Differentially Private Video Activity Recognition

Zelun Luo [1] Yuliang Zou [2] Yijin Yang [3] Zane Durante [1] De-An Huang [4]
Zhiding Yu [4] Chaowei Xiao [4,5] Li Fei-Fei [1] Animashree Anandkumar [4,6]

[1] Stanford University [2] Virginia Tech [3] Arizona State University [4] NVIDIA
[5] University of Wisconsin–Madison [6] Caltech

Abstract

In recent years, differential privacy has seen significant advancements in image classification; however, its application to video activity recognition remains under-explored. This paper addresses the challenges of applying differential privacy to video activity recognition, which primarily stem from: (1) a discrepancy between the desired privacy level for entire videos and the nature of input data processed by contemporary video architectures, which are typically short, segmented clips; and (2) the complexity and sheer size of video datasets relative to those in image classification, which render traditional differential privacy methods inadequate. To tackle these issues, we propose Multi-Clip DP-SGD, a novel framework for enforcing video-level differential privacy through clip-based classification models. This method samples multiple clips from each video, averages their gradients, and applies gradient clipping in DP-SGD without incurring additional privacy loss. Moreover, we incorporate a parameter-efficient transfer learning strategy to make the model scalable for large-scale video datasets. Through extensive evaluations on the UCF-101 and HMDB-51 datasets, our approach exhibits impressive performance, achieving 81% *accuracy with a privacy budget of* $\epsilon = 5$ *on UCF-101, marking a* 76% *improvement compared to a direct application of DP-SGD. Furthermore, we demonstrate that our transfer learning strategy is versatile and can enhance differentially private image classification across an array of datasets including CheXpert, ImageNet, CIFAR-10, and CIFAR-100.*

1. Introduction

The rising adoption of machine learning in privacy-sensitive sectors has intensified the need for privacy-preserving machine learning [8, 74, 75]. Within these realms, videos constitute a substantial portion of the data, such as surveillance streams used for patient monitoring [23, 25, 43]. This underscores the criticality of developing video models that are robust against privacy attacks inherent to machine learning [54, 63, 64]. Notable advancements in image classification models have been achieved through the employment of differential privacy [1, 10, 35, 48, 55], furnishing probabilistic privacy assurances by minimizing the impact of data point substitutions within the dataset. With DP, it is mathematically infeasible to reverse engineer the model to recover the original training data through the model inversion attack [20]. This prevents model inversion attacks that attempt to recover training videos or identify individuals in the training set. DP also provides robustness against membership inference attacks [64], where an attacker tries to determine if a specific data point was used for training. However, the field has yet to see breakthroughs in video classification comparable to those in image classification, creating uncertainties about their transferability to videos.

We pinpoint two characteristics inherent to video classification that hinder the direct adaptation of differential privacy to videos. Firstly, existing private image classification models primarily assume *per-sample* differential privacy, offering probabilistic guarantees for each individual sample (*i.e.* image, video clip) fed into the model during training. This becomes an issue for state-of-the-art video classification models that usually process multiple short clips instead of full videos. Under per-sample differential privacy, these models only assure privacy for the short clips used in training, not for the entire videos. This poses a challenge for state-of-the-art video classification models [18, 41, 42, 45] which, in various training iterations, sample multiple short *clips* from the same video instead of processing the entire video as a single sample. Under per-sample differential privacy, privacy is assured only for these individual short clips, and not for the whole video from which they are sampled. Secondly, as illustrated in Table 1, video datasets that are commonly used, such as UCF-101 [66], HMDB-51 [33], and Kinetics [4], are substantially larger in scale compared to the image classification datasets like MNIST and CIFAR-10, which are traditionally employed in DP research [1, 55]. Directly employing DP algorithms like DP-SGD on these video datasets is impractical due to its detrimental impact

	Method	Architecture(s)	Pre-training	MNIST	F-MNIST	CIFAR-10	ImageNet
w/o pre-train	DP-SGD [1]	2-layer NN	-	95.00@2.00	-	-	-
	DPNAS [7]	DPNASNet	-	98.57@3.00	88.09@3.00	68.33@3.00	-
	Tempered Sigmoid [56]	6/6/9-layer CNN	-	98.10@2.93	86.10@2.70	66.20@7.53	-
	DP-ScatterNet [68]	ScatterNet	-	98.70@2.93	89.70@3.00	69.30@3.00	-
	Norm-DP-SGD [9]	LeNet-5/VGG-16	-	98.18@3.00	-	77.40@2.00	-
	Private-kNN [80]	5-layer CNN	-	98.80@0.47	-	-	-
	AdaCliP [58]	2-layer NN	-	95.56@2.00	-	-	-
	DDP-SGD [13].	6-layer CNN	-	96.34@1.20	83.81@2.00	-	-
w/ pre-train	DP-SGD [1]	4-layer CNN	CIFAR-100	-	-	67.00@2.00	-
	Private-kNN [80]	4-layer CNN	CIFAR-100	-	-	70.80@2.92	-
	Scalable [44]	ResNet-18	ImageNet	-	-	81.57@1.50	-
	DP-ScatterNet [68]	ResNet-50	ImageNet	-	-	92.70@2.00	-
	DP-ImageNet [35]	ResNet-18	Places365	-	-	-	47.90@10.00

Table 1. **State-of-the-Art Differential Privacy Methods.** Evaluation of differential privacy methods on diverse vision datasets using the acc@ϵ metric, selecting results nearest to $\epsilon = 3$ for each method. The table emphasizes the limited exploration of large-scale vision tasks and privacy-sensitive domains in current literature.

on large-scale training.

