End-to-End Deblending of Simultaneous Source Data Using Transformer

Shaohuan Zu, Chaofan Ke, Chengzhi Hou, Junxing Cao, and Hongjing Zhang

***Abstract*— Simultaneous source acquisition is becoming more promising than the traditional seismic acquisition by firing multiple sources with a short interval time, which improves acquisition efficiency and enhances data quality. However, the blended interference severely obscures the coherent signal, challenging the conventional seismic data processing methods. Recently, convolution neural network (CNN) has been successfully implemented to address the blended interference. Different from CNN, the self-attention mechanism-based transformer neural network is good at capturing the global features. In this letter, we propose a deblending transformer (DT) based on the transformer module to separate the simultaneous source data. The DT architecture mainly includes linear embedding operation, patch partition-based transformer block, and output projection layer. The patch partition algorithm is embedded into the multihead self-attention (MHSA) module, which extracts the vertical, horizontal, and local information. In addition, with the help of linear embedding operation and output projection algorithm, the DT can easily extract the global features from the input. Experiments on synthetic and field data demonstrate that the proposed method has better deblending performance than the U-net and curvelet-based methods.**

***Index Terms*— Deblending, self-attention mechanism, simultaneous source, transformer.**

I. Introduction

CONVENTIONAL seismic acquisition is usually designed with a large time interval to avoid overlap in record, which enhances the acquisition efficiency. Different from the traditional seismic acquisition, simultaneous source acquisition utilizes two or more sources to obtain dense sampling data or wide-azimuth data [1]–[4], which brings higher acquisition efficiency and better data quality [5]. However, these benefits are hindered by the blended interference. For better imaging, smart survey design and intelligent deblending algorithms are urgently required [6], [7].

Many methods are developed to deal with simultaneous source data. Some authors directly migrate the blended record using reverse time migration [8] or least-squares migration algorithm [9]. Dai *et al*. [10] proposed a least-squares reverse time migration (LSRTM) algorithm to reduce crosstalk noise. With additional regularization terms, LSRTM can perform well in more complex blended data [11]. Besides, there are the other two categories of deblending methods that can be used to separate the blended data. The first category considers deblending as a denoising problem, which is realized using filter [12] or transform method [13], [14]. The second category is the inversion, which aims to estimate the unblended signal and iteratively subtract the blended noise [15]. Deblending by inversion is the state-of-the-art method but suffers from great computation cost.

Recently, convolution neural network (CNN) has shown great advantages in the seismic exploration field [16]–[18]. For CNN, the convolution operation is subject to the size of the local receptive field, which makes CNN difficult to extract the global representations. In contrast, the self-attention mechanism-based transformer is capable of global perception [19]. Transformer network and its improvement version achieve extraordinary performance in many language tasks. Furthermore, transformer has been applied in the field of computer vision as a large-scale pretrain model [20]. Those transformer-based networks achieve state-of-the-art results in different vision tasks, demonstrating the effectiveness of transformer.

In this letter, we propose the deblending transformer (DT) network to suppress the blended interference. In the DT network, the patch partition-based multihead self-attention (MHSA) module can capture the horizontal, vertical, and local features. Besides, with the linear embedding and output projection operations, the DT network is easy to study the global features. In order to alleviate the burden of massive samples, we split the input of every MHSA module in three different ways to provide prior information for the transformer. By splitting the input horizontally, vertically, and locally, we can build an end-to-end transformer network with relatively few samples. Compared with curvelet-based method [21] and U-net [22], DT obtains the better deblending performance evaluated by the signal-to-noise ratio (SNR) in the processing of the two synthetic examples and the field example.

II. Theory

A. Deblending Transformer (DT)

The deblending performance of CNN has been confirmed by many researchers [17], [23]. Due to the local receptive field, CNN is difficult to capture the global clues. Many CNN models use large receptive field or dilated convolution to

Manuscript received January 28, 2022; revised April 6, 2022 and May 1, 2022; accepted May 6, 2022. Date of publication May 10, 2022; date of current version May 19, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42030812, Grant 42004104, and Grant 41974160; and in part by the Sichuan Science and Technology Program under Grant 2020YJ0119. *(Corresponding author: Shaohuan Zu.)*

Shaohuan Zu and Junxing Cao are with the State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, the Key Laboratory of Earth Exploration and Information Technology of Ministry of Education, and the College of Geophysics, Chengdu University of Technology, Chengdu 610059, China (e-mail: zushaohuan19@cdut.edu.cn).

