VideoSSL: Semi-Supervised Learning for Video Classification

Longlong Jing[1] * Toufiq Parag[2] Zhe Wu[2] Yingli Tian[1] Hongcheng Wang[2]
[1]The City University of New York, [2]Comcast Applied AI Research

Abstract

We propose a semi-supervised learning approach for video classification, VideoSSL, using convolutional neural networks (CNN). Like other computer vision tasks, existing supervised video classification methods demand a large amount of labeled data to attain good performance. However, annotation of a large dataset is expensive and time consuming. To minimize the dependence on a large annotated dataset, our proposed semi-supervised method trains from a small number of labeled examples and exploits two regulatory signals from unlabeled data. The first signal is the pseudo-labels of unlabeled examples computed from the confidences of the CNN being trained. The other is the normalized probabilities, as predicted by an image classifier CNN, that captures the information about appearances of the interesting objects in the video. We show that, under the supervision of these guiding signals from unlabeled examples, a video classification CNN can achieve impressive performances utilizing a small fraction of annotated examples on three publicly available datasets: UCF101, HMDB51, and Kinetics.

1. Introduction

Video understanding has been a topic of interest in computer vision community for many years. Although video understanding and analytic tasks such as action recognition have been pioneered by early classical vision studies [27, 36], the more recent methods have gained much success with CNNs [45, 46]. Among many CNN based algorithms for video classification exploiting different types of information extracted from the video (RGB values or optical flow) and various network architectures (Two stream [42], LSTM [4, 16, 34], 3D CNN [18]), the variants of 3D CNNs utilizing the spatiotemporal features have produced the state of the art results [45, 13, 38, 2, 46, 9].

Similar to other machine learning problems, a large annotated dataset is critical for training CNNs (comprising millions of parameters) to achieve good performance for video classification. In spite of seemingly unlimited number of videos available on the internet, categorizing and curating these videos to create a useful video dataset such as [21, 10, 20] is still expensive and tedious [37]. The labels associated with the videos from social media are often noisy and need to be corrected manually. In addition, some videos require trimming as the action or video event often does not span through the video length [21].

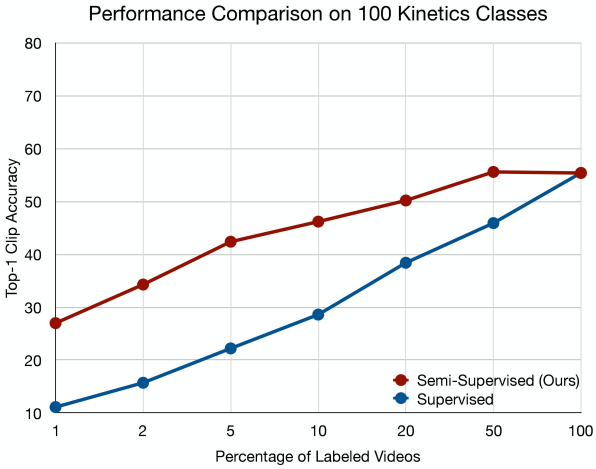

Figure 1: Video classification accuracy as a function of the fraction of labeled videos. With a small percentage of labeled examples, the 3D CNN trained by our proposed semi-supervised method significantly outperforms that trained in supervised setting.

In order to reduce the dependence on annotated datasets, several studies have investigated pretraining features with millions of web videos in a weakly supervised fashion where the video labels are noisy [7, 8, 20]. After feature learning, these methods finetune the overall network on the target dataset in a fully supervised fashion. Others have employed self-supervision for video feature learning [12, 31].

However, both finetuning (after pertaining) and self-supervised methods assume the existence of a high quality labeled dataset which incurs the aforementioned costs. A semi-supervised learning (SSL) method, on the other hand, can reduce these costs by requiring fewer annotated training examples from target dataset. Several methods for semi-supervised learning in 2D image domain have reported very

*The work was partially done at Comcast Applied AI Research, Washington, DC.

Figure 2: A single frame from some selected videos in Kinetics dataset. For these categories, object appearance in one single frame provides sufficient information to categorize them as playing instrument or sport (top row) and eating (bottom row) [41, 17].

promising results [44, 32]. A recent survey by [37] compares the performances of these methods as well as suggests scenarios where SSL is a better choice than pretraining or self-supervised methods. However, the 3D video domain has not observed significant interest/number of works in semi-supervised setting. Yalniz et al. [50] utilized pseudo-labels for 3D classification although the algorithm heavily relies on a large annotated dataset (e.g., Kinetics [21]) to train a strong teacher model.

In this paper, we propose a semi-supervised method, VideoSSL, for video classification with spatiotemporal networks. Given a small fraction of the annotated training samples, our proposed method leverages two supervisory signals extracted from the unlabeled data to enhance classifier performance. As the first supervisory signal, we use pseudo-labels [29] of the unlabeled data – a technique that has been demonstrated to be highly effective on 2D images – for semi-supervised learning of 3D video clips. We utilize the appearance cues of objects of interest, distilled by the prediction of a 2D image classifier CNN on a random video frame, as the second regularizer for VideoSSL.