Summary of Contributions: (1) We achieve the first substantial result on differentially private video action classification ($+76\%@\epsilon = 5$ on UCF-101) by simultaneously addressing the two aforementioned challenges. (2) We establish a framework for video-level differential privacy and introduce a novel multi-clip method (Multi-Clip DP-SGD) that allows clip-based video classification models to attain video-level differential privacy without demanding additional privacy budgets. (3) Through an exhaustive analysis, we pinpoint the most effective parameter-efficient transfer learning strategy for the application of differentially private training to large-scale video datasets.

Firstly, we recognize a disparity between the level of privacy sought (encompassing an entire video) and the nature of inputs processed by modern video classification architectures [18, 41] (which are short clips of frames). Conventional training methods focus on clips, and employing differentially private training algorithms like DP-SGD under this setup merely achieves clip-level privacy, compromising the privacy of the full videos. To tackle this issue, we put forward a video-based, multi-clip differential privacy training scheme termed Multi-Clip DP-SGD (Figure 1). For each selected video, our approach samples several clips per iteration. Rather than directly clipping gradients for each clip, which would inflate the privacy budget in proportion to the number of clips and thus impede the balance between privacy and utility, we average the gradients among clips from the same video before applying the clipping. Our method harnesses the rich information in each video without escalating the privacy budget.

Secondly, we recognize that video classification models and datasets considerably exceed the size of their image classification counterparts typically examined in differential privacy research (UCF-101 [66] with 2,916,000 frames compared to MNIST/CIFAR-10 with 60,000 images). Transfer learning has been pivotal in successfully scaling differential privacy to large datasets like ImageNet [61] in the realm of image classification [10, 35, 48]. By publicly pre-training models on even more extensive datasets like JFT-3B [67], the efficacy of private fine-tuning on target datasets is significantly enhanced. Yet, the optimal fine-tuning strategy, especially for video architectures, remains elusive. To address this, we undertake an exhaustive analysis of transfer learning strategies that are parameter-efficient for differential privacy, as depicted in Figure 2. Our study reveals two key insights: (1) fine-tuning normalization layers along with the final linear layer yields the best performance with a negligible increase in trainable parameters; (2) incorporating additional parameters from adapters [28] proves advantageous when there is a substantial domain disparity in pre-training. Leveraging these insights along with our multi-clip video training, we conduct experiments on UCF-101 and HMDB-51 and achieve unprecedented results in video classification. In particular, the LayerNorm [3] architecture (in MViT [17]), parameter-efficient transfer learning, and our multi-clip differential privacy training (Multi-Clip DP-SGD) collectively account for our substantial gains ($+76\%@\epsilon = 5$ compared to training from scratch).

Finally, we extend our insights to large-scale image classification. Our analysis reveals that contemporary architectures employing LayerNorm [3], such as ViT-S [12] and ConvNeXt-T [40], outperform the traditionally used ResNet-50-GroupNorm [35, 71] in privacy-utility tradeoffs, despite comparable parameter counts and ImageNet-1K [61] performance. These discoveries propel the state-of-the-art in differential privacy methods, showing substantial gains on standard benchmarks (e.g., $+21.7\%@\epsilon = 4$ on CIFAR-10, $+71.1\%@\epsilon = 4$ on CIFAR-100) and opening avenues for applying differential privacy to previously unexplored privacy-sensitive datasets like CheXpert [30].

2. Related work

Differentially-Private SGD (DP-SGD). (ϵ, δ)-differential privacy has become the gold standard for database and model privacy due to its probabilistic guarantees of minimal data leakage [14, 15]. The most common method for

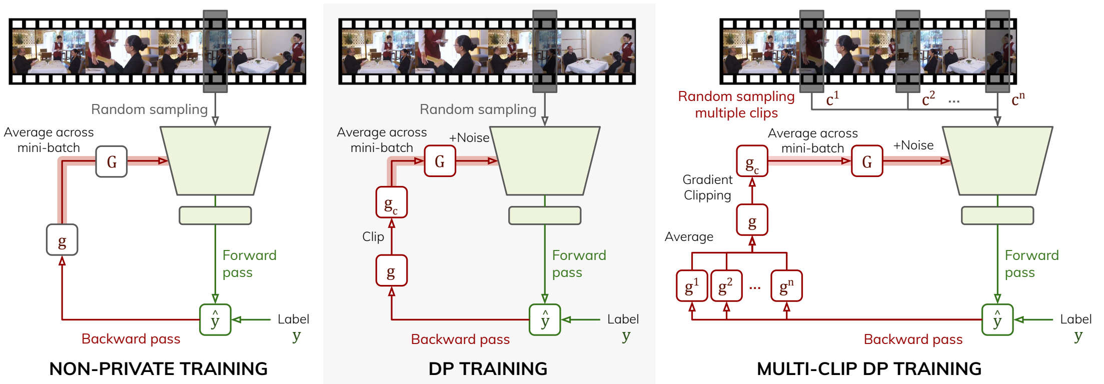


Figure 1. **Comparison of Training Methods for Videos.** We compare non-private training (left), DP-SGD (middle), and Multi-Clip DP-SGD (right). DP-SGD utilizes gradient clipping and adds Gaussian noise to gradients per video clip. Multi-Clip DP-SGD selects several clips per iteration and combines gradients within a video prior to clipping. By ensuring that each video is observed only once per training step, the performance is enhanced without compromising privacy.

ensuring differential privacy for deep learning models is DP-SGD proposed in Abadi *et al.* [1]. This modification on standard stochastic gradient descent (SGD) uses gradient clipping and noise additions to mitigate the effects of individual data points [1]. There have been several follow-up works [11, 24, 70] on DP-SGD. Some explore adaptive methods for clipping gradients to help learning [58]. Others propose methods for loss function smoothing to mitigate the effects of the noise added to the gradient [70]. Our work uses the accounting method based on Rényi differential privacy [50], which provides a tighter bound on the privacy-utility estimate than the accounting method used in Abadi *et al.* [1]. There have also been a series of works exploring the effect of neural architecture choices on the performance obtained when using DP-SGD [9, 51, 56].

