



Adversarial Self-supervised Learning for Semi-supervised 3D Action Recognition

Chenyang Si^{1,2,3} , Xuecheng Nie³, Wei Wang^{1,2}, Liang Wang^{1,2},
Tieniu Tan^{1,2}, and Jiashi Feng³

¹ University of Chinese Academy of Sciences, Beijing, China

² CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences,
Beijing, China

{wangwei,wangliang,tnt}@nlpr.ia.ac.cn, chenyang.si@cripac.ia.ac.cn

³ Department of ECE, National University of Singapore, Singapore, Singapore
nioxuecheng@u.nus.edu, elefjia@nus.edu.sg

Abstract. We consider the problem of semi-supervised 3D action recognition which has been rarely explored before. Its major challenge lies in how to effectively learn motion representations from unlabeled data. Self-supervised learning (SSL) has been proved very effective at learning representations from unlabeled data in the image domain. However, few effective self-supervised approaches exist for 3D action recognition, and directly applying SSL for semi-supervised learning suffers from misalignment of representations learned from SSL and supervised learning tasks. To address these issues, we present Adversarial Self-Supervised Learning (ASSL), a novel framework that tightly couples SSL and the semi-supervised scheme via neighbor relation exploration and adversarial learning. Specifically, we design an effective SSL scheme to improve the discrimination capability of learned representations for 3D action recognition, through exploring the data relations within a neighborhood. We further propose an adversarial regularization to align the feature distributions of labeled and unlabeled samples. To demonstrate effectiveness of the proposed ASSL in semi-supervised 3D action recognition, we conduct extensive experiments on NTU and N-UCLA datasets. The results confirm its advantageous performance over state-of-the-art semi-supervised methods in the few label regime for 3D action recognition.

Keywords: Semi-supervised 3D action recognition · Self-supervised learning · Neighborhood consistency · Adversarial learning

1 Introduction

Recently, 3D action recognition (a.k.a. skeleton-based action recognition) has made remarkable progress through learning discriminative features with effective networks [7, 12, 18, 29, 30, 44, 47]. However, these methods heavily rely on the available manual annotations that are costly to acquire. Techniques requiring less or no manual annotations are therefore developed, and among them a

powerful approach is semi-supervised learning. It is aimed at leveraging unlabeled data to enhance the model’s capability of learning and generalization such that the requirement for labeled data can be alleviated. It has been widely applied in the image domain [14–16, 24, 25, 27, 34]. Compared with these methods, [45] has recently proposed a more efficient way of feature learning from unlabeled data, namely self-supervised semi-supervised learning (S^4L), that couples self-supervision with a semi-supervised learning algorithm. It employs the self-supervised technique to learn representations of unlabeled data to benefit semi-supervised learning tasks. Self-supervised learning is very advantageous in making full use of unlabeled data, which learns the representations of unlabeled data via defining and solving various pretext tasks. Thus in this work we exploit its application to semi-supervised 3D action recognition, which has little previous investigation.

As images contain rich information that is beneficial to feature extraction, many effective SSL techniques [5, 37, 42] are image-based. Comparatively, for tasks over skeleton data which represent a person by 3D coordinate positions of key joints, it becomes very challenging to leverage SSL techniques to learn discriminative motion representation. Therefore, how to learn motion representation with SSL technique is an urgent problem for this task. Recently, [48] proposes a SSL method to learn temporal information of unlabeled sequence via skeleton inpainting. This SSL treats each sample as an *individual* such that it ignores the shared information among samples with the same action class. As a result, semi-supervised 3D action recognition has derived little benefit from the representations learned by skeleton inpainting.

Moreover, we also find that, directly applying SSL for semi-supervised learning suffers from misalignment of representations learned from self-supervised and supervised learning tasks. As shown in Fig. 1, labeled and unlabeled samples are enforced with supervised and self-supervised optimization objectives respectively. Though both sampled from the same data distribution, their feature distributions are misaligned. This misalignment would weaken the generalization of semi-supervised 3D action recognition models to unseen samples. A task with similar problem as ours is unsupervised domain adaptation (UDA) that matches the feature distributions from *different domains*. While their problem is quite similar to ours, there exist important differences between UDA and our task. In UDA, the discrepancy of feature distributions is caused by different

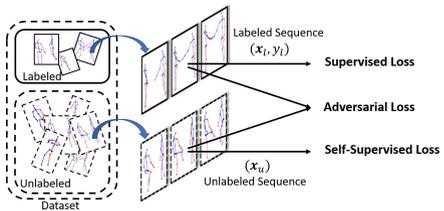


Fig. 1. Illustration of our main idea. We design an effective SSL scheme to capture the discriminative motion representations of unlabeled skeleton sequences for 3D action recognition. Since directly applying SSL to semi-supervised learning suffers from misalignment of representations learned from SSL and supervised learning tasks, we further pioneer to align their feature distributions via adversarial learning

domains. Our problem is the misalignment of representations learned from SSL and supervised learning tasks in semi-supervised 3D action recognition. One line of research in UDA is adversarial-based adaptation methods [9, 20, 35] that have shown promising results in domain adaptation. These methods seek to minimize an approximate domain discrepancy distance through an adversarial objective with respect to a domain discriminator. Hence, inspired by the alignment effect of adversarial learning in UDA, we exploit its application to couple the self-supervision method into a semi-supervised learning algorithm.

