Machine Learning-Based VHF Lightning Radiation Sources Identification

Shuang-Jiang Du, Yun Li, Zheng Sun, Shi Qiu, and Li-Hua Shi, *Member, IEEE*

***Abstract*—Identifying the validity of the location result is an important step in lightning radiation source mapping, which can eliminate the interference of noise location results, retain the real radiation source, and obtain a clear and continuous lightning channel development map. The localization methods, such as electromagnetic time reversal and multiple signal classification have high location accuracy, but the validity identification of their location result depends on the subjectively set threshold, which makes it hard to accurately distinguish the location results of weak radiation source and noise. In order to retain the weak radiation sources as much as possible and eliminate the noise interference, this article proposes two machine learning-based validity identification methods, namely, the continuous wavelet transform-based convolutional neural network model (CWT-CNN) and the spatiotemporal clustering algorithm. The CWT-CNN model can learn the time–frequency characteristics of the sliding window data to identify the lightning radiation source in advance and only retain the data containing useful signals. The spatiotemporal clustering algorithm can adaptively adjust the clustering parameters by learning the spatial and temporal distribution properties of the known location results to restore weak radiation sources that were incorrectly eliminated by former criteria. Experiments and analysis show that compared with the previous validity identification methods, the two methods proposed in this article are good at separating location results of weak radiation source from noise points, can obtain more continuous lightning maps without noise interference, and find some additional lightning branches.**

***Index Terms*—Clustering, convolutional neural network (CNN), lightning mapping, locating result identification.**

I. Introduction

FINE localization algorithms of very-high frequency (VHF) lightning radiation sources can effectively help analyzing the development process of lightning and the spatial structure distribution of radiation sources. Compared with the locating process, the validity identification of location results is also a very important process, that is, to determine whether the obtained location results are reasonable and credible.

The VHF-based lightning mapping array (LMA) [1], [2], [3], [4] usually uses the time of arrival to locate 3-D coordinates of radiation sources by fitting multidegree-of-freedom equations. The validity of the solution is determined by a fitting error chi-square and setting a threshold based on the statistic analysis. Continuous broadband interferometer [5], [6], [7], [8] uses the cross-correlation between waveforms to obtain the relative phase difference or time delay, and then perform 2-D localization. These methods generally use sliding window to extract data, and locate each group of data in turn. Not all the locating results of the sliding windows are valid, to eliminate the noise points, Stock et al. [7] proposed four metrics, namely, closure delay, standard deviation, multiplicity of contributing windows, and correlation amplitude. Among them, the first three are not associated with the strength of the signal, whereas the correlation amplitude does.

To pursue high-localization accuracy and improve the performance under noise, electromagnetic time reversal (EMTR) [9], [10], [11], [12], multiple signal classification (MUSIC) [13], and orthogonal propagator method (OPM) [14] are paid more and more attention in this field recently, due to their superior performance and capability of capturing the weak radiation sources. In these localization methods, the time domain waveforms are also segmented and extracted by sliding window and processed in turn. At the same time, the global maximum search method is used to locate the coordinates of radiation sources. To identify the valid location results and remove noise points, Wang et al. [15] first proposed the coherence ratio (CR) and energy ratio (ER) metrics. In later studies, Chen et al. [16] proposed the MUSIC with uniform L-shaped array, Chouragade et al. [17] proposed Real-valued MUSIC (RV-MUSIC), Liu et al. [18] used EMTR for 3-D coordinate localization, and Li et al. [14] used OPM for fine lightning radiation sources. All of these localization methods adopt these two metrics and can obtain fine and clear lightning development structures.

The CR metric filters out noise points by limiting the variation of the spatial position of multiple adjacent radiation sources in the time sequence, but cannot completely remove the noise points. Because when the sliding window is too dense or the variation threshold is relatively loose, the noise points cannot be completely filtered by CR metric only. In order to obtain a clear lightning map without noise interference, the ER metric can be used for a second filtering. The ER metric distinguishes the radiation source from noise points by calculating the ratio of the maximum energy value to the average energy value during the global search. This method is obviously related to the energy of the radiation. The problem is that the ER values of many weak radiation sources are not high and are comparable to

Received 14 April 2024; revised 30 June 2024; accepted 5 September 2024. Date of publication 8 October 2024; date of current version 18 December 2024. This work was supported by the National Science Foundation of China under Grant 51977219 and Grant 42105077. *(Corresponding author: Li-Hua Shi.)*

The authors are with the National Key Laboratory on Electromagnetic Environmental Effects and Electro-Optical Engineering, Army Engineering University of PLA, Nanjing 210007, China (e-mail: shuangjiangdu@163.com; lihuashi@aliyun.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TEMC.2024.3466962.

Digital Object Identifier 10.1109/TEMC.2024.3466962

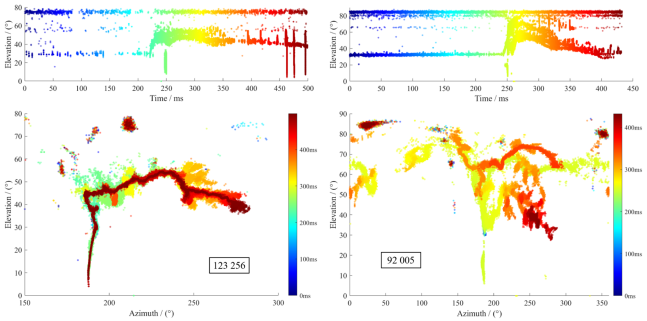

Fig. 1. Two VHF lightning mapping results filtered by the CR/ER metric: (a) Mapping result of Trig230553 with CR≥0.3/ER≥0.5. (b) Mapping result of Trig231042 with CR≥0.3/ER≥0.35.

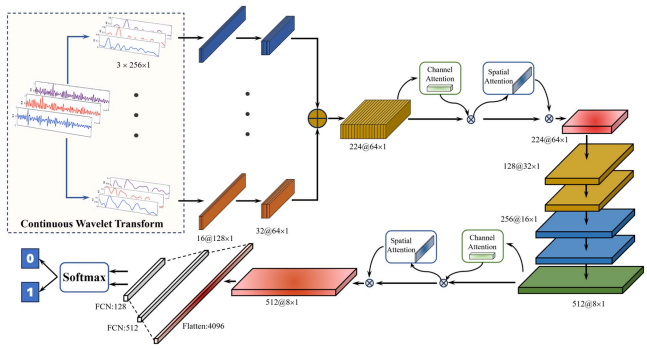

Fig. 2. CWT-CNN-based lightning radiation sources identification.

or smaller than those of noise. To get a clear lightning map, many weak sources of radiation must be filtered out. On the contrary, in order to obtain more comprehensive information about radiation sources, noise interference must inevitably be introduced. Two lightning maps of artificially triggered lightning VHF data located by EMTR with specific CR/ER thresholds reveal this problem, as shown in Fig. 1. The above problem arises from the empirically selected CR and ER thresholds and the incompleteness of these two metrics.

In this article, considering localization algorithms, such as EMTR and MUSIC that use sliding windows to extract data and global maximum search for location, two location validity identification methods based on machine learning are proposed, namely, a continuous wavelet transform-based convolutional neural network (CWT-CNN) and adaptive spatiotemporal clustering, which try to objectively distinguish radiation source and noise location, and retain weak radiation source as much as possible while avoiding noise interference. By means of supervised learning, the CWT-CNN model learns the time–frequency characteristics of the time-domain signal to distinguish the radiation source from noise, and then judges the validity of the corresponding location result. By learning the spatiotemporal distribution properties of pre-existing radiation sources, the spatiotemporal clustering algorithm is able to further identify noise point while restoring potential weak radiation sources filtered out by the CR/ER metric.