Large-Scale Differential Privacy. Recent works have shown that differential privacy is possible even on large-scale datasets like ImageNet by leveraging transfer learning, but requires significant pre-training [10, 35]. Prior work shows that using regular DP-SGD to train on ImageNet achieves an accuracy of $0-1\%$ [35]. However, pre-training image models on large-scale internet data has led to reasonable performances on ImageNet, with Kurakin *et al.* [35] getting 47.9% image classification accuracy with $\epsilon = 10$, $\delta = 8 \cdot 10^{-7}$. In natural language processing (NLP), recent work has shown that large pre-trained language models can be effective differentially private learners [2, 37].

Transfer Learning. In *transfer learning*, a model is first pre-trained on an initial dataset and then fine-tuned on a downstream dataset [46, 53, 81]. There are many ways to fine-tune models. The simplest way is full fine-tuning, which trains the entire model on the downstream dataset [34]. Another way to fine-tune is to freeze the entire model except the last layer and use it as a feature extractor, and then train a linear probe classifier using the feature extractor embedding space [5, 21]. There have also been ongoing research threads using parameter efficient fine-tuning in order to save computation costs. Prompt-tuning has recently become a common strategy in natural language processing (NLP) for efficient fine-tuning of large language models [36, 39], and adapter modules have also recently become a popular way to efficiently fine-tune large models [22, 27]. In alternative approaches to parameter efficient transfer learning, the normalization layers or weights discovered through the lottery ticket hypothesis [19] are fine-tuned to enable efficient transfer learning [38, 49].

Video and Privacy. As videos may contain personal information, there have been several works on privacy-preserving video processing framework. Possible methods include anonymizing faces in videos [60], cryptographic approach [57], decreasing the video resolution [62], and learning anonymization [73]. While previous research has explored the application of differential privacy to videos, its primary emphasis lies in directly introducing differential privacy to the entire video [69]. We propose a multi-clip differential privacy training algorithm for video classification, and achieve the first non-trivial result on commonly used video classification datasets.

3. Differentially Private Video Classification

We pinpoint two challenges in differentially private video classification: (1) a discrepancy between model inputs (clips) and privacy objectives (videos); (2) adapting differential privacy for large-scale video datasets. In Section 3.1, we lay the groundwork with an overview of differential privacy (DP). In Section 3.2, we delineate video-level DP, the crux of this study. In Section 3.3, we introduce our solution, Multi-Clip DP-SGD, for video-level DP. In Section 3.4, we outline parameter-efficient transfer learning strategies for scaling differential privacy to videos. Lastly, we examine suitable network architectures in Section 3.5.

3.1. Preliminary

Differential Privacy. Differential privacy [14] provides a formal privacy guarantee to prevent information leakage of

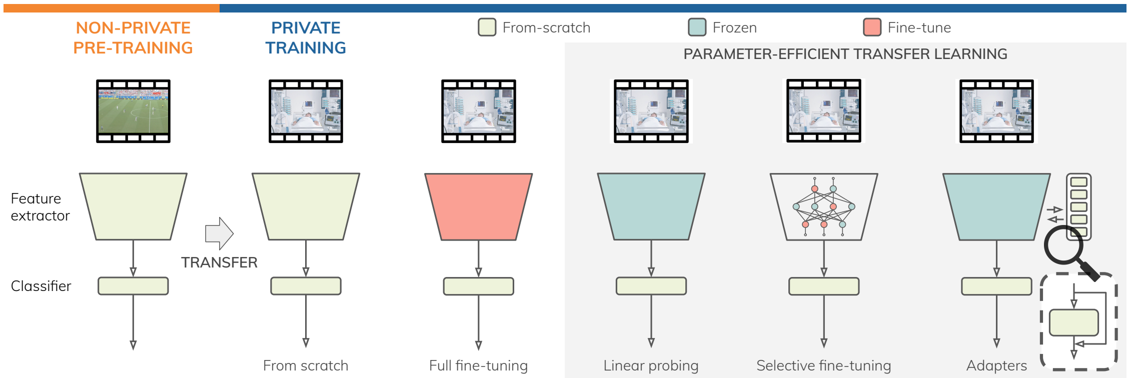


Figure 2. **Transfer Learning Paradigms.** This figure compares four transfer learning paradigms: full fine-tuning, linear probing, selective fine-tuning, and adapters. Notably, the latter three paradigms are characterized as parameter-efficient as they focus on training smaller subnetworks as opposed to the entire network.

each individual data point within a dataset. By adding randomization to computation over a dataset (*e.g.* training a ML model), the influence of each data point is bounded, and thus the instance-level privacy is protected. There are two important parameters controlling the strength of differential privacy guarantee: $\epsilon > 0$ and $\delta \in [0, 1]$. And the privacy guarantee becomes stronger as both parameters get smaller. Formally, we have the following definition, which makes use of the concept of *adjacent datasets* (*i.e.* two datasets than only differ by one data entry).

Definition 1 *Given a randomized model training algorithm $M : D \rightarrow R$ and two adjacent datasets, $d_1, d_2 \in D$, then we say that M satisfies (ϵ, δ)-differential privacy if $\forall S \subseteq R$:*

$$P[M(d_1) \in S] \leq e^{\epsilon} P[M(d_2) \in S] + \delta.$$

Here, D represents the set of all potential datasets, and R encompasses all possible models that could be trained on any dataset within D. $P[.]$ indicates the likelihood of obtaining model R when training a dataset using the randomized algorithm M. The essence of this definition lies in the concept that if d_1 and d_2 are adjacent datasets (differ by only one data example), an effective randomized algorithm should create a challenge in distinguishing between the two models trained on d_1 and d_2 separately.