Many, if not all, actions can be decomposed as one or more objects (noun) performing an activity (verb) [7]. Consequently, a hint about the object (noun) appearance can offer a very strong indication of the actions being performed in the video clip [17, 41]; we illustrate this insight with examples of actions in Figure 2. Girdhar *et al.* [8] harnessed the appearance information in the form of the output probabilities of a 2D image classifier for *pretraining* the spatiotemporal feature representation. Our algorithm proposes to use the predictions of 2D image classifiers as regulatory information for *semi-supervised* training of 3D CNNs or their variants. In addition, we show that the capability of the video classifier can be further magnified by the incorporation of a semi-supervised technique, in particular the pseudo-label method [29].

We have tested our method on three most widely used datasets UCF101 [43], HMDB51 [25] and Kinetics [21]. On all the datasets, our proposed algorithm consistently outperforms those trained by the supervised algorithm from a small fraction of annotated examples. The video classifiers learned by the proposed method can attain up to 20% higher accuracy than those of the classifiers trained by a fully supervised approach from limited data. Figure 1 depicts a sample comparison between the performances of two networks trained by the proposed and supervised strategies. More interestingly, our proposed method needs only $10 \sim 20\%$ of the labeled data to produce a 3D CNN to match or supersede the accuracy of another network with the same architecture but trained from the whole dataset in a previous study [13]. Our proposed technique can be generally applied to learn any 3D CNN variants for video classification.

This work contributes to the overall effort of video event recognition in multiple directions. We propose an accurate and robust semi-supervised training algorithm for 3D CNNs (or its variants) for video classification. We experimentally demonstrate that a straightforward execution of semi-supervised method does not yield a 3D video classifier with satisfactory performance. On the other hand, a calibrated utilization of the object appearance cues for semi-supervised learning profoundly improves the accuracy of the resulting model. We validate the utility and consistency of our technique by reporting improved performances on different public datasets through rigorous testing under many different configurations.

2. Related Work

Video Classification: Early studies on action recognition relied on hand designed features and models [36, 35, 27, 28, 30, 48]. Recently various networks have been proposed to capture both the spatial and temporal information for video classification tasks including: 2D CNN-based methods [20, 42, 49], RNN-based methods [4], and 3D CNN-based methods [45, 38, 2, 46, 47, 9]. Some interesting analytical studies have recently investigated which categories of videos require temporal information for recognition [41, 17].

The 3D CNNs and their variants have made significant progress in video classification by simultaneously capturing spatial and temporal information [2, 45, 38, 46, 9, 5]. However, due to the extra temporal dimension, the 3D CNNs usually have millions of parameters which may leads overfitting when trained on small datasets. For that, in addition to learning from larger datasets like Kinetics [21], there have been multiple efforts to pretrain the feature representations from millions of weakly annotated videos [8, 7].

Semi-Supervised Learning: Semi-supervised learning is a technique to train the network both with labeled and unlabeled data [26, 40, 32, 44, 29, 37]. Recently, several semi-

supervised learning methods have been proposed for image classification. Considering the different random data augmentations to input data and CNN configurations under dropout selection as noise to the learning process, [26, 40] introduced a consistency loss between the network outputs from the same input sample at different training iterations, or their moving averages, as a regularization term for semi-supervised learning. In addition, Tarvainen and Valpola [44] proposed to utilize a teacher model obtained from moving averages of past network weights to calculate a more 'stable' prediction. VAT is proposed by Miyato *et al.* to model the perturbations that added to the data which most significantly affect the output of the prediction function [32]. Grandvalet and Bengio suggested minimizing entropy of the model predictions to generate more confident predictions [11] whereas pseudo-label proposed to use the label predicted with highest confidence as the true label of the example for training [29]. Most of these methods have been tested on small datasets including CIFAR10 [24] and SVHN [33], but their ability to adapt to large datasets has not been investigated yet.

Semi-supervised learning of CNNs for 3D tasks has not yet received considerable interest in the community. A preliminary study by Zeng *et al.* [51] employed an encode-decoder framework for action recognition but tested only on toy datasets containing few tens of images. The work of [1] pretrains the feature representation through adversarial training and fine-tunes the discriminator on the target dataset; it does not learn the CNN in a semi-supervised manner. Yalniz *et al.* proposed to employ pseudo-label methods for semi-supervised learning while the teacher network is trained on large-scale weakly labeled videos or images to obtain better performance [50]. In our work, we experimentally demonstrate that the 2D semi-supervised learning techniques do not yield a satisfactory performance when directly extended to 3D network and therefore not useful.

Self-Supervised Learning: Self-supervised learning is another trend of approach to learn visual features from unlabeled data [23, 19, 31, 22]. For learning video features from unlabeled videos, a network is trained to solve a pretext task and the label for pretext tasks are generated based on the attribute of the data. Various pretext tasks have been proposed to learn visual features from videos. Misra *et al.* [31] proposed to train a network to verify whether the input frame sequence is in correct temporal order or not. Korbar *et al.* [23] proposed to train a network by verifying whether the input video segment and audio segment are temporally correspondent or not. A recent study by Zhai *et al.*, combines the self-supervision with semi-supervised learning [52]. However, this method was designed for and tested on 2D images only.