II. Theories

In this section, two machine learning methods, the CWT-CNN model and spatiotemporal clustering algorithm, are proposed to address the locating results identification.

A. CR and ER Metrics for Fault Points Filtering

Wang et al. [15] explained the specific meaning of CR and ER metrics in detail, and here we add some additional analysis.

1) CR Metric:

$$\mathrm{CR} = \frac{N(\phi_i \le \phi_{\mathrm{thr}})}{N} \tag{1}$$

where N is the number of location points adjacent to each other in the time series. ϕ is the spherical center angular distance of the 2-D coordinates of the two location points projected onto the 3-D spherical surface. The numerator of the formula represents the number of paired location points whose distance is less than a specified threshold. Obviously, the choice of N and the setting of the threshold have an influence on the final effect.

2) ER Metric:

$$\mathrm{ER} = \ln\left\{\frac{\max[P(\varpi_i)]}{\frac{1}{W}\sum_{i=1}^{W} P(\varpi_i)}\right\} \tag{2}$$

where $P(\varpi_i)$ is the energy or amplitude distribution in the whole space. ER represents the ratio of the maximum energy value, that is, the energy value of the radiation source, to the spatial average energy value. Based on the view that the energy of the radiation source is higher than that of the noise, the noise points can be removed by setting an appropriate threshold.

However, experiments show that the ER metric cannot completely separate the noise and radiation sources. To completely remove noise, the ER threshold is set to a high value. Although a clear lightning map is obtained in this way, many weak radiation sources are incorrectly filtered out, resulting in some channels not being continuous. On the contrary, if the ER threshold is lowered, the noise interference will not be eliminated and there will still be many noise points in the lightning map.

B. CWT-CNN for Radiation Source Identification

In order to solve the problem in the abovementioned method, we propose to automatically identify useful data segments obtained by sliding windows, in advance of further processing. A classification model of seven-channel 1-D convolutional neural network (1DCNN) is designed as shown in Fig. 2. The VHF time-domain waveforms received by the three antennas are transformed by continuous wavelet transform (CWT) to obtain seven groups of wavelet coefficient envelopes in different frequency bands, which are used as the input of the model. The coefficient envelopes from three antennas in each subfrequency band are combined to one channel of the CNN and there are seven parallel channels in the first processing stage. In each channel, the time

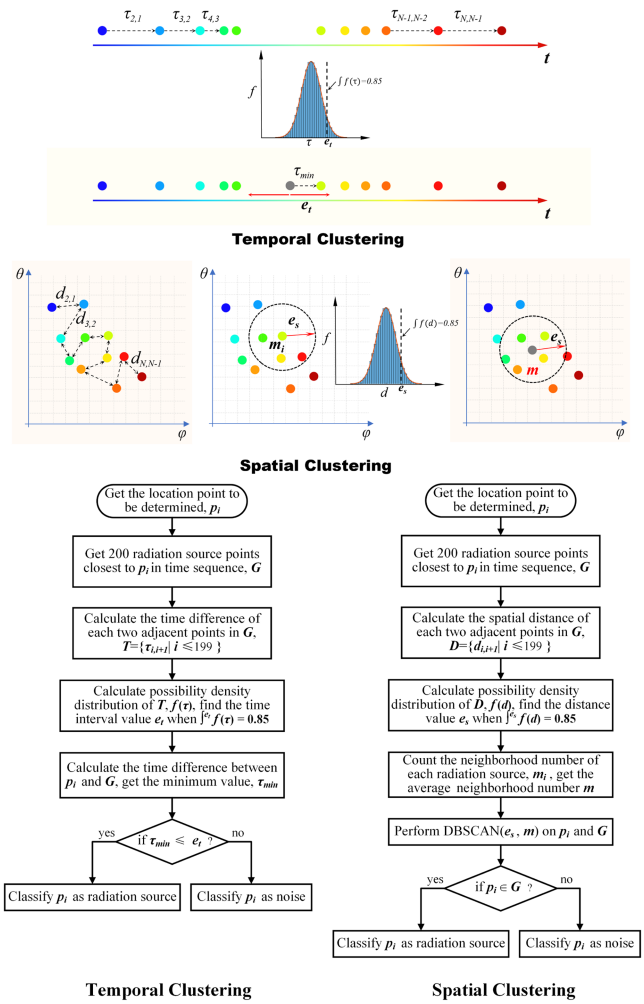

Fig. 3. Adaptive spatiotemporal clustering model and its flowchart.

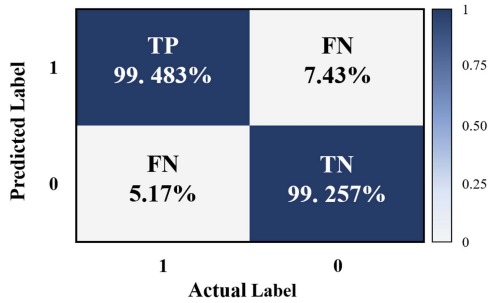

Fig. 4. Confusion matrix predicted by the CWT-CNN model.

series data are processed by two 1DCNN modules, and their output feature vectors are integrated through concatenation. The channel and spatial attention mechanisms [19] are introduced and adjusted into 1-D attention mechanisms in the initial and final stages of the model to enhance the features at specific times and channels. After several convolutional layers, the feature vectors are transferred to a fully connected three-layer network, and the binary classification result is output by the softmax

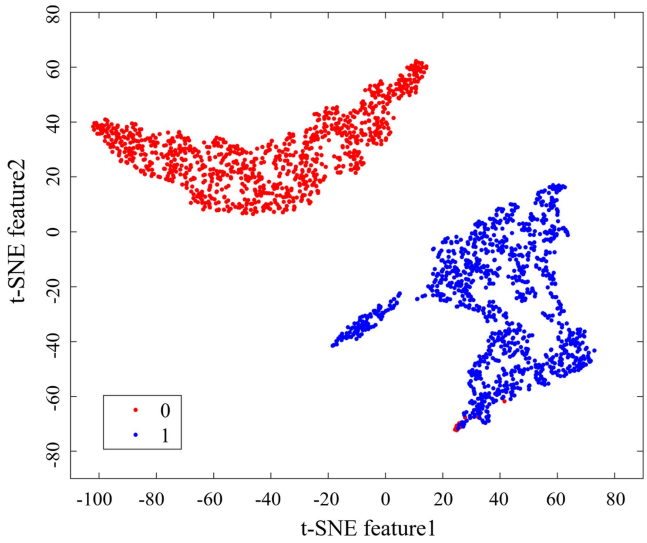

Fig. 5. Feature visualization of positive and negative samples by t-SNE.

function at the end of the model, where a value of 1 indicates that the data in the sliding window is the radiation source, and a value of 0 means noise.

By learning the time–frequency information in different frequency bands of the signals received by different antennas, the CNN model can know whether the sliding window contains a radiation source, so as to judge whether the locating result corresponding to the sliding window is valid or not.

C. Adaptive Spatiotemporal Clustering

The development of lightning channels is a constantly changing process in time and space. Therefore, the location results of radiation sources should have a certain spatial-temporal continuity compared with noise points. On this basis, we propose a spatiotemporal clustering method with adaptive parameters to find more valid low energy location results.

This method first obtains an initial clear lightning map by the CR/ER metric. In order to make up for the weak radiation sources filtered out by the ER metric, the rest location points are clustered in time and space successively to determine whether they belong to the same class as the initially retained points. If so, they are regarded as valid location results. Fig. 3 demonstrates the schematic diagram and flowchart of temporal and spatial clustering.

