IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 January 2023, accepted 4 February 2023, date of publication 13 February 2023, date of current version 16 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244683

RESEARCH ARTICLE

Understanding the Effect of Different Prior Knowledge on CNN Fault Interpreter

YU AN AND RUIHAI DONG
School of Computer Science, University College Dublin, Belfield, Dublin 4, D04 V1W8 Ireland

Corresponding author: Ruihai Dong (ruihai.dong@insight-centre.org)

This work was supported in part by the Science Foundation Ireland (SFI) under Grant SFI/12/RC/2289_P2, and in part by the Beijing Dublin International College Fund.

ABSTRACT As deep learning (DL) models have been successfully applied to various image processing tasks, DL models, particularly convolutional neural networks (CNN), have been introduced into the geosciences to assist geologists in faster seismic interpretation. However, the generalization of DL-based fault interpretation is a challenge. When applied to seismic data with different characteristics, their performance degrades significantly. Several recent studies have proposed transfer learning techniques, in which similar but different source tasks are assumed to benefit the target task. However, it is unclear which source datasets would be most beneficial for this particular task (i.e. fault interpretation). In this paper, we first demonstrate through a systematic literature review that synthetic seismic datasets are the most popular source datasets in this area. Further, previous studies have not compared them with other types of datasets. Then, we demonstrate experimentally that the choice of source dataset should be influenced by the amount of annotation available in the target dataset. In addition, normalisation appears to be an essential factor in fine-tuning techniques, particularly when interpreting faults. Finally, state-of-the-art performance was achieved on the ThebeFault dataset (0.903 for AP, 0.849 for OIS and 0.845 for ODS). Our code is publicly available at: https://github.com/anyuzoey/pretrain.

INDEX TERMS Deep learning, fault, seismic interpretation, transfer learning.

I. INTRODUCTION

Seismic interpretation, a process of analysing and interpreting seismic data, is essential for obtaining subsurface geological information such as geological structures and natural resources. In recent years, seismic interpretation algorithms based on DL have gradually emerged due to the high sensitivity of seismic datasets, the high variability of geological structures, the time-consuming and labour-intensive manual annotation, and the excellent performance of DL algorithms in other fields. However, DL algorithms are also data-hungry, leading to most DL-based seismic interpretation literature using synthetic seismic datasets.

Recent work such as [1] and [2] has argued that DL models trained directly using synthetic datasets show unsatisfactory performance on field seismic datasets because synthetic datasets are very different from real datasets. Even if a field seismic dataset is used to train a DL model, the trained DL model may still perform poorly on the new field seismic dataset due to the significant characteristic differences between different seismic datasets [3]. Therefore, transfer learning techniques have been proposed to alleviate this problem.

Transfer learning solves the problem of insufficient training datasets by transferring knowledge from a source domain (e.g. a synthetic dataset) to a target domain (e.g. a new target dataset) [4]. A common assumption of transfer learning is that the accuracy of the target dataset will benefit from the knowledge learned from similar but not identical datasets. It is difficult to define the similarity between two datasets, especially for pixel semantic segmentation tasks. In computer vision, ImageNet (a large-scale image classification dataset) is often chosen as the default pre-training dataset, and most achieve satisfactory performance regardless of the specific

The associate editor coordinating the review of this manuscript and approving it for publication was Joey Tianyi Zhou.

vision task. This choice implies an assumption that image datasets are all similar. Nevertheless, [5] mentioned that several medical image processing pieces of literature believe non-medical source datasets may be too different from the target medical dataset and will not provide useful prior knowledge [6], [7], [8]. Reference [5] then performed several experiments and concluded that ImageNet is the best source dataset for medical datasets and that the size of the source dataset is disproportionate to its performance on the target medical dataset.

Like medical image datasets, seismic datasets are signal-imaged and differ significantly from optically imaged natural image datasets such as ImageNet. Whether this difference affects the effectiveness of the prior knowledge provided by natural image datasets on seismic datasets still needs to be explored. Besides, much literature in seismic data interpretation that adopted transfer learning uses seismic datasets as source datasets by default without a doubt of whether it is the best for this specific task. Thus, we are motivated to compare the popular computer vision datasets with seismic datasets and investigate which type would be the best source dataset for seismic interpretation tasks.

The highlights of this paper are as follows:

- Provided the first systematic literature review on deep transfer learning and seismic fault interpretation. Revealed synthetic seismic data is almost the default source dataset in the domain literature.
- Conducted the first investigation regarding the influence of different source datasets on the effectiveness of DL fault interpreters when transfer learning was used.
- Demonstrated that applying transfer learning does not necessarily lead to a better outcome; seismic source data is not always the best option; and it is important to consider the number of annotations available in the target dataset when choosing the source dataset.
- Analyzed the impact of normalisation and outliers on DL fault interpreters.
- Achieved state-of-the-art performance on the Thebe-Fault dataset.

The remainder of this paper is organised as follows. Section II presents our systematic literature review of the source datasets used for DL-based fault interpreters. Section III describes the datasets involved. Section IV documents the methodology. Section V provides experimental results while Section VI gives a discussion of experimental results, limitations and future works. Finally, our paper is summarised in Section VII.

II. SYSTEMATIC LITERATURE REVIEW

Geological faults are planar fractures in the Earth's crust that are often associated with the accumulation of natural resources such as oil and gas [2]. Fault interpretation is one of the most critical components of seismic interpretation, focusing on locating and annotating geological faults (i.e. the black lines in the Fig. 3 ThebeFault and FaultSeg annotations) on seismic data. Fault interpretation is particularly challenging due to the thinness of the target. It was not until 2018, as shown in Fig. 1, that DL techniques were successfully incorporated into fault interpretation. DL-based fault interpretation models have evolved rapidly since then, demonstrating encouraging results on seismic datasets [2], [9]. In response to the demands of big data-based DL algorithms, two seismic datasets have been made publicly available in 2019 and 2021 respectively. Nevertheless, some recent literature has highlighted the fact that most supervised trained CNN fault interpreters are unable to correctly connect newly acquired seismic data with the training data, resulting in unsatisfactory generalization performance on new seismic data [1]. This phenomenon is primarily associated with two challenges in this task: 1. the huge variation in seismic data characteristics. 2. DL methods' high data volume requirement while obtaining seismic fault annotations is difficult.

Transfer learning techniques, which focus on improving the learning of new tasks by transferring knowledge of known tasks, are particularly well suited to deal with the two challenges mentioned above. Transfer learning on DL methods is called deep transfer learning. The most popular class of deep transfer learning is network-based deep transfer learning methods. Its basic principle is that shallow convolutional layers learn features of low-level generic visual cues such as angles and edges. In contrast, the deeper convolutional layers gradually move towards complex visual cues focused on objects and highly relevant to the target task. This technique directly reuses partial or all of the pre-trained weights and fine-tunes the weights to suit the new task [4]. With the success of transfer learning in other DL tasks, network-based deep transfer learning has been adopted recently in geoscience. It relaxes the required training data volume and alleviates the poor generalisation ability of current CNN fault interpreters to new seismic datasets [1].

Although much literature has successfully applied the network-based deep transfer learning technique in the fault interpretation task, there is a lack of systematic review on what source dataset is often used for the specific seismic fault interpretation task and for what reason. Here, a systematic literature review was conducted in accordance with Kitchenham's best practice systematic review guidelines [10]. The detailed flowchart is shown in Fig. 2. In response to the research questions we want to answer, we focus on papers using neural networks and involving pre-training and fine-tuning for the fault interpretation task. On 2 April 2022, we searched articles that include "seismic", "neural network", and "fault" in the title or abstract and contain the keywords "pre-train" or "fine-tune" in the full text. Articles should be peer-reviewed journal articles or conference research articles as they are of high quality.

Two well-known digital libraries (IEEE Xplore and Science Direct) were searched, and a total of 13 research papers were retrieved. One was excluded because it was irrelevant to seismic data, and four were excluded because keywords only appeared in related work. Besides, one paper,

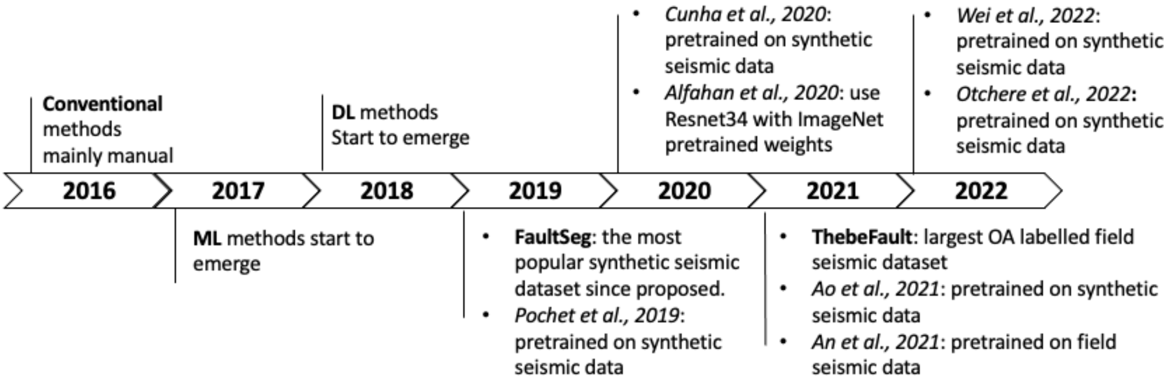

Conventional methods mainly manual
DL methods Start to emerge
Cunha et al., 2020: pretrained on synthetic seismic data
Alfahan et al., 2020: use Resnet34 with ImageNet pretrained weights
Wei et al., 2022: pretrained on synthetic seismic data
Otchere et al., 2022: pretrained on synthetic seismic data
2016
2017
2018
2019
2020
2021
2022
ML methods start to emerge
FaultSeg: the most popular synthetic seismic dataset since proposed.
Pochet et al., 2019: pretrained on synthetic seismic data
ThebeFault: largest OA labelled field seismic dataset
Ao et al., 2021: pretrained on synthetic seismic data
An et al., 2021: pretrained on field seismic data

FIGURE 1. The developmental trajectory of fault interpretation. Among them, we indicate the occurrence time of two seismic datasets and seven selected literature.

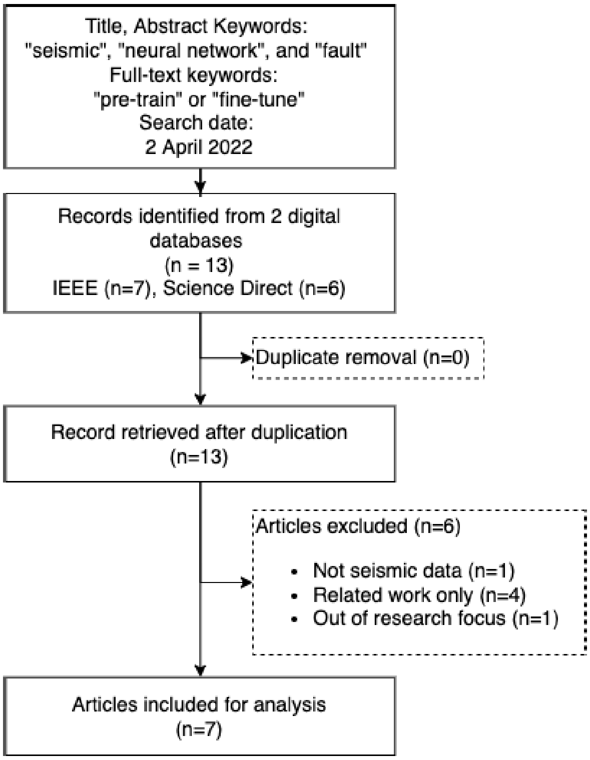

Title, Abstract Keywords: "seismic", "neural network", and "fault"
Full-text keywords: "pre-train" or "fine-tune"
Search date: 2 April 2022
Records identified from 2 digital databases (n = 13)
IEEE (n=7), Science Direct (n=6)
Duplicate removal (n=0)
Record retrieved after duplication (n=13)
Articles excluded (n=6)
Not seismic data (n=1)
Related work only (n=4)
Out of research focus (n=1)
Articles included for analysis (n=7)

FIGURE 2. Flowchart showing our literature search inclusion process.

which uses pre-trained weights to build a complex ensemble network, is excluded as it is irrelevant to our research focus. As a result, seven research papers were involved in the analysis.

The most popular source dataset category is synthetic seismic datasets, with five of the seven (71.4%) papers selected falling into this category [1], [11], [12], [13], [14]. A typical background is that synthetic seismic datasets are often used to train the DL-based fault interpreters, and the trained models are generalised poorly on field seismic datasets. Fine-tuning techniques are, therefore, used to address this problem. In addition to the above five articles, [2] mentioned using fine-tuning techniques to improve their DL-based fault interpretation models trained from field seismic datasets.

For both types of articles, the fine-tuning technique was only used to improve the performance of DL models on new field seismic datasets without considering the differences in the source datasets and their effectiveness. Other than the above six articles, [15] discusses the difference between having a source dataset and not. This article used the ImageNet dataset (default choice in computer vision) as the source dataset and significantly improved the prediction results. However, it did not discuss the difference between ImageNet, a non-seismic dataset, and seismic datasets regarding the target seismic interpretation problem. Among them, only [1] mentioned comparing ImageNet with synthetic seismic data as one of the potential future works. Therefore, in this work, we compare several popular computer vision datasets, including ImageNet, with seismic datasets for the first time to investigate which can provide better prior knowledge for the target seismic fault interpretation task.

III. DATASETS

In this paper, we involved five different datasets, two (i.e. ThebeFault, FaultSeg) of which are seismic datasets and three (i.e. ImageNet, COCO, BSDS500) of which are non-seismic datasets. All datasets involved are publicly available datasets. Table 1 presents the summary of the five datasets including their source links. Fig. 3 shows some visual examples of each dataset.

ThebeFault [2], [3], [19], [20] is a large geological fault dataset obtained by experts from the Fault Analysis Group at University College Dublin, annotated on the seismic data allocated from the ThebeFault gas field, located on the North West Shelf of Australia. The size of this dataset is $1803[inline] \times 1537[sample] \times 3174[crossline]$. As described in [2], the dataset contains three subsets: the training set, the validation set and the test set. The first 900 inline sections were divided as the training set, followed by 200 inlines as the validation set, and the final 703 inlines as the test set. The processed training and validation sets are 181,029 and 64,317 patches of size 96×96 pixels, respectively. This is

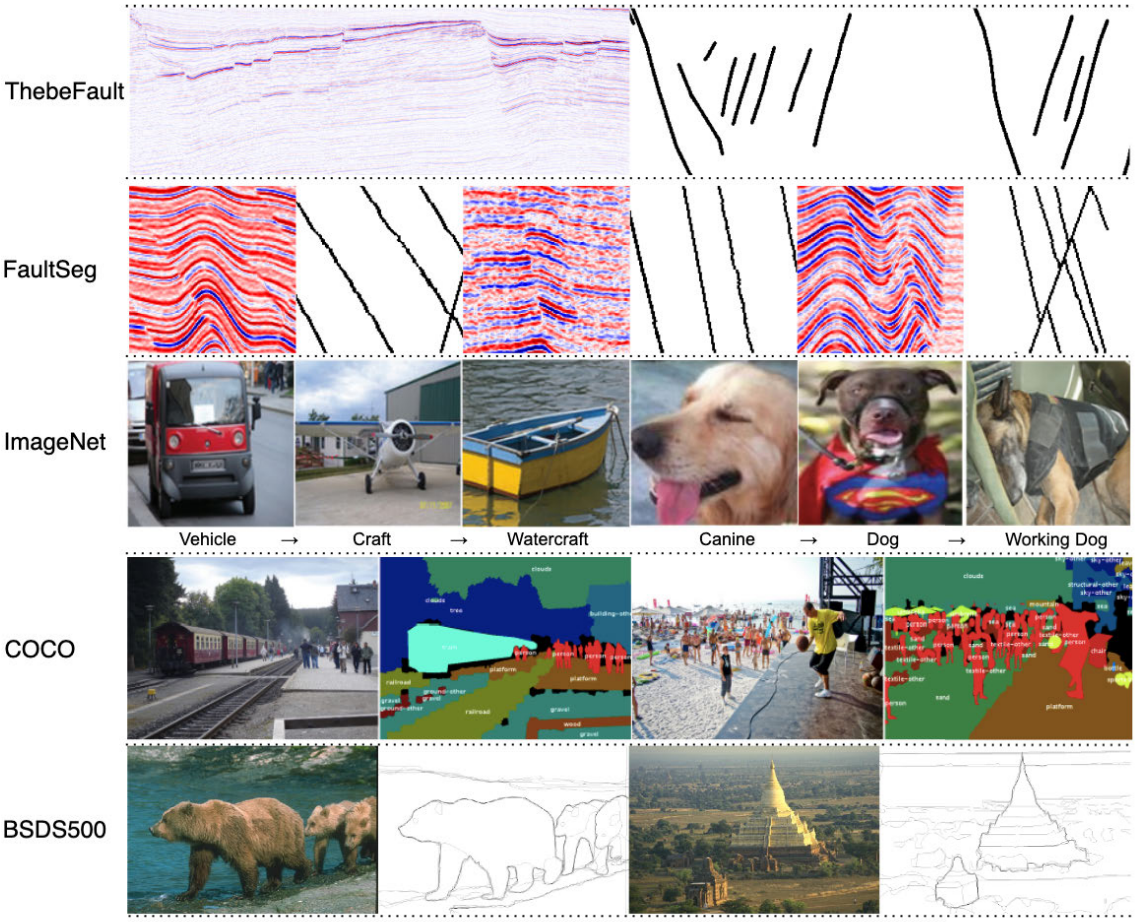


FIGURE 3. Illustration of involved dataset: ThebeFault [3] (left is seismic data, right is fault annotation), FaultSeg [9] (three image-annotation pairs, left are synthetic seismic data, right are fault annotations), ImageNet [16] (six image-annotation pairs, upper are image data, lower are corresponding image-level label/annotations.), COCO [17] (two image-annotation pairs, left are image data, right are semantic segmentation annotations.), BSDS500 [18] (two image-annotation pairs, left are image data, right are edge/contour annotations).

TABLE 1. Details of the datasets involved in this paper.

Name	Origin	Training size	Item size	Channel	Min	Max	Mean	STD
ThebeFault[a]	Seismic	900	1537x3174	1	-8.394	9.433	0.000	0.124
FaultSeg[b]	Seismic	200	128x128x128	1	-7.892	8.168	0.000	1.052
BSDS500[c]	Non-seismic	200	321x481	3	0	1	[0.432, 0.436, 0.364]	[0.252, 0.236, 0.245]
COCO[d]	Non-seismic	95,279	640x480	3	0	1	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]
ImageNet[e]	Non-seismic	1,281,167	469x387 (average)	3	0	1	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]

[a] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YBYGBK [b] https://github.com/xinwucwp/faultSeg
[c] https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/ [d] https://cocodataset.org/ [e] https://www.image-net.org/

the largest publicly available expert-annotated field seismic dataset, to the best of our knowledge.

FaultSeg is a well-known 3D synthetic seismic data proposed by [9] to train a CNN fault interpreter. The paper proposes a seismic data-generating approach that can generate seismic images and corresponding fault annotations with various characters by customising different combinations of parameters (e.g. seismic fold level and fault structure, wavelet peak frequency magnitude and noise intensity). This approach generates datasets that theoretically cover enough geological structures and should perform well on different field seismic datasets. Based on the above assumption and the fact that it is so far the most popular synthetic seismic baseline dataset in this area, we believe this dataset can provide sufficient prior experience related to geological fault recognition. This dataset has 220 seismic cubes of size 128 × 128 × 128. Among them, 200 belong to a training set, and the other 20 are split into a test set.

ImageNet dataset proposed in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [16] is the most popular image classification benchmark in the image processing domain. It provides the first large-scale open-source dataset with 1000 object classes and 1,281,167 training images, 50,000 validation images and 100,000 test images.

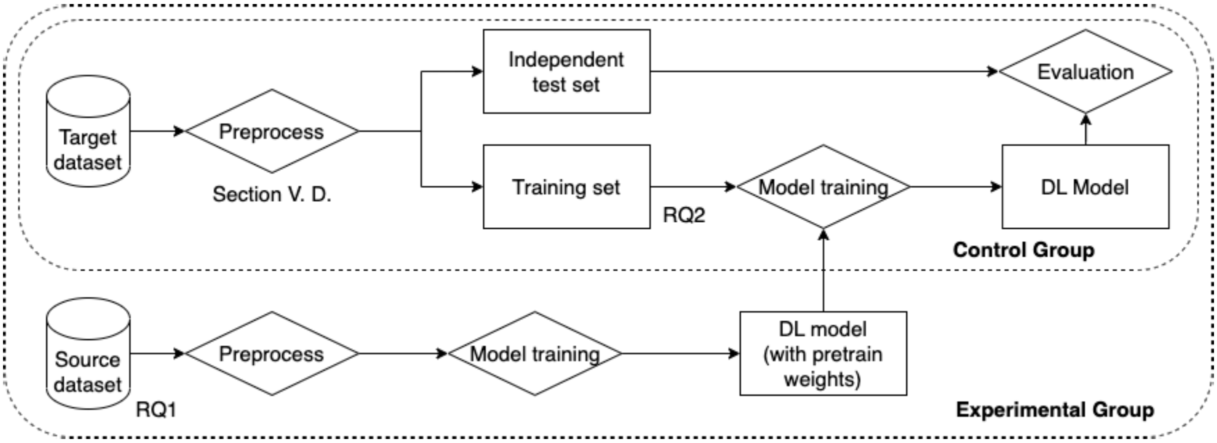


FIGURE 4. Overview of the methodology. The experimental group has an additional component where models are pre-trained on the source dataset than the control group. There are three main areas of interest in this paper. 1. the impact of source dataset selection, 2. the amount of annotations in the target dataset that can be involved in fine-tuning, and 3. the impact of image processing methods, especially normalisation.

ImageNet is therefore often considered one milestone in the history of DL development and often used as the default pre-training source dataset for different image processing tasks [21]. Images in ImageNet are all coloured images captured by commercial cameras with an average size of 469 × 387 pixels.

The Common Objects in Context (COCO) Dataset [17] is another popular computer vision baseline for object detection, segmentation and captioning tasks and a popular source dataset choice. In this paper, we focus on how this dataset will contribute to the task of fault interpretation. As fault interpretation is often considered an image segmentation task, we hypothesise this dataset may be more similar to seismic fault interpretation datasets than ImageNet. Here we use the pre-trained weights trained from the COCO dataset provided by PyTorch for its ease of use [22]. In PyTorch, the weights were trained using 20 of the 91 stuff categories from the COCO train2017 version. The corresponding training set was 95,279 coloured images with an image size of 640 by 480 pixels.

Berkeley Segmentation Data Set 500 (BSDS500) [18] is a classic benchmark for the edge detection task. Annotations include object contours, interior object boundaries and background boundaries. Multiple annotators annotate each image. As faults yield otherwise continuous rock layers discontinuous, traditional fault interpretation usually involves using edge detection algorithms. We hypothesise that this popular edge detection dataset may be closer to the seismic fault interpretation task and provide useful prior knowledge. Therefore, we selected and introduced this dataset in our experiments. The dataset consists of 500 colour images divided into three parts, of which 200 are used for training, 100 for validation and the remaining 200 for testing.

IV. METHODOLOGY

In the systematic literature review described above, we demonstrate that seismic datasets, particularly synthetic datasets, are often used as source datasets. And, there is a lack of literature exploring the effects of source datasets on the CNN fault interpreter. Therefore, this paper focuses on filling this research gap. More specifically, we aimed to answer the following two research questions. RQ1: When using transfer learning, how do different source datasets affect the effectiveness of CNN fault interpreters? RQ2: Is the amount of annotations within the target set relevant to the choice of source dataset?

For RQ1, we can break it down into two sub-questions: What source datasets (i.e. seismic or non-seismic) are most beneficial for fault interpretation? Do different source datasets always lead to positive results? To investigate these questions systematically, we created a control group and an experimental group, as shown in Fig. 4. The control group represents DL models trained from scratch (i.e. random initialisation of weights) using only the training set of the target dataset. The experimental group represents fault interpreters trained with a priori knowledge. Specifically, the DL models were trained using one of the source datasets and then fine-tuned using the target dataset. We had two goals in mind for this experiment. As a first step, we examined whether the experimental group consistently outperforms the control group when the source datasets are different. Next, we assessed whether seismic source datasets have more positive effects on evaluation results than non-seismic sources.

Building on RQ1, we also investigated whether the number of annotations available for fine-tuning within the target set influenced the selection of the source dataset (i.e. RQ2). Since it is difficult to obtain fault annotations in practice, it is rare to find target datasets with as many annotations as ThebeFault. In order to simulate scenarios with different amounts of annotation in real applications, we gradually reduced the amount of annotations in the target dataset involved in the fine-tuning process (i.e. 10%, 1%, 0.1%).

In order to simulate real-world usage, the ThebeFault dataset was chosen as the target dataset since it is the only real seismic dataset available. The other four datasets were selected as source datasets to provide a priori knowledge. Thus, the control group is trained by supervised learning using the entire ThebeFault training set. The experimental group are models that were pre-trained on one of the four source datasets and fine-tuned on the ThebeFault dataset using its training set.

DeepLabV3 with ResNet101 [23] backbone (DR) was chosen as the primary model architecture for two main reasons. First, it is the best-performing model for the segmentation task offered by the PyTorch library [22]. Secondly, the PyTorch library offer pre-trained weights learnt from ImageNet or COCO for this dataset, which can greatly save the effort of repeated training. In addition, we compared our optimal solution with several benchmarks including a state-of-the-art solution on the ThebeFault dataset. Included benchmark models are: UNet [24], DeepLabV3-MobileNet (DeeplabM) [25], [26], [27], HED [28] and RCF [29].

We followed the same evaluation methods proposed by [2] and [3]: average precision (AP) and two F1 scores based on the optimal threshold per image (OIS) and the optimal threshold for the dataset (ODS). AP provides a comprehensive evaluation result without considering threshold selection and is chosen as the primary evaluation metric compared to ODS and OIS.

V. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

The parameter settings for all experiments were kept as constant as possible and close to baseline to ensure a fair comparison. The input patches are all 96 by 96 pixels in size. Adam [30] optimiser with an initial learning rate of 0.01 was set for UNet and DeepLabM and 0.001 for HED, RCF and DR. The batch size was set to 64 (as literature [2]) for the four baseline models and 32 for model DR. The model architecture for DR is more complex and requires more GPU memory. The training and validation iterations were set to 100 and 20 per epoch, respectively. Learning rate and early stopping scheduler were used to avoid overfitting. All experiments in this paper were conducted on a GeForce GTX 1080 Ti graphics card. Following PyTorch's recommendation of normalizing the images, we first normalized the target dataset to [0, 1] based on its maximum and minimum values, and then z-score normalized the target dataset using the mean and variance of the source dataset to ensure a consistent distribution between the two datasets.

B. RQ1

In the "our solutions" section of Table. 2, we have named the control group, i.e. the fault interpreters that did not apply transfer learning, "DR_DN". The interpreter for the experimental group is then denoted by the suffix "*ft". When using the full ThebeFault training set, the model with prior knowledge showed comparable or higher performance than the control group model. While seismic source datasets are generally considered more beneficial to fault interpreters, ImageNet, a non-seismic dataset, provided a surprising optimal boost. The boosts for AP, OIS and ODS were 1.53%, 1.51% and 1.50% respectively. This is followed by another non-seismic dataset, COCO, and a seismic dataset, FaultSeg. As a result of using the COCO dataset as the source dataset, AP, OIS, and ODS gain 0.15%, 1.07%, and 1.11%, respectively. Using the FaultSeg dataset as the source dataset gives a 0.51%, 0.43%, and 0.33% improvement in AP, OIS, and ODS. Finally, the non-seismic dataset, BSDS500, achieved a slightly lower (−0.16% to −0.39%) result than the control group.

TABLE 2. SOTA comparisons on the ThebeFault dataset.

Model	AP	OIS	ODS
Benchmarks			
UNet	0.757	0.769	0.766
DeeplabM	0.784	0.759	0.756
HED	0.823	0.811	0.806
RCF	0.794	0.806	0.800
Ablation study			
UNet_IN	0.838	0.829	0.825
UNet_ClipDN	0.833	0.836	0.886
UNet_DN	0.883	0.842	0.837
DeeplabM_DN	0.868	0.818	0.815
HED_DN	0.887	0.830	0.827
RCF_DN	0.839	0.833	0.828
Our solutions			
DR_DN	0.887	0.834	0.830
DR_FaultSegft	0.892	0.838	0.833
DR_BSDS500ft	0.886	0.830	0.826
DR_COCOft	0.889	0.844	0.841
DR_ImageNetft	**0.903**	**0.849**	**0.845**

This experiment demonstrates that applying transfer learning does not necessarily lead to a better outcome. Contrary to popular belief, the seismic source dataset is not necessarily the best choice. With the current findings in hand, we will continue to explore the factors that influence both findings and attempt to explain them.

In addition, we present some baseline methods and their performance data in Table. 2. Our method significantly improved fault interpretation, with the most accurate fault interpreter, DR_ImageNetft, showing an improvement of 7.95%, 3.78%, and 4.49%, respectively, over the previous SOTA results (i.e. HED [2]) in the evaluation metrics AP, OIS, and ODS.

C. RQ2

As shown in Fig. 5, the performance of the CNN fault interpreter on the target ThebeFault test set shows a high correlation with the volume of training annotations involved in control group training or experimental group fine-tuning. For the pure supervised model DR_DN in the control group, the test set's performance drops as expected when trained using fewer samples. When too few samples are involved in the training process, the performance of model DR_DN drops dramatically. By introducing prior knowledge learned from different source datasets, models in the experimental group generally alleviate the requirement for the amount

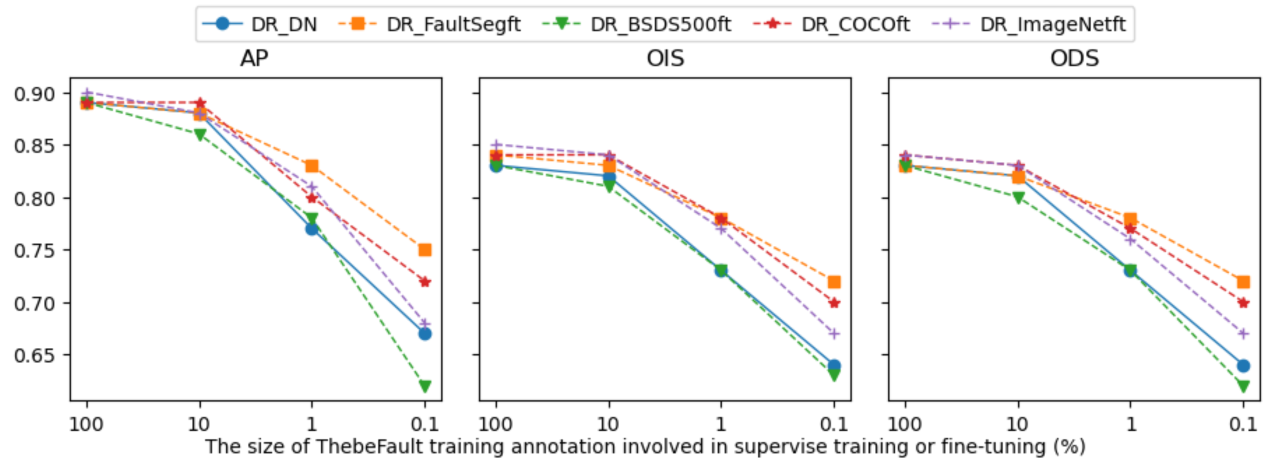


FIGURE 5. Performance of the CNN fault interpreter on the ThebeFault test set with different amounts of ThebeFault training annotations involved in the supervised learning or fine-tuning. Here, 100%, 10%, 1% and 0.1% are the four experimental annotation volumes involved in supervised training of the DR_DN baseline or fine-tuning from the four source datasets. For example, 0.1% represents using only 0.1% of the ThebeFault training annotations to simulate the use case that only a limited number of annotations are available on the new seismic dataset. Y-axis labels are placed as subtitles.

of training data and slow down the rate of performance degradation.

Interestingly, we found that as fewer labels were used in fine-tuning, the lead of DR_ImageNetft at 100% and DR_COCOft at 10% was replaced by DR_FaultSegft. Furthermore, DR_COCOft consistently outperforms DR_ImageNetft when viewed from the 10% point to the right. As a result of this phenomenon, it would be prudent to take into consideration the number of annotations available in the target dataset when selecting the source dataset. When only a limited number of fault annotations are available in the target dataset, the FaultSeg dataset is better than the popular image pre-training dataset and can alleviate the need for large amounts of labelled data. Some visual examples are provided in Fig. 6.

D. NORMALISATION

As Table. 2 shows, our methods show a significant performance improvement over the baseline models. Referring to the training process of the control group model in Fig. 4, there are three main possible elements, i.e. data processing method, model selection, and training method. The impact of the latter two is more intuitive, and in this section, we focus on the impact of the data processing method. We note that the main difference between our methods and the benchmark methods in terms of data handling is the data normalisation method. We use the z-score normalisation (i.e. "_DN") method suggested by the PyTorch library, whereas the benchmark methods all use min-max normalisation (i.e. "_IN"). Their formulas are listed in 1 and 2, where x_{min} and x_{max} represent the image-wise minimum and maximum values of the input pixels x. μ and σ represent the mean and standard deviation of the source dataset. Moreover, we found that of the seven relevant papers found through the systematic literature review, three did not provide details of the normalisation process, three used min-max normalisation and only one used z-score normalisation.

$$x_{IN} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

$$x_{DN} = \frac{x - \mu}{\sigma} \tag{2}$$

To focus on the differences introduced by the normalisation methods, we designed ablation experiments. As shown in Table 2, we use the suffixes "_IN" and "_DN" to denote the image-wise min-max normalisation method and z-score normalisation, respectively. The suffix "_ClipDN" denotes clipping of outliers prior to z-score normalisation.

As Table. 2 shows, our version of baseline models all show a significant improvement in effectiveness over the original baseline models. Here, to isolate the impact of the training method, we add two more sets of UNet-based experiments, namely UNet_IN and UNet_ClipDN. it can be observed that the z-score normalisation approach leads to a 4.41% 1.27% 1.21% improvement in the AP, OIS, and ODS metrics, respectively, for the same model under the same training conditions.

We suspect that this performance improvement may be related to the considerable outliers in the field seismic dataset. Here, the outliers are defined by the interquartile range (IQR) method [31]. As shown in Fig. 7, applying min-max normalisation to the seismic images changes the distribution of the original data. In addition, outliers in seismic data provide valuable geological information that can be used to visualise different horizons (i.e. rock layers), as shown in Fig: 8. And the experimental findings also verify that the fault interpreter is less effective if the outliers are clipped (i.e. UNet_ClipDN and UNet_DN). This again demonstrates the uniqueness of seismic datasets compared to natural image datasets. In this regard, we recommend not clipping seismic outliers and instead using z-score normalisation, which is more tolerant of outliers.

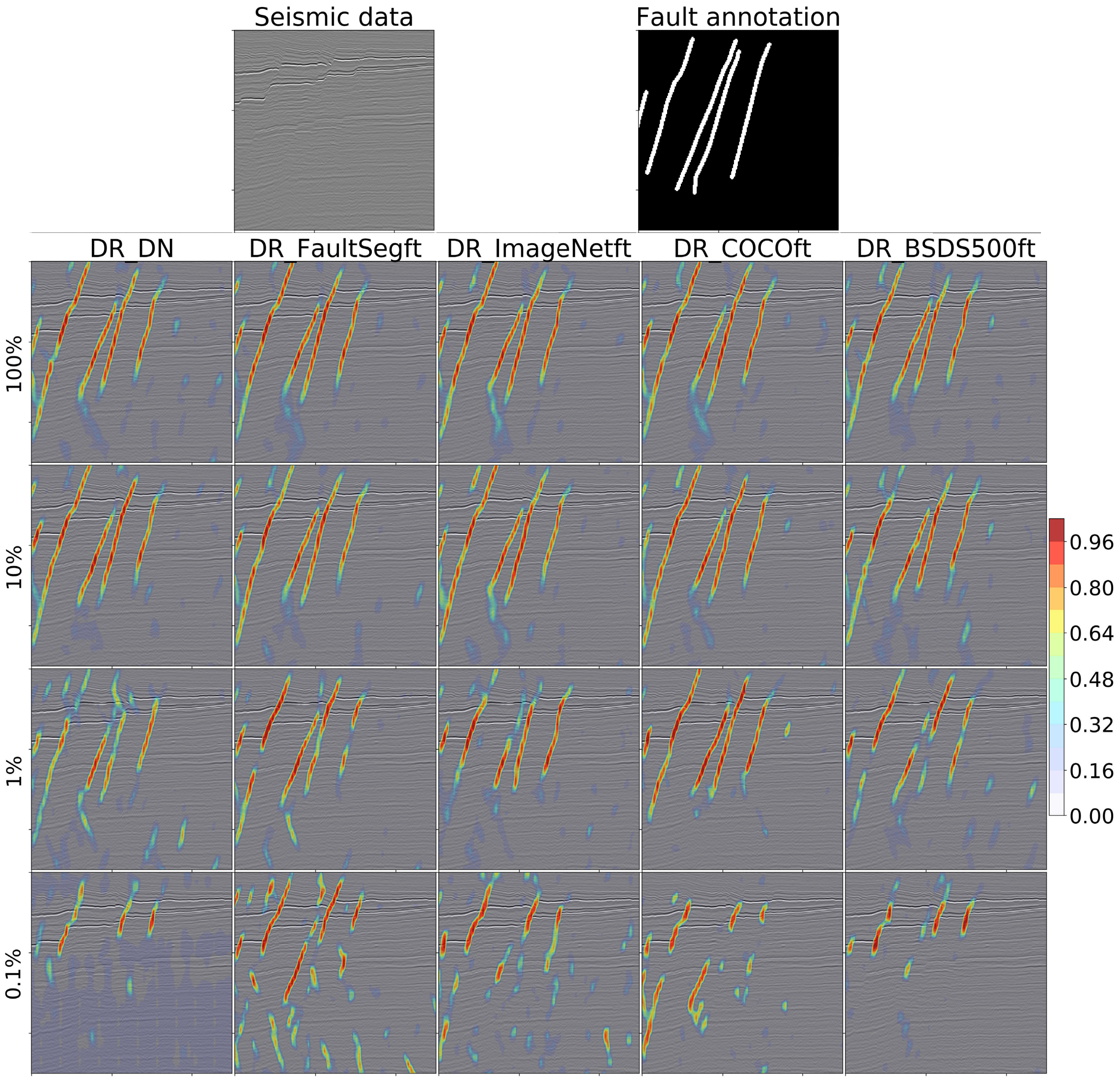


FIGURE 6. Visual examples of the CNN fault interpreter predictions on the ThebeFault test set. Seismic data is a 500 × 500 patch randomly cropped from the Thebe test set. Fault annotations are annotations labelled by fault experts. The coloured lines are predictions by CNN models, overlaid on seismic data for reference.

VI. DISCUSSION

This paper is the first to focus on the impact of the choice of source dataset on the final DL fault interpreters, when applying deep transfer learning. This issue has been neglected in previous studies. Transfer learning literature typically recommends using a dataset that is similar to the target dataset. However, it is difficult to measure the distance between datasets and to demonstrate a correlation between distance and performance. Additionally, previous studies have typically relied on transfer learning to enhance the generalisation of fault interpreters trained from synthetic seismic datasets only, without examining the impact of the choice of source datasets. Our results can assist geologists and DL experts in selecting and utilizing existing datasets more rationally, thus improving the performance of DL fault interpreters.

Based on the results of RQ1 and RQ2, we find that using transfer learning does not necessarily produce more positive results for seismic datasets. Moreover, unlike common assumptions in previous studies, seismic datasets do not necessarily outperform non-seismic datasets as source datasets. In particular, when the quantity of target dataset annotations is large enough, the large-scale ImageNet dataset will provide the most effective a priori experience. However, when the target seismic dataset has few annotations, the synthetic seismic dataset will provide the best a priori experience. Reviewing the dataset comparisons shown in Table 1, ImageNet is the largest dataset, followed by COCO, Thebefault, FaultSeg, and finally BSDS500. Based on task similarity, ThebeFault is most similar to FaultSeg, followed by BSDS500, COCO and finally ImageNet. Combining the above experimental results and the above dataset ranking,

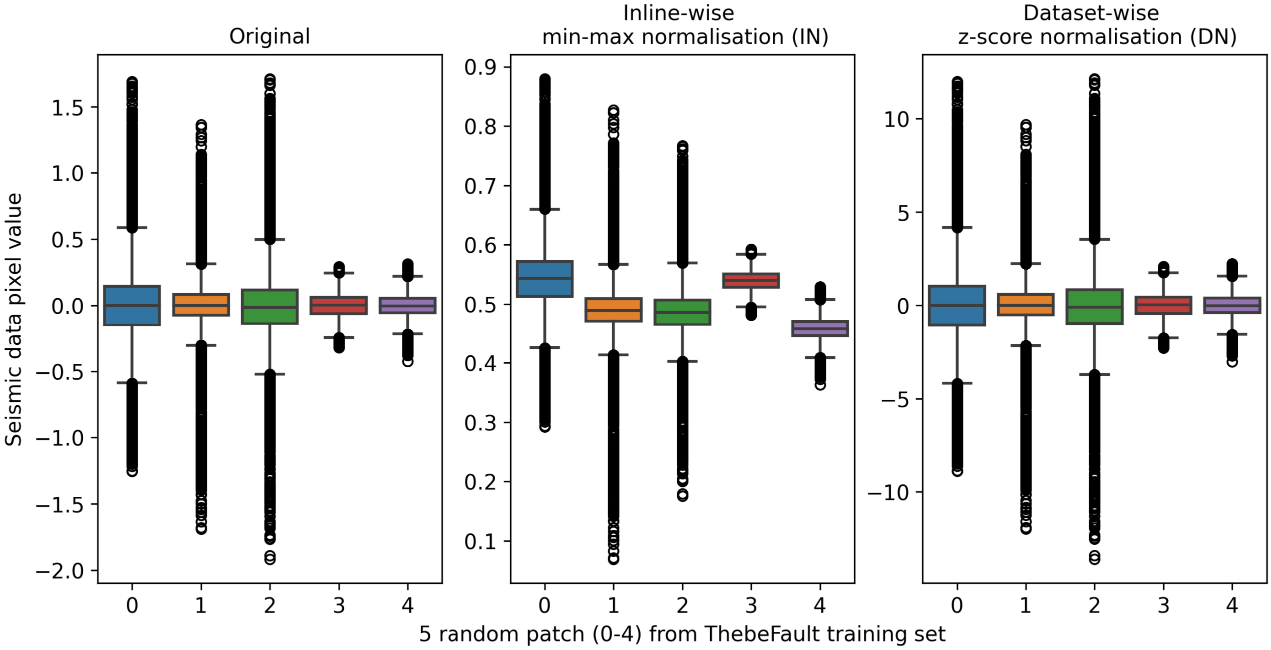


FIGURE 7. Box plots of 5 random training patches with different normalisation. Left: original seismic data without normalisation. Middle: inline normalisation (IN) proposed by [2]. Right: dataset-wise z-score normalisation (DN). The outliers are indicated by black circles. Because of the large number of outliers, they are superimposed on each other and are not clearly visible.

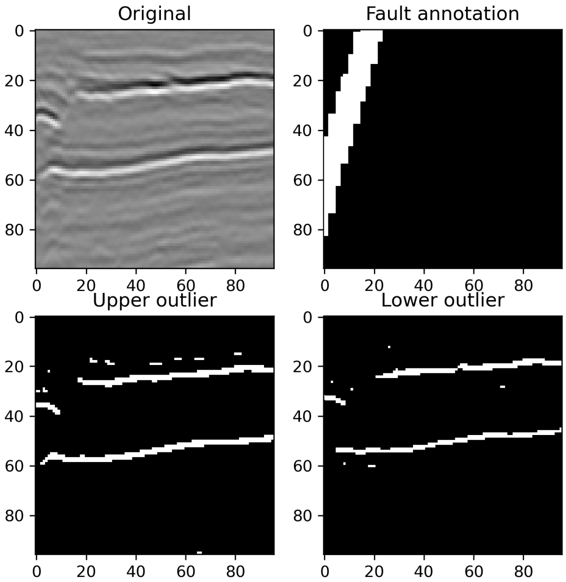


FIGURE 8. Outlier illustration of patch #2 in Fig. 7. Top left: original seismic data. Top right: corresponding fault annotation. Lower left: pixels with values above the upper boundary. Lower right: pixels with values below the lower boundary. There is a clear association between the fault annotation and the outliers.

it can be speculated that when there are only a limited number of labels, the FaultSeg dataset is relevant enough to provide valid a priori knowledge of the target task, despite its small size. In cases where there are many labels for the target dataset, a larger dataset would provide more general purpose prior experience which may be of some benefit to the model.

In order to explain the above experimental results, here we introduce a loss landscape visualization method. The loss landscape visualisation method was initially developed to explain why specific networks or setups were easier to train than others [32]. They concluded that a visually wider and flatter loss landscape usually represents a good DL model, implying a better convergence ability and generalisation. With a wider and flatter loss landscape, the network weights ideally should update to global minima more easily with less likely to land in chaotic regions (i.e. a place that contains many local minima and has a high loss barrier to global minima) [32], [33], [34].

As shown in Fig. 9, all models trained with prior knowledge (i.e. the right four columns in Fig. 9) produced flatter loss landscapes than models trained using just the target dataset (i.e. the left-most column in Fig. 9). This suggests that the source dataset may affect convergence and generalisation ability. Once again, the benefits of having source datasets are demonstrated. Of the four source datasets, ImageNet and COCO provided the two widest and flattest loss landscapes, followed by FaultSeg and finally BSDS500. This order is almost identical to the size order of the source datasets, which may indicate that a larger training dataset may provide a model with better generalisation. In addition, it appears that having more labels from the target dataset during supervised training or fine-tuning produces a flatter loss landscape and lower error values (i.e. the top row 100% ThebeFault training set vs. 0.1% ThebeFault training set).

This paper is also the first to investigate the impact of data normalisation processing on DL seismic interpreters. Previous work has mostly ignored this issue, commonly

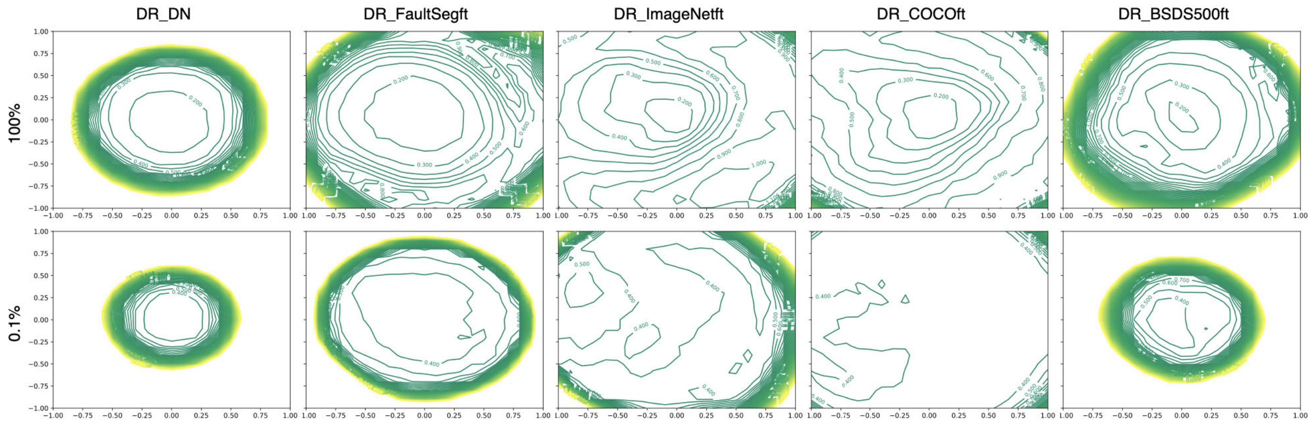


FIGURE 9. loss landscape visualisation of the trained CNN fault interpreter. From left to right: DR_DN (model trained using only ThebeFault training set) and four models pre-trained using one of the four source datasets and then fine-tuned on the ThebeFault training set. The top row uses the entire ThebeFault training set, and the lower row uses 0.1% ThebeFault training set.

TABLE 3. GPU time cost (hours) for each solution.

Model	100%[f]	10%[f]	1%[f]	0.1%[f]	Pretrain[g]
UNet_IN	0.78				0
UNet_clipDN	1.21				0
UNet_DN	1.14				0
DeeplabM_DN	0.80				0
HED_DN	9.13				0
RCF_DN	11.44				0
DR_DN	8.02	5.93	2.98	1.18	0
DR_FaultSegft	8.09	3.84	2.41	0.68	12.76
DR_BSDS500ft	8.18	4.32	3.20	2.38	6.31
DR_COCOft	7.85	5.06	3.23	0.77	0[h]
DR_ImageNetft	8.10	4.25	2.97	0.76	0[h]

[f] % represents the percentage of ThebeFault training set involved for fine-tuning.
[g] Time cost (hours) for pretraining.
[h] ImageNet, COCO pretrained weights are provided by PyTorch.

using simple min-max normalisation methods to rescale values to [0, 1]. However, we noticed that seismic datasets have very different values from natural image datasets. For natural image datasets taken with a standard optical camera, the images are stored as integer values from 0 to 255. The standard normalisation for image datasets is min-max normalisation, which divides all values by 255. Seismic datasets, on the other hand, are images formed by processing signals captured by remote geophones/sensors and have no fixed min and max values. Besides, the extreme values of the seismic dataset contain critical geological information. Therefore seismic datasets without effective processing can lead to reduced performance of DL seismic interpreters. We recommend that interested researchers use z-score normalisation when processing seismic datasets and do not clip outliers in seismic datasets.

The study involved a considerable amount of time and effort in pre-training and fine-tuning each dataset. In Table. 3, we list the time costs associated with each solution. Clearly, model training time is positively correlated with the amount of training data. Model architecture also has a significant impact. Compared to the SOTA solution (i.e. HED), our optimal model requires less training time. Furthermore, this model has pre-training models for ImageNet and COCO, which significantly reduces pre-training time. Thus, when choosing the source dataset, it is essential to also take into account the time required for pre-training. However, transfer learning will generally have a faster convergence rate when pre-training time is not taken into account.

Due to limitations in computing resources, only five datasets were analysed. However, if more datasets are involved, it is possible to find more suggestions for source dataset selection. As fault annotations is often intellectual property in the field of seismic data interpretation, there are very few relevant publicly available datasets. Due to this constraint, only two seismic datasets are used in this paper. We, therefore, encourage future researchers to publish their data, code, and pre-trained models to facilitate subsequent research. Despite the fact that we analyzed the impact of the source datasets using loss surfaces, we did not find any significant correlation between the accuracy of the ThebeFault test set and the loss landscape. Moreover, future work could focus on methods that measure the distance between datasets, so that a suitable source dataset could be recommended.

VII. CONCLUSION

In this paper, a systematic review and extensive experiments were presented to understand the effect of different prior knowledge on CNN fault interpretation models. Based on the numerical and visual results, we recommend that domain researchers consider the available annotations in the target dataset and then decide on the source dataset accordingly. Neural networks loss landscape visualisations demonstrate that having a source dataset and involving more training data could help result in a CNN model with better generalisation. Even though there aren't enough explanations about "why and what makes certain source datasets suitable?", we hope our paper will assist our colleagues in selecting the appropriate source datasets. Additionally, our approach achieves state-of-the-art performance on the ThebeFault test set.

REFERENCES

[1] A. Cunha, A. Pochet, H. Lopes, and M. Gattass, ‘‘Seismic fault detection in real data using transfer learning from a convolutional neural network pre-trained with synthetic seismic data,’’ *Comput. Geosci.*, vol. 135, Feb. 2020, Art. no. 104344, doi: 10.1016/j.cageo.2019.104344.

[2] Y. An, J. Guo, Q. Ye, C. Childs, J. Walsh, and R. Dong, ‘‘Deep convolutional neural network for automatic fault recognition from 3D seismic datasets,’’ *Comput. Geosci.*, vol. 153, Aug. 2021, Art. no. 104776, doi: 10.1016/j.cageo.2021.104776.

[3] Y. An, J. Guo, Q. Ye, C. Childs, J. Walsh, and R. Dong, ‘‘A gigabyte interpreted seismic dataset for automatic fault recognition,’’ *Data Brief*, vol. 37, Aug. 2021, Art. no. 107219, doi: 10.1016/j.dib.2021.107219.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, ‘‘A survey on deep transfer learning,’’ in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2018, pp. 270–279.

[5] I. Van Den Brandt, F. Fok, B. Mulders, J. Vanschoren, and V. Cheplygina, ‘‘Cats, not CAT scans: A study of dataset similarity in transfer learning for 2D medical image classification,’’ 2021, *arXiv:2107.05940*.

[6] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, ‘‘Transfusion: Understanding transfer learning for medical imaging,’’ 2019, *arXiv:1902.07208*.

[7] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, ‘‘Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data,’’ in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1251–1255.

[8] E. W. Teh and G. W. Taylor, ‘‘Learning with less data via weakly labeled patch classification in digital pathology,’’ in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 471–475.

[9] X. Wu, L. Liang, Y. Shi, and S. Fomel, ‘‘FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation,’’ *Geophysics*, vol. 84, no. 3, pp. 35–45, May 2019.

[10] S. Keele, ‘‘Guidelines for performing systematic literature reviews in software engineering,’’ Softw. Eng. Group, School Comput. Sci. Math., Keele Univ., Keele, U.K., Dept. Comput. Sci., Univ. Durham, Durham, U.K., Tech. Rep., 2.3, 2007.

[11] A. Pochet, P. H. B. Diniz, H. Lopes, and M. Gattass, ‘‘Seismic fault detection using convolutional neural networks trained on synthetic poststacked amplitude maps,’’ *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 352–356, Mar. 2019, doi: 10.1109/LGRS.2018.2875836.

[12] Y. Ao, W. Lu, B. Jiang, and P. Monkam, ‘‘Seismic structural curvature volume extraction with convolutional neural networks,’’ *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7370–7384, Sep. 2021, doi: 10.1109/TGRS.2020.3042098.

[13] D. A. Otchere, B. N. Tackie-Otoo, M. A. A. Mohammad, T. O. A. Ganat, N. Kuvakin, R. Miftakhov, I. Efremov, and A. Bazanov, ‘‘Improving seismic fault mapping through data conditioning using a pre-trained deep convolutional neural network: A case study on Groningen field,’’ *J. Petroleum Sci. Eng.*, vol. 213, Jun. 2022, Art. no. 110411.

[14] X. L. Wei, C. X. Zhang, S. W. Kim, K. L. Jing, Y. J. Wang, S. Xu, and Z. Z. Xie, ‘‘Seismic fault detection using convolutional neural networks with focal loss,’’ *Comput. Geosci.* vol. 158, Jan. 2022, Art. no. 104968, doi: 10.1016/j.cageo.2021.104968.

[15] M. Alfarhan, M. Deriche, A. Maalej, G. AlRegib, and H. Al-Marzouqi, ‘‘Multiple events detection in seismic structures using a novel U-Net variant,’’ in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2900–2904.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ‘‘ImageNet: A large-scale hierarchical image database,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[17] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, ‘‘Microsoft COCO: Common objects in context,’’ 2014, *arXiv:1405.0312*.

[18] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, ‘‘Contour detection and hierarchical image segmentation,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[19] Y. An, Q. Ye, J. Guo, and R. Dong, ‘‘Overlap training to mitigate inconsistencies caused by image tiling in CNNs,’’ in *Proc. Int. Conf. Innov. Techn. Appl. Artif. Intell.* (Lecture Notes in Computer Science), vol. 12498. Cham, Switzerland: Springer, 2020, pp. 35–48, doi: 10.1007/978-3-030-63799-6_3.

[20] H. Du, Y. An, Q. Ye, J. Guo, L. Liu, D. Zhu, C. Childs, J. Walsh, and R. Dong, ‘‘Disentangling noise patterns from seismic images: Noise reduction and style transfer,’’ *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2022.3219117.

[21] M. Huh, P. Agrