A review of speech enhancement based on self-supervised learning

Sisi Du
College of Information Engineering
Jiangxi University of Science and Technology
Ganzhou, Jiangxi, China
2923283504@qq.com

Huifang Lai*
Academic Administration
Jiangxi University of Science and Technology
Ganzhou, Jiangxi, China
252071256@qq.com

Fengpei Ge
Beijing University of Posts and Telecommunications
Beijing, China
gefengpei@bupt.edu.cn

Chundong Xu
College of Information Engineering
Jiangxi University of Science and Technology
Ganzhou, Jiangxi, China
xuchundong@jxust.edu.cn

Abstract

The Speech enhancement based on deep learning has become one of the mainstream research directions. The quality of speech enhancement is improved by training the correspondence between noisy speech and clean speech. However, in practical applications, it is not easy to simultaneously collect clean speech data. The self-supervised learning mechanism allows models to learn from data without explicit labels without relying on label data, so it has become one of the research hotspots in the field of deep learning. In self-supervised learning, the model learns the representation of data by solving prediction tasks related to the input data, which are usually generated by the data itself rather than externally provided labels. The method based on self-supervised learning does not need to use noisy speech-clean speech data pairs, but by mining the semantic and other related information of noisy speech data to obtain the goal of advancing speech quality. In order to fully elaborate the speech enhancement algorithm based on self-supervised learning, the model based on deep learning is firstly reviewed, and the self-supervised learning methods are classified according to different auxiliary tasks. Secondly, several instance models are introduced in detail according to different model training methods. Then, the advantages and disadvantages of several instance models are analyzed in detail according to different model training methods.Finally, the existing research work in this sub-field is summarized in detail, and the potential development trend in the future is discussed.

CCS Concepts

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Neural networks**;

*Corresponding author

AIBDF 2024, Ganzhou, China

ACM ISBN 979-8-4007-1086-5/2024/12
https://doi.org/10.1145/3718491.3718551

Keywords

Speech enhancement, Deep learning, Self-supervised learning

ACM Reference Format:
Sisi Du, Huifang Lai, Fengpei Ge, and Chundong Xu. 2024. A review of speech enhancement based on self-supervised learning. In *The 4th Asia-Pacific Artificial Intelligence and Big Data Forum (AIBDF 2024), December 27–29, 2024, Ganzhou, China*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3718491.3718551

1 Introduction

Speech enhancement (SE) technology is for improving the quality and intelligibility of perceptual speech. It is the front-end preprocessing module of Automated Speech Recognition (ASR) [1], Speaker Recognition [2], Hearing Aids [3], etc. It is also one of the important research directions within the realm of speech processing.

In contrast to the traditional speech enhancement technology [4], the speech enhancement technology on the basis of DNN can greatly improve the speech enhancement effect. Diffrent from the traditional speech enhancement model based on digital signal processing, SE model based on deep learning can better deal with non-stationary noise problems and extract deeper semantic features. Nowadays, most of the speech enhancement research on deep learning focuses on supervised learning and semi-supervised learning scenarios. In these scenarios, a large number of noisy speech-clean speech pairs are needed to train the speech enhancement task model. Although these studies have achieved good performance, the model relies heavily on clean speech data, and its performance is not ideal when the sample data is small. The cost of collecting clean voice is high, especially in some complex and changeable environments. Acquiring a vast amount of clean voice data is unfeasible; supervised learning scenarios only train part of noisy speech-clean speech data, which is prone to overfitting the model, resulting in poor generalization of the model. This phenomenon is particularly serious when clean speech of noisy is scarce. The fully supervised speech enhancement deep learning model is vulnerable to suffer Overfitting, which is related to clean speech, resulting in poor robustness of the speech enhancement model.

In order to solve the deficiencies of supervised learning, Self-supervised Leaning (SSL) offers a promising paradigm to learning, which Lowering the reliance on manual labels. In SSL, the model learns by training hand-crafted auxiliary tasks, in which the supervisory signal is automatically obtained from the data itself without manual annotation. SSL allows the model to learn more information representations from unlabeled data with well-designed excuse tasks. Thereby it can get high accuracy, generalization, and robustness on various tasks.