Chaofan Ke and Chengzhi Hou are with the College of Geophysics, Chengdu University of Technology, Chengdu 610059, China.

Hongjing Zhang is with the School of Geophysics and Measurement-control Technology, East China University of Technology, Nanchang, Jiangxi 330013, China.

Digital Object Identifier 10.1109/LGRS.2022.3174106

alleviate this limitation, but may imply lower spatial resolution and network deterioration. In contrast, the transformer model, containing feed-forward neural network block and MHSA module, has the ability of global features interaction. The self-attention mechanism in an MHSA module can be defined as follows:

$$[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = P_{\mathrm{QKV}}(\mathbf{z}_t) \tag{1}$$

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{Q}$ denotes the information query set, $\mathbf{K}$ is the corresponding similarity set, and $\mathbf{V}$ denotes the value set. The matrices of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are projected by the input sequence $\mathbf{z}_t \in \mathbb{R}^{N\times D}$ in self-attention module using a fully connected layer as expressed in (1), where N represents the length of the input sequence and D stands for the dimension of the sequence. A scale factor $1/\sqrt{D}$ is applied to norm the gradient. The function of softmax makes the sum of $\mathbf{Q}\mathbf{K}^T/\sqrt{D}$ equal to one. The attention weights are expressed as $\mathbf{W} = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{D})$, where $\mathbf{W}_{i,j}$ are based on the pairwise correlation between the query $\mathbf{Q}_i$ and key $\mathbf{K}_j$ representations. With the attention weights, we can extract the relevant features through the value set $\mathbf{V}$. More details about transformer can be referred to [19].

As described by Dosovitskiy *et al*. [20], transformer can obtain better performance than CNN when training with large datasets. However, the high-quality field data for training is not easy to obtain, which hinders the application in the field of seismic survey. In order to apply transformer in seismic data processing, a special tokenization method is adopted. In this letter, we alternately split the 2-D input to feed every MHSA module, vertically, horizontally, and locally to get a set of patches. The input for the DT Block is $\mathbf{z} \in \mathbb{R}^{(H,W,C)}$. For self-attention module, the input $\mathbf{z} \in \mathbb{R}^{(H,W,C)}$ should be transformed to $\mathbf{z}_t \in \mathbb{R}^{N\times D}$, where $t \in (h, v, l)$. In every horizontal-MHSA module, we reshape it into $\mathbf{z}_h \in \mathbb{R}^{N\times(1\cdot W\cdot C)}$, where $(1, W)$ represents the patches extracted horizontally. In the module of vertical-MHSA, we reshape it into $\mathbf{z}_v \in \mathbb{R}^{N\times(H\cdot 1\cdot C)}$, where $(H, 1)$ represents the patches extracted vertically. In the module of local-MHSA, we reshape it into $\mathbf{z}_l \in \mathbb{R}^{N\times(L^2\cdot 1)}$, where (L, L) denotes the patch size. Each patch is treated as a "token," which is flattened to a 1-D sequence to feed the transformer network. These special tokens can be considered as prior information, which aims to improve the deblending performance and alleviate the burden of massive computation cost.

The overall architecture of DT is shown in Fig. 1. Our designed network consists of three parts, input linear embedding, DT block, and output projection layer. Following vision transformer (ViT) [20], the linear embedding operation is added at the beginning to extract 16 feature maps, which is a 3×3 convolution layer. The main part of DT contains two repeated DT blocks with a different number of self-attention heads M. MHSA module divides $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into M groups. Each group of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ individually extracts different representations. In our network, the number of self-attention heads increases with depth from 8 to 16 to learn the different combinations of global features. Each DT block includes three