In this work, we propose an Adversarial Self-Supervised Learning (ASSL) Network for semi-supervised 3D action recognition. As shown in Fig. 1, our model leverages (i) self-supervised learning to capture discriminative motion representation of unlabeled skeleton sequences, and (ii) adversarial regularization that allows to align feature distributions of labeled and unlabeled sequences. More specifically, in addition to a self-inpainting constraint [48] for learning temporal information of each *individual* unlabeled sample, we propose a new perspective of consistency regularization within the neighborhood to explore the data relationships. Neighborhoods can be considered as tiny sample-anchored clusters with high compactness and class consistency. Consistency regularization within the neighborhood further reveals the underlying class concept of the self-supervised motion representation. Such discriminative motion representations significantly improve the performance of semi-supervised 3D action recognition. Moreover, considering that adversarial learning can minimize the discrepancy between two distributions, we also propose a novel adversarial learning strategy to couple the self-supervision method and a semi-supervised algorithm. The adversarial regularization allows the model to align the feature distributions of labeled and unlabeled data, which boosts the capability of generalization to unseen samples for semi-supervised 3D action recognition.

We perform extensive studies for semi-supervised 3D action recognition on two benchmark datasets: NTU RGB+D [28] and N-UCLA [39] datasets. With the proposed ASSL network, we establish new state-of-the-art performances of semi-supervised 3D action recognition. Summarily, our main contributions are in three folds:

1. We present an Adversarial Self-Supervised Learning (ASSL) framework for semi-supervised 3D action recognition, which tightly couples SSL and a semi-supervised scheme via adversarial learning and neighbor relation exploration.
2. We offer a new self-supervised strategy, *i.e.*, neighborhood consistency, for semi-supervised 3D action recognition. By exploring data relationships within the neighborhood, our model can learn discriminative motion representations that significantly improve the performance of semi-supervised 3D action recognition.
3. We identify that directly applying SSL for semi-supervised learning suffers from the representation misalignment of labeled and unlabeled samples. A novel adversarial regularization is pioneered to couple SSL into a semi-supervised algorithm to align their feature distributions, which further boosts the capability of generalization.

2 Related Work

2.1 3D Action Recognition

Human action recognition is one of important computer vision tasks. Due to the informative representation for the action, skeleton-based action recognition has been examined thoroughly in past literature. Previously, the traditional approaches [11, 36–38] try to design various hand-crafted features from skeleton sequences to represent human motion, *e.g.*, relative 3D geometry between all pairs of body parts [36]. Recently, deep learning has also been applied to this task due to its wide success. To model temporal dependencies, many methods leverage and extend the recurrent neural networks (RNNs) to capture the motion features for skeleton-based action recognition, *e.g.*, HBRNN [7] and VA-LSTM [47]. Based on Convolutional Neural Networks (CNNs) that are powerful at learning hierarchical representations, spatio-temporal representations are extracted for action recognition in [6, 12, 18, 41]. For graph-structured data, graph-based approaches [19, 31, 32] are popularly adopted for skeleton-based action recognition, *e.g.*, ST-GCN [44] and AGC-LSTM [30]. Though successful, these supervised methods highly rely on massive data samples with annotated action labels, which are expensive to obtain. Semi-supervised approaches are thus developed to alleviate this data annotation limitation, and in this paper, we apply it to learning motion representation for 3D action recognition.

2.2 Semi-supervised Learning

Semi-supervised learning algorithms learn from a data set that includes both labeled and unlabeled data, usually mostly unlabeled. For a comprehensive review of semi-supervised methods, we refer readers to [3]. Recently, there is increasing interest in deep learning based semi-supervised algorithms. One group of these methods is based on generative models, *e.g.*, denoising autoencoders [26], variational autoencoders [14] and generative adversarial networks [25, 27]. Some semi-supervised methods add small perturbations to unlabeled data, and require similar outputs between them by enforcing a consistency regularization, *e.g.*, Π -Model [15], Temporal Ensembling [15], Mean Teacher [34] and Virtual Adversarial Training [24]. There are also some other works. To name a few, Lee *et al.* [16] pick up the class with maximum predicted probability as pseudo-labels for unlabeled data, and use them to train the models. [10] presents a conditional entropy minimization for unlabeled data, which encourages their predicted probability to bias some class. The work most related to ours is [45] which proposes a new technique for semi-supervised learning by leveraging SSL techniques to learn representation of unlabeled images. Their work enlarges the generalization of semi-supervised learning methods. In this work, we exploit effective SSL to learn discriminative motion representation for semi-supervised 3D action recognition. Moreover, we further propose a novel adversarial regularization to couple SSL into the semi-supervised algorithm.