In recent years, SSL has obtained great result in image processing and other fields. Because it can reduce the target data required by deep learning models and generate objective functions independently [5] [6]. Summary articles on SSL have also increased.CHang [7] published a review of SSL for speech recognition in 2021. They introduced the commonly used SSL methods in speech recognition in detail, and focused on the general application of self-supervised speech representation pre-trained by advanced end-to-end automated speech recognition (E2E-ASR) model. AbdelraHman [8] published a review on self-supervised speech representation in 2022. This paper introduces the theory and method of SSL representation model in detail, and summarizes the framework, classification and data set of speech SSL model. It holds a pivotal position in the application of SSL in Speech field, and also establish a theoretical foundation for the understanding of speech enhancement SSL in this paper. In the same year, Huang [9] studied the SSL of speech enhancement and separation. In this paper, the effects of 13 SSL upstream approaches on SE and speech separation tasks were evaluated. The focus was on exploring the adaptability of SSL upstream models to downstream tasks, without systematically introducing the composition and classification of SE self-supervised learning models.Wu [10] published the generation and comparison methods of SSL in 2023. They introduced in detail the empirical applications of SSL methods in computer vision, natural language processing and graphics learning, focusing on how SSL methods are applied to downstream tasks. The central content is SSL methods. However, due to the huge amount of training data of the self-supervised model, the sharp elevated parameter numbers entail more sophisticated equipment in application, and different self-supervised methods have not sufficient theoretical basis for improving the veracity and intelligibility of speech, and the work in SE is still relatively small. In order to take advantage of SSL methods to enhance the performance of SE models and tap their application potential in SE, this paper systematically combs the speech enhancement models based on SSL on the basis of summarizing the opportunities and challenges of SE. According to the different classifications of model training methods, the speech enhancement models on the basis of SSL are introduced, and their performance is compared comprehensively. Finally, the development direction of SE based on SSL is prospected.

2 SE model based on deep learning

2.1 TRADITIONAL SE MODEL

In traditional speech enhancement, noisy speech can be provided by the following formula (1) :

$$\mathrm{y(n)} = x(n) + d(n) \tag{1}$$

Where $x(n)$ denotes clean speech data,$d(n)$denotes additive noise data, and$y(n)$denotes noisy speech signal. Additive noise has a serious impact on speech quality, and non-additive noise such as reverberation noise can be converted into additive noise in some ways.

The process of SE needs to expect the clean speech data from the noisy speech one$y(n)$, so that the difference with$x(n)$is as small as possible. Because the noise superimposed in the speech has different types and different signal-to-noise ratios, SE needs to have good generalization performance for noise, that is, it needs to have the ability to remove different types and different signal-to-noise ratios.

2.2 ENHANCEMENT MODEL BASED ON DEEP LEARNING

Traditional speech enhancement techniques usually need to model the noisy speech signal based on setting constraints or assumptions to solve the estimated clean speech signal. When these restrictions and supposition are not established, the performance of SE will be poor. SE model on the basis of deep learning does not straightly calculate the estimated clean speech data, but gets the arguments of the optimal methods on the dataset depending on the configuration of objective function, resulting implicitly excavate the nonlinear mapping relationship$f(\bullet)$between the noisy and the clean speech data, and realize the mapping from the noisy voice data$y(n)$to the clean one$x(n)$ [11].

SE model based on deep learning is shown in Figure 1. The deep neural network can be DNN [12], Wave-Unet [13], DCCRN, DEMUCS [14] and other different network structures. The formal description is shown in Formula (2) :

$$\tilde{x} = F(y, \theta) \tag{2}$$

where$y$and$\tilde{x}$Stand for the noisy and estimated clean speech data. They can be both time-domain waveforms and time-frequency domain transform features. Also it can be a time-frequency domain mask estimate. SE model based on deep learning transforms the problem of SE into the problem of finding the optimal solution of parameterθ.

3 SSL model

As described in the previous chapter, the SE model deep learning-based approaches require refinement to boost SE effectiveness by training a large number consisting of noisy speech samples and their clean counterparts. In practical environmental applications, the acquisition cost of clean speech signals is high, and noisy speech signals are easy to collect. SSL model can train the model without additional clean speech, aiming to automatically discover supervised signals by constructing auxiliary tasks. These supervised signals only come from the data itself without external annotation. The core idea of SSL is that one part of the data can be used to predict another part, so that the model can be trained without explicit supervision.

3.1 SSL paradigm

SSL is a one of unsupervised learning, also known as pretext task. SSL mainly take advantage of pretext tasks to excavate its own

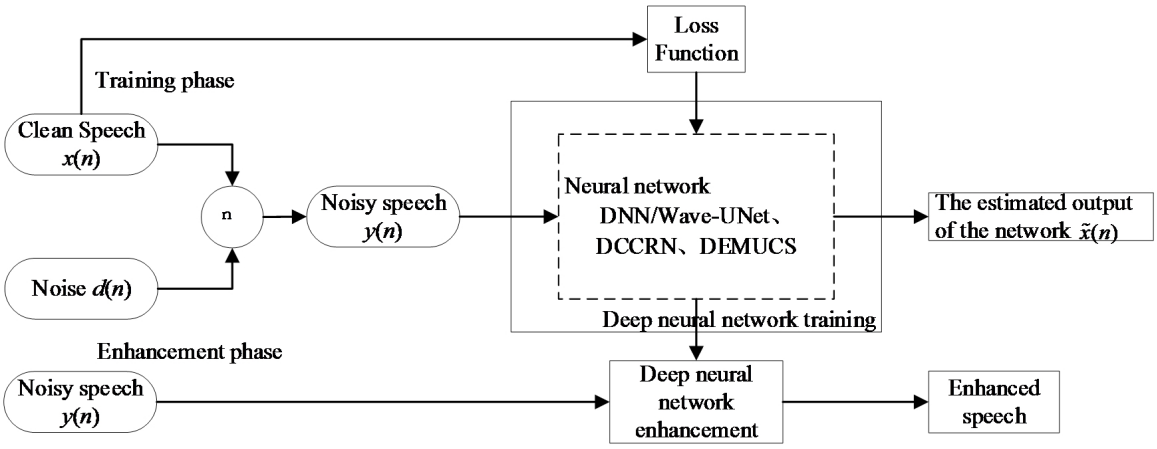

Figure 1: Block diagram of deep learning-based speech enhancement model

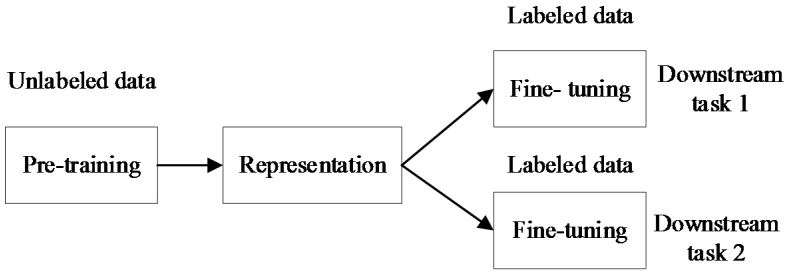

Figure 2: Self-supervised model

supervision information from massive unsupervised data, and trains the model through this structured supervision information, so as to it can learn worthwhile representation of SE tasks.

The training of self-supervised models is usually splited into two stages. The first phase is called pre-training, which uses a large amount of unlabeled data to train without specific tasks ; the second phase can be called fine-tuning. According to different speech tasks, a small amount of labeled data is used to train the model, as shown in Figure 2.

3.2 SSL method

According to the different auxiliary tasks, the self-supervised model can be divided into three categories : generation, comparison and prediction. The generative model focuses on the information embedded in the data, based on auxiliary tasks such as reconstruction, including predicting future input from past input, predicting original input from hidden input or from other corrupted views. These tasks use the attributes and structures of the data itself as self-supervised signals. At present, the generative models used for speech include autoregressive prediction [15] and masking reconstruction [16]. Speech includes a lot of complex features. Learning to redevelop the original speech data may not be the optimal method to find potential changes in speech context [8]. Therefore, this problem is solved by distinguishing the target sample (positive) and the interference sample (negative) learning representation of the given anchor representation.wav2vec [17] and so on are currently used in speech contrast models. The difficulty of this method lies in how to construct positive and negative samples. In order to avoid this problem, the predictive model usually uses a completely independent model, which does not use the contrast loss, but uses the loss function, such as square error. At present, predictive self-supervised models for speech include discrete BERT [16],WavLM [18] and data2vec [19].

4 SE model based on SSL

The supervised deep learning SE model network directly establishes a relational correspondence between noisy speech and clean speech data, while SSL method uses pretext tasks to pre-train the neural network on noisy speech data. Most of the existing self-supervised SE models need to use a limited amount of clean speech data to fine-tune the model, and a small part can directly use the SSL method to retrieve the clean speech data underlying the noisy one without the corresponding clean speech label data. Table 1 shows the existing SE work using SSL. These models improve the enhancement performance of the original system to varying degrees.

According to the way in which the self-supervised model is applied between speech enhancement tasks, the training schemes of the existing self-supervised SE models can be divided into two types : Pre-Training and Fine-Tuning, Joint Learning [26]. In the PF

Table 1: Analysis of speech enhancement model using self-supervised learning

Model	Merit	Defect
SS-DAELD [20]	**Joint learning**, the first attempt to introduce SSL training in supervised speech augmentation models. Given that the training dataset encompasses numerous noise scenarios and different SNR ranges, the SS-DAELD tends to achieve higher performance.	In a stationary noise environment, the performance will be worse than that of the supervised speech enhancement model.
K-SENet [21]	**Joint learning**, In this paper, the joint training of self-supervised feature models is introduced into the basic model Wave-U-Net, and a perceptual loss function optimization model combined with self-supervised features is proposed, which effectively improves the model performance.	The model processes a large amount of data and is incompatible with real-time speech enhancement .
DeVo [22]	**Joint training**, the combination of the self-supervised model and the vocoder HiFiGAN, can synthesize clean speech by only inputting noisy speech, and the subjective evaluation is better than the foregone state-of-the-art SE network, and real-time SE can be carried out.	The objective evaluation was slightly inferior to the supervised enhancement model.
SSF-CVAE [23]	**Pre-training-fine-tuning**, the introduction of clean self-supervised features in the Conditional Variational Autoencoder (CVAE) framework for latent spatial modeling, does not increase the computational effort, but outperforms the basic model in both subjective and objective evaluation.	SSL is not used to learn more feature information that is conducive to speech enhancement, and the performance improvement is limited.
SSSR-BLSTM [24]	**Pre-training-fine-tuning**, the introduction of self-supervised feature coding of clean speech and noisy speech effectively improves the intelligibility of the model and MOS evaluation index.	SSL is not used to learn more feature information that is conducive to speech enhancement, and the performance improvement is limited.
WavLM-DEMUCS [25]	**Pre-training-fine-tuning**, using pre-trained WavLM model to initialize the DEMUCS speech enhancement model.	The self-supervised model is used as a high-level feature extractor, and SSL is not used to learn more feature information that is conducive to speech enhancement, and the performance improvement is limited.

scheme, based on the basic model of speech enhancement, the encoder of the self-supervised model is first pre-trained on a multitude of original noisy speech signal sets. Then, the encoder is employed for training the SE model and fine-tune the clean speech data sets. It can be seen as the initialization of SE model encoder parameters. The data sets used for PF can be arbitrary (as shown in Figure 3). In the JL scheme, the pretext task of the self-supervised model and the speech enhancement task are jointly trained. The loss function of model is composed of the self-supervised and speech enhancement task loss functions, and a hyperparameter is used to control the contribution of the self-supervised model [26]. as shown in Figure 4.

5 Conclusion

This paper summarizes two training methods based on self-supervised speech enhancement, namely, Pre-training and fine-tuning and Joint Learning schem. Compared with the supervised speech enhancement task, the self-supervised speech enhancement task is more in line with the reality of life and is favored by more and more researchers.

The construction of a limited label dataset model is a necessary prerequisite for promoting the development of SE. With the development of SSL, speech enhancement based on self-supervised learning has also experienced a blowout development. This is also accompanied by new problems that cannot be effectively solved by current methods. New problems inspire researchers to continuously explore new models and promote the development of this field.

In this paper, the following future research directions are proposed for the difficulties in the development of self-supervised speech enhancement :

(1)In view of the large demand for model training data, we can try to optimize the model structure and decrease the number of parameters while remaining the model capability. For example, using more efficient neural network architectures or using techniques such as pruning and quantization to reduce model complexity.

(2)The adaptability of the self-supervised learning model to different data sets can be improved by introducing multi-task learning, transfer learning and other technologies.

(3)Self-supervised learning tasks that are closer to the requirements of speech enhancement tasks can be designed to extract more targeted features.

Acknowledgments

This work was Supported by National Natural Science Foundation of China (12204062, 11864016), and Jiangxi Province Key Laboratory

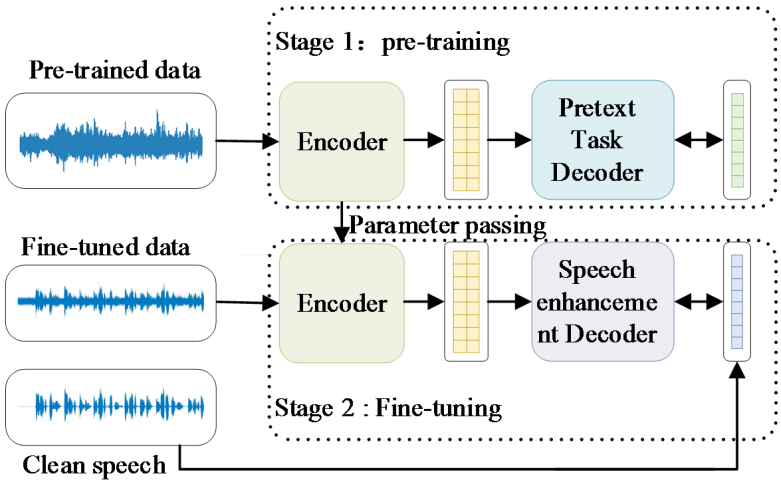

Figure 3: Pre-training and fine-tuning scheme

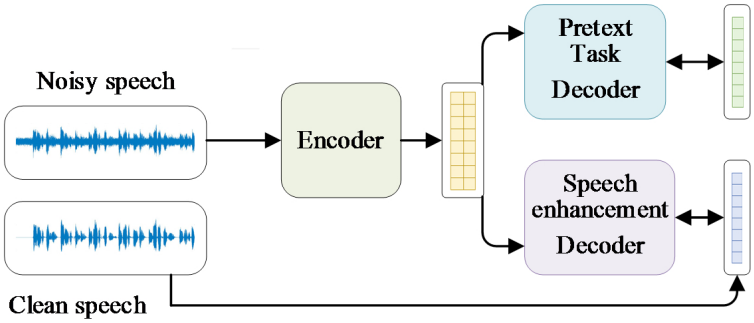

Figure 4: Joint Learning schem

of Multidimensional Intelligent Perception and Control, China (No. 2024SSY03161)

References

[1] XIANG Y, BAO C C. Speech enhancement via generative adversarial LSTM networks[C]//2018 16th International Workshop on Acoustic Signal Enhancement . Tokyo: IEEE, 2018: 46-50.

[2] Hua Long, Linpu Zhang, Yubing Shao, *et al.* Speaker feature constrained multi-task convolutional network speech enhancement[J]. Mini-Micro Systems, 2021,42(10):2178-2183.

[3] Yatao Zhu, Fei Chen, Yuchen Zhang, *et al.* Speech enhancement algorithm for binaural hearing aids based on recurrent neural network[J]. Journal of Sensing Technology, 2021,34(09):1165-1172.

[4] Zhi Tao, Heming Zhao, Chenghui Gong. Speech enhancement based on auditory masking effect and barklet transform [J]. Journal of Acoustics, 2005, 30(4): 367-372.