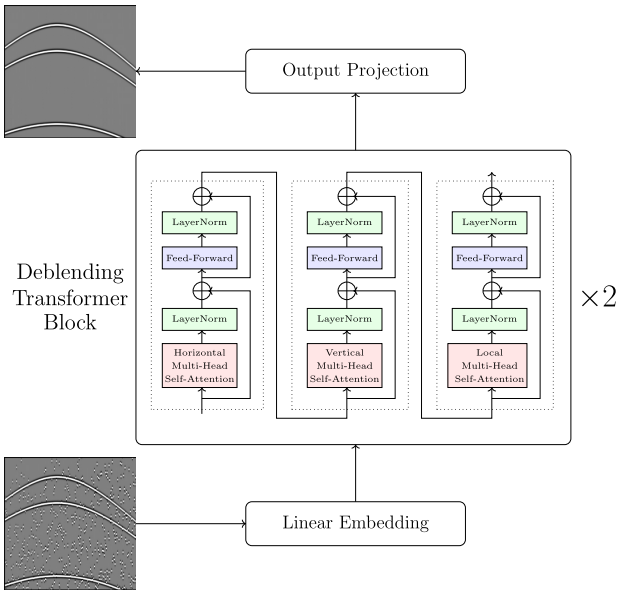

Fig. 1. Architecture of deblending transformer. The deblending transformer block contains three successive transformer modules with horizontal, vertical, and local multihead self-attention parts.

successive transformer layers, as shown in Fig. 1. The three transformer layers with different MHSA modules can extract horizontal, vertical, and local features, respectively. With these different types of partitions, DT can easily capture hidden features from the seismic data. As shown in Fig. 1, the designed transformer layer includes a feature partition-based MHSA module, followed by a fully connected feed-forward network (FCN) module (contains dimension scale up fully connection layer and scale down fully connection layer) with Gaussian error linear unit (GELU) nonlinearity activation function. A LayerNormalization (LN) layer is applied after an MHSA module and an FCN module. There is a residual connection between each MHSA+LN sublayer and FCN+LN sublayer [24]. In the output projection, a convolution layer instead of the fully connected layer is used to enhance the computational efficiency.

We estimate the network parameters by minimizing the mean-square-error (MSE) loss function, which can be expressed as

$$L(\Theta) = \frac{1}{U}\sum_{i=1}^{U} \|\mathcal{F}(\mathbf{x}_i; \Theta) - \mathbf{y}_i\|_F^2 \tag{3}$$

where $\mathbf{x}_i$ denotes the input patch, U denotes the patch number, $\|\cdot\|_F$ stands for the Frobenius norm, $\mathbf{y}_i$ denotes the ground truth patch, Θ represents the network parameters that need to be updated, and $\mathcal{F}(\mathbf{x}_i; \Theta)$ denotes the estimated output. Then, we use the AdamW (Adam optimizer with correct weight decay algorithm) optimizer to train our model.

B. Iterative Deblending Strategy

Taking two sources as example, the blending progress can be expressed as

$$\mathbf{b}_{\mathrm{ble}} = \mathbf{b}_1 + \Gamma_2\mathbf{b}_2 \tag{4}$$

where $\mathbf{b}_i$ denotes the common receiver gather (CRG) of the ith source, Γ_2 denotes the blending operator for the second source, and $\mathbf{b}_{\text{ble}}$ denotes the blended CRG. In (4), the first source is the reference source, which has no delay time, namely, $\Gamma_1 = \mathbf{I}$. Thus, the blended CRG $\mathbf{b}_{\text{ble}}$ can be viewed as the pseudo-deblended data of the first source. The pseudo-deblended data of the second source can be obtained by the following:

$$\Gamma_2^{-1}\mathbf{b}_{\text{ble}} = \Gamma_2^{-1}\mathbf{b}_1 + \mathbf{b}_2. \tag{5}$$

Combining (4) and (5), we can obtain

$$\mathbf{B} = \mathbf{FD} \tag{6}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_{\text{ble}} \\ \Gamma_2^{-1}\mathbf{b}_{\text{ble}} \end{bmatrix}, \mathbf{F} = \begin{bmatrix} \mathbf{I} & \Gamma_2 \\ \Gamma_2^{-1} & \mathbf{I} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}. \tag{7}$$