Differentially Private SGD (DP-SGD). DP-SGD [1] is one of the most common strategies to protect the privacy of training data. With slight modifications upon the traditional SGD algorithm, DP-SGD effectively provides instance-level privacy protection. More specifically, it has the following modifications. First, instead of computing mini-batch level gradients at each iteration, DP-SGD computes the per-sample gradient and clips gradient values greater than a constant C. Then, these clipped gradient values are averaged over the training mini-batch and Gaussian noises $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ are added. DP-SGD was designed to integrate seamlessly with deep neural networks due to its simplicity of implementation and no additional assumptions about the data. Figure 1 compares training with and without DP-SGD for a clip-based video classification model.

Privacy Accountant. A privacy accountant keeps track of the privacy spendings during model training. The momentum accountant [1] was first proposed to compute the privacy loss for DP-SGD. Later, Rényi differential privacy was proposed to relax differential privacy by using the Rényi divergence [50]. It provides a tighter privacy-utility bound and thus can allow for greater accuracy with the same probabilistic guarantees as standard DP-SGD.

3.2. Video-Level Differential Privacy

In this work, our goal is differentially private video classification. We aim to train a video classifier $f(\cdot)$, which takes a video v_i as input and outputs its class. In addition, $f(\cdot)$ should be trained with privacy guarantees. More specifically, we are interested in training $f(\cdot)$ with *video-level* differential privacy. Recall that Definition 1 builds on the concept of adjacent datasets. By video-level DP, we mean that each data entry in the adjacent datasets is a video, and the two datasets only differ by one video.

While our goal is to classify the entire video, modern video classification architectures are often designed to take multiple *clips* consisting of several frames as input during training [17, 41]. Each video can be temporally chunk into N_i clips: $v_i = [c_i^1 \dots c_i^{N_i}]$. For simplicity, we assume that the clips are consecutive and non-overlapping, and by concatenating all the c_i^j by order we have the original video v_i. Clip-based architecture is a popular and effective design for video classification. However, this creates problems for applying differential privacy to video classification.

Since the video models are trained on clips, if we directly apply privacy-preserving algorithms in training, such as DP-SGD, then each data entry in the adjacent datasets of Definition 1 is no longer a video, but instead a clip. In this case, the privacy gaurantees are given for clips instead of videos. We refer to this as *clip-level* differential privacy. It is important to note that clip-level differential privacy does not imply video-level differential privacy for the same (ϵ, δ). Video-level differential privacy considers adjacent datasets differ by one video v_i, which are in fact N_i clips. This is beyond what is covered by clip-level differential privacy (*i.e.* adjacent datasets differ by only one clip). Since a video $v_i = [c_i^1 \dots c_i^{N_i}]$ consists of a group of clips c_i^j, one possible approach is to apply differential privacy's property of *group privacy* to translate clip-level differential privacy to video-level differential privacy [14]. However, this would simply multiply the privacy budget by the number of clips in videos, which significantly harms the privacy-utility tradeoff.

3.3. Multi-Clip DP-SGD for Videos

We have discussed that naively applying privacy-preserving training algorithms to video model training only achieves clip-level DP, which does not imply our desired video-level DP. Now we discuss our solution, Multi-Clip DP-SGD, which builds on DP-SGD [1] to fully exploit information in videos during training, while not increasing their sensitivity in DP.

The first step is to sample by videos instead of clips in training. Next, one naive way to process the sampled video is to sample only one clip per video in each mini-batch during the training time, since the architectures are designed to take a clip as input. However, this severely wastes the privacy budget, since we only utilize a small portion of information but with the price of increased privacy loss on the whole video. Repeating this process in training does not help with the privacy-utility tradeoff.

Inspired by the user-level privacy protection [47], we propose a new multi-clip scheme for video-level differential privacy training. The core idea is to increase the usage of each video, while not increasing the sensitivity/influence of each video. Figure 1 illustrates our Multi-Clip DP-SGD for videos. First, we sample multiple clips from a video, and compute the per-sample gradient for each clip as in normal DP-SGD. However, instead of clipping the gradient for each clip, we first *average* the gradients from all the clips before clipping the gradient (and adding Gaussian noise).

Since there is no information exchange between different videos, the computation above satisfies video-level DP. And since we average multiple clips from the same video before gradient clipping, we can regard that each video is seen only once in each forward pass. And thus the privacy loss does not increase. As we will show in the experiments, Multi-Clip DP-SGD better leverages the information in each video without incurring additional privacy costs and improves the privacy-utility tradeoff of video classification models.

3.4. Parameter-Efficient Transfer Learning

While we have proposed Multi-Clip DP-SGD for videos, by itself it is still not sufficient to apply differential privacy to commonly used video datasets and models because of their scales. Even comparatively smaller video datasets like UCF-101 [66] and HMDB-51 [33] are still much larger than CIFAR datasets used in differential privacy literature. Recent works have shown that transfer learning provides a significant benefit to DP-SGD in terms of privacy-utility tradeoff and scaling to large image classification tasks [37, 68]. Building on this strategy, we identify *parameter-efficient* transfer learning as a promising direction for scaling differential privacy to video datasets and conduct a comprehensive study along this direction.

Specifically, DP-SGD aims to minimize the influence of the training data during the training process by (1) clipping the ℓ_2 norm of per-sample gradient $\mathbf{g}$ at $\mathbf{g}/\max(1, \|\mathbf{g}\|_2/C)$, and (2) adding Gaussian noise sampled from $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$, where σ and C are gradient norm bound and noise scale respectively. The above analysis indicates that DP-SGD favors models with smaller number of trainable parameters: the isotropic Gaussian noise applied to training gradients has a greater impact on models with more trainable parameters because the expected norm of the noise increases with the number of parameters [37]. On the other hand, modern deep learning models tend to be highly overparameterized [26, 32] and researchers have shown that increasing the number of parameters leads to improved generalization [52]. Parameter-efficient transfer learning provides an opportunity to use these large models in DP, while not increasing the number of trainable parameters. Next, we discuss the (parameter-efficent) transfer learning schemes we study in this paper.