Knowledge Distillation: Hinton *et al.* [15] originally proposed to transfer the knowledge from several deep networks to one smaller network by optimizing the KL divergence of the distributions of the networks. Radosavovic *et al.* [39] proposed to distill knowledge from unlabeled data by using the prediction of a network whereas Garcia *et al.* [6] propose to jointly transfer knowledge of different modalities to one modality. The work of [8] suggested distilling the appearance information of the objects of interest in the video through the output of a 2D image classification network for pretraining the 3D features of a video classifier. The 3D classifier is then finetuned on the target dataset using all its annotation. The proposed algorithm, on the other hand, uses the appearance information for semi-supervised training with a small fraction annotated samples from a dataset – it does not require the target dataset to be exhaustively annotated. Such an approach could be beneficial for scenarios where collecting and annotating data is difficult and costly [37].

3. VideoSSL Training

Our proposed algorithm VideoSSL trains a 3D CNN for video classification in a semi-supervised fashion. Motivated by the impressive performance of spatiotemporal 3D CNNs and their variants [46, 38, 2, 13, 9], we used a 3D ResNet [13] that computes the (softmax) probabilities of different video classes. It is worth pointing out that VideoSSL method can be used to learn any 3D CNN and its variants. In our semi-supervised setting, the softmax probabilities from a 2D image classifier are utilized as a teaching signal to the training of 3D CNN. In the learning phase, the 3D CNN is designed to produce another output, which we also referred to as an embedding, with the same dimensionality as the 2D network output.

The 3D CNN is trained by jointly minimizing three loss functions. The cross-entropy loss with respect to the labels of a small percentage of data points is backpropagated to update the weights of the 3D CNN. In addition, we also backpropagate the loss against the pseudo-labels [29] computed by the 3D CNN on unlabeled examples. The third loss, which facilitates the knowledge distillation, is computed between the 2D image network prediction and the embedding from the 3D CNN computed for both labeled and unlabeled data. A schematic diagram of the whole training process is presented in Figure 3 and we describe the losses used in VideoSSL in the following sections.

3.1. Learning from Labeled Data

Let $X = \{x_1, \ldots, x_K\}$ denote the annotated video clips with corresponding category indicators $\{y_1, \ldots, y_K\}$ and $Z = \{z_1, \ldots, z_U\}$ be the unlabeled data in a batch of training examples. If there are C video categories, i.e., $y_i \in \{0, 1\}^C$, for any input video clip x_i, the 3D network produces a softmax probability $p(x_i) \in \mathbf{R}^C$ for x_i to be-

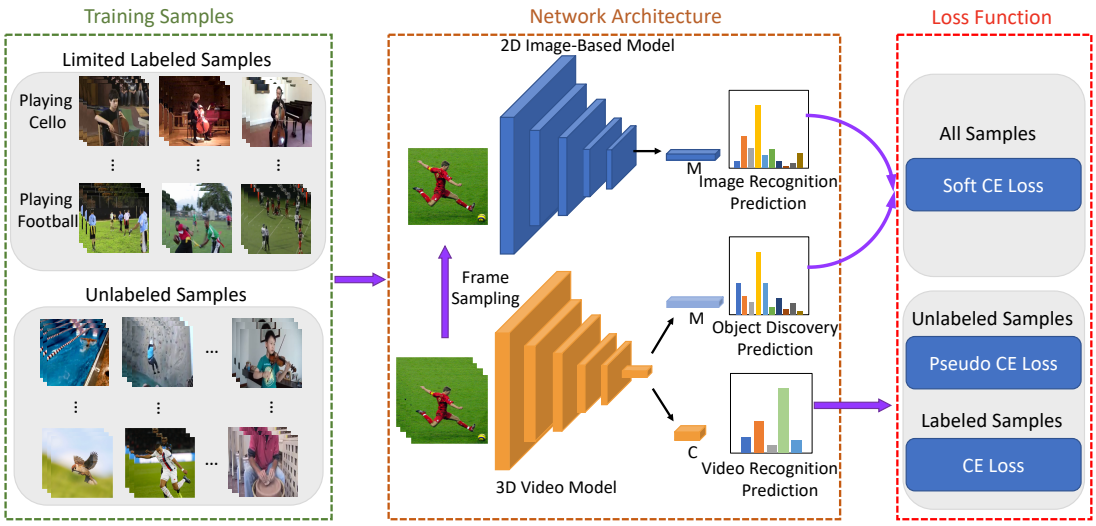

Figure 3: The framework of the proposed video semi-supervised learning approach. The 3D network is optimized with three loss functions: 1) **CE Loss**: the video cross-entropy (CE) loss on the labeled data which paired with human-annotated labels, 2) **Pseudo CE Loss**: the pseudo cross-entropy loss on the pseudo-labels of unlabeled data, and 3) **Soft CE Loss**: the soft cross-entropy loss on both unlabeled and labeled data to teach the video classification network to capture the appearance information.

long to any of the C classes. Given the small set of examples X, the first loss 3D ResNet training minimizes is the cross-entropy loss.

$$L_s = -\sum_{x_i \in X} \sum_{c} y_i^c \log p^c(x_i) \tag{1}$$

Here we omit the weight/parameter variables from the loss functions for better readability.

