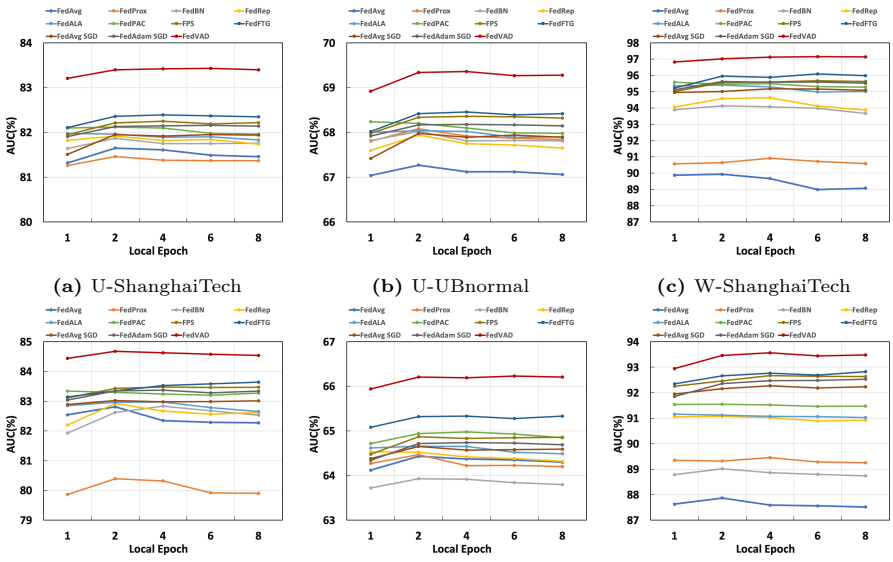
(a) U-ShanghaiTech **(b)** U-UBnormal **(c)** W-ShanghaiTech **(d)** W-UCF-Crime

(e) W-XD-Violence **(f)** U-Random-Mix **(g)** W-Random-Mix **(h)** W-Random-Mix

Fig. 4: Comparison of communication efficiency of four datasets in two settings. U and W denote unsupervised and weakly supervised settings, respectively.

4.2 Performance Comparison

Comparison with existing methods. To demonstrate the effectiveness of our proposed method, we compare it with two categories of methods, including: a) centralized [13, 29], a single model that is trained with the combination of all the local data, which serves as the upper-bound of FL models; and b) ten state-of-the-art FL algorithms, encompassing traditional federated learning methods like FedAvg [24] and FedProx [18], personalized federated learning methods such as FedBN [19], FedRep [5], FedALA [51], and FedPAC [43], and pre-trained federated learning methods including FedAvg SGD [26], FedAdam SGD [26], FPS [3], and FedFTG [52]. Nguyen et al. [26] explore the impact of pre-training on the performance of federated learning based on the FedOPT framework. We select FedAvg SGD and FedAdam SGD settings of their work. To ensure the fairness of the comparison, we use the same public dataset to pre-train the



(a) U-ShanghaiTech **(b)** U-UBnormal **(c)** W-ShanghaiTech

(d) W-UCF-Crime **(e)** W-XD-Violence **(f)** W-Random-Mix

Fig. 5: Effect of the local epoch number per round.

global models of these two methods. FedAvg is chosen as the underlying federated learning framework. We report the overall performance on two settings in Tab.1. We can observe that our proposed FedVAD achieves superior performances over all competitors on the four datasets. Our method outperforms FedAvg, showing an increase of 1.77% and 2.09% in unsupervised settings for the ST and UB datasets, respectively. In weakly supervised settings, it achieves gains of 8.04% on ST, 1.87% on UCF, and 1.78% on XD, demonstrating superior performance compared with other federated learning approaches. In addition, we show the performance of the ST dataset for each client under two different settings in Tab.2. FedVAD consistently outperforms other federated learning methods in the majority of clients, highlighting the pivotal role of the Adaptive Semantic-Enhanced Distillation module in its enhanced performance.

Communication Efficiency. Fig.4 shows the AUC score of training different models on different datasets. It is shown that FedVAD trains much faster than the baseline algorithms, and it reaches higher accuracy in a shorter period. Even when compared with pre-trained federated learning methods, FedVAD also demonstrates more robust performance. From the results in Fig.4a to Fig.4b, we can clearly observe a remarkable phenomenon: most of the methods show a decreasing trend in performance as the number of communication rounds increases, especially after reaching a particular number of communication rounds. FedVAD, on the other hand, exhibits distinctive characteristics. In consecutive communication rounds, FedVAD consistently shows an increasing trend of performance, which undoubtedly proves its stability and superiority in long-time communica-

tion. Although FedVAD's performance may be slightly inferior to other methods in some specific communication rounds, as shown in Fig.4c to Fig.4e, it consistently maintains a high level of performance in terms of the overall trend.