Full Fine-Tuning. All model parameters are trained on the target data for full fine-tuning, which makes it not parameter-efficient. We select this approach to compare with other parameter-efficient approaches. Although full fine-tuning is probably the most common approach in non-DP transfer learning scenarios, it usually does not work well under differential privacy settings. Differential privacy struggles with more trainable parameters since DP-SGD (clipping/noise) negatively affects models with more trainable parameters.

Linear Probing. In this setting, all model parameters are fixed except the last linear layer. The frozen part can be regarded as a feature extractor, and we train a liner classification on top of the extracted features. Since the last linear layer only contains a small portion of parameters (compared to the full model), DP-SGD can be well combined with linear probing to provide a decent privacy guarantee. However, the feature extractor is trained on the source domain, and thus training only a linear layer may be sub-optimal due to the existence of domain shifts.

Selective Fine-Tuning. As a trade-off between full fine-tuning and linear probing, one can choose to train a small portion of parameters in the feature extractor, in addition to the last linear layer. This strategy allows a model to adapt itself to the target distribution while still maintaining a reasonable amount of trainable parameters. And thus, DP-SGD can be effectively applied under such a setting to provide a good privacy-utility trade-off. Previous work has used Lottery Ticket Hypothesis [19] to select the parameters to train [44]. In this paper, we choose to train the normalization layers in addition to the last linear layer.

Adapter Training. Similar to selective fine-tuning, adapter training also aims to train a small number of model parameters in addition to the last linear layer. However, instead of updating the parameters in the original model architecture, we incorporate additional parameters inside the current network and train them together with the last linear layer. Similar to adapter training in non-DP settings [22, 28, 78], we design lightweight 2-layer MLPs with skip-connection and insert these modules inside the pre-trained networks. These modules are initialized as identity mappings and gradually trained to stabilize model training.

3.5. Architecture

Finally, we further investigate model architectures that are suitable for training with DP. In particular, we find that architectures with different types of normalization layers are worth further investigation. BatchNorm [29] is empirically proved to stabilize model training and speed up convergence [72], and thus it has been an essential component in most modern CNNs. However, BatchNorm layers are not supported in differentially-private models. The training time behavior of BatchNorm requires computation of mean and variance of each input mini-batch, creating a dependency between samples, which violates the formulation of differential privacy. There are two ways to resolve this issue. The first one is to replace BatchNorm with other normalization layers (*e.g.* GroupNorm [71], LayerNorm [3]). This solution seems to be sub-optimal, since the new normalization layers are not designed for these CNNs. The second option is to use models without BatchNorm by its original design. We find that the recent vision transformers are good fits in this case. We further study a new CNN architecture, ConvNeXt. In this paper, we study two different types of network architectures: convolutional neural networks (CNN) and vision transformers. For CNNs, we adopt ResNet [26] and ConvNeXt [40]. For vision transformers, we adopt ViT [12] in image experiments and MViT [17] in video experiments.

4. Experiments

Now we evaluate our proposed Multi-Clip DP-SGD and parameter-efficient transfer learning for video classification. In addition, we further evaluate our parameter-efficient transfer learning on scaling image classification and compare it with state-of-the-art approaches in DP.

4.1. Differentially Private Video Classification

Experimental Setup. We use UCF-101 [66] and HMDB-51 [33] for our evaluation of differentially private video classification. UCF-101 is a widely used action recognition dataset that contains 133,20 action videos annotated into 101 action classes. HMDB-51 is a video classification dataset of human motion, containing 6849 clips divided into 51 action categories, each containing a minimum of 101 clips. Unlike image classification, video classification [18, 65] of human activities typically only samples a clip of frames from each video in a training batch. We report top-1 accuracy as the utility metric. To the best of our knowledge, this is the first study of differentially private video classification on a large-scale video dataset. In our video experiments, we utilize MViT [17] as the backbone architecture. The configurations "Scheme=Full fine-tune/From scratch" and "#Clips=1" indicate the direct application of DP-SGD [1] to the MViT model.

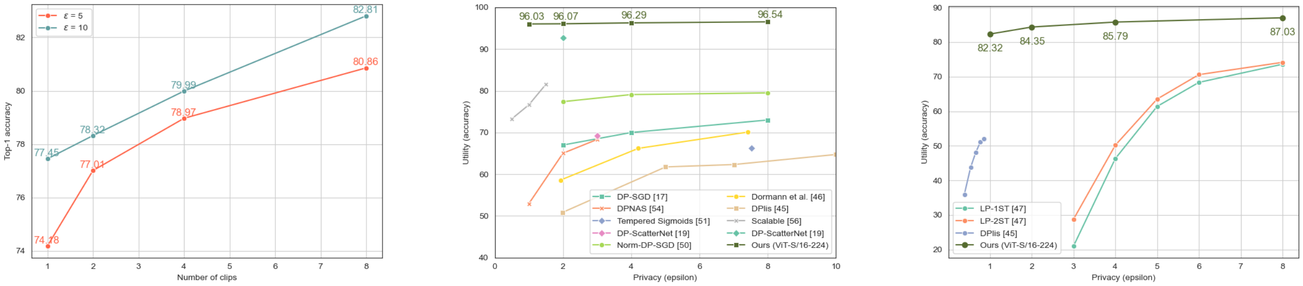
Scheme	#Params.	#Clips	$\epsilon = 5$	$\epsilon = 10$
From scratch	36.4M	1	4.93	6.60
Full fine-tune	36.4M	1	29.15	38.58
Linear probe	77.7K	1	75.05	78.17
Linear probe	77.7K	8	77.96	80.49
Adapter	1.8M	8	79.19	81.73
Selective fine-tune	110K	8	**80.86**	**82.81**

Table 2. **Kinetics-400 → UCF-101.** We adopt $\delta = 10^{-5}$ and report the top-1 accuracy. We conduct experiments using an MViT-B/16×4 model. "Scheme=Adapter/Selective fine-tune" refers to the use of parameter-efficient transfer learning, and "#Clips=8" refers to the use of multi-clip DP-SGD.