3.2. Learning with Pseudo-Labels of Unlabeled Data

Given a set of unlabeled examples Z, the method of pseudo-label computes an estimate of their true labels from the prediction of a classifier and use it to train the classifier itself [29]. In our proposed training, the estimated label $\hat{y}_i^c$ of z_i for class c is assigned 1 if the prediction confidence $p^c(z_i)$ from 3D CNN on unlabeled sample z_i exceeds δ. A large δ enforces the algorithm to select highly confident samples; for such samples, predictions for less confidence classes become extremely small. As explained later (Section 3.4), we learn the network for a sufficient number of iterations before using its predictions for the pseudo-label approach. The resulting cross-entropy loss against the pseudo-labels can be formulated as follows.

$$\hat{y}_i^c = \begin{cases} 1, & \text{if } p^c(z_i) \geq \delta \\ p^c(z_i), & \text{otherwise} \end{cases} \tag{2}$$

$$L_u = -\sum_{z_i \in Z} \sum_{c} T\ \hat{y}_i^c \log p^c(z_i). \tag{3}$$

In Equation 3, T is a predefined weight we use to emphasize on the impact of confident samples (i.e., $p^c(z_i) \geq \delta$). We randomly select half of the examples in a batch from annotated examples and remaining half from examples without annotation.

3.3. Knowledge Distillation for All Data

As several studies have already reported, appearance information can provide a strong cue for video/action recognition [17, 41, 8]. Our method seeks to distill the information about the appearances of the objects of interest in the video by exploiting the softmax predictions of a 2D ResNet [14] image classifier. The 2D ResNet we apply has already been trained on the ImageNet dataset [3] and its weights stay fixed throughout training and testing. For our VideoSSL approach, we distill the appearance information from both labeled and unlabeled video clips.

Given an image (or frame) a from any video, let us denote the output of the 2D ResNet as $h(a) \in \mathbf{R}^M$, where $M = 1000$ for networks trained on ImageNet. In our experiments, we have randomly selected the frame a from a video clip, both for training and testing.

For each video $v \in \{X \cup Z\}$, the 3D ResNet also produces another embedding $q(v) \in \mathbf{R}^M$ whose dimension matches that of the output of $h(a)$. During training we enforce the embedding from video classifier $q(v)$ to match the output of image classifier $h(a)$ when a is a frame selected from v. The distillation loss utilized for this purpose is a soft cross-entropy loss that treats the 2D ResNet predictions as soft labels.

$$L_d = -\sum_{\substack{v \in \{X \cup Z\} \\ a \in v}} \sum_{l=1}^{M} h^l(a) \log q^l(v) \tag{4}$$

We are using a knowledge distillation formula similar to that employed in [8]. However, as we explain in Section 3.4, our proposed VideoSSL method learns the overall 3D CNN (not just the features) by minimizing the distillation loss in conjunction with the supervised and pseudo-label losses in a semi-supervised fashion. This approach is fundamentally different from the feature learning of [8] for pretraining video classifiers.

3.4. Combined Loss Function

The overall training process trains the 3D network with a combined loss.

$$L = L_s + \lambda_u L_u + \lambda_d L_d. \quad (5)$$

The balancing weight for the pseudo-labels uses warm-up so that $\lambda_u = 1$ after a certain number of training iterations τ. With a sufficiently large τ, we can train the 3D CNN long enough to produce some meaningful predictions for pseudo-labels. The $\lambda_d = 1$ for all our experiments.

4. Experimental Results

In this section, we conduct extensive experiments to evaluate the proposed approach and compare with other semi-supervised learning methods from 2D image domain applied to video data. Our semi-supervised learning framework is trained and tested on several widely used datasets for video classification including: UCF101 [43], HMDB51 [25], Kinetics [21]. In the following, we first describe our experimental setting and network architecture & training before reporting performances on these 3 datasets.

4.1. Implementation Details

We have used 3D ResNet-18 [13] as a video classifier in all our experiments. This 3D ResNet architecture is very similar to the 2D ResNet [14], except all the convolutions are performed in 3D. That is, it has 4 convolutional blocks with different numbers of 3D convolutions (within the block) based on the ResNet. We have primarily experimented to 3D ResNet-18 (each block with two 3D convolutions) with 64, 128, 256, 512 feature maps. The 3D ResNet-18 has a C class output for video categories. During training, it also produces a $M = 1000$ length embedding for each video. The 2D ResNet-50 image classifier is collected from the Pytorch repository. Our implementation was built around the code released by [13].

