A Weakly Supervised Learning Method for Recognizing Childhood Tic Disorders

Ruizhe Zhang[1], Xiaojing Xu[2], Zihao Bo[1], Junfeng Lyu[1], Yuchen Guo[3], and Feng Xu[1(✉)]

[1] School of Software and BNRist, Tsinghua University, Beijing, China
feng-xu@tsinghua.edu.cn
[2] China-Japan Friendship Hospital, Beijing, China
[3] BNRist, Tsinghua University, Beijing, China

Abstract. So far, the number of individuals with Tic disorder worldwide has reached 59 million, and the prevalence of the disorder is rapidly increasing globally. In this work, we focus on weakly supervised learning methods for recognizing childhood tic disorders. In situations with limited data availability, we design a relative probability metric based on the characteristics of the data and a multi-phase learning algorithm is proposed based on relative probability in order to efficiently utilize coarse-labeled data in a "from easy to difficult" manner. Furthermore, the effectiveness of our method is validated through ablation experiments. Through extensive experiments on the test dataset, we demonstrate that our method behaves extraordinarily compared to baseline approaches, improving AUC by 3.0%, and facilitating expedited diagnostic assessment for medical practitioners.

Keywords: tic disorders · facial data processing · weakly supervised learning

1 Introduction

Tic disorder [1, 2] is a motor or vocal muscle spasm characterized by symptoms such as frequent eye blinking, head jerking, facial distortions, repetitive coughing, and throat clearing. Diagnosing Tic disorder in clinical settings is typically a complex process, further complicated by the fact that the majority of affected individuals are children, who often have low cooperation, leading to diagnostic challenges. Research [10, 11] has primarily focused on pathology and clinical aspects over the past few decades, with limited studies on the identification and detection of tic disorder symptoms in patients.

In recent years, machine learning has been widely applied to medical problems, particularly in the areas of disease diagnosis and classification. Some studies have employed video-based action recognition to diagnose diseases. The mainstream approach for video action recognition is based on Convolutional Neural Networks (CNNs) [3, 5]. One popular approach is the two-stream architecture [16–18]. Another approach is the use of 3D CNNs [19–24] that can directly capture spatiotemporal information from video sequences. Furthermore, attention mechanisms [4] allow the model to allocate more attention to relevant parts of the video, improving both accuracy and efficiency. These

L. Fang et al. (Eds.): CICAI 2023, LNAI 14474, pp. 100–112, 2024.
https://doi.org/10.1007/978-981-99-9119-8_10

networks above need fully-supervised data but for the problem of recognizing tic disorder, data annotation requires professional doctors, which incurs high manpower costs and poses challenges in annotation. Moreover, our available labeled data is limited. Therefore, the fully supervised methods are not suitable for our research problem.

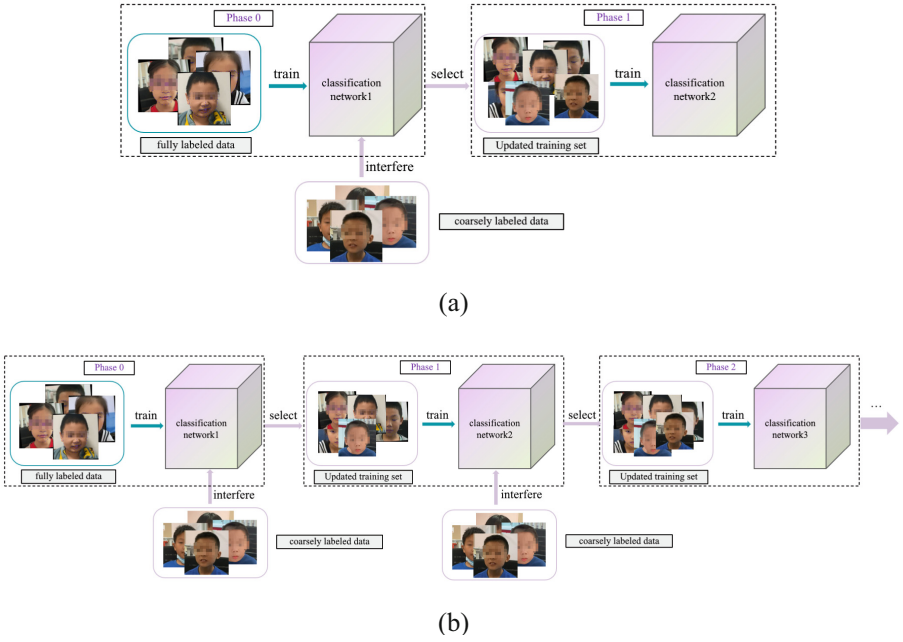

Fig. 1. Frameworks of one-phase training (a) and our multi-phase training method (b).