2.3 Self-supervised Learning for Action Recognition

Self-supervised learning for action recognition aims to learn motion representations from the unlabeled data by solving the pretext tasks. Recently, a stream of studies [1, 8, 17, 23, 33, 43] design various temporal-related tasks to learn the temporal pattern from the unlabeled RGB videos. For example, a sequence sorting task is introduced in [17]. [21, 40] propose to learn the video representation by predicting motion flows. Note that, these methods are for learning representations from RGB videos and not applicable to long-term skeleton sequences. For 3D action recognition, Zheng *et al.* [48] propose a conditional skeleton inpainting architecture to learn the long-term dynamics from unlabeled skeleton data. However, this SSL ignores the shared information among samples with the same action class and therefore may yield less discriminative feature representations. Hence, we propose an effective self-supervised strategy to learn discriminative representation that is beneficial for semi-supervised 3D action recognition.

3 Method

3.1 Problem Formulation

Instead of relying on massive labels in existing methods, we use only a few labeling data in semi-supervised 3D action recognition. Formally, let \mathcal{X} be the training set. The training samples $\mathbf{x}_i \in \mathcal{X}$ are skeleton sequences with T frames, and $\mathbf{x}_i = \{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,T}\}$. At each time t , the $\mathbf{x}_{i,t}$ is a set of 3D coordinates of body joints, which can be obtained by the Microsoft Kinect and the advanced human pose estimation algorithms [2, 46]. In contrast to supervised 3D action classification, training samples are split to two subsets in our task here: a labeled training set denoted as $\mathcal{X}_L = \{\mathbf{x}_1, \dots, \mathbf{x}_L\}$ and an unlabeled training set denoted as $\mathcal{X}_U = \{\mathbf{x}_1, \dots, \mathbf{x}_U\}$. The training samples $\mathbf{x}_l \in \mathcal{X}_L$ have annotated labels $\{y_1, \dots, y_L\}$ with $y_l \in \mathcal{C}$, where $\mathcal{C} = \{1, \dots, C\}$ is a discrete label set for C action classes. The training samples $\mathbf{x}_u \in \mathcal{X}_U$ are unlabeled. Usually, L is smaller than U ($L \ll U$).

Inspired by S^4L [45], we propose an Adversarial Self-Supervised Learning framework to learn discriminative motion representations from \mathcal{X}_L and \mathcal{X}_U . It couples self-supervised techniques into the semi-supervised scheme via adversarial learning and neighbor relation exploration. Detailed descriptions of ASSL are described in the following subsections.

3.2 Neighborhood Consistency for Semi-supervised 3D Action Recognition

Semi-supervised 3D action recognition aims to learn discriminative motion representation from massive unlabeled sequences. However, this is difficult over succinct 3D human poses. To tackle this challenge, we propose an effective SSL strategy, neighborhood consistency, that enhances the underlying class semantics

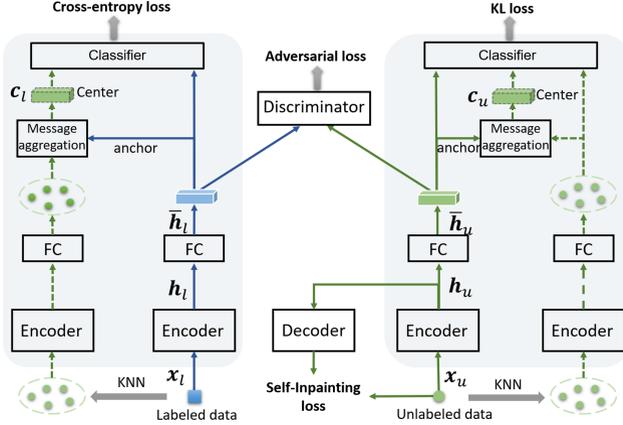


Fig. 2. Framework of Adversarial Self-Supervised Learning (ASSL). The ASSL leverages SSL and adversarial regularization for semi-supervised 3D action recognition. For SSL techniques, in addition to a self-inpainting constraint [48] for learning temporal information of each individual unlabeled sample, we propose to apply a new consistency regularization within the neighborhood to explore data relations. The adversarial training with a feature discriminator is used to align feature distributions of labeled and unlabeled samples, which further boosts generalization of semi-supervised models to unseen samples

of motion representation by exploring data relations within the neighborhoods, so as to improve recognition performance.

As shown in Fig. 2, we first employ skeleton inpainting [48] to learn temporal information for each unlabeled sequence. Specifically, an encoder network *Enc* takes an input skeleton sequence \mathbf{x}_u from training set \mathcal{X}_U and produces a vector as the temporal features $\mathbf{h}_u \in \mathbb{R}^d$. Conditioned on the learned representation \mathbf{h}_u , a decoder network *Dec* aims to fill the masked regions in the input sequence. Due to the difference between the action classification (discrimination) and skeleton inpainting (regression) tasks, we use a translation layer *i.e.*, a linear layer, to bridge the gap between the feature spaces of both tasks. The output of linear layer is denoted as $\bar{\mathbf{h}}_u$ for the sample \mathbf{x}_u . Then, in this feature space, we employ K-nearest neighbor [4] to select K nearest neighbors from unlabeled training set \mathcal{X}_U . The neighbor set of \mathbf{x}_u is denoted as $\Omega_{x_u} = \{\mathbf{x}_u^1, \dots, \mathbf{x}_u^K\}$. A message aggregation module is proposed to produce the local center vector. We use a multilayer perceptron to assign a weight for each neighbor sample, which evaluates their similarities as the anchor. The weights α_k are computed as follows:

$$\alpha_k = \frac{\exp\left(MLP\left(|\bar{\mathbf{h}}_u - \bar{\mathbf{h}}_u^k|\right)\right)}{\sum_{k=1}^K \exp\left(MLP\left(|\bar{\mathbf{h}}_u - \bar{\mathbf{h}}_u^k|\right)\right)}, \quad (1)$$

where $\bar{\mathbf{h}}_u^k$ is the translated feature of neighbor sample $\mathbf{x}_u^k \in \Omega_{x_u}$, $MLP(\cdot)$ denotes the multilayer perceptron in message aggregation module. According to the

computed weights $\{\alpha_1, \dots, \alpha_K\}$, the local class center \mathbf{c}_u can be aggregated with the neighbor set Ω_{x_u} as follows:

$$\mathbf{c}_u = \sum_{k=1}^K \alpha_k \bar{\mathbf{h}}_u^k. \quad (2)$$

Considering the high compactness and class consistency within neighborhoods, we require that the samples within neighborhoods achieve a similar prediction with the local center \mathbf{c}_u . However, for a sample \mathbf{x}_u , its neighbor samples either share the class label (positive) with \mathbf{x}_u or not (negative). To minimize the impact of negative neighbors, we introduce a simple selecting criterion: we get the 1-nearest labeled neighbor from the labeled training set \mathcal{X}_L for the anchor \mathbf{x}_u and the neighbor \mathbf{x}_u^k . If the labeled neighbors of the anchor \mathbf{x}_u and the neighbor \mathbf{x}_u^k have the same label, \mathbf{x}_u^k is regarded as the positive neighbor. The set of selected positive neighbor for sample \mathbf{x}_u is denoted as $\Omega_{x_u}^p$. Finally, the loss of consistency regularization within neighborhood is defined as follows:

$$\mathcal{L}_{KL} = \sum_{\mathbf{x}_u \in \mathcal{X}_U} \left(KL(f_c(\mathbf{c}_u), f_c(\bar{\mathbf{h}}_u)) + \sum_{\mathbf{x}_u^k \in \Omega_{x_u}^p} KL(f_c(\mathbf{c}_u), f_c(\bar{\mathbf{h}}_u^k)) \right), \quad (3)$$

where $f_c(\cdot)$ is the classifier that outputs the predictions, $KL(\cdot)$ denotes Kullback-Leibler divergence.

Like consistency regularization for unlabeled samples $\mathbf{x}_u \in \mathcal{X}_U$, the neighbor sets of labeled examples $\mathbf{x}_l \in \mathcal{X}_L$ are also selected from the unlabeled set \mathcal{X}_U . which are denoted as Ω_{x_l} . Similarly, we use the feature $\bar{\mathbf{h}}_l$ of \mathbf{x}_l as the anchor to estimate its local center representation \mathbf{c}_l with its neighbors set Ω_{x_l} as the Eq. (1)–(2) (shown in Fig. 2). Under the assumption that the anchor shares the same class semantic with the local center, we use a cross-entropy loss $CE(\cdot)$ for the center \mathbf{c}_l :

$$\mathcal{L}_{CE}^c = \sum_{\mathbf{x}_l \in \mathcal{X}_L} (CE(f_c(\mathbf{c}_l), y_l)), \quad (4)$$

where y_l is the class label of \mathbf{x}_l .

Overall, the optimization objectives of unlabeled samples can be formulated as follows:

$$\mathcal{L}_U = \mathcal{L}_{KL} + \mathcal{L}_{CE}^c + \mathcal{L}_{inp}, \quad (5)$$

where \mathcal{L}_{inp} denotes the skeleton inpainting loss that is the L_2 distance between the inpainted sequence and the original input sequence. Minimizing this optimization objective \mathcal{L}_U encourages the model to enhance the underlying class concept of the self-supervised motion representation and yield discriminative feature representations.

3.3 Adversarial Learning for Aligning Self-supervised and Semi-supervised Representations

According to the training of existing semi-supervised learning methods, the labeled and unlabeled samples are enforced with supervised and SSL objectives, respectively. In this work, Eq. (5) is used for the unlabeled samples. Although our proposed SSL technique is quite effective for semi-supervised 3D action recognition, we identify that the representations learned with supervised and SSL task are misaligned. As shown in the Fig. 3, with the benefit of SSL technique, the features of *Sup. + Sel.* present a more compact distribution than *Sup.*. However, in contrast to the intra-class compactness of labeled data (the squares with black border), there are scattering distributions for the unlabeled data in Fig. 3(b).