Scheme	#Params.	#Clips	$\epsilon = 5$	$\epsilon = 10$
From scratch	36.4M	1	3.62	4.13
Full fine-tune	36.4M	1	17.93	21.21
Adapter	1.8M	1	50.34	53.37
Selective fine-tune	110K	1	51.68	56.48
From scratch	36.4M	8	6.87	7.24
Full fine-tune	36.4M	8	29.20	33.75
Adapter	1.8M	8	53.62	56.23
Selective fine-tune	110K	8	**57.49**	**60.52**

Table 3. **Kinetics-400 → HMDB-51.** We adopt $\delta = 10^{-5}$ and report the top-1 accuracy. We conduct experiments using an MViT-B/16×4 model. "Scheme=Adapter/Selective fine-tune" refers to the use of parameter-efficient transfer learning, and "#Clips=8" refers to the use of multi-clip DP-SGD.

Parameter-Efficient Transfer Learning. The results are shown in Table 2 and Table 3. We find that full fine-tuning significantly improves training from scratch (+24%@$\epsilon = 5$ in UCF-101 and +14%@$\epsilon = 5$ in HMDB-51). In addition, linear probing further improves full fine-tuning by limiting the number of trainable parameters (+46%@$\epsilon = 5$ in UCF-101 and +32%@$\epsilon = 5$ in HMDB-51). Among different parameter-efficient transfer learning schemes, selective fine-tuning improves linear probing by 2.9%@$\epsilon = 5$ on UCF-101 with a slight increase of trainable parameters. Adapter's improvement is less than selective fine-tuning in this scenario. As we will see in later on image classification, the relative performance differences of these



(a) UCF-101

(b) CIFAR-10

(c) CIFAR-100

Figure 3. **Quantitative Analysis on UCF-101, CIFAR-10, and CIFAR-100.** (a) evaluates Multi-Clip DP-SGD for video classification on UCF-101 by experimenting with varying numbers of clips $(1, 2, 4, 8)$ and reporting the top-1 accuracy. (b) and (c) offers a comparative analysis on CIFAR-10 and CIFAR-100 against state-of-the-art differential privacy methods, reporting top-1 accuracy with $\delta = 10^{-5}$.

Architecture	Scheme	#Trainable Params.	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
ResNet-50-GN	From scratch	23.7M	4.16	7.29	11.83	17.25	-
	Full fine-tune	23.7M	52.96	59.63	66.92	71.32	82.40
	Linear probe	204K	62.14	65.89	68.71	70.28	-
	Selective fine-tune	258K	**62.33**	**68.93**	**72.32**	**74.69**	-
ViT-S/16-224	From scratch	21.7M	13.76	16.22	19.67	22.43	-
	Full fine-tune	21.7M	79.36	83.64	85.67	86.59	90.27
	Linear probe	38.5K	71.64	74.81	76.47	77.03	-
	Selective fine-tune	57.7K	**84.73**	**86.90**	**87.68**	**88.43**	-
	Adapter	928K	82.32	84.35	85.79	87.03	-
ConvNeXt-T	From scratch	27.9M	9.20	12.93	15.97	18.45	-
	Full fine-tune	27.9M	65.93	74.00	79.17	81.22	89.98
	Linear probe	76.9K	76.59	78.82	80.32	80.91	-
	Selective fine-tune	93.2K	**83.11**	**84.69**	**86.06**	**86.67**	-
	Adapter	1.7M	79.37	83.68	85.11	86.05	-

Table 4. **Ablation Study on ImageNet-1k → CIFAR-100.** The models are trained for 50 epochs with $\delta = 10^{-5}$, and the top-1 accuracy is reported to validate the effectiveness of the design choices.

parameter-efficient schemes (linear, adapter, selective) are dataset dependent. For transfer learning from Kinetics to UCF-101 and HMDB-51, linear probing is already effective because of the small domain gap.

Multi-Clip DP-SGD. We study the effect of our Multi-Clip DP-SGD in Figure 3a, where we compare results with varying numbers of clips. With only 1 clip, the result is reduced to standard DP-SGD. We use selective fine-tuning for this analysis. With the increasing number of clips in Multi-Clip DP-SGD, the utlity continues to improve given a fixed privacy budget. Overall, our multi-clip approach leads to 6.7%@$\epsilon = 5$ compared to standard DP-SGD, which is larger than the difference between different parameter-efficient schemes. This shows the importance of our multi-clip approach for video classification under DP.

4.2. Differentially Private Image Classification

To underscore the potency of our parameter-efficient transfer learning strategies, we put them to the test on large-scale image classification tasks using two datasets: ImageNet and CheXpert. ImageNet is widely regarded as the benchmark dataset, while CheXpert offers insights into the application of differential privacy in privacy-sensitive scenarios with large domain gaps.

ImageNet. ImageNet [61] has been widely used for image classification. It consists of 1,000 classes of 1,281,167 training images and 50,000 validation images. We report results on ImageNet using two ViT [12] archiectures and two different pre-training datasets. As shown in Table 5, full fine-tuning results in worse performance than the parameter-efficient strategies. We note that using the pre-training dataset from CLIP [59] results in substantial performance gains, highlighting the importance of large-scale pre-training for transfer learning in the differentially private setting. We significantly bridge the gap with supervised performance, with the ViT-B/16 architecture achieving 85.49% top-1 accuracy on ImageNet with public training [12].