1) Temporal Clustering: In a certain period of time, the occurrence time of real radiation sources shows a specific statistical distribution, and the average time interval of two adjacent radiation sources in the time series can reflect the development time characteristics of the lightning. To determine the validity of a specific location result, we select the 200 real radiation sources that are closest to it in a time series. For these real radiation sources, the time interval between two temporally adjacent sources is calculated in order to obtain a time difference

set as follows:

$$\tau_{i,j} = t_i - t_j$$
$$T = \{\tau_{i+1,i} | \, i \in [1, N-1]\}. \quad (3)$$

Typically, the time intervals present a specific statistical distribution. We take the corresponding value at 85% probability in its probability density function as its average time interval value e_t, which is similar to a normal distribution, with 85% confidence level that the time intervals of adjacent radiation sources are within this range. For the locating point to be determined, the time interval of the closest radiation source point on its timing sequence, $\tau_{\min}$, is calculated, compared with the average time interval e_t. If $\tau_{\min} \leq e_t$, it is considered that the locating point satisfies the temporal distribution of the radiation source and belongs to the same class as the real radiation source in the temporal domain.

2) Spatial Clustering: The spatial distribution is not unidirectional, so density-based spatial clustering of applications with noise (DBSCAN) [20] is used. Neighborhood radius, e_s, and minimum number of points, m, are two basic parameters of DBSCAN. Similarly, over a period of time, the radiation source will develop into a certain spatial extent. The spatial distance of two temporal adjacent radiation sources can reflect the development speed of lightning channels. The Euclidean distance of all two adjacent points on the time series is calculated, and a distance difference set is obtained. Similar to the time difference set, this set also has a specific distribution, and we take the value of the probability density function at 85% probability as the average distance, e_s. With each radiation source as the center, the number of other radiation sources, m_i, in its neighborhood radius e_s is calculated. Average over m_i yields the final minimum number of points m. The expression is as follows:

$$d_{i,j} = \text{norm}(d_i - d_j)$$
$$m_i = \text{card}(\{d_{i,j} | \, d_{i,j} \leq e_s, i, j \in [1, N], i \neq j\}) \quad (4)$$

where norm denotes the Euclidean distance between two points, and card denotes the number of elements in the set.

For the locating result to be determined, if the point keeps the same class with the real radiation sources in time and space, it is considered as a valid locating result. It is worth noting that the larger the confidence level is set, the larger the corresponding distance threshold is, and more located points will be retained, but the noise points will also be retained with a greater probability. On the contrary, if the probability value is set to be small, the number of retained location results will be reduced, but the noise points will not be retained. Different from the CR metric, the parameters of the spatiotemporal clustering algorithm are adjusted adaptively according to the distribution characteristic of the radiation sources, and the confidence level is also a value set in accordance with the statistical law, so this method is relatively objective.

III. Experiments on VHF Lightning Data

In this section, we use two sets of classical artificial triggered lightning VHF data, Trig230553 and Trig231042, to test the performance of our proposed methods. The two data are recorded by MARCOS uniform L-shaped VHF array [14] at different times, and digitized at a rate of 500 MHz, lasting for 500 ms.

In the time domain, the time reversal technique (TR) method works like a delay-and-sum operation for the time reversed time series. This algorithm is more robust than the equation-solving and the matrix-transforming algorithm but more time-consuming. In our previous work [12], we try to accelerate it by GPU processing. By introducing the CNN preprocessing to identify useful time windows and eliminate the unnecessary ones, the operation time will be further saved. The reason we choose frequency domain TR (FDTR) is the time delay can be replaced by phase shift without time-domain interpolation and some unwanted frequency bands can be removed. Therefore, in this article, it is continued adopted to locate the lightning radiation sources. For the time-domain waveform, the sliding window is used to obtain the data segments, and the length of the sliding window is 256 sampling points with an overlap rate of 0.75.

A. Training and Testing of CNN Model

For the CWT-CNN model, we use three steps to test its performance. In the first step, a classical artificial triggered lightning VHF data, Trig230553, is used as a training and test dataset for the CWT-CNN model. We use FDTR to obtain the initial location results of the VHF lightning and then set the threshold CR = 0.3 and ER = 1.1 to filter the noise points to get a clear lightning map. The retained location points and their corresponding sliding window data are set as positive samples. Then, we set CR = 0.0, to get the noise points as negative samples. To ensure data balance, 200 000 positive and negative samples were randomly selected.

These positive and negative samples are divided equally into test samples and training samples. The training samples were fed into the CWT-CNN model for training for 40 epochs, and the validation loss of the model gradually converged. Then, the test samples are sent to the trained model for testing to determine the test accuracy. Fig. 4 is the confusion matrix predicted by the CWT-CNN model. In binary classification problems, we often divide the samples into positive and negative samples, where TP means that the positive sample is actually predicted. TN means actual predicted negative sample, FP means a negative sample that was incorrectly predicted as positive, and FN means a positive sample that was incorrectly predicted as negative.

The accuracy of the binary classification model can be calculated using the following formula, and in order to eliminate the influence of the imbalance of the number of positive and negative samples, it can also be calculated using the F1-Measure

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{99483+99257}{200,000} = 99.37\% \quad (5)$$

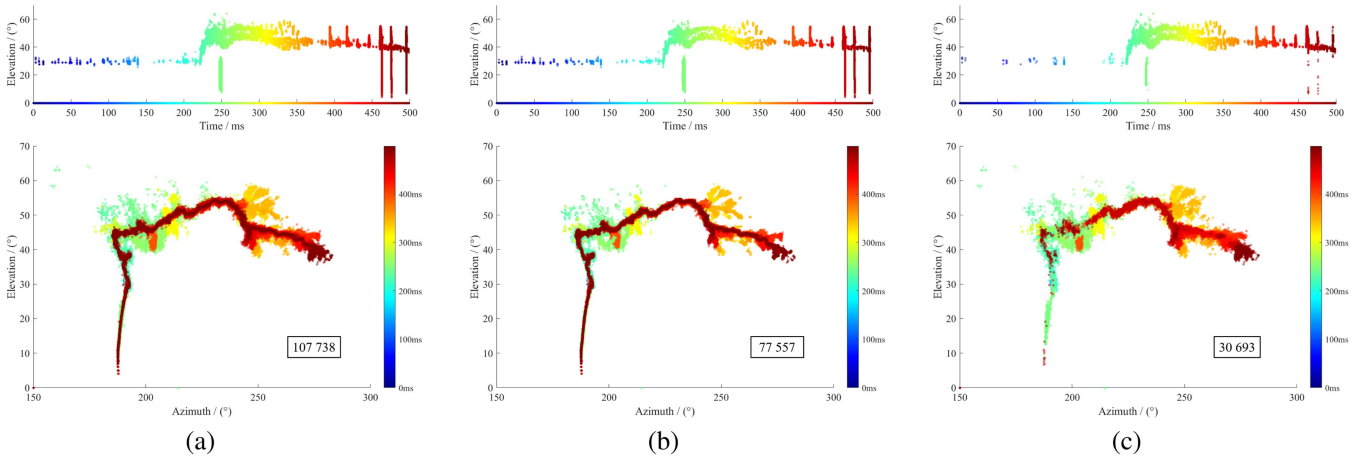

Fig. 6. Comparison of lightning maps of Trig230553 identified by (a) CWT-CNN and (b) CR/ER metric. (c) Locating results that the CWT-CNN model keeps more than the CR/ER metric.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{99483}{99483 + 743} = 99.26\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{99483}{99483 + 517} = 99.48\% \quad (6)$$

$$F1 - \text{Measure} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0.9926 \times 0.9948}{0.9926 \times 0.9948} = 99.37\%. \quad (7)$$

The final testing accuracy of the model is 99.37%.