Thus, although both sequences are sampled from the same data distribution, their feature distributions are misaligned due to different optimization objectives. To tackle this problem, we propose a novel adversarial training strategy to couple SSL method with the semi-supervised 3D action recognition. Specifically, a discriminator Dis is trained to distinguish the unlabeled features from the labeled features. And the model is trained simultaneously to confuse the discriminator Dis . Hence, the adversarial loss is defined as follows:

$$\mathcal{L}_{adv} = \frac{1}{L} \sum_{x_l \in \mathcal{X}_L} (\log(Dis(\bar{h}_l))) + \frac{1}{U} \sum_{x_u \in \mathcal{X}_U} (\log(1 - Dis(\bar{h}_u))). \quad (6)$$

The adversarial regularization allows the model to align the feature distributions of labeled and unlabeled data. Therefore, like the labeled data, the feature distribution of unlabeled data becomes more intra-class compactness, which boosts the capability of generalization to unseen samples. More analyses about adversarial regularization are reported in Sect. 4.3.

3.4 Model Architecture and Optimization

Unlike the existing 3D action recognition method [7, 12, 18, 29, 30, 44, 47] learning the discriminative features through the designed effective networks, the goal of this work is to explore effective semi-supervised scheme for 3D action recognition. Therefore, this work adopts a universal architecture. In order to effectively capture the motion dynamics, we use three bidirectional GRU layers to encode

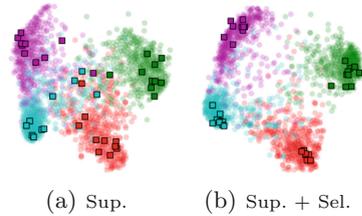


Fig. 3. The t-SNE visualization of motion features learned by *Sup.* and *Sup. + Sel.* (a) *Sup.* is trained with the supervised objective for the labeled samples. (b) *Sup. + Sel.* is trained through optimizing the supervised and SSL objectives (Eq. (5)) for the labeled and unlabeled samples, respectively. Different colors indicate different classes. Best viewed in color. The squares with black border denote the labeled data, and others are unlabeled data (Color figure online)

the input skeleton sequence in the *Enc*. The decoder consists of two unidirectional GRU layers. We use 4 linear layers and 3 linear layers in the discriminator and the multilayer perceptron of message aggregation, respectively. The classifier is a two-layer perceptron.

During training, our ASSL network is learned by minimizing the following loss on the training data:

$$\mathcal{L} = \mathcal{L}_L + \lambda_1 \mathcal{L}_U + \lambda_2 \mathcal{L}_{adv}. \quad (7)$$

where \mathcal{L}_L is a cross-entropy loss of all labeled examples in \mathcal{X}_L , λ_1 and λ_2 are non-negative scalar weights. Note that, we always sample the same number labeled and unlabeled samples in mini-batches.

4 Experiments

In this section, we evaluate and compare our work with previous semi-supervised methods and also conduct detailed component analysis.

4.1 Experimental Setup

Datasets. Two popular benchmark datasets, NTU RGB+D dataset [28] and Northwestern-UCLA dataset [39], are used for our experiments.

NTU RGB+D dataset [28] contains 56,880 samples covering 60 different classes of human actions performed by 40 distinct subjects. These videos are collected with three cameras simultaneously in different horizontal views. Two evaluation protocols are provided: Cross-Subject (CS) and Cross-View (CV). For CS protocol, skeleton sequences performed by 20 subjects are used for training, and the rest for testing. For CV protocol, all videos from Camera 2 and 3 are used for training while those from Camera 1 are used for testing. For semi-supervised 3D action recognition, 5%, 10%, 20% and 40% of training sequences of each class are labeled on the training set.

Northwestern-UCLA dataset [39] has 1,494 samples performed by 10 different subjects belonging to 10 action classes. Each action sample is captured by three Kinect cameras simultaneously from a variety of viewpoints. Its training set consists of samples from the first two cameras and the rest from the third camera form the testing set. For semi-supervised 3D action recognition, we use 5%, 15%, 30% and 40% labels of training sequences of each class on the training set.

Baselines. There is no available semi-supervised baseline for 3D action recognition, so we use following methods as baselines that achieve state-of-the-art performances in the RGB domain:

- 1) Supervised-only (Sup.), training with labeled skeleton sequences only.
- 2) Pseudo labels [16], leveraging the idea that the predicted labels of unlabeled samples are used for training. First, train a model with the labeled data, then predict the classes of unlabeled samples. These pseudo labels are used to retrain the network in a supervised fashion with labeled and unlabeled data simultaneously.

- 3) Virtual Adversarial Training (VAT) [24], training with unlabeled data to make the model robust around input data point against local perturbation. It generates small adversarial perturbations for unlabeled samples, which greatly alter the output distribution; then consistency loss is applied over unlabeled training data to encourage consistency of predictions for input data and its adversarially perturbed version.
- 4) Conditional Entropy Minimization (EntMin) [10], minimizing the entropy of prediction over unlabeled training data as a regularization for model training. Predicted class probabilities are encouraged to be near a one-hot vector via training with unlabeled data.
- 5) Self-Supervised Semi-Supervised Learning (S^4L) [45], the most related method to ours. It trains the model on self-supervised and semi-supervised tasks in a multi-task fashion. For 3D action recognition, we use the skeleton inpainting framework [48] as the pretext task for self-supervised learning.