CheXpert (Medical Image Classification). CheXpert [31] is a large-scale dataset of chest X-rays annotated by professional radiologists. It consists of 224,316 chest radiographs of 65,240 patients. Following the standard setting [31, 77], we report the AUC score as the utility metric on five selected diseases, i.e., Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion. Results are in Table 6. Adapter outperforms selective fine-tuning with the large domain gap in this case. In addition, for the ViT backbone, when we have sufficient privacy budget, DP-SGD is sufficient to fine-tune the full model and full fine-tuning per-

Architecture	Pre-train	Batch size	Scheme	#Trainable Params.	$\epsilon = 10$
ViT-B/16-224	CLIP [59]	65,536	Full fine-tune	86.6M	66.09
			Linear probe	769K	79.36
			Selective fine-tune	808K	**79.86**
ViT-S/16-224	Places365 [79]	1,024	Full fine-tune	22.1M	29.34
			Linear probe	385K	26.53
			Selective fine-tune	404K	**40.78**
			Adapter	1.3M	40.47
ViT-S/16-224	Places365 [79]	65,536	Full fine-tune	22.1M	30.21
			Selective fine-tune	404K	**45.12**
			Adapter	1.3M	41.79

Table 5. **Large-Scale Differential Privacy Experiments on ImageNet-1k.** We adopt $\delta = 10^{-6}$ and report the top-1 accuracy.

Architecture	Scheme	#Trainable Params.	Pre-train	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
ViT-S/16-224	From scratch	21.7M	-	74.82	74.93	75.13	75.60
	Full fine-tune	21.7M	IN-21k→IN-1k	82.57	83.50	**84.98**	**85.78**
	Selective fine-tune	21.1K	IN-21k → IN-1k	81.76	83.64	83.71	84.46
	Adapter	892K	IN-21k→IN-1k	**83.31**	**83.69**	83.75	84.77
ConvNeXt-T	From scratch	27.8M	-	76.93	77.19	77.51	79.08
	Full fine-tune	27.8M	IN-21k → IN-1k	85.23	85.37	86.83	87.21
	Selective fine-tune	20.2K	IN-21k → IN-1k	85.06	85.29	85.92	86.56
	Adapter	1.7M	IN-21k → IN-1k	**86.77**	**87.06**	**87.78**	**88.05**

Table 6. **Large-Scale Differential Privacy Experiments on CheXpert.** We adopt $\delta = 10^{-5}$ and report the AUROC. All models were trained for 10 epochs.

forms the best. However, parameter-efficient schemes still outperform with less privacy budgets ($\epsilon = 1, 2$).

4.3. Comparison with State-of-the-Art Methods

In this section, we conduct an ablation study and compare our approach with state-of-the-art differential privacy methods on CIFAR-100, as shown in Table 4.

Parameter-Efficient Transfer Learning. We still find that transfer learning outperforms training from scratch by a large margin on datasets commonly used in differential privacy literature. For CNN models (ResNet, ConvNeXt), linear probing outperforms full fine-tuning when using a small ϵ. Selective fine-tuning is consistently the best-performing method on CIFAR-100 despite only having a small increase of trainable parameters compared to linear probing. The improvement is more significant for LayerNorm-based architectures (ViT and ConvNeXt).

LayerNorm-Based Architectures. We find that using recent architectures based on LayerNorm leads to better performance gains compared to the commonly used ResNet-50, despite having a similar amount of trainable parameters and similar performance when publicly trained. For ResNet-50, the performances of linear probing and selective fine-tuning are much closer. In contrast, on ViT and ConvNeXt, selective fine-tuning leads to more significant improvement. This suggests that selective fine-tuning might synergize better with LayerNorm compared to GroupNorm in our experiments.

Comparison with SOTA. We compare our models with state-of-the-art results on CIFAR-10 and CIFAR-100. Results on the more challenging CIFAR-100 are more limited compared to CIFAR-10. As shown in Figure 3, under our parameter-efficient transfer learning scheme based on the previous study, our method consistently outperforms the existing approaches by a large margin, across all measured ϵ values and across both datasets.

4.4. Implementation Details

We use the implementation from the Opacus library [76] to apply DP-SGD for model training. We utilize distributed training and model checkpointing [6] to increase the batch size. For video experiments, we use the PyTorchVideo library [16]. All the experiments are conducted on a Google Cloud instance with 8 Nvidia A100 GPUs. While hyperparameters play a critical role in differentially private training, they are notoriously difficult to tune. In order to facilitate comparison between methods, we pre-define the training epochs for each dataset. And we fix the clipping norm as $C = 1$. We search for the optimal learning rate on the CIFAR-100 dataset with a fixed $\epsilon = 1$. We perform a grid search over the learning rate between $[10^{-4}, 10^{-2}]$.

5. Conclusion and Limitation

We introduce the pioneering application of differential privacy to video classification on the UCF-101 and HMDB-51 datasets. Our integrated approach, featuring the innovative Multi-Clip DP-SGD and parameter-efficient transfer learning tailored for modern video architectures under DP, achieves a remarkable 76% accuracy improvement at $\epsilon = 5$ compared to the direct application of DP-SGD. One limitation of our study is the use of (ϵ, δ)-differential privacy as the sole standard for evaluating privacy. The lack of established criteria for practical ϵ settings necessitates future research to assess models against pragmatic privacy metrics.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, 2016. 1, 2, 3, 4, 5, 6

[2] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*, 2021. 3

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2, 6

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[6] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. 8

[7] Anda Cheng, Jiaxing Wang, Xi Sheryl Zhang, Qiang Chen, Peisong Wang, and Jian Cheng. Dpnas: Neural architecture search for deep learning with differential privacy. *arXiv preprint arXiv:2110.08557*, 2021. 2

[8] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019. 1

[9] Ali Davody, David Ifeoluwa Adelani, Thomas Kleinbauer, and Dietrich Klakow. On the effect of normalization layers on differentially private training of deep neural networks. *arXiv preprint arXiv:2006.10919*, 2020. 2, 3