In addition, we visualize the learned features of the neural network. We obtain the 512-D feature vector learned from the fully connected layer of the penultimate layer of the CWT-CNN model and visualize it in two dimensions by t-SNE [21] dimensionality reduction. In total, 1000 positive and negative samples each are fed into the model to obtain the corresponding learned feature vectors, as shown in Fig. 5. As shown in the figure, the feature vectors learned by the CWT-CNN model can well distinguish between positive and negative samples.

B. Prediction by CNN Model

In the second step, the data of 3.9 million sliding windows in Trig230553 are input into the trained model in turn to predict whether each sliding window contain radiation sources. The locating results corresponding to the sliding window data containing the radiation source judged by the CWT-CNN model are retained. Fig. 6(a) is the final retained locating result of Trig230553 by the CWT-CNN model. This triggered lightning consists of three leader and return stroke (RS) process. Fig. 6(b) is the retained locating result using the CR/ER metric, where CR $\geq$ 0.3 and ER $\geq$ 1.1. Fig. 6(c) is the locating results that the CWT-CNN model keeps more than the CR/ER metric. The CWT-CNN model retains 107 738 valid locating results, much more than the CR/ER metric and gets a more continuous lightning map without introducing additional noise interference.

In the last step, the generalization and transfer ability of the model is tested. Trig231042 is another lightning VHF data, on which the CWT-CNN model has not been trained. We use the model trained in the first step to predict this data, and then, use FDTR to locate the sliding window data, which identified as radiation source. As comparison, we also use FDTR to locate the whole data and use the CR/ER metric to eliminate the noise points. The final result is shown in Fig. 7, different from Trig230553, this triggered lightning is a slow discharge process, which does not have the leader and RS process. The lightning map stored by CWT-CNN still contains much more radiation sources and the structure is more continuous than that of the CR/ER metric. It can be seen that the model is able to predict untrained data well, which shows a good classification performance and generalization ability of the CWT-CNN model.

C. Iterative Spatiotemporal Clustering to Restore Weak Radiation Sources

We use the iterative spatiotemporal clustering to restore the valid location results filtered by ER metric, of which the flowchart is shown in Fig. 8.

Step 1: Get a clear lightning map by CR/ER metric as the initial retained located results, also as the known valid location results set G.

Step 2: Each locating result to be determined, p_i, is spatial-temporal clustered with the known valid location results set G.

Step 3: Find out the location results to be determined p_i, which belong to the same class as the known valid location results set G, and classify them as the valid location results.

Step 4: Upgrade the known valid locating results set G', and repeat the Steps 2–4 until the set G' do not increase.

Following the above steps, we perform spatiotemporal clustering on the location results of Trig230553. We use the location results retained by the CR/ER metric as the initial lightning map, where CR $\geq$ 0.3 and ER $\geq$ 1.1. After several iterations of spatiotemporal clustering, the valid location results do not increase, and the final clustering result is shown in Fig. 9. Spatiotemporal clustering retains 123 313 location results, 59.0% improvement over the CR/ER metric. Comparing the location results in Fig. 1, it can be seen that the CWT-CNN model and

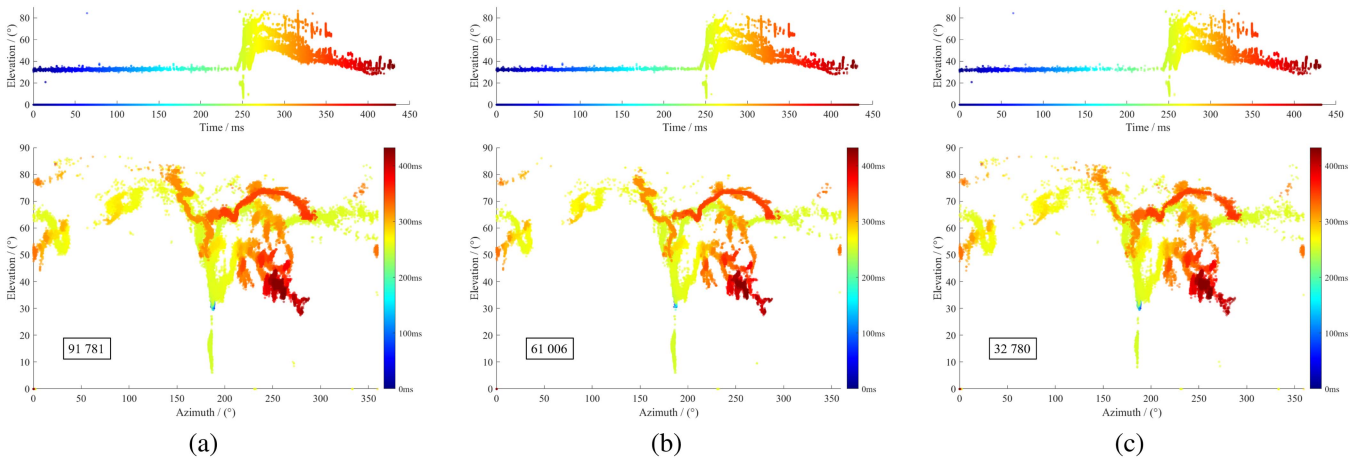

Fig. 7. Comparison of lightning maps of Trig231042 identified by (a) CWT-CNN and (b) CR/ER metric. (c) Locating results that the CWT-CNN model keeps more than the CR/ER metric.

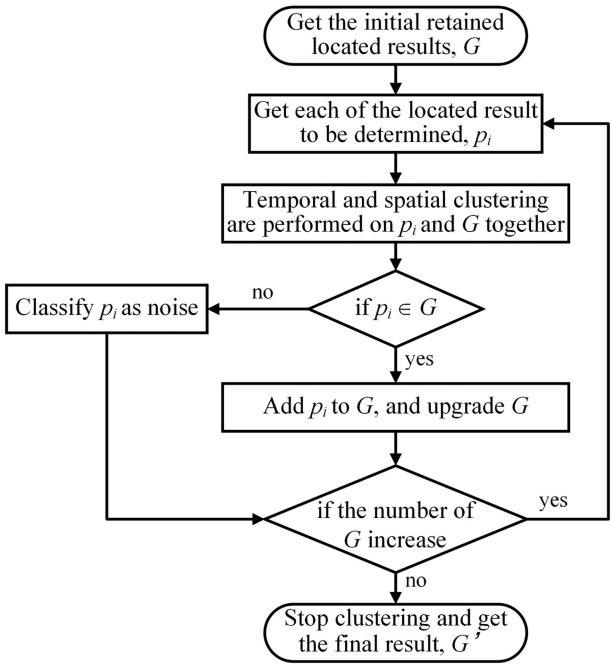

Fig. 8. Flowchart of iterative spatiotemporal clustering.

the spatiotemporal clustering model can effectively eliminate the noise interference and obtain a clearer lightning map under the condition of retaining the same number of radiation sources.

IV. Further Analysis

A. Detail Comparison of Lightning Structure

For the artificially triggered lightning data, Trig230553, we select some zoom-in views of the development structure from Figs. 6 and 9, and compare their location results retained by CR/ER metric, CWT-CNN, and spatiotemporal clustering, respectively, in detail.