Implementation. All comparisons with semi-supervised baselines are made under the same setting to be fair. In all experiments, the dimension of hidden states in the GRU and bidirectional GRU is set to 512. On both datasets, we randomly sample $T = 40$ frames from each skeleton sequence as input during training and testing. We train all networks by the ADAM optimizer [13]. The learning rate, initiated with 0.0005, is reduced by multiplying it by 0.5 every 30 epochs. We set $\lambda_1 = 1$ and $\lambda_2 = 0.1$ in Eq. (7). Our experiments are all implemented with PyTorch and 1 Titan Xp GPU.

4.2 Comparison with Semi-supervised Methods

We evaluate our method by comparing it with baselines for semi-supervised 3D action recognition and show results on NTU and N-UCLA datasets respectively in Tables 1 and 2.

Table 1. Test accuracy (%) on NTU dataset (Cross-Subject (CS) and Cross-View protocols (CV)) with 5%, 10%, 20 and 40% labels of training set. *v./c.* denotes the number of labeled videos per class

Method	5%		10%		20%		40%	
	CS (33 <i>v./c.</i>)	CV (31 <i>v./c.</i>)	CS (66 <i>v./c.</i>)	CV (62 <i>v./c.</i>)	CS (132 <i>v./c.</i>)	CV (124 <i>v./c.</i>)	CS (264 <i>v./c.</i>)	CV (248 <i>v./c.</i>)
Supervised-only	47.2	53.7	57.2	63.1	62.4	70.4	68.0	76.8
Pseudolabels [16]	50.9	56.3	58.4	65.8	63.9	71.2	69.5	77.7
VAT [24]	51.3	57.9	60.3	66.3	65.6	72.6	70.4	78.6
VAT + EntMin [10]	51.7	58.3	61.4	67.5	65.9	73.3	70.8	78.9
S^4L (Inpainting) [45]	48.4	55.1	58.1	63.6	63.1	71.1	68.2	76.9
ASSL (ours)	57.3	63.6	64.3	69.8	68.0	74.7	72.3	80.0

As seen from tables, with the proposed ASSL network, we establish new state-of-the-art performances of semi-supervised 3D action recognition. To be specific, S^4L (Inpainting) performs worse than *Pseudolabels*, *VAT* and *VAT + EntMin*, suggesting it is inefficient to learn discriminative representation via

Table 2. Test accuracy (%) on N-UCLA dataset with 5%, 15%, 30% and 40% labels of training set. *v./c.* denotes the number of labeled videos per class

Method	5% (5 <i>v./c.</i>)	15% (15 <i>v./c.</i>)	30% (30 <i>v./c.</i>)	40% (40 <i>v./c.</i>)
Supervised-only	34.1	37.9	48.9	58.8
Pseudolabels [16]	35.6	48.9	60.6	65.7
VAT [24]	44.8	63.8	73.7	73.9
VAT + EntMin [10]	46.8	66.2	75.4	75.6
S^4L (Inpainting) [45]	35.3	46.6	54.5	60.6
ASSL (ours)	52.6	74.8	78.0	78.4

skeleton inpainting and thus semi-supervised 3D action recognition has derived little benefit from self-supervised representations. S^4L (Inpainting), though an advanced semi-supervised approach, requires an effective self-supervised representations that are difficult to learn in this task. Compared with these semi-supervised methods, our benefit is larger when the number of labels is reduced. For example, with 5% labels of training set on NTU dataset, the results of our ASSL present greater improvement compared with *VAT + EntMin*. This clearly demonstrates the power of the proposed ASSL.

4.3 Ablation Study

We then investigate effectiveness of the neighborhood consistency and adversarial training in our proposed ASSL on NTU and N-UCLA datasets. We also analyze effects of different neighborhood sizes and Neighborhood quality.

Neighborhood Consistency. We evaluate the effects of the proposed self-supervised strategy, neighborhood consistency, upon the discriminativeness of motion representations that is shown in final performance of semi-supervised 3D action recognition. In Table 3, the model *Sup. + Inp.* is trained with a cross-entropy loss for labeled data and a self-inpainting loss \mathcal{L}_{inp} for unlabeled data. Instead of self-inpainting loss, *Sup. + Nei.* explores the data relations within neighborhoods by enforcing the consistency regularization (Eq. (3), (4)) for unlabeled data. We can see that *Sup. + Nei.* significantly outperforms the *Sup. + Inp.*. The comparison results justify that our neighborhood consistency can learn more discriminative motion representations that are more beneficial for semi-supervised 3D action recognition.

Moreover, the self-inpainting constraint [48] aims at learning temporal information of each *individual* unlabeled sequence. The goal of our neighborhood consistency regularization is to explore inter-sample relations within neighborhoods. We jointly learn the two features in *Sup. + Inp. + Nei.*. It can be seen compared with *Sup. + Inp.* and *Sup. + Nei.*, *Sup. + Inp. + Nei.* achieves better performance on both datasets for semi-supervised 3D action recognition. This illustrates that the representations learned by our neighborhood consistency are complementary to those learned with self-inpainting. Therefore, the benefits of

Table 3. Ablation study on self-supervised learning methods, skeleton inpainting (*Inp.*) [48] and neighbor consistency (*Nei.*). Classification accuracy (%) is reported on NTU with 5% labels and N-UCLA with 15% labels.