To assess FedVAD's generalization capability, we conduct comparative experiments with a random selection of clients, detailed in Fig.4f to Fig.4h. The results demonstrate that FedVAD marginally surpasses other methods in both experimental setups. Its consistent performance improvement and remarkable generalization capacity distinctly position it as a robust solution in the federated learning landscape.

4.3 Ablation Study and Analysis

In this section, we investigate the effects of key designs on the performance of the proposed FedVAD framework.

The effect of key modules. We conduct an ablation study to demonstrate the effectiveness of the main components of our method: Visual Distillation (VD), Federated Visual Consistency Clustering (FVCC), and Adaptive Semantic-Enhanced Distillation (ASED). In Tab.3a, we show the experimental results on four different datasets under the two settings, through which we observe that each independent module plays a key role in the overall performance improvement. Specifically, compared with the baseline algorithm FedAvg, VD provides only a limited performance improvement on each dataset if it is only on the server side. More notably, in the weakly supervised learning setting, the FVCC module improves the AUC on the ST, UCF, and XD datasets by significant percentages from 90.54% to 94.46%, from 82.74% to 83.24%, and from 64.52% to 65.26%, respectively. In addition, when the ASED module is added, we observe a further improvement in performance, culminating in AUC of 97.07%, 84.68%, and 66.21%. These significant performance improvements on the dataset clearly demonstrate the effectiveness of both the FVCC and ASED modules.

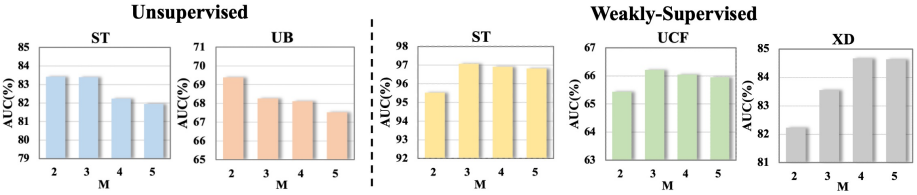
The effect of public datasets size. For constructing a server-side public dataset, we select 10% to 20% of the training set from each client to form this public dataset. We test our FedVAD model across various sizes of these public datasets, enabling an evaluation of its scalability and adaptability in diverse data environments. Specifically, we chose 20%, 50%, 80%, and 100% of the total proportion of public datasets to participate in each round of federated learning. As shown in Fig.6a, as the proportion increases, the global model of each group is better optimized and the performance of FedVAD on each dataset is significantly improved.

4.4 Hyperparameter Analysis

We further analyze the influence of three hyperparameters: the number of local epochs, the number of clustering groups, weighting effects of distillation loss.

Number of Local Epochs. We investigate the effect of the number of local epochs for four datasets in two settings. The total epoch number is set to $G \times L$ where G and L denote the number of global and local epochs, respectively. By

(a) FedVAD results on public datasets of different sizes.



(b) Effect of the cluster groups number per round.

Fig. 6: Effects of different parameters.

Table 3: The effect of key modules on the performance of the proposed FedVAD framework and a comparison of VAD **accuracy** between GPT-4V and conventional VAD methods.

(a) The effect of key modules on FedVAD.

Variants			Datasets and Settings				
			Unsupervised		Weakly-Supervised		
VD	FVCC	ASED	ST	UB	ST	UCF	XD
✔	✘	✘	81.86	68.13	90.54	82.74	64.52
✔	✔	✘	82.09	68.22	94.46	83.34	65.26
✔	✔	✔	83.42	69.36	97.07	84.68	66.21

(b) Comparison of VAD and GPT-4V methods.

	ST	UB	UCF	XD
MTDNAD [39]	97.02	86.72	90.27	87.46
PEL4VAD [29]	97.63	87.39	92.32	89.74
GPT-4V [45]	**49.25**	**48.34**	**55.37**	**66.28**
FedVAD	94.38	85.27	89.42	85.84

keeping the total epoch number unchanged, the FedVAD is shown in Fig.5 with the number of local epochs growing. We observe that in the non-independent identically distributed (non-IID) setting, the performance of both traditional and personalized federated learning methods degrades when the number of local training epochs reaches a certain level. This indicates that with the increase in the number of local training epochs, models tend to seek locally optimal solutions based on their specific data distributions rather than the global optimum. In contrast, within the pre-trained federated learning approach, the model undergoes initial training on a public dataset and achieves a more optimized parameter initialization. Therefore, when the number of local training epochs per communication round increases, the model consistently demonstrates performance improvement. We choose the best number of local epochs in the paper.