As shown in Fig. 10, the subfigures in the first column are the location results by CR/ER metric, the second column corresponds to the results by CWT-CNN, and the third column belongs to the results by spatiotemporal clustering. We compared the location results of the three methods for the four processes. The black dot circle in the subfigures of the first column represents the missing location results of the CR/ER metric compared with the other two methods. The subfigures in the first row represent the location results of one K process, and those in the second row represent the location results of the initial stage of the first dart-leader. The subfigures in the third and the last rows are the location results of the first and second RS processes. It can be found that the lightning structures in the localization map retained by CWT-CNN and spatiotemporal clustering are more continuous. In the K process and the first dart-leader process, the spatiotemporal clustering algorithm retains the most radiation sources, whereas the CWT-CNN model retains the most radiation sources in the two RS processes. This is because in the first two processes, the development of radiation sources is relatively slow and the distribution of radiation sources is concentrated, so the spatiotemporal clustering algorithm can well retain weak radiation sources in the adjacent area. In the latter two processes, the radiation sources develop faster and have weaker energy, the initial retained location points are rare and sparsely distributed, and many missed radiation sources are relatively isolated, so they cannot be retained by the spatiotemporal clustering algorithm.

In the first RS, CWT-CNN and spatiotemporal clustering can both retain more lightning radiation sources at the end of the process. In the second RS, the CR/ER metric fundamentally does not preserve the radiation source because the radiation is weak and filtered by the ER metric. Instead, CWT-CNN and spatiotemporal clustering can well compensate for this problem, they maintain some relatively continuous radiation sources on the RS channel, which will be helpful for estimating the RS velocity for further analysis.

For the second artificially triggered lightning data, Trig231042, we also select one K process in lightning maps from Fig. 7 and compare the location results retained by the

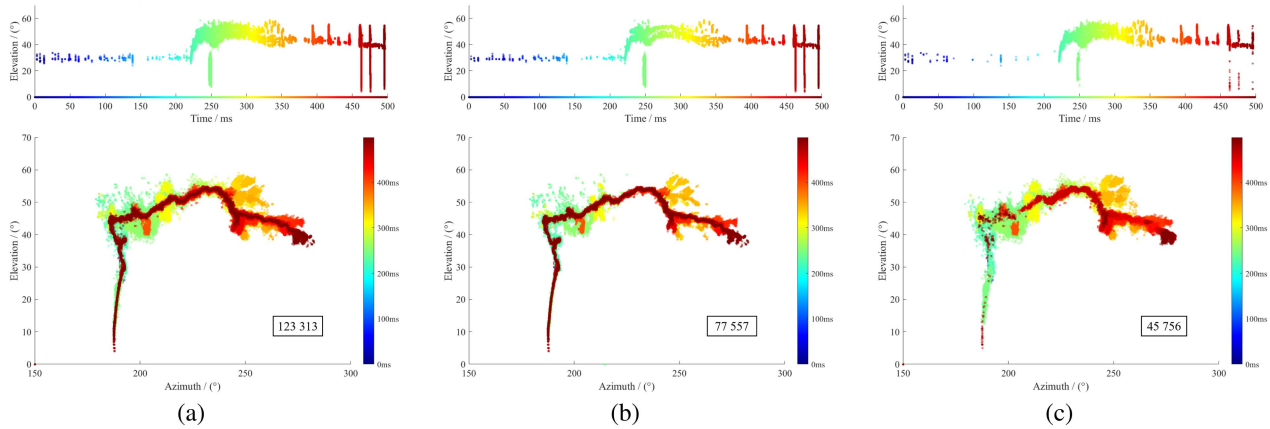

Fig. 9. Comparison of lightning maps of Trig230553 identified by (a) spatiotemporal clustering and (b) CR/ER metric. (c) Locating results that the spatiotemporal clustering keeps more than the CR/ER metric.

CR/ER metric and the CWT-CNN model. As shown in Fig. 11, the CWT-CNN model maintains more radiation sources at the beginning of the K process compared with the CR/ER metric. Based on the location results, we can conclude the initial development direction and start time of the process. We can see that the start time obtained by CWT-CNN is earlier than that of the CR/ER metric, and the start direction of the former is different from that of the latter.

B. Statistic Analysis

The energy distributions of the radiation source signals corresponding to the location results retained by these methods are calculated. The normalized voltage amplitude and the ER of the time domain signal of each sliding window are calculated, where the ER represents the ratio of the radiation frequency band energy to the noise frequency band energy. Here, we set radiation frequency band ranges in 25–88 MHz and 108–200 MHz and the noise frequency band ranges in 88–108 MHz. The statistical histograms of location results of Trig230553 retained by the three methods are shown in Fig. 12. For both amplitude and ER statistical histograms, in the high-energy part, the three methods retain almost the same number of location points. In the low-energy part, the CWT-CNN model and the spatiotemporal clustering algorithm retain much more numbers of location point than the CR/ER method, which proves that the two proposed methods could retain much more weak radiation sources. The spatiotemporal clustering algorithm retains the most radiation location points, it is because, for some weak radiation sources with low energy, their time–frequency characteristics are not obvious, and the CWT-CNN model cannot accurately identify them, whereas the clustering algorithm can retain these radiation sources through the spatiotemporal distribution characteristics. Therefore, the latter is able to retain more radiation sources than the former, especially in the very low-energy part.

For the triggered lightning Trig231042, the statistical histograms of location results retained by the CWT-CNN model and CR/ER metric are also shown in Fig. 13 and its result is similar to Trig230553. From the result analysis of the two sets of triggered lightning data, we can conclude that the CWT-CNN model and spatiotemporal clustering algorithm maintain much more radiation sources with lower energy than the CR/ER metric.

The CR metric takes into account the spatial and temporal distribution continuity between radiation sources and is not coupled to the energy of radiation sources, therefore, this method makes sense to some extent. However, the ER metric takes into account the energy of the radiation, so it judges the low energy radiation source as noise, which makes no sense, especially because EMTR itself can locate the low energy radiation source, but the ER metric filters out the weak radiation source, so it cannot fully reflect the locating performance of EMTR.

C. Feature Analysis

In order to further analyze the correctness of the CNN prediction results, we selected the time-domain waveforms corresponding to the location results in Fig. 10(d) and (e), and analyzed the output eigenvectors of these time-domain waveforms predicted by the CWT-CNN model using the same method as those used in Fig. 5. We input the time-domain waveforms corresponding to the three types of location results, including the noise points (label-1), the results retained by the CR/ER metric (label 0), and the additional location results retained by the CWT-CNN model compared with the CR/ER metric (label 1) forward to the CWT-CNN model and get the feature vector output from the second to last fully connected layer. We then use t-distributed stochastic neighbor embedding (t-SNE) to reduce the dimensionality of these feature vectors, visualize them, and compare the spatial distributions of the three different data types, as shown in Fig. 14. As can be seen from the figure, the feature vectors of the time-domain waveform corresponding to the location points retained by the CR/ER metric and CWT-CNN model have a large spatial overlap and are significantly different from the feature vectors of the noise points, which confirmed the effectiveness of the CWT-CNN model.

The same method as described above is used to analyze the correctness of the clustering algorithm results. We select the time-domain waveforms corresponding to the location results in Fig. 10(d) and (f), and analyze the output eigenvectors of these time-domain waveforms predicted by the CWT-CNN model.

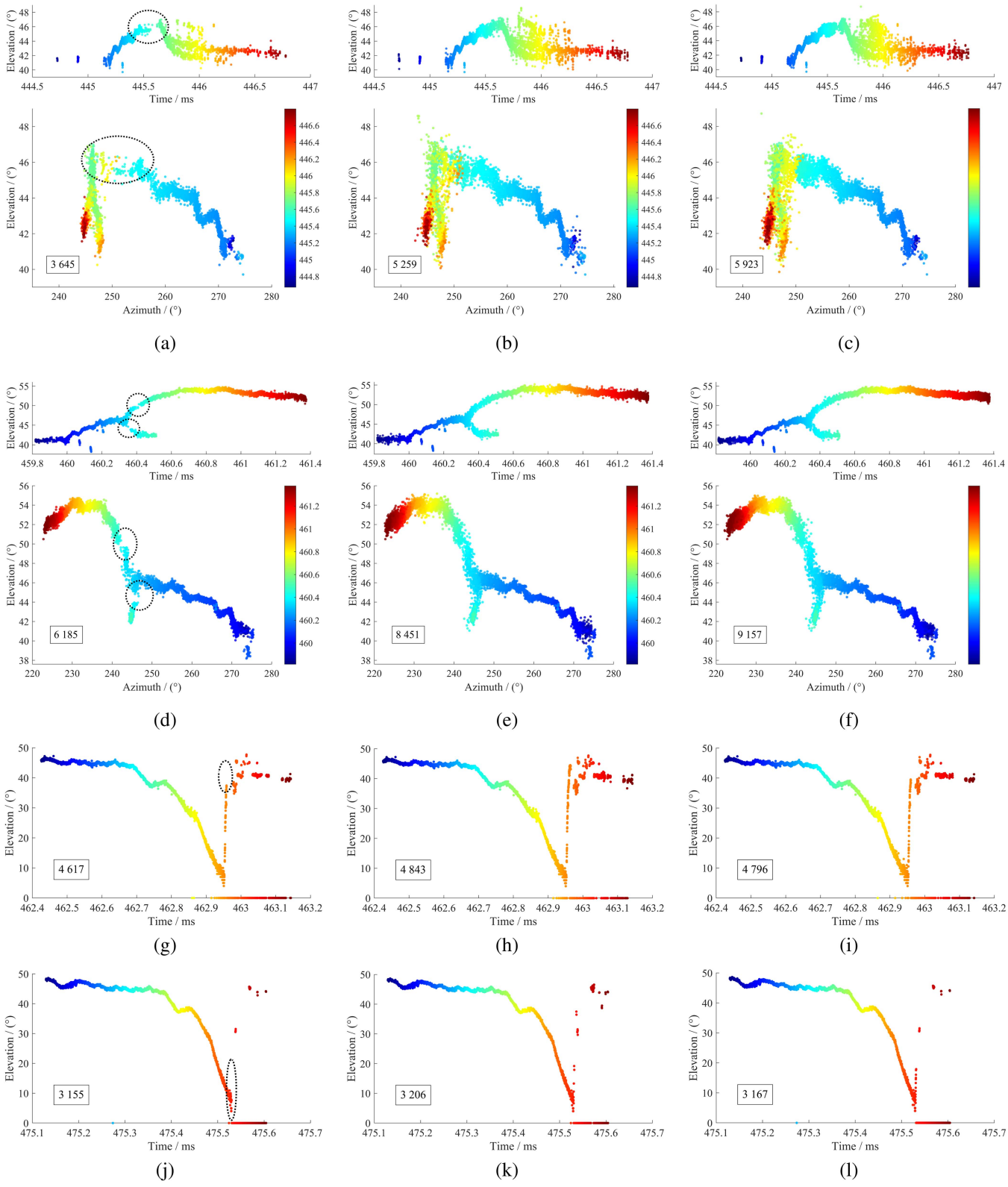

Fig. 10. Comparisons of Trig230553 lightning structures in detail identified by the CR/ER metric, CWT-CNN, and spatial-temporal clustering. (a), (d), (g), and (j) are the results of the CR/ER metric. (b), (e), (h), and (k) are the results of the CWT-CNN. (c), (f), (i), and (l) are the results of the spatial-temporal clustering. Dot circle illustrates the missing structure by the CR/ER metric. (a)–(c) is the K process. (d)–(f) is the initial process of the first dart-leader. (g)–(i) is the first RS process. (j)–(l) is the second RS process.

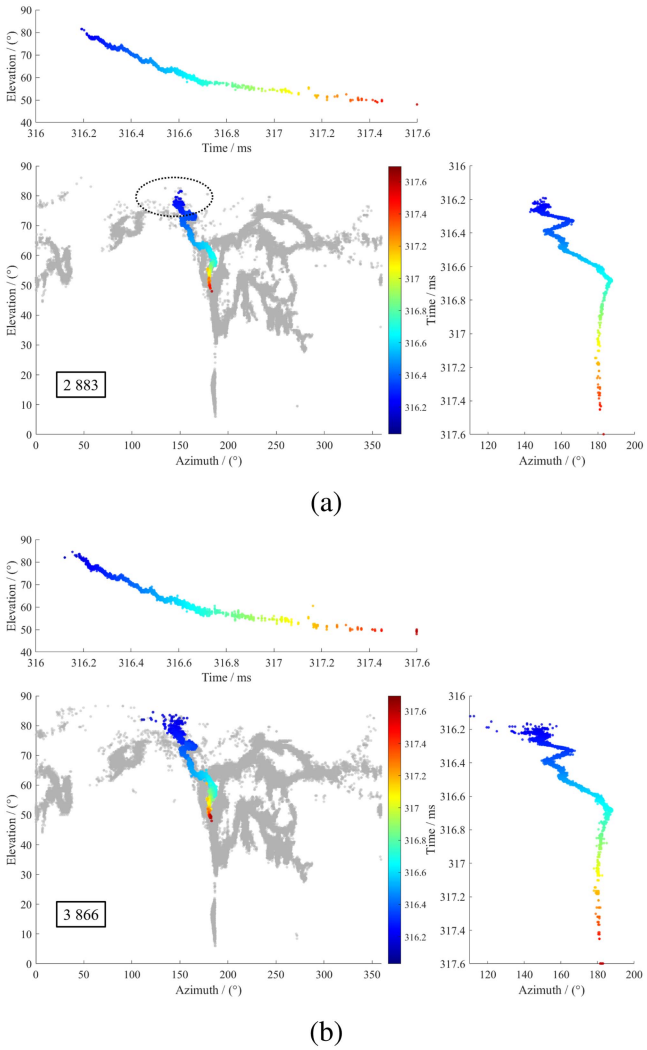

Fig. 11. (a) Locating result comparison of the K process in lightning maps of Trig231042 by CR/ER metric and (b) CWT-CNN.

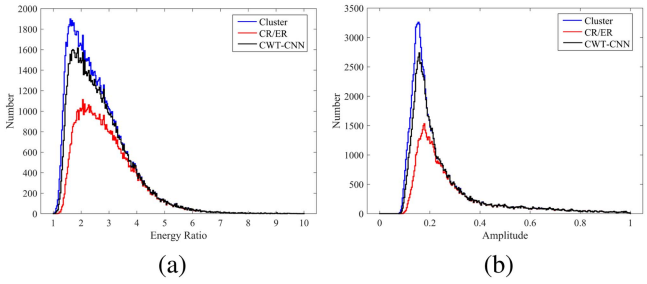

Fig. 12. (a) Statistical histogram of the frequency band ER and (b) normalized voltage amplitude of the sliding window signal corresponding to the locating result of Trig230553.

Fig. 15 shows the corresponding feature distribution. Similar to the results of CWT-CNN, the feature vectors corresponding to the additional retained results of the spatiotemporal clustering algorithm almost all overlap with those corresponding to the retained results of the CR/ER metrics in spatial distribution,

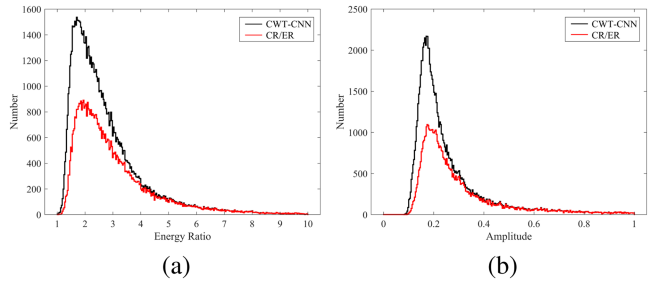

Fig. 13. (a) Statistical histogram of the frequency band ER and (b) normalized voltage amplitude of the sliding window signal corresponding to the locating result of Trig231042.

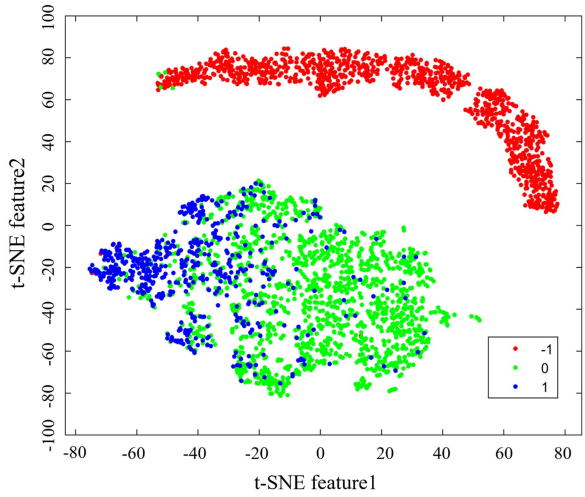

Fig. 14. Feature visualization of CWT-CNN location results by t-SNE. Label -1 represents noise. Label 0 represents location results retained by the CR/ER metric. and Label 1 represents the additional retained location results of the CWT-CNN model compared with the CR/ER metric.

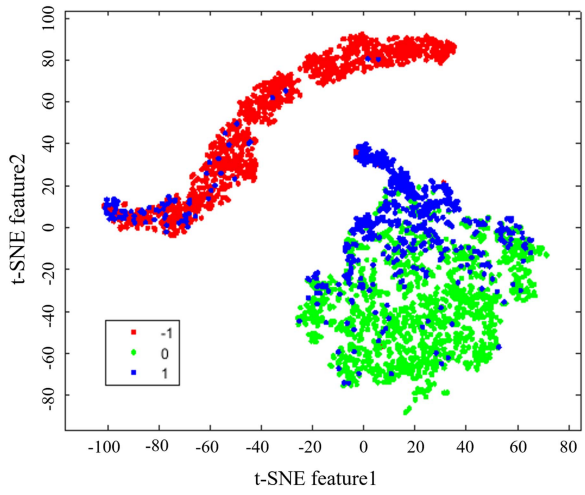

Fig. 15. Feature visualization of spatiotemporal clustering location results by t-SNE. Label-1 represents noise. Label 0 represents location results retained by the CR/ER metric. and Label 1 represents the additional retained location results of the spatiotemporal clustering method compared with the CR/ER metric.

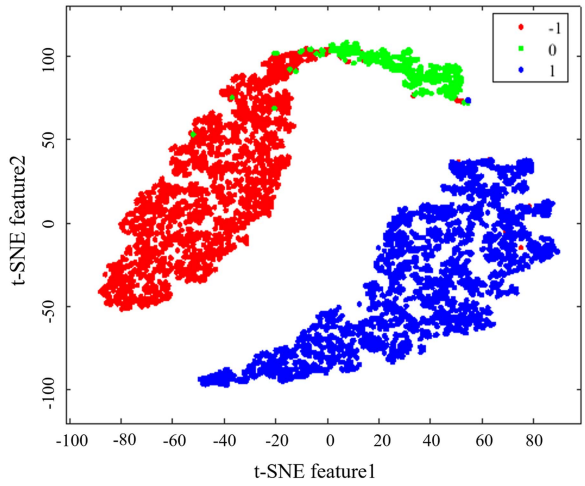

Fig. 16. Feature visualization of spatiotemporal clustering and CWT-CNN location results by t-SNE. Label-1 represents noise. Label 0 represents additionally retained location results of the spatiotemporal clustering method compared with the CWT-CNN model. Label 1 represents the location results retained by the CWT-CNN model.

which also shows that the location results retained by the spatiotemporal clustering algorithm are basically real radiation sources.

In addition, we compare the distribution relationship of the feature vectors corresponding to the location result retained by the two methods, CWT-CNN and spatiotemporal clustering. We select the location results in Fig. 10(e) and (f) for comparison and analyze their output eigenvectors in the same way as mentioned above. The distribution of feature vectors are shown in Fig. 16. The red points (label -1) represent the noise feature vectors, the blue points (label 1) represent feature vectors of the results retained by the CWT-CNN model, and the green points (label 0) are feature vectors of the additional retained results of the spatiotemporal clustering algorithm compared with the CWT-CNN model. It can be seen from the distribution that for these feature vectors of additionally retained results by the spatiotemporal clustering, its distribution is between the noise feature vectors and the feature vectors retained by the CWT-CNN model, and does not overlap with them too much. According to the analysis in Part B of this section, these location points are all radiation sources with very weak energy, which are judged as noise by CWT-CNN model, therefore, their distribution should be close to the noise points. However, there is an obvious boundary between them and noise in the feature distribution. This shows that the extra retained location results by the spatiotemporal clustering algorithm may be real and effective radiation sources.

V. Discussion

Here, we discuss the respective characteristics of the two methods, CWT-CNN model and spatial-temporal clustering algorithm, and also give an outlook on the problem we are studying in this article.

The spatiotemporal clustering algorithm must rely on a preexisting location map and expand based on that map. It can compensate for the weak radiation sources on the development channel filtered out by the ER metric, so that the entire location map is more continuous and the number of radiation sources is more. However, it is unlikely that this method can discover unknown spatiotemporal location-independent branches of the lightning structure.

Different from the spatial-temporal clustering algorithm, the CWT-CNN model does not rely on the spatial-temporal position relationship of the radiation source, but starts from the time-frequency characteristics of the radiation source signal, and judges whether the locating results are valid through prior knowledge. In this way, it cannot only make up for the weak radiation source, but also find the spatiotemporal location independent branch of the lightning structure.

Furthermore, the CNN method can be combined with EMTR, MUSIC, and OPM methods. These methods have high-localization accuracy, but it takes a long time to locate the single sliding window data based on the global maximum search. The CNN method can determine in advance whether there is a radiation source in the sliding window, so a lot of noise interference can be eliminated preliminarily, thus reducing the calculation of the localization methods. It can be seen from the previous experiment results that this method of eliminating noise interference in advance not only does not decrease the mapping quality in the later stage, but will retain more weak radiation sources.

Besides, the current TR method uses sliding windows of fixed length to segment the VHF data. However, during the lightning development, there will be many mixed simultaneous radiations occurring in one sliding window, and it is very important but very hard to distinguish these mixtures. Ismail et al. [22] suggested that these mixed simultaneous radiations of regular VHF radiations may become an irregular pulse train. Considering that voice identification uses a time–frequency map to recognize human voice, the artificial neural network may find some new information in our mixed signals. Therefore, it is possible to train a proper neural network to not only recognize the radiation source but also separate the mixed simultaneous signals.

Last but not the least, by using different set of lightning data for training and testing, it is proved that our model could be applied to an unseen lightning. In other words, we can say the CWT-CNN model has the ability to recognize radiation sources before further processing. However, the two examples uses the same measurement equipments and the same site background. Fortunately, CNN has the ability of evolution and continuous learning. In this way, in further study, through constant training and learning and combined with the lightning database, the model can gradually improve the discrimination ability, and theoretically can well identify more different unseen lightning data.