Weakly supervised learning [25–36] (WSL) is the method to solve this issue, which aims at improving the performance of models by exploiting many unlabeled data. Among various techniques in weakly supervised learning, pseudo-labeling methods have gained significant attention due to their effectiveness in leveraging unlabeled data. Pseudo-labeling is a technique that assigns labels to unlabeled data based on the predictions of a trained model. These assigned labels are considered "pseudo-labels" and are then used to augment the training set for further model refinement. Typically, a one-phase learning scheme in Fig. 1(a) is adopted.

However, it is insufficient for knowledge excavation to exploit the unlabeled data only once, so multi-phase learning comes out further enhances the performance of weakly supervised learning methods. Multi-phase learning, divides the weakly supervised learning process into multiple stages, each with a specific objective or set of labeled and unlabeled data. In each phase, the model is trained and pseudo-labels are generated based on the current phase's predictions. These pseudo-labels are then used as training data for the next stage, enabling the model to learn progressively and capture more complex patterns over successive stages.

But issues arise as a result of these methods. Firstly, As the model is trained with these pseudo-labels, it becomes biased towards making predictions that align with the labels generated in the previous phase, in which way, increasing the number of phases becomes meaningless for model performance improvement. Moreover, if the pseudo-labels are noisy or incorrect, the model's predictions may be influenced by these errors and hinder further learning progress, leading to degraded performance. Secondly, the method to generate pseudo-labels plays a crucial role. Common methods for generating pseudo-labels include thresholding and Top-K selection. However, challenges remain in selecting appropriate thresholds or K values and handling noisy or uncertain samples.

To address these problems, we design a metric called "relative probability" (RPr) based on the characteristics of the annotations. We not only use this metric for generating pseudo-labels but also involve it in multi-phase training to measure the learning difficulty of positive samples. An RPr-guided method is proposed, and during the multi-phase learning process, the RPr threshold decreases by phase so that easy samples can be selected in early phases. This multi-phase learning with thresholds decrease strategy (MPLTD) allows the model to initially learn from easy or simple data to enhance its performance, and subsequently, in later phases, tackle more challenging or difficult data (see Fig. 1(b)). The main contributions of this paper can be summarized as follows:

- We propose a facial data processing and dimensionality reduction method. In the case of limited data, this dimensionality reduction method reduces the training time and training difficulty of the model while achieving better accuracy.
- We design a relative probability metric that balances the accuracy of pseudo-label generation and the number of positive samples obtained. It effectively improves the learning performance of the model on coarsely labeled data.
- We propose a multi-phase learning process that implements a "from easy to hard" weakly supervised learning approach. This method is relatively universal and applicable.

2 Related Work

Tic disorder diagnosis has no great progress made in this area until the 2010s. In 2010, Bernabei et al. [10] conducted a study using wearable devices with accelerometers to detect twitching movements in the limbs and trunks of Tourette syndrome patients, achieving an accuracy of 80.5%. In 2016, Shute et al. [11] conducted research based on brain electrical stimulation and observed low-frequency central medial-prefrontal (CM-PF) activity to detect tic symptoms in patients.

Facial landmark detection is the process of automatically locating and identifying key points or landmarks on a human face. Cootes et al. [12] proposed the Active Appearance Models (AAM) which model the shape and texture of the face as random variables and estimate them through optimization methods, thus achieving facial landmark detection. Kazemi et al. [13] introduced a fast and accurate method for facial landmark detection based on ensemble models of regression trees, enabling rapid detection of facial landmarks. In 2015, Yang et al. [14] proposed a cascaded regression approach for robust facial landmark tracking. This method progressively improves the localization accuracy of facial landmarks by training a series of regressors in a cascade. Bulat et al. [15]

provided a review of 2D and 3D facial landmark detection problems and presented a large-scale 3D facial landmark dataset.