Methods	NTU 5%		N-UCLA 15%
	CS (33 <i>v./c.</i>)	CV (31 <i>v./c.</i>)	(15 <i>v./c.</i>)
Supervised-only (Sup.)	47.2	53.7	37.9
Sup. + Inp	48.4	55.1	46.6
Sup. + Nei	52.1	57.8	60.0
Sup. + Inp. + Nei	55.2	61.1	66.4
ASSL	57.3	63.6	74.8

combining these two SSL techniques to capture discriminative representation from unlabeled sequences in our final model are verified (seen Eq. (5)).

Neighborhood Size. We assume that the larger neighborhood size imposes stronger regularization and gives better performance. In order to justify this hypothesis, we investigate the effects of different neighborhood sizes in Fig. 4. As neighborhood size increases, the performance is improved and then becomes saturated. This implies that more discriminative representations can be learned with a larger size. But, if using too large a size, the model will cover distant data points that have weak semantic consistency within the neighborhood, and hence the performance becomes saturated.

Neighborhood Quality. We further examine effects of the class consistency of anchor Neighborhood, *i.e.*, Neighborhood quality. In Fig. 5, we report the progress of the ratio of neighbor samples sharing the same action label as the anchor throughout training. We can observe the ratio of class consistent neighborhoods increases, and then becomes saturated. This indicates exploring data relations is helpful to inferring underlying class semantics, thus facilitating the clustering of samples with the same action labels.

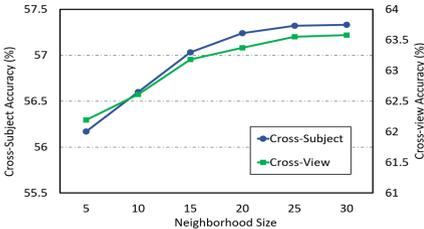


Fig. 4. Classification accuracy (%) with different neighborhood size on NTU dataset with 5% labels

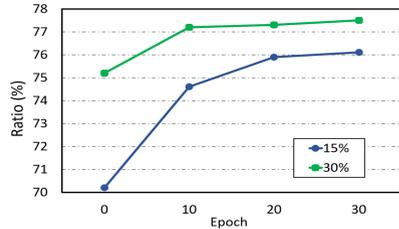
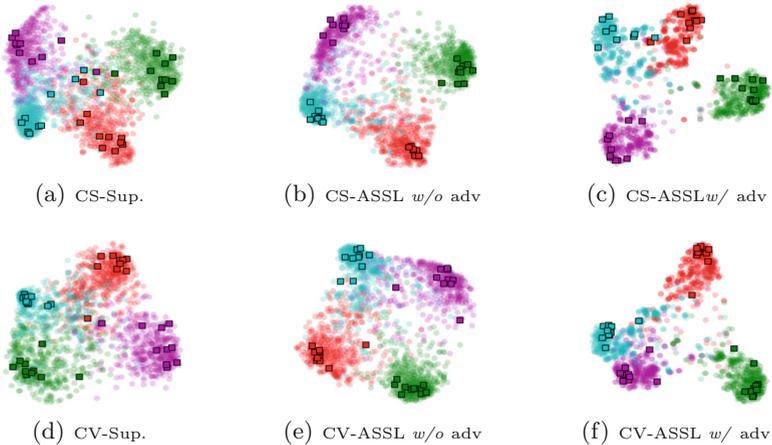


Fig. 5. The ratio of neighbor samples sharing the same action label as the anchor throughout training on N-UCLA dataset

Table 4. Ablation study on adversarial training. Classification accuracy (%) is reported on NTU with 5% labels and N-UCLA with 15% labels.

Methods		NTU 5%		N-UCLA 15%
		CS (33 <i>v./c.</i>)	CV (31 <i>v./c.</i>)	(15 <i>v./c.</i>)
Sup. + Inp.	<i>w/o</i> adv	48.4	55.1	46.6
	<i>w/</i> adv	51.2	57.1	52.4
Sup. + Nei.	<i>w/o</i> adv	52.1	57.8	60.0
	<i>w/</i> adv	53.4	59.1	68.5
ASSL (Sup. + Inp. + Nei.)	<i>w/o</i> adv	55.2	61.1	66.4
	<i>w/</i> adv	57.3	63.6	74.8

Adversarial Training. The adversarial alignment is proposed to mitigate the gap between representations learned from supervised and self-supervised tasks. To evaluate effectiveness of adversarial training for coupling self-supervision methods with the semi-supervised 3D action recognition, we train several self-supervised models with or without adversarial regularization. The results are reported in Table 4. It is obvious that all models with adversarial regularization achieve better performances than those without. For example, on N-UCLA dataset, the result of *ASSL w/adv* is 74.8%, outperforming *ASSL w/o adv* by 8.4%. The improved performance in Table 4 demonstrates that it is an effective strategy to couple self-supervision with semi-supervised algorithms by adversarial training.