Number of clustering groups. Considering the difference in the number of clients in different datasets, for fewer clients, we choose M as 2, 3, 4, and 5, and for more clients, we add M as 8, and 10 for a better evaluation of the effectiveness of our method. In Fig.6b, we show the performance of the FedVAD method when

the number of sets clustered varies. We observe that in the unsupervised setting, the performance shows a slight decrease as the number of clusters increases due to the fact that the training data contains only normal samples. On the contrary, under the weakly supervised setting, good performance is presented as the number of clusters increases, but after a certain number is reached, the abnormal samples may be overly subdivided, resulting in the model not being able to identify these abnormalities efficiently, and also presenting a decreasing trend. As a whole, clustering positively affects FedVAD on all datasets.

4.5 Exploratory Analysis

How does GPT perform in video anomaly detection? To substantiate the effectiveness of our proposed method, we execute comparative analyses across four distinct datasets against the well-acknowledged GPT-4V, as shown in Tab.3b. With customized textual prompts, we enable the **GPT-4V** framework to categorize image sequences as normal or anomalous, employing accuracy as the evaluation metric. The findings prominently showcase a marked discrepancy in the performance of GPT-4V when contrasted with both the traditional centralization approach and our FedVAD method. Furthermore, feeding video captions into **GPT-4** may overlook critical visual nuances, as this approach primarily captures textual information, potentially omitting significant visual details.

What distinguishes our FedVAD from pre-trained model-based FL? Pre-trained FL involves initializing each client with model parameters trained on large-scale datasets. According to Fig.4 and Tab.1, such approaches can accelerate the convergence of the global model and do not improve the VAD performance. Despite the general knowledge embedded in robust pre-trained models like GPT-4V, they exhibit notable constraints when applied to particular scenarios and tasks. Our approach crafts a learning strategy aided by GPT-4's inferential capabilities, specifically devised for the unique requirements of federated anomaly detection, thereby securing superior performance.

5 Conclusion & Limitations

In this work, we introduce a novel federated video anomaly detection framework designed to address the challenges of data heterogeneity. We design the Federated Visual Consistency Clustering strategy and incorporate an Adaptive Semantic-Enhanced Distillation process to fuse semantic contextual information, thereby significantly enhancing the generalizability and robustness of the global model across various client clusters. However, it is important to acknowledge that our experiments utilized simulated datasets, not real-world systems. Future research should aim to apply and validate our approach in actual system environments, taking into account factors like bandwidth and poisonous clients. This could lead to further refinements and performance enhancements of our method.

Acknowledgements

This work was supported by NSFC (No.62206200, 62206137, 62036012, 62376196, U23A20387), and Tianjin Natural Science Foundation (No.22JCQNJC00940, 22JCYBJC00030).

References

1. Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 481–490 (2019)
2. Acsintoae, A., Florescu, A., Georgescu, M.I., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: New benchmark for supervised open-set video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20143–20153 (2022)
3. Chen, H.Y., Tu, C.H., Li, Z., Shen, H.W., Chao, W.L.: On the importance and applicability of pre-training for federated learning. In: The Eleventh International Conference on Learning Representations (2022)
4. Chow, K.H., Liu, L., Wei, W., Ilhan, F., Wu, Y.: Stdlens: Model hijacking-resilient federated learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16343–16351 (2023)
5. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: International conference on machine learning. pp. 2089–2099. PMLR (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Doshi, K., Yilmaz, Y.: Federated learning-based driver activity recognition for edge devices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3338–3346 (2022)
8. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
9. Gan, Y., Pan, M., Zhang, R., Ling, Z., Zhao, L., Liu, J., Zhang, S.: Cloud-device collaborative adaptation to continual changing environments in the real-world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12157–12166 (2023)
10. Ghosh, A., Chung, J., Yin, D., Ramchandran, K.: An efficient framework for clustered federated learning. Advances in Neural Information Processing Systems **33**, 19586–19597 (2020)
11. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1705–1714 (2019)
12. Guo, T., Guo, S., Wang, J.: pfedprompt: Learning personalized prompt for vision-language models in federated learning. In: Proceedings of the ACM Web Conference 2023. pp. 1364–1374 (2023)
13. Hirschorn, O., Avidan, S.: Normalizing flows for human pose anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13545–13554 (2023)

14. Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16312–16322. IEEE (2023)
15. Jiang, M., Roth, H.R., Li, W., Yang, D., Zhao, C., Nath, V., Xu, D., Dou, Q., Xu, Z.: Fair federated medical image segmentation via client contribution estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16302–16311 (2023)
16. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
17. Li, M., Li, Q., Wang, Y.: Class balanced adaptive pseudo labeling for federated semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16292–16301 (2023)
18. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)
19. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)
20. Liu, B., Lv, N., Guo, Y., Li, Y.: Recent advances on federated learning: A systematic survey. arXiv preprint arXiv:2301.01299 (2023)
21. Liu, Y., Yang, D., Wang, Y., Liu, J., Song, L.: Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. arXiv preprint arXiv:2302.05087 (2023)
22. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE international conference on computer vision. pp. 341–349 (2017)
23. Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., Zhang, H.: Unbiased multiple instance learning for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8022–8031 (2023)
24. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
25. Miao, J., Yang, Z., Fan, L., Yang, Y.: Fedseg: Class-heterogeneous federated learning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8042–8052 (2023)
26. Nguyen, J., Wang, J., Malik, K., Sanjabi, M., Rabbat, M.: Where to begin? on the impact of pre-training and initialization in federated learning. arXiv preprint arXiv:2210.08090 (2022)
27. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article (2023)
28. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14372–14381 (2020)
29. Pu, Y., Wu, X., Wang, S.: Learning prompt-enhanced context features for weakly-supervised video anomaly detection. arXiv preprint arXiv:2306.14451 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

31. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
32. Sada, A.B., Bouras, M.A., Ma, J., Runhe, H., Ning, H.: A distributed video analytics architecture based on edge-computing and federated learning. In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). pp. 215–220. IEEE (2019)
33. Sapkota, H., Yu, Q.: Bayesian nonparametric submodular video partition for robust anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3212–3221 (2022)
34. Sato, F., Hachiuma, R., Sekii, T.: Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6471–6480 (2023)
35. Sattler, F., Müller, K.R., Samek, W.: Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. IEEE transactions on neural networks and learning systems **32**(8), 3710–3722 (2020)
36. 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. pp. 6479–6488 (2018)
37. Sun, S., Gong, X.: Hierarchical semantic contrast for scene-aware video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22846–22856 (2023)
38. Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., Zhang, C.: Fedproto: Federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8432–8440 (2022)
39. Wan, B., Jiang, W., Fang, Y., Luo, Z., Ding, G.: Anomaly detection in video sequences: A benchmark and computational model. IET Image Processing **15**(14), 3454–3465 (2021)
40. Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., Khazaeni, Y.: Federated learning with matched averaging. arXiv preprint arXiv:2002.06440 (2020)
41. Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q., Poor, H.V.: Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security **15**, 3454–3469 (2020)
42. Xiang, A.: Being'seen'vs.'mis-seen': Tensions between privacy and fairness in computer vision. Harvard Journal of Law & Technology **36**(1) (2022)
43. Xu, J., Tong, X., Huang, S.L.: Personalized federated learning with feature alignment and classifier collaboration. arXiv preprint arXiv:2306.11867 (2023)
44. Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., Yu, H.: Federated transfer learning. In: Federated Learning, pp. 83–93. Springer (2020)
45. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 **9**(1), 1 (2023)
46. Yang, Z., Liu, J., Wu, Z., Wu, P., Liu, X.: Video event restoration based on keyframes for video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14592–14601 (2023)
47. Yu, F., Zhang, W., Qin, Z., Xu, Z., Wang, D., Liu, C., Tian, Z., Chen, X.: Fed2: Feature-aligned federated learning. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 2066–2074 (2021)

48. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y.: Bayesian nonparametric federated learning of neural networks. In: International conference on machine learning. pp. 7252–7261. PMLR (2019)
49. Zhang, C., Li, G., Qi, Y., Wang, S., Qing, L., Huang, Q., Yang, M.H.: Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16271–16280 (2023)
50. Zhang, C., Li, G., Xu, Q., Zhang, X., Su, L., Huang, Q.: Weakly supervised anomaly detection in videos considering the openness of events. IEEE transactions on intelligent transportation systems **23**(11), 21687–21699 (2022)
51. Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., Guan, H.: Fedala: Adaptive local aggregation for personalized federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11237–11244 (2023)
52. Zhang, L., Shen, L., Ding, L., Tao, D., Duan, L.Y.: Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10174–10183 (2022)
53. Zhang, Z., Zhong, S.h., Fares, A., Liu, Y.: Detecting abnormality with separated foreground and background: Mutual generative adversarial networks for video abnormal event detection. Computer Vision and Image Understanding **219**, 103416 (2022)
54. Zhao, Y., Liu, H., Li, H., Barnaghi, P.M., Haddadi, H.: Semi-supervised federated learning for activity recognition. ArXiv **abs/2011.00851** (2020), `https://api.semanticscholar.org/CorpusID:226226498`
55. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. arXiv preprint arXiv:2302.05160 (2023)