The videos from all the datasets are resized to a spatial resolution at 136×136. During training, 16 consecutive frames are randomly selected from each video as a training clip and a 112×112 patch is randomly cropped from each frame to form an input clip. The size of the input becomes 3 channels $\times$ 16 frames $\times$ 112 $\times$ 112 pixels. The input to the 3D ResNet-18 was also normalized by the mean and variance of the sport-1M dataset. We used random crop and temporal jittering for data augmentation in all our experiments. The input size and data preprocessing strategies are very similar to existing studies [46, 8, 7].

All the models are trained on different percentage of labeled data. We have randomly selected different percentages P of labeled examples from each of the datasets, e.g., $P \in \{5, 10, 20, 50\}$. In our VideoSSL training, we used P percentage of labeled data to compute the supervised loss in Equation 1. Annotations for all remaining examples were ignored in the semi-supervised setting and treated as unlabeled examples. Given the split of annotated and unannotated examples, our VideoSSL learning minimizes the joint loss in Equation 5 to learn a 3D CNN from scratch. For all the experiments on the same dataset, the same testing splits are used for fair comparison.

We have used the Stochastic Gradient Descent (SGD) with momentum 0.9 and weight decay 0.001 as a minimizer for the joint loss. The initial learning rate during learning was set to 0.01 and was decreased by a factor of 10 every 40000 iterations. The batch size for every optimization step was 128 distributed among multiple GPUs. For pseudo-label technique, T and τ were set to 10 and $\frac{2}{3}$ of the total iterations respectively. Also, the prediction threshold was set to $\delta = 0.95$ based on the suggestion of [37]. In practice, any value ≥ 0.9 produced very similar accuracy values.

For all the experiments below, we report the Top-1 clip and video accuracy values on the validation or test datasets. After training, the prediction of the 3D ResNet on the center video clip (both spatial and temporal) is reported as the clip Top-1 accuracy. The video accuracy is the average of the classifier confidences on all consecutive non-overlapping clips within the video.

4.2. Baseline Methods

In all our experiments on different datasets, we have compared the performance of the CNN trained by the proposed algorithm to those trained by different methods as well as their combinations listed below. Unless otherwise mentioned, the same experimental setup was maintained for all the experiments.

1. Supervised baseline (Supervised) learns the 3D Resnet18 [13] from only the labeled examples.
2. MeanTeacher (MT) applies the method of [44] on video data.
3. PseudoLabel (PL) applies the technique of [29] on video data.
4. Supervised with Distillation (SD) uses the knowledge distillation loss, as described in Section 3.3, along with the supervised loss for the training.
5. The self-supervised and semi-supervised learning method (S^4L) of [52] extended to video data. We adopt the S^4L-rotate strategy originally proposed for 2D images for 3D videos. In particular, we minimize

Table 1: The performance comparison on UCF-101 dataset. All values reported are Top-1 accuracy values. The proposed method consistently outperforms all the other methods.

%Label	Supervised[13]		PL[29]		MT[44]		SD		MT+SD		S^4L [52]		Ours	
	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video
5	15.1	16.9	17.2	17.6	15.3	17.5	29.3	31.2	28.4	30.3	21.0	22.7	30.9	**32.4**
10	21.6	24.0	23.5	24.7	24.0	25.6	38.6	40.7	37.5	40.5	27.1	29.1	40.2	**42.0**
20	30.0	32.2	33.9	37.0	33.4	36.3	42.1	45.4	41.7	45.5	34.7	37.7	46.2	**48.7**
50	35.1	38.3	43.9	47.5	42.5	45.8	49.8	53.9	49.2	53.0	44.9	47.9	51.5	**54.3**

the cross-entropy loss on labels and rotations of the annotated and unlabeled videos respectively in our experiments.

For supervised learning, we used only the labeled examples, as given by the percentage P, to train the CNN from scratch.

4.2.1 Results on UCF101 Dataset

Dataset: UCF101 is a widely used dataset for human action recognition [43]. It consists of $13,320$ videos belong to 101 action classes and contains approximately 130 videos for each class. Although relatively small in size, it is a balanced dataset and each class has around 100 videos for training. Videos have the spatial resolution of 240 pixels and 25 FPS frame rate. There are three training/testing splits available for this dataset, and the split 1 is used for all the experiments in our paper.

Performance Comparison: Table 1 shows the clip and video Top-1 accuracy of our proposed method and the baselines for video classification with 3D ResNet-18. As shown in the table, our proposed strategy amplifies the video Top-1 accuracy of the 3D ResNet-18 by more than 16% with $\{5\%, 10\%, 20\%, 50\%\}$ annotated samples. Across all percentages of labeled data, our algorithm produces the most accurate classifier among all other techniques.

These experiments also suggest that the straightforward application of the existing semi-supervised methods PL [29] and MT [44] to 3D video classifier is not beneficial. It is interesting to observe that the accuracy of MT is similar or worse than PL, which contrasts the findings of [37] albeit for 2D images. However, as [52] points out, such an outcome has been observed in practice before. The adaptation of knowledge distillation [8] is instrumental in achieving good performances for semi-supervised learning from a limited percentage of data. The combination of the semi-supervised PL technique to knowledge distillation further improves the accuracy of the resulting 3D CNN by contributing additional information to the training process.