**Fig. 6.** The t-SNE visualization of motion features learned by *Supervised Baseline* (Sup.), *ASSL w/o adv* and *ASSL w/adv* (ours) on NTU dataset. Different colors indicate different classes. Best viewed in color. The squares with black border denote the labeled data, and others are unlabeled data (Color figure online)

To further explore this scheme, we visualize the feature distributions of *Sup.*, *ASSL w/o adv* and *ASSL w/adv* by using t-SNE [22] in Fig. 6. For the model *Sup.* trained with only supervised objective on labeled data, the decision boundaries of its feature distributions are very ambiguous. The model *ASSL w/o adv* is trained with supervised and self-supervised objectives for labeled and unlabeled data, respectively. Compared with *Sup.*, the features of *ASSL w/o adv* present tighter distributions, which benefit from self-supervised learning. But, long-tail distributions still exist for unlabeled samples (circles). Figure 6(c) and (f) show clearly the alignment between feature distributions of labeled and unlabeled data for *ASSL w/adv*, *i.e.*, the proposed ASSL. Overall, the comparison results prove the effectiveness of adversarial training for coupling self-supervision with semi-supervised action recognition. And this drives our model to learn more discriminative features that have desired intra-class compactness and inter-class separability.

5 Conclusions

In this paper, we consider the semi-supervised learning scheme for 3D action recognition task. The proposed ASSL effectively couples SSL into semi-supervised algorithm via neighbor relation exploration and adversarial learning. Exploring data relations with neighborhood consistency regularization encourages the model to learn discriminative motion representations that significantly improve the performance of this task. Moreover, we introduce a novel adversarial regularization to couple SSL method into a semi-supervised algorithm. This allows the model to align the feature distributions of labeled and unlabeled samples and boosts the capability of generalization to unseen samples. Our experiments verify that the proposed neighbor relation exploration and adversarial learning are strongly beneficial for semi-supervised 3D action recognition. With the proposed ASSL network, we establish news state-of-the-art performances of semi-supervised 3D action recognition.

Acknowledgements. This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61420106015, 61976214, 61721004), Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO. 2019JZZY010119). Jiashi Feng was partially supported by MOE Tier 2 MOE2017-T2-2-151, NUS_ECRA_FY17_P08, AISG-100E-2019-035. Chenyang Si was partially supported by the program of China Scholarships Council (No. 201904910608). We thank Jianfeng Zhang for his helpful comments.

References

1. Büchler, U., Brattoli, B., Ommer, B.: Improving spatiotemporal self-supervision by deep reinforcement learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 797–814. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_47
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
3. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised Learning. MIT Press, Cambridge (2006)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.* **13**, 21–27 (1967)
5. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
6. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: ACPR (2015)
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR (2015)
8. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: CVPR (2017)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015)
10. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NIPS (2005)
11. Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M.: Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In: IJCAI (2013)
12. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: CVPR (2017)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
14. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS (2014)
15. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
16. Lee, D.H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: ICML (2013)
17. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: ICCV (2017)
18. Li, C., Zhong, Q., Xie, D., Pu, S.: Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: IJCAI (2018)
19. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: CVPR (2019)
20. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: NIPS (2018)
21. Luo, Z., Peng, B., Huang, D.A., Alahi, A., Fei-Fei, L.: Unsupervised learning of long-term motion dynamics for videos. In: CVPR (2017)
22. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)

23. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 527–544. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_32
24. Miyato, T., Maeda, S., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1979–1993 (2018)
25. Odena, A.: Semi-supervised learning with generative adversarial networks. arXiv preprint [arXiv:1606.01583](https://arxiv.org/abs/1606.01583) (2016)
26. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: NIPS (2015)
27. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS (2016)
28. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: CVPR (2016)
29. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR (2019)
30. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: CVPR (2019)
31. Si, C., Jing, Y., Wang, W., Wang, L., Tan, T.: Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 106–121. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_7
32. Si, C., Jing, Y., Wang, W., Wang, L., Tan, T.: Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network. *Pattern Recogn.* **107**, 107511 (2020)
33. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: ICML (2015)
34. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: NIPS (2017)
35. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
36. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3D skeletons as points in a lie group. In: CVPR (2014)
37. Vemulapalli, R., Chellappa, R.: Rolling rotations for recognizing human actions from 3D skeletal data. In: CVPR (2016)
38. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR (2012)
39. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning, and recognition. In: CVPR (2014)
40. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: CVPR (2019)
41. Wang, P., Li, Z., Hou, Y., Li, W.: Action recognition based on joint trajectory maps using convolutional neural networks. In: ACM MM (2016)
42. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
43. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR (2019)
44. Yan, S., Xiong, Y., Lin, D., Xiaoou Tang: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)

45. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4L: self-supervised semi-supervised learning. In: ICCV (2019)
46. Zhang, J., Nie, X., Feng, J.: Inference stage optimization for cross-scenario 3D human pose estimation. arXiv preprint [arXiv:2007.02054](https://arxiv.org/abs/2007.02054) (2020)
47. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: ICCV (2017)
48. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: AAAI (2018)