Following the denoising principle, the deblending can be realized by

$$\hat{\mathbf{D}} = \mathcal{F}(\mathbf{B}, \Theta_{\text{trained}}) \tag{8}$$

where Θ_{trained} denotes the trained network, $\hat{\mathbf{D}}$ denotes the deblended data, and function $\mathcal{F}$ represents the network test process. Another solution considers deblending as an inversion problem, where the blended interference can be predicted and subtracted iteratively. Following Zu *et al.* [17], an iterative framework is incorporated into our test process. The iterative deblending equation can be described as

$$\hat{\mathbf{D}}_i = \mathcal{F}(\mathbf{B} - (\mathbf{F} - \mathbf{I})\hat{\mathbf{D}}_{i-1}, \Theta_{\text{trained}}) \tag{9}$$

where $\hat{\mathbf{D}}_i$ is the deblended result in the ith iteration. In the first iteration, $\hat{\mathbf{D}}_1 = \mathcal{F}(\mathbf{B}, \Theta_{\text{trained}})$, which is the same with equation 8. In the second iteration, the input are $\mathbf{B} - (\mathbf{F} - \mathbf{I})\hat{\mathbf{D}}_1$, where the blended interference is estimated and subtracted from the blended records to some extent. The rest can be done in the same manner. After several iterations, the deblended results can be obtained.

III. Numerical Experiments

In this section, we apply the proposed DT method to synthetic and field data examples to test the deblending performance and compare with curvelet-based method and U-net method. We simulate 200 blended CRGs to train all the networks. The size of each CRG is 1024×512. For a fair comparison, all networks are trained by the same training scheme.

To quantitatively analyze the deblended results, the SNR is used to evaluate the test performance

$$\text{SNR (dB)} = 10\log_{10}\frac{\|\mathbf{s}\|^2}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2} \tag{10}$$

where $\mathbf{s}$ denotes the ground truth and $\hat{\mathbf{s}}$ denotes the estimated output. The closer recovered data are to the ground truth, the bigger SNR is.

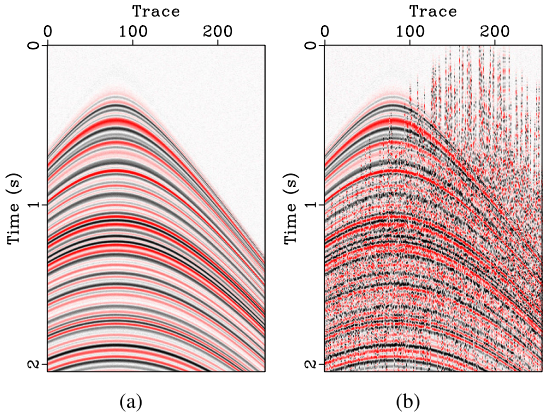

Fig. 2. Simple synthetic CRG. (a) Clean seismic record. (b) Pseudo-deblended record.

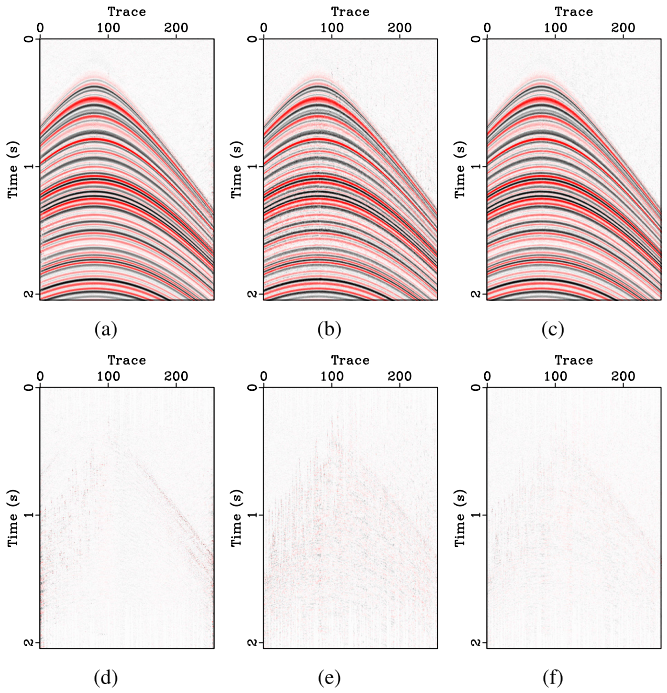

Fig. 3. Deblended results by (a) curvelet-based method, (b) U-net method, and (c) proposed method. Deblended errors corresponding to (d) curvelet-based method, (e) U-net method, and (f) proposed method.