[5] T Chen, S Kornblith, M Norouzi, *et al.* A simple framework for contrastive learning of visual representations[C]// in Proc. Int. Conf. Mach. Learn., 2020, pp. 1597-1607.

[6] P Liu, W Yuan, J Fu, *et al.* Pre-train prompt and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Comput. Surv., Aug., 2022.

[7] X Chang *et al.* An Exploration of Self-Supervised Pretrained Representations for End-to-End Speech Recognition[J]. 2021 IEEE Automatic Speech Recognition and Understanding Workshop, Cartagena, Colombia, 2021, pp. 228-235.

[8] A Mohamed *et al.* Self-Supervised Speech Representation Learning: A Review[J]. in IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1179-1210, Oct. 2022.

[9] Z Huang, S Watanabe, S. -w. Yang, *et al.* Investigating Self-Supervised Learning for Speech Enhancement and Separation[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 2022, pp. 6837-6841.

[10] X Liu *et al.* Self-Supervised Learning: Generative or Contrastive[J]. in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 857-876, 1 Jan. 2023.

[11] Y. He. Research on Intelligent Information Collection of Business Negotiation System based on Speech Rate Change Recognition Algorithm[C]//2022 International Conference on Inventive Computation Technologies, Nepal, 2022, pp.

[12] XU Y, DU J, DAI L, *et al.*A regression approach to speech enhancement based on deep neural networks[J].IEEE/ACM Transactions on Audio Speech and Language Processing,2015,23(1):7-19.

[13] Guimarães H R, Nagano H, Silva D W. Monaural speech enhancement through deep wave-U-net[J]. Expert Systems with Applications,2020, 158:113582.

[14] Alexandre Defossez, Gabriel Synnaeve, Yossi Adi. Real Time Speech Enhancement in the Waveform Domain[J].arXiv:2006.12847,2020.

[15] Y.-A Chung, W.-N Hsu, H. Tang, *et al.* An unsupervised autoregressive model for speech representation learning[C]//in Proc. Interspeech, 2019, pp. 146-150.

[16] J. Devlin, M.-W Chang, K. Lee, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding[C]//in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics:Hum. Lang. Technol,2019,pp.4171-4186.

[17] S. Schneider, A. Baevski, R. Collobert, *et al.* wav2vec: Unsupervised pre-training for speech recognition[C]//in Proc. Interspeech, 2019,pp. 3465-3469.

[18] S. Chen *et al.* WavLM: Large-scale self-supervised pre-training for full stack speech processing[C]//IEEE J. Sel. Topics Signal Process., early access, Jul. 4, 2022.

[19] A. Baevski, W. Hsu, Q. Xu, *et al.* data2vec: A general framework for self-supervised learning in speech, vision and language[C]//2022, arXiv:2202.03555.

[20] R. E. Zezario, T. Hussain, X. Lu, *et al.* Self-Supervised Denoising Autoencoder with Linear Regression Decoder for Speech Enhancement[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 2020, pp. 6669-6673.

[21] T. Sun *et al.* Boosting the Intelligibility of Waveform Speech Enhancement Networks through Self-supervised Representations[C]//2021 20th IEEE International Conference on Machine Learning and Applications, Pasadena, CA, USA, 2021, pp. 992-997.

[22] B. Irvin, M. Stamenovic, M. Kegler, *et al.* Self-Supervised Learning for Speech Enhancement Through Synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 2023, pp. 1-5.

[23] Y. Lee, K. Jung. Boosting Speech Enhancement with Clean Self-Supervised Features Via Conditional Variational Autoencoders[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 12396-12400.

[24] G. Close, W. Ravenscroft, T. Hain, *et al.* Perceive and Predict: Self-Supervised Speech Representation Based Loss Functions for Speech Enhancement[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece,2023,pp.1-5.

[25] X. -Y. Zhao, Q. -S. Zhu, J. Zhang. Speech Enhancement Using Self-Supervised Pre-Trained Model and Vector Quantization[C]//2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Chiang Mai, Thailand, 2022, pp. 330-334.

[26] Y. Liu *et al.* Graph Self-Supervised Learning: A Survey[C]// in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 6, pp. 5879-5900, 1 June 2023.