Weakly supervised learning focuses on developing algorithms and techniques to address the challenges of training machine learning models with limited or noisy supervision. The most important line of research in WSL explores methods for generating pseudo-labels [31, 33], which are inferred labels assigned to unlabeled data based on some heuristics or assumptions. These pseudo-labels are used to train the model in a semi-supervised [32–36] or self-supervised manner. Many works are devoted to semi-supervised these years, such as self-training, label propagation [29], and so on.

3 Method

In this section, we first define the problem of tic disorder recognition and classification. To address privacy concerns, we employ dimensionality reduction techniques to convert facial images into facial landmark points, thus preserving the privacy of the patients. Firstly, we train an initial model with fully labeled data. Then we use this model to generate pseudo-labels for the coarse labeled data, selecting reliable positive samples to be added to the training set. We retrain the model and repeat this process iteratively. In the pseudo-label generation step, we introduce the concept of relative probability, which ensures that the selected positive samples exhibit similar features to the most prominent movements in the long segments. For the iterative part, we propose a method to gradually decrease the threshold value so that the model initially learns from simple samples to improve accuracy and then focuses on difficult samples to enhance generalization.

3.1 Data Description

We collected a total of 129 videos from children with tic disorders. Based on the level of annotation detail, we divided all the videos into two categories: fully labeled videos (42 videos) and coarsely labeled videos (87 videos). The fully labeled videos consist of short segments, where each annotated segment has a length of 2 s. On the other hand, the coarsely labeled videos consist of long segments, where each segment has a length ranging from 3 to 10 s.

3.2 Tic Disorder Recognition Problem Definition

Let $\aleph$ be the set of all videos, where each video $X \in \aleph$ consists of several short segments $x_1, x_2, x_3, ..., x_N \in X$. Each short segment x_i is composed of several frames $a_1, a_2, a_3, ..., a_N \in x_i$ (usually 48 frames). Each frame a_i is the basic unit of our data processing, but not the basic unit for model prediction and tic recognition. The smallest unit of tics is the short segment x_i. In a video, each short segment can be one of the following: eye tic, mouth tic, nose tic, or normal. Among them, the first three can occur simultaneously, while the normal class can only occur alone.

We define the tic recognition task to determine whether a short segment x is a tic segment. In this task, we combine the tics in the eye, mouth, and nose regions as the positive class for binary classification, while the normal segments are the negative class. For convenience, we refer to it as "face binary classification" in the subsequent tables. Additionally, we define three tic disorder classification tasks to differentiate the tic regions. In each task, the tic region of interest is considered the positive class, while normal segments are the negative class. For example, in the eye tic disorder classification task, the positive class is eye tics, and the negative class is normal actions.

In summary, we define four binary classification tasks, where the positive and negative class samples are composed of multiple short segments. The labels for these samples are $y_i \in \{0, 1\}$, where 0 represents the negative class and 1 represents the positive class. Our goal is to achieve high classification accuracy (ACC) and area under the ROC curve (AUC) for these tasks.

3.3 Feature Extraction in Facial Data

In the context of limited data, to enhance the generalization of the algorithm, we perform feature point extraction, facial segmentation, and face alignment on each frame of the video segments. The overall process is in Algorithm 1 and visualized results are presented in Fig. 2.

Specifically, in step 3 of the algorithm, the method for calculating the rotation matrix is as follows: first, calculate the center coordinates of the left and right eyes $(centerX, centerY)$. Then calculate the angle between the line connecting the left and right eyes and the horizontal line. This angle represents the rotation angle θ. Finally, we can calculate the rotation matrix M as follows:

$$M = \begin{bmatrix} \cos\theta & -\sin\theta & (1-\cos\theta) \times centerX + \sin\theta \times centerY \\ \sin\theta & \cos\theta & (1-\cos\theta) \times centerY - \sin\theta \times centerX \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

By utilizing the rotation matrix M, we can transform the coordinates of any point (x, y) in the original image into the coordinates (x', y') of the corresponding point in the new image. The transformation relationship between them is given by:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = M \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

Algorithm 1 Facial Landmark Alignment

Input:

Single frame

Output:

Aligned single frame with facial landmark coordinates

1. Apply a face detector to detect facial landmark.

in the video.

3. Calculate the rotation matrix based on the coordinates of the left and right eyes.

4. Perform an affine transformation on the image to obtain the rotation-aligned image and its corresponding landmark.

Fig. 2. Face alignment algorithm flow. Our method ultimately compresses the (1080, 1920) image into a facial landmark sequence of size (68, 2).

3.4 Relative Probability Guided Multi-phase Learning

Relative Probability (RPr). To proceed with our method, we propose the concept of Relative Probability. For each long segment, where PR_{cut} is our defined relative probability indicator, it represents the model's confidence in predicting the current segment. $PR_{\max}$ is the maximum value of confidence scores among all short segments cut in the long segment, and $PR_{\min}$ is the minimum. The calculation formula for relative probability is as follows:

$$PR_{relative} = \frac{PR_{cut} - PR_{\min}}{PR_{\max} - PR_{\min}} \tag{3}$$

Two thresholds $thsd_1$ and $thsd_2$ are set in advance and a short segment is marked as a positive sample in the following condition:

$$PR_{relative} > thsd_1 \&\& PR_{cut} > thsd_2 \tag{4}$$

Through this approach, we effectively exploit the prior information inherent in the coarse annotations, assuming that the short segment with the highest confidence score corresponds to the most salient movement within the given long segment. Considering

the inherent similarity of movement patterns within each long segment, our objective is to select positive samples that not only surpass the confidence threshold but also exhibit a high degree of resemblance to the most prominent movement feature present in the segment. This strategic selection process aims to mitigate the risk of false positives, thereby enhancing the reliability and precision of our approach.

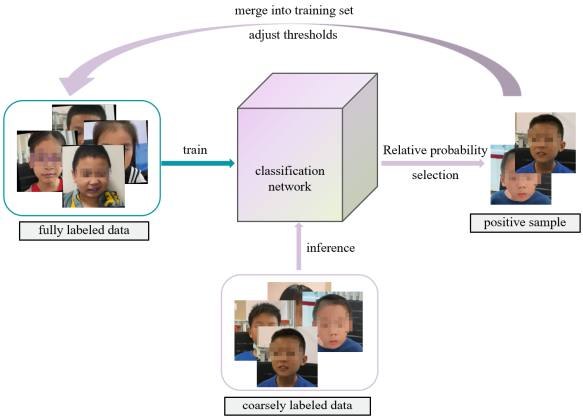

Fig. 3. Relative Probability guided multi-phase learning flowchart

Multi-phase Learning with Thresholds Decrease. Although we have introduced the concept of relative probability to improve the accuracy of generating pseudo-labels, it is inevitable that erroneous pseudo-label noise may still occur, potentially misleading the model during training. Additionally, the integration of rough labeled data into the training set requires careful consideration of techniques and strategies. If we simply incorporate all positively labeled samples into the training set, in the subsequent iterations, the model may tend to assign high confidence scores to these selected positive samples, resulting in pseudo-labels that are nearly identical to those of the previous round. Consequently, this iterative process can become stagnant, hindering any improvement in model accuracy.

Considering these problems, we propose a multi-phase learning algorithm with a threshold decrease. The overall process is shown in Algorithm 2.

We have designed multiple expressions to update the threshold with respect to the number of phases and we discover that the easiest and most effective method is a linear decay strategy (see Eq. (5)). And after sufficient experiments, we find that when $step_1 = step_2 = 0.05$ and thresholds stop decreasing in 4^{th} phase, the proposed method get the best result.

$$thsd_i = thsd_i - step_i \; i = 1, 2 \tag{5}$$

Algorithm 2 Multi-phase learning with threshold decrease

Input:

Video data (including coarse and fine annotations), the number of training phases N

Output:

Trained model

Training:

For j=0; j<=N; j++ **do**

1. Train the model on the training set (initially fully labeled) until convergence, and record the model's accuracy.
2. Perform inference on all coarse annotated data using the model.
3. Apply equation (4) to generate pseudo-labels based on the current thresholds.
4. Incorporate all selected positive samples that are not already in the training set.
5. Decrease the thresholds by equation (5).

End for

Testing:

Feed the test dataset into the trained model to obtain test results.

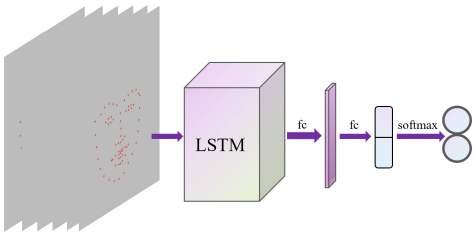

Fig. 4. The architecture of our classification network

As for the classification network in Fig. 1 and Fig. 3, we find that Long Short-Term Memory (LSTM) networks can capture the temporal motion features of facial landmarks, leading to superior classification performance.

4 Experiments

In this section, we train our models on our train dataset (which contains 2436 short segments) and coarse labeled data (87 videos). We evaluate the proposed method on our independent test dataset (which contains 833 short segments). For the methods that use the sequence of facial landmark points as input, we calculate the displacement between consecutive frames, resulting in a sequence of displacement vectors representing the motion of facial landmarks. For the methods that use original images as input, we perform facial segmentation to isolate the face region and apply grayscale normalization to enhance the consistency of the input data. All models are trained in RTX 3090.

4.1 Comparisons to Existing Methods

To the best of our knowledge, the field of tic disorder recognition lacks publicly available datasets, and there is a scarcity of relevant research with no established state-of-the-art (SOTA) method. Considering this, we conduct experiments using ResNet-3D and I3D, which are widely adopted methods in the domain of action recognition. We employ these models to tackle the task of movement disorder recognition, aiming to assess their performance and suitability. Considering the limited amount of data, we encounter challenges in evaluating transformer-based methods. For our facial feature extraction and privacy preservation method, which generates facial landmark points as input, we explore the performance of traditional machine learning methods.

As shown in Table 1, after face alignment, our method achieves an average AUC of 95.1%, 1.9% higher than ResNet-3D. The methods that use original images as input behave poorly even though they have much more parameters. I believe that the insufficient training data is one of the reasons. Additionally, simple LSTM or MLP models are already sufficient to capture the features of tic behaviors and our facial feature extraction method not only preserves privacy but also leads to better classification results.

Table 1. Our facial feature extraction and multi-phase learning based LSTM method vs. current method for video action detection. The numbers in the table represent AUC (%).

method	face	eye	mouth	nose	avg
ResNet-3D	94.7	92.9	93.1	92.0	93.2
I3D	92.6	88.5	87.9	88.6	89.4
MLP	95.0	93.4	93.2	90.9	93.1
MLP w/o alignment	93.3	92.0	92.5	90.1	92.0
RF [7, 8]	89.1	88.3	86.6	86.9	87.7
RF w/o alignment	88.0	88.1	85.6	86.7	87.1
LSTM (ours)	**97.0**	**96.2**	**93.5**	**93.8**	**95.1**
LSTM w/o alignment	95.7	94.9	92.9	92.8	94.1

For our best method LSTM in Table 1, we conduct complete experiment and find that while other methods may have higher AUC in the first phase, our method gradually surpasses them in subsequent stages and converges around four phases, demonstrating clear advantages compared to one-phase methods (Fig. 5).

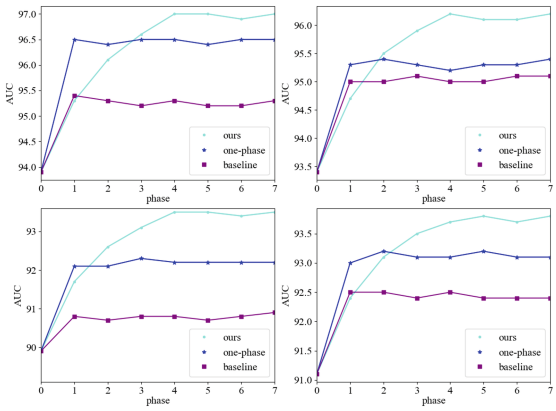

Fig. 5. AUC's change curve with phase on four binary classification tasks. From left to right, they are: face, eyes, mouth, nose. Baseline refers to the one-phase method with Top2 selection. One-phase changes Top2 selection into RPr selection.

4.2 Ablation Study

In this subsection, we evaluate the effect of proposed RPr selection and MPLTD algorithm. We conduct our experiment on ResNet-3D and LSTM. The results are in Table 2.

For LSTM, our RPr and MPLTD methods achieve the average AUC of 95.1%, outperforming baseline by 3%. For ResNet-3D, our method achieves the average AUC of 93.2%, 4.2% higher than baseline. Moreover, we discover that both RPr and MLTD methods lead to a significant increase in AUC. Through these proposed methods, we leverage both coarse-labeled and fine-labeled data in a comprehensive manner and achieve a more robust and effective training process for our model.

Table 2. Quantitive evaluation of our proposed method RPr and MLTD. The effectiveness is tested both in LSTM and ResNet-3D.

model	RPr	MPLTD	avg. AUC (%)
LSTM			92.1
	√		94.3
		√	93.9
	√	√	**95.1**
ResNet-3D			89.0
	√		90.5
		√	92.1
	√	√	**93.2**

5 Conclusion

In this work, we proposed a framework for facial feature extraction, suitable for weakly supervised learning with a limited amount of data and privacy preservation. Furthermore, based on the characteristics of our data, we introduced the concept of relative probability (RPr) and developed a multi-phase learning with threshold decrease (MPLTD) algorithm, achieving higher AUC than baseline. At last, we conducted ablation experiments to validate the effectiveness of each algorithm and achieved an ideal result. Our high-accuracy model not only assists doctors in diagnosis but also has the potential to be applied throughout the entire treatment process. It can be used to monitor and analyze the recovery and treatment progress of patients, providing guidance on medication and treatment approaches. In the future, we will continue to explore the tic disorder in limbs and address the multi-modal problem incorporating speech input.

Acknowledgements. This work was supported by Beijing Natural Science Foundation (M22024).

References

1. Leckman, J.F., Bloch, M.H.: Tic disorders. In: Rutter's Child and Adolescent Psychiatry, pp. 757–773 (2015)
2. Cohen, S.C., Leckman, J.F., Bloch, M.H.: Clinical assessment of Tourette syndrome and tic disorders. Neurosci. Biobehav. Rev. **37**(6), 997–1007 (2013)
3. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015)
4. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
5. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press (1995)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Liaw, A., Wiener, M.: Classification and regression by random forest. R News **2**(3), 18–22 (2002)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Bernabei, M., et al.: Automatic detection of tic activity in the Tourette Syndrome. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 422–425. IEEE, August 2010
11. Shute, J.B., et al.: Thalamocortical network activity enables chronic tic detection in humans with Tourette syndrome. NeuroImage Clin. **12**, 165–172 (2016)
12. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
13. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1867–1874 (2014)
14. Yang, H., Liu, H.: Cascaded regression based landmark localization for robust facial feature tracking. IEEE Trans. Image Process. **24**(8), 2479–2490 (2015)
15. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1021–1030 (2017)

16. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
17. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634(2015)
18. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 791–800, October 2016
19. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Hara, K., Kataoka, H.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018)
23. Jiang, B., Zhang, L., Zhang, D., Zhang, M., Yang, H., Guo, Y.: T3D: temporal 3D ConvNet for real-time action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12309–12316 (2020)
24. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5534–5542 (2017)
25. Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1796–1804 (2015)
26. Zhang, Z., Xu, J., Yang, L., Xiong, Y.: Deep learning based intervertebral disc segmentation from weakly labeled training data. J. Med. Syst. **42**(6), 100 (2018)
27. Durand, T., Mordan, T., Thome, N.: Weakly supervised object detection: a survey. Int. J. Comput. Vision **127**(9), 1191–1234 (2019)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
29. Zhuĺ, X., Ghahramaniĺн, Z.: Learning from labeled and unlabeled data with label propagation (2002)
30. Ma, X., et al.: Dimensionality-driven learning with noisy labels. In: International Conference on Machine Learning, pp. 3355–3364. PMLR, July 2018
31. Lee, D.H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, no. 2, p. 896, June 2013
32. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. **32**(2), 523–534 (2020)
33. Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S.: Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4568–4577 (2021)
34. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Trans. Knowl. Data Eng. (2022)

35. Huynh, T., Nibali, A., He, Z.: Semi-supervised learning for medical image classification using imbalanced training data. In: Computer Methods and Programs in Biomedicine, p. 106628 (2022)
36. Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C.: SimMatch: semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14471–14481 (2022)