Perhaps the most compelling outcome of our experiments is, with only 10% of annotated data the proposed method can achieve the same video Top-1 accuracy of the 3D ResNet-18 trained from scratch in a fully supervised manner in [13]. With 50% labeled examples, the proposed approach produces a 12% more accurate CNN.

4.2.2 Results on Kinetics Dataset

Dataset: Kinetics is a large-scale dataset for video understanding tasks [21]. The Kinetics-400 version provides $306,245$ 10-second training videos for 400 action classes. Since many videos are not available on the YouTube any more, we were able to download $226,127$ and $18,613$ videos for training and validation respectively. This dataset is significantly larger than UCF-101 and has become increasingly popular in the action recognition community [7, 13, 8, 5].

The distribution of videos across different categories is not balanced in Kinetics-400. Some classes in this dataset contain over 900 videos whereas more than 80 classes contain less than 300 videos. We compared the performances of the proposed and the baseline methods in two settings. The first experiment attempts to create a data subset where each activity class has at least 700 training videos. Consequently, the training methods working on this (more) balanced subset will have access to substantial amount labeled examples. This subset of Kinetics dataset contains 100 classes and is referred to as Kinetics-100 in this paper. The second experiment compares the performances of the proposed algorithm and baselines on the whole Kinetics-400 dataset.

Performance Comparison: As shown in Table. 2, our method consistently improve the accuracy of the 3D ResNet over that trained by the supervised method by a significant amount on the Kinetics-100 subset. The improvement over the supervised method reduces from roughly 20% to 10% in video Top-1 accuracy when the labeled data increases from 5% to 50%. It is expected that the difference in accuracy between semi and fully supervised methods will decrease with the increase of labeled data. The results suggest that the off the shelf application of the existing semi-supervised methods (PL and MT) offer little benefit to video classification of Kinetics dataset as well.

The proposed method can achieve a higher video Top-1 accuracy of the 3D ResNet-18 trained by fully supervised training in [13] with only 20% of annotated data in the Kinetics dataset. The accuracy of the proposed method is higher than all the other baselines evaluated on Kinetics-100 on all fraction of labels used.

Table 3 shows that our proposed semi-supervised train-

Table 2: The performance comparison on Kinetics-100 dataset. The proposed method consistently improves both the clip and video Top-1 classification accuracy and outperforms all other methods.

%Label	Supervised[13]		PL[29]		MT[44]		SD		MT+SD		S^4L [52]		Ours	
	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video
5	23.6	27.2	24.8	27.8	23.8	27.8	40.2	45.2	40.8	46.6	29.6	33.0	43.1	**47.6**
10	31.2	36.3	34.6	38.9	31.5	36.4	44.7	49.8	43.9	49.4	37.5	43.3	48.4	**52.6**
20	40.7	46.8	41.8	48.0	40.8	47.1	49.8	55.6	50.0	55.3	44.7	51.1	51.3	**57.7**
50	49.6	55.5	51.2	59.0	51.2	59.3	57.3	63.8	57.6	63.9	49.1	54.6	58.2	**65.0**

Table 3: The performance comparison on the whole Kinetics-400 dataset. The proposed method outperforms multiple baselines.

%	Supervised[13]		MT[44]		SD		MT+SD		Ours	
	clip	video	clip	video	clip	video	clip	video	clip	video
10	17.3	20.7	16.2	19.5	25.1	31.5	26.3	30.6	30.0	**33.8**
20	24.2	29.6	23.3	20.0	30.4	35.9	31.9	37.0	33.3	**38.5**
50	34.8	41.8	34.3	41.8	37.0	46.6	39.4	46.1	40.2	**47.0**

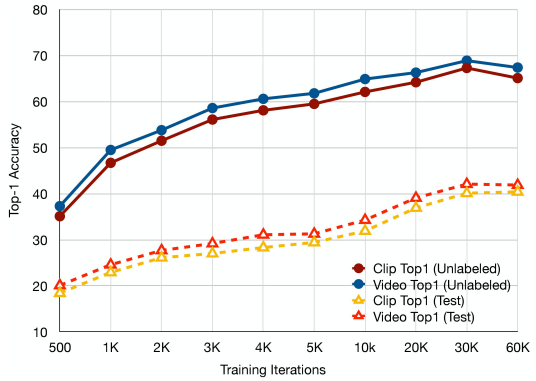

Figure 4: Progression of clip and video Top-1 accuracy of 3DCNN trained by the proposed algorithm on the unlabeled training samples (solid lines) and test split (dashed lines) of UCF101 dataset with training iterations.

ing algorithm can improve the performance of the 3D CNN over those trained by baseline techniques on the whole Kinetics-400 as well. Since many categories in the full Kinetics-400 dataset contain fewer than 300 videos, we conduct our experiments on $\geq 10\%$ samples. With the experimental setup unchanged, the PL [29] strategy produced unacceptably low accuracy on Kinectics-400. Consequently, we excluded PL results in Table 3 as we cannot explain this outcome.