[10] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022. 1, 2, 3

[11] Friedrich Dörmann, Osvald Frisk, Lars Nørvang Andersen, and Christian Fischer Pedersen. Not all noise is accounted equally: How differentially private learning benefits from large sampling rates. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2021. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 6, 7

[13] Jian Du, Song Li, Moran Feng, and Siheng Chen. Dynamic differential-privacy preserving sgd. *arXiv preprint arXiv:2111.00173*, 2021. 2

[14] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, 2008. 2, 3, 5

[15] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. 2

[16] Haoqi Fan, Tullie Murrell, Heng Wang, Kalyan Vasudev Alwala, Yanghao Li, Yilei Li, Bo Xiong, Nikhila Ravi, Meng Li, Haichuan Yang, Jitendra Malik, Ross Girshick, Matt Feiszli, Aaron Adcock, Wan-Yen Lo, and Christoph Feichtenhofer. PyTorchVideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. `https://pytorchvideo.org/`. 8

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2, 4, 6

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 2, 6

[19] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019. 3, 6

[20] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015. 1

[21] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021. 3

[22] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3, 6

[23] Sara Gerke, Serena Yeung, and I. Glenn Cohen. Ethical and legal aspects of ambient intelligence in hospitals. *JAMA*, 323(7):601, feb 2020. 1

[24] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34:27131–27145, 2021. 3

[25] Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, et al. Towards vision-based smart hospitals: a system for tracking and monitoring hand hygiene compliance. In *Machine Learning for Healthcare Conference*, pages 75–87. PMLR, 2017. 1

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[27] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021. 3

[28] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019. 2, 6

[29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6

[30] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019. 2

[31] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019. 7

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[33] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 5, 6

[34] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 3

[35] Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022. 1, 2, 3

[36] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[37] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021. 3, 5

[38] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 3

[39] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 3

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2, 6

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 1, 2, 4

[42] Zelun Luo, Zane Durante, Linden Li, Wanze Xie, Ruochen Liu, Emily Jin, Zhuoyi Huang, Lun Yu Li, Jiajun Wu, Juan Carlos Niebles, et al. Moma-lrg: Language-refined graphs for multi-object multi-actor activity parsing. *Advances in Neural Information Processing Systems*, 35:5282–5298, 2022. 1

[43] Zelun Luo, Jun-Ting Hsieh, Niranjan Balachandar, Serena Yeung, Guido Pusiol, Jay Luxenberg, Grace Li, Li-Jia Li, N Lance Downing, Arnold Milstein, et al. Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring. *Machine Learning for Healthcare (MLHC)*, 2(1), 2018. 1

[44] Zelun Luo, Daniel J Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *CVPR*, 2021. 2, 6

[45] Zelun Luo, Wanze Xie, Siddharth Kapoor, Yiyun Liang, Michael Cooper, Juan Carlos Niebles, Ehsan Adeli, and Fei-Fei Li. Moma: Multi-object multi-actor activity parsing. *Advances in neural information processing systems*, 34:17939–17955, 2021. 1

[46] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. *Advances in neural information processing systems*, 30, 2017. 3

[47] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *ICLR*, 2018. 5

[48] Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022. 1, 2

[49] Rahul Mehta. Sparse transfer learning via winning lottery tickets. *arXiv preprint arXiv:1905.07785*, 2019. 3

[50] Ilya Mironov. Renyi differential privacy. *CoRR*, abs/1702.07476, 2017. 3, 4

[51] Felix Morsbach, Tobias Dehling, and Ali Sunyaev. Architecture matters: Investigating the influence of differential privacy on neural network design. *arXiv preprint arXiv:2111.14924*, 2021. 3

[52] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018. 5

[53] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 3

[54] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *CCS*, apr 2017. 1

[55] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018. 1

[56] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Ulfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *arXiv preprint arXiv:2007.14191*, page 10, 2020. 2, 3

[57] Sikha Pentyala, Rafael Dowsley, and Martine De Cock. Privacy-preserving video classification with convolutional neural networks. In *International conference on machine learning*, pages 8487–8499. PMLR, 2021. 3

[58] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019. 2, 3

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7, 8

[60] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018. 3

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2, 7

[62] Michael S Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 3

[63] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed System Security Symposium*, 2019. 1

[64] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, may 2017. 1

[65] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 6

[66] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2, 5, 6

[67] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 2

[68] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020. 2, 5

[69] Han Wang, Shangyu Xie, and Yuan Hong. Videodp: A flexible platform for video analytics with differential privacy. *Proc. Priv. Enhancing Technol.*, 2020(4):277–296, 2020. 3

[70] Wenxiao Wang, Tianhao Wang, Lun Wang, Nanqing Luo, Pan Zhou, Dawn Song, and Ruoxi Jia. Dplis: Boosting utility of differentially private deep learning via randomized smoothing. *arXiv preprint arXiv:2103.01496*, 2021. 3

[71] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 2, 6

[72] Yuxin Wu and Justin Johnson. Rethinking" batch" in batchnorm. *arXiv preprint arXiv:2105.07576*, 2021. 6

[73] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[74] Liang Xiao, Xiaoyue Wan, Xiaozhen Lu, Yanyong Zhang, and Di Wu. Iot security techniques based on machine learning: How do iot devices use ai to enhance security? *IEEE Signal Processing Magazine*, 35(5):41–49, 2018. 1

[75] Serena Yeung, Francesca Rinaldo, Jeffrey Jopling, Bingbin Liu, Rishab Mehra, N Lance Downing, Michelle Guo, Gabriel M Bianconi, Alexandre Alahi, Julia Lee, et al. A computer vision system for deep learning-based detection of patient mobilization activities in the icu. *NPJ digital medicine*, 2(1):1–5, 2019. 1

[76] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021. 8

[77] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 2020. 7

[78] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 6

[79] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 8

[80] Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang. Private-knn: Practical differential privacy for computer vision. In *CVPR*, 2020. 2

[81] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 3