VI. Conclusion

The CWT-CNN model and spatiotemporal clustering algorithm proposed in this article overcome the disadvantage that the CR/ER metric cannot properly distinguish the weak radiation source and noise interference. Not only can they maintain a continuous lightning map, conserve much more weak radiation sources but also avoid noise interference, compared with the latter. Both methods use feature learning to determine the validity of radiation source locating results. The CWT-CNN

method determines the validity by learning the time-frequency characteristic of the individual radiation source, whereas the spatiotemporal clustering algorithm achieves this by learning the spatiotemporal distribution characteristic of a part of radiation sources. Compared with the previous methods, their decision criteria are objective and their high effectiveness has also been proven experimentally.

REFERENCES

[1] W. Rison, R. J. Thomas, P. R. Krehbiel, T. Hamlin, and J. Harlin, "A GPS-based three-dimensional lightning mapping system: Initial observations in Central New Mexico," *Geophysical Res. Lett.*, vol. 26, no. 23, pp. 3573–3576, 1999.

[2] W. J. Koshak, R. J. Solakiewicz, and R. J. Blakeslee, "North Alabama lightning mapping array (LMA): VHF source retrieval algorithm and error analyses," *J. Atmos. Ocean. Technol.*, vol. 21, no. 4, pp. 543–558, Apr. 01, 2004.

[3] B. Liu et al., "Fine three-dimensional VHF lightning mapping using waveform cross-correlation TOA method," *Earth Space Sci.*, vol. 7, no. 1, Jan. 2020, Art. no. e2019EA000832.

[4] Z. F. Chen et al., "A method of three-dimensional location for LFEDA combining the time of arrival method and the time reversal technique," *J. Geophysical Res.: Atmos.*, vol. 124, no. 12, pp. 6484–6500, Jun.27, 2019.

[5] S. Qiu, B. H. Zhou, L. H. Shi, W. S. Dong, Y. J. Zhang, and T. C. Gao, "An improved method for broadband interferometric lightning location using wavelet transforms," *J. Geophysical Res.: Atmos.*, vol. 114, no. D18, 2009, Art. no. D18211.

[6] X. M. Shao, D. N. Holden, and C. T. Rhodes, "Broad band radio interferometry for lightning observations," *Geophysical Res. Lett.*, vol. 23, no. 15, pp. 1917–1920, 1996.

[7] M. G. Stock, P. R. Krehbiel, J. Lapierre, T. Wu, M. A. Stanley, and H. E. Edens, "Fast positive breakdown in lightning," *J. Geophysical Res.: Atmos.*, vol. 122, no. 15, pp. 8135–8152, 2017.

[8] M. G. Stock et al., "Continuous broadband digital interferometry of lightning using a generalized cross-correlation algorithm," *J. Geophysical Res.: Atmos.*, vol. 119, no. 6, pp. 3134–3165, Mar., 2014.

[9] N. Mora, F. Rachidi, and M. Rubinstein, "Application of the time reversal of electromagnetic fields to locate lightning discharges," *Atmos. Res.*, vol. 117, pp. 78–85, Nov.1, 2012.

[10] G. Lugrin, N. M. Parra, F. Rachidi, M. Rubinstein, and G. Diendorfer, "On the location of lightning discharges using time reversal of electromagnetic fields," *IEEE Trans. Electromagn. Compat.*, vol. 56, no. 1, pp. 149–158, Feb. 2014.

[11] T. Wang, S. Qiu, L. H. Shi, and Y. Li, "Broadband VHF localization of lightning radiation sources by EMTR," *IEEE Trans. Electromagn. Compat.*, vol. 59, no. 6, pp. 1949–1957, Dec. 2017.

[12] S.-J. Du, L.-H. Shi, Y.-C. Liu, and S. Qiu, "Accelerate the time-reversal computing of lightning location using GPU and phase-difference filter," *IEEE Trans. Electromagn. Compat.*, vol. 65, no. 2, pp. 496–506, Apr. 2023.

[13] T. Wang, L. Shi, S. Qiu, Z. Sun, and Y. Duan, "Continuous broadband lightning VHF mapping array using MUSIC algorithm," *Atmos. Res.*, vol. 231, 2020, Art. no. 104647.

[14] S. L. Li, S. Qiu, L. H. Shi, Y. Li, and Y. T. Duan, "Broadband very high frequency localization of lightning radiation sources based on orthogonal propagator method," *Acta Physica Sinica*, vol. 68, no. 16, pp. 165202-1–165202-9, 2019.

[15] T. Wang et al., "Multiple-antennae observation and EMTR processing of lightning VHF radiations," *IEEE Access*, vol. 6, pp. 26558–26566, 2018.

[16] H. Chen et al., "A low computational cost lightning mapping algorithm with a nonuniform L-shaped array: Principle and verification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4106410.

[17] J. Chouragade and R. K. Muthu, "Continuous mapping of broadband VHF lightning sources by real-valued MUSIC," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4102707.

[18] B. Liu, L. H. Shi, S. Qiu, H. Y. Liu, Z. Sun, and Y. F. Guo, "Three-dimensional lightning positioning in low-frequency band using time reversal in frequency domain," *IEEE Trans. Electromagn. Compat.*, vol. 62, no. 3, pp. 774–784, Jun. 2020.

[19] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[20] L. Wang, B. M. Hare, K. Zhou, H. Stocker, and O. Scholten, "Identifying lightning structures via machine learning," *Chaos, Solitons Fractals*, vol. 170, 2023, Art. no. 113346.

[21] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[22] M. M. Ismail, M. Rahman, V. Cooray, M. Fernando, P. Hettiarachchi, and D. Johari, "On the possible origin of chaotic pulse trains in lightning flashes," *Atmosphere*, vol. 8, no. 2, 2017, Art. no. 29.

Shuang-Jiang Du was born in Hubei, China, in 1996. He received the B.S. degree in aircraft engineering from the National University of Defense Technology, Changsha, China, in 2018, and the M.S. degree in electronic science and technology in 2021 from Army Engineering University, Nanjing, China, where he is currently working toward the Ph.D. degree in electromagnetic environmental effects and protection engineering.

His research interests include lightning detection and protection and machine learning.

Yun Li was born in Hubei, China, in 1990. He received the B.S. degree in meteorological radar engineering, the M.S. degree in measurement technology and instruments, and the Ph.D. degree in disaster prevention and reduction engineering and protective engineering from the PLA University of Science and Technology, Nanjing, China, in 2012, 2015, and 2019, respectively.

His research interests include lightning detection and protection, and theoretical and computational electromagnetics.

Zheng Sun received the B.S. degree in automatic control from Southeast University, Nanjing, China, in 2009, and the Ph.D. degree in electrical engineering from the PLA University of Science and Technology, Nanjing, in 2014.

He is currently working as a Lecturer with the PLA Army Engineering University, Nanjing. His main research interests include computing electromagnetics and lightning protections.

Shi Qiu was born in Shandong, China, in 1984. He received the B.S. degree in atmospheric physics and atmospheric environmental and the Ph.D. degree in disaster prevention and reduction engineering and protective engineering from the PLA University of Science and Technology, Nanjing, China, in 2005 and 2012, respectively.

He is currently a Research Scientist with the National Key Laboratory on Electromagnetic Environmental Effects and Electro-Optical Engineering, Army Engineering University of PLA, Nanjing. His research interests include lightning physics and measurement techniques and lightning electromagnetic environment and propagation.

Li-Hua Shi (Member, IEEE) was born in Hebei, China, in 1969. He received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 1990, the M.S. degree in electrical engineering from the Nanjing Engineering Institute, Nanjing, China, in 1993, and the Ph.D. degree in instrument science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1996.

He is currently a Professor and the Director of the National Key Laboratory on Electromagnetic Environmental Effects and Electro-Optical Engineering, Army Engineering University of PLA, Nanjing. His research interests include time-domain measurement and signal processing technology.

Dr. Shi is an HPEM Fellow of the Summa Foundation, USA.