A. Simple Example

Fig. 2(a) shows the conventional CRG, whose size is 512×256 and the time sampling is 4 ms. Since the first source is the reference source, which does not have the dithering time, the blended record shown in Fig. 2(b) can be viewed as the pseudo-deblended record for the first source. To demonstrate the advantages of the proposed method, the curvelet-based method, the U-net method, and the proposed method are implemented to separate the pseudo-deblended record. The separation results and the corresponding residuals are shown in Fig. 3. It can be observed that the three methods can remove most incoherent interference. However, the damage to signal of the proposed method is the least, as shown in Fig. 3(f). Note that for the proposed DT and U-net methods, the iteration is 10, however, for the curvelet-based method, we enlarge the iteration to 40 for obtaining the result. Fig. 4

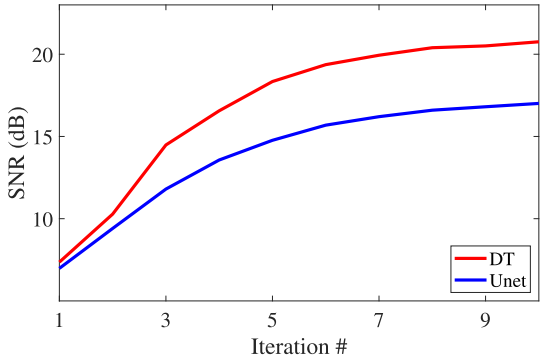

Fig. 4. Comparison of different methods performance on the simple example, where the red line denotes the proposed DT method and the blue line denotes the U-net method.

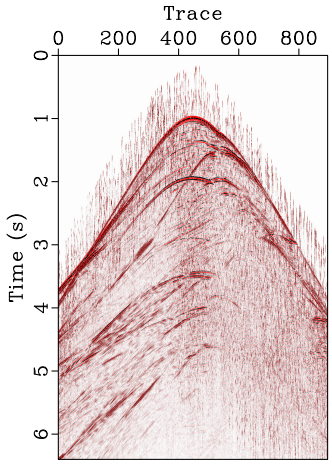

Fig. 5. Complex synthetic pseudo-deblended CRG.

shows the comparison of SNRs, where the red line corresponds to the DT method that converges to 20.7 dB and the blue line corresponds to the U-net method that converges to 17.0 dB. Since the iteration of the curvelet-based method is different from the proposed and U-net methods, we do not plot it. The SNR corresponding to the curvelet-based method is 17.9 dB.

B. Complex Example

Fig. 5 shows the complex pseudo-deblended CRG, where the dithering range changes from −0.8 to 0.8 s. To testify the deblending performance of the proposed algorithm on the complex synthetic pseudo-deblended data, we also test the curvelet-base and U-net methods. The deblended results and corresponding residuals of the three methods are shown in Fig. 6. It is obvious that the three methods can separate the pseudo-deblended record and the proposed DT has the best performance as shown in Fig. 6(c) and the least error as shown in Fig. 6(f). The SNRs changing with iterations of U-net and the proposed method are shown in Fig. 7, where the red line related to the proposed method converges to 17.5 dB and the blue line related to U-net method converges to 14.7 dB. For the curvelet-based method, the SNR converges to 13.4 dB. It can be seen that the proposed DT method has the highest SNR values and the full amplitude-preserving ability among other methods.

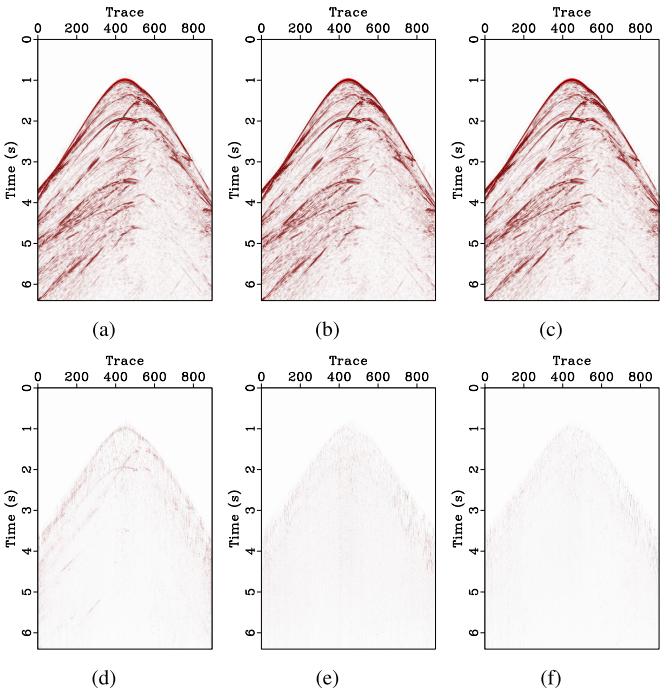

Fig. 6. Deblended results of the complex example by (a) curvelet-based method, (b) U-net method, and (c) the proposed method. Residuals corresponding to (d) curvelet-based method, (e) U-net method, and (f) the proposed method.

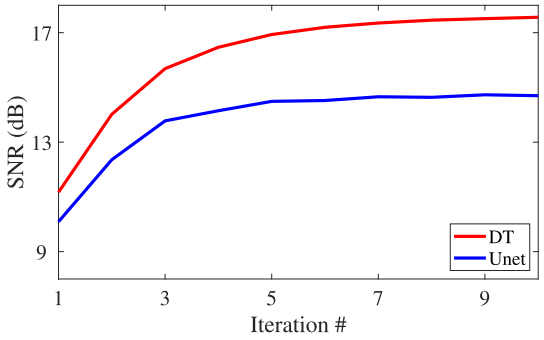

Fig. 7. Comparison of different methods performance on the complex example, where the red line denotes the proposed DT method and the blue line denotes the U-net method.

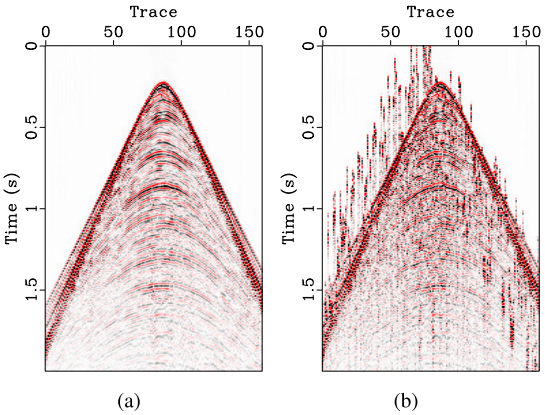

Fig. 8. Field CRG. (a) Clean seismic record. (b) Pseudo-deblended record.

C. Field Examples

To further demonstrate the advantages of the proposed method, we conduct the experiment with real seismic data.

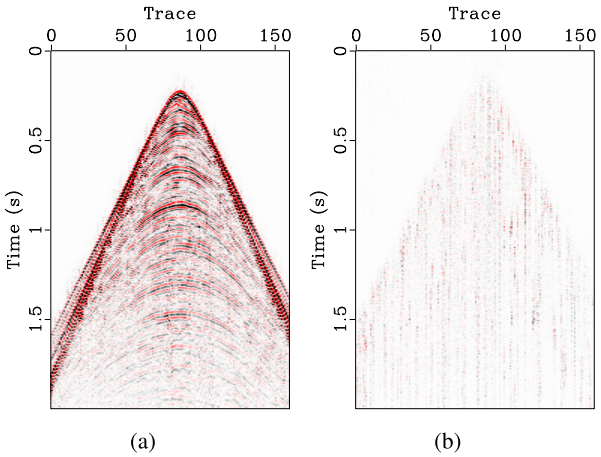

Fig. 9. Deblended results of the field pseudo-deblended record by the proposed method: (a) deblended result, (b) residual corresponding to (a).

The selected field CRG is shown in Fig. 8(a), which consists of 160 traces, and each trace has 500 samples with 4 ms time sampling interval. As shown in Fig. 8(b), the pseudo-deblended record is generated using the random dithering time varying from -0.4 to 0.4 s. Before separating the field data, we use a field data gathered from another zone to fine-tune the network parameters. The deblended result and residual of field pseudo-deblended record by the proposed method are shown in Fig. 9, and the recovered SNR of field pseudo-deblended data converges to 14.92 dB indicates that the proposed method achieves the good deblending performance and preserves the event energy to some extent.

IV. Conclusion

This letter presents DT, a self-attention module-based network, focusing on deblending. DT attends to separate the blended seismic data through learning the global representation of horizontal, vertical, and local variation rules. The self-attention mechanism allows global information interaction between data, improving the deblending performance. Experiments on synthetic and field examples confirm that the proposed DT network can deal with the blended interference robustly and effectively. Furthermore, the transformer network has been proven its excellent generalization ability and the potential to handle the multimodality task. However, lacking some inductive biases leads to the demand of a large dataset. Appropriate constraints, more smooth loss function, or more intelligent design are welcomed to improve the transformer, which needs to be delved deeper.

References

[1] A. J. Berkhout, "Changing the mindset in seismic data acquisition," *Lead. Edge*, vol. 27, pp. 924–938, Jul. 2008.

[2] G. Hampson, J. Stefani, and F. Herkenhoff, "Acquisition using simultaneous sources," in *Proc. 78th Annu. Int. Meeting, SEG Expanded Abstr.*, 2008, pp. 2816–2820.

[3] G. Blacquière, G. Berkhout, and E. Verschuur, "Survey design for blended acquisition," in *Proc. SEG Tech. Program Expanded Abstr.*, Jan. 2009, pp. 56–60.

[4] S. Gan, S. Wang, Y. Chen, and X. Chen, "Simultaneous-source separation using iterative seislet-frame thresholding," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 197–201, Feb. 2016.

[5] C. Bagaini, "Overview of simultaneous vibroseis acquisition methods," in *Proc. 76th Annu. Int. Meeting, SEG Tech. Program Expanded Abstr.*, Jan. 2006, pp. 70–74.

[6] Y. Xue, F. Chang, D. Zhang, and Y. Chen, "Simultaneous sources separation via an iterative rank-increasing method," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1915–1919, Dec. 2016.

[7] W. Huang, R. Wang, X. Gong, and Y. Chen, "Iterative deblending of simultaneous-source seismic data with structuring median constraint," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 58–62, Jan. 2018.

[8] T. W. Fei, Y. Luo, S. Aramco, and G. T. Schuster, "De-blending reverse-time migration," in *Proc. SEG Tech. Program Expanded Abstr.*, Jan. 2010, pp. 3130–3134.

[9] Y. Tang and B. Biondi, "Least-squares migration/inversion of blended data," in *Proc. SEG Tech. Program Expanded Abstr.*, Jan. 2009, pp. 2859–2863.

[10] W. Dai, P. Fowler, and G. T. Schuster, "Multi-source least-squares reverse time migration," *Geophys. Prospecting*, vol. 60, no. 4, pp. 681–695, 2012.

[11] Z. Xue, Y. Chen, S. Fomel, and J. Sun, "Seismic imaging of incomplete data and simultaneous-source data using least-squares reverse time migration with shaping regularization," *Geophysics*, vol. 81, no. 1, pp. S11–S20, Jan. 2016.

[12] Y. Chen, "Deblending using a space-varying median filter," *Explor. Geophys.*, vol. 46, no. 4, pp. 332–341, 2014.

[13] L. Zhang, Y. Wang, Y. Zheng, and X. Chang, "Deblending using a high-resolution radon transform in a common midpoint domain," *J. Geophys. Eng.*, vol. 12, no. 2, pp. 167–174, Apr. 2015.

[14] M. Bai and Y. Chen, "Least-squares Gaussian beam transform for deblending distance-separated simultaneous sources," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5280–5292, Jun. 2021.

[15] Y. Chen, S. Fomel, and J. Hu, "Iterative deblending of simultaneous-source seismic data using seislet-domain shaping regularization," *Geophysics*, vol. 79, no. 5, pp. V179–V189, Sep. 2014.

[16] J. Sun, S. Slang, T. Elboth, T. Larsen Greiner, S. McDonald, and L.-J. Gelius, "A convolutional neural network approach to deblending seismic data," *Geophysics*, vol. 85, no. 4, pp. WA13–WA26, Jul. 2020.

[17] S. Zu, J. Cao, S. Qu, and Y. Chen, "Iterative deblending for simultaneous source data using the deep neural network," *Geophysics*, vol. 85, no. 2, pp. V131–V141, Mar. 2020.

[18] X. Liu *et al.*, "Deep classified autoencoder for lithofacies identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5909914.

[19] A. Vaswani *et al.*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, pp. 1–5, Jun. 2017.

[20] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, pp. 1–22, Feb. 2020.

[21] T. T. Y. Lin and F. J. Herrmann, "Designing simultaneous acquisitions with compressive sensing," in *Proc. 71st Int. Conf. Exhib. (EAGE)*, 2009, p. 127.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[23] B. Wang, J. Li, J. Luo, Y. Wang, and J. Geng, "Intelligent deblending of seismic data based on U-Net and transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8885–8894, Oct. 2021.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.