4.2.3 Results on HMDB51 Dataset

Dataset: HMDB51 is another widely used dataset for human action recognition [25]. It consists of $6,770$ videos belong to 51 action classes and each class has roughly 70 videos for training. There are three splits available for this dataset and we used split 1 for all our experiments. In spite of the smaller size compared to UCF101 and Kinetics, the performances of the existing techniques have been lower than those on the other two datasets [13, 15, 45]. This implies a higher complexity to deal with HMDB51 with respect to the other datasets.

Performance Comparison: Due to the relatively small size of HMDB51 dataset, the performances of our proposed method compared against the baseline methods on $\{40\%, 50\%, 60\%\}$ annotated examples instead.

Table 4 compares the performances of the proposed algorithm and the baseline methods. The findings from this experiment conform almost exactly to those from the UCF101 and Kinetics – our VideoSSL trained 3D CNNs from different percentages of annotations that are consistently superior to those trained by the supervised, exiting semi-supervised and also the self semi-supervised techniques. Likewise, our approach produced a 3D ResNet-18 more accurate than that trained by [13] with only 50% of annotations.

4.3. Analysis of Training

The success of a semi-supervised method relies heavily on how well it learns to classify the unlabeled samples during the training process. In Figure 4, we plot the accuracy progression of the CNN under training on the unlabeled training data as well as the test data at different training iterations. This experiment was performed on 10% labeled examples of UCF101 dataset. The plot clearly illustrates how the performance of the CNN was improved by the proposed method over the training process on both clip and video classifications.

Figure 6 plots the category-wise increase the classification accuracy (clip Top-1) of the network trained by our method compared to that trained by the supervised approach with 10% labels of UCF101. As seen on the plot, the performance of the 3D ResNet learned by our method improved for 90% of the categories. Example classes such as Boxing Speed Bag (+48.6), Playing Tabla (+48.4), Sumo Wrestling (+44.1), Rafting(+42.3), Bench Press (+39.6) imply the appearance information of the objects of interest in the video played a major role in this improvement.

There are categories in both the UCF101 and Kinetics100 datasets where the proposed VideoSSL obtained better classifiers (with 10% labels) than the SD method that

Table 4: The performance comparison on HMDB51 dataset. The proposed method consistently improves both the clip and video Top-1 classification accuracy.

%Label	Supervised[13]		PL[29]		MT[44]		SD		MT+SD		S^4L [52]		Ours	
	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video	clip	video
40	17.1	18.0	26.3	27.3	26.4	27.2	31.6	32.6	32.1	32.3	28.8	29.8	32.6	32.7
50	29.1	30.7	30.9	32.4	29.2	30.4	34.1	35.1	30.8	33.6	28.9	31.0	34.9	**36.2**
60	30.0	31.2	31.4	33.5	31.1	32.2	35.4	36.3	34.5	35.7	32.5	35.6	35.7	**37.0**

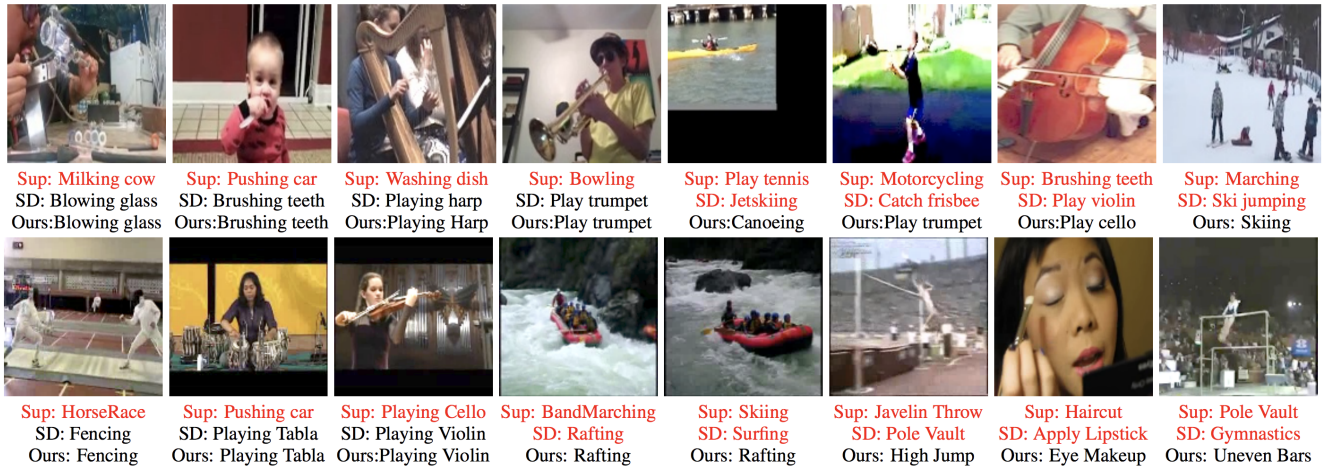

Figure 5: Qualitative comparison of our algorithm with baseline methods, top row: Kinetics100, bottom row: UCF101. Each image is a frame from a video that was correctly classified by the 3D CNN learned by the proposed method from 10% examples. Sup, SD, and ours refer to the predictions of the supervised [13], SD and proposed method respectively. The predictions from supervised CNN appear to be arbitrary compared to the video category whereas those from SD seem to capture and exploit the scene characteristics.

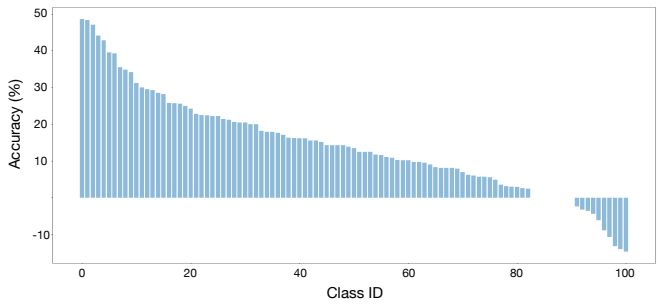

Figure 6: Per-class (x-axis) accuracy improvements (y-axis) on UCF101 dataset between the network trained by the baseline supervised and our proposed method. For 90% of the classes, the proposed algorithm can significantly boost the performance of the learned model.

partially utilizes the object appearance cues. Figure 5 shows some representative frames from these classes from both these datasets. As can be expected, the mis-classification of the supervised method [13] appears to be rather arbitrary with respect to the actual categories. SD, on the other hand, classifies these video into (wrong) categories with very similar scene characteristics. Examples of SD misclassification predict throwing frisbee for passing football or jetskiing for canoeing. The proposed technique utilized additional knowledge supplied by the pseudo-label method to resolve the confusion and achieve a superior performance on these categories.

5. Conclusion

This study proposes a new algorithm for semi-supervised learning of video classifier. We show in this work that a straightforward application of the existing semi-supervised methods (that are originally developed for 2D images) cannot achieve satisfactory performance for 3D video classification. The proposed method exploits the appearance information of the object of interest in video to produce highly accurate 3D classifiers given limited annotated examples. From only 20 $\sim$ 50% annotated samples, the proposed approach can learn CNNs that can potentially outperform those trained in a fully supervised manner. We have tested the accuracy and robustness of our algorithm on three most widely used datasets with different percentages of training labels and compared against the several baseline combinations. We hope that our proposed learning strategy will be useful for reducing the costs for creating a training dataset for video understanding and will instigate more efforts on semi-supervised video training.

Acknowledgement. This material is partially based upon the work supported by National Science Foundation (NSF) under award numbers IIS-1400802 and IIS-2041307. The authors also thank Shing Chau, Chinar Dingankar, Kenneth Tran, Ruth Dawson and William McMaster from Comcast Labs for their help with resources necessary to run the experiments.

References

[1] Unaiza Ahsan, Chen Sun, and Irfan Essa. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv preprint arXiv:1801.07230*, 2018.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[6] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[7] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.

[8] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[9] Felix Gonda, Donglai Wei, Toufiq Parag, and Hanspeter Pfister. Parallel separable 3d convolution for video and volumetric data understanding. In *BMVC*, 2018.

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 3, 2017.

[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[12] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. *arXiv preprint arXiv:1909.04656*, 2019.

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *CVPR*, pages 18–22, 2018.

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

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2013.

[19] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint arXiv:1902.06162*, 2019.

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[22] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. *arXiv preprint arXiv:1811.09795*, 2018.

[23] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NIPS*, pages 7774–7785, 2018.

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[25] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *HPCSE*, pages 571–582. Springer, 2013.

[26] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

[27] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[28] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[29] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

[30] J. Liu, Jiebo Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[31] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *ECCV*, 2016.

[32] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE*

transactions on pattern analysis and machine intelligence, 41(8):1979–1993, 2018.

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[34] Joe Yue-Hei Ng, Matthew J. Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond Short Snippets: Deep Networks for Video Classification. In *CVPR*, 2015.

[35] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[36] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[37] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018.

[38] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[39] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2018.

[40] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016.

[41] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. *arXiv preprint arXiv:1907.08340*, 2019.

[42] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*. 2014.

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR*, 12-01, 2012.

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015.

[46] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[47] G"ul Varol, Ivan Laptev, and Cordelia Schmid. Long-term Temporal Convolutions for Action Recognition. *TPAMI*, 2017.

[48] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC* , London, United Kingdom, 2009.

[49] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *ECCV*, 2016.

[50] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[51] Ming Zeng, Tong Yu, Xiao Wang, Le T Nguyen, Ole J Mengshoel, and Ian Lane. Semi-supervised convolutional neural networks for human activity recognition. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 522–529. IEEE, 2017.

[52] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. *arXiv preprint arXiv:1905.03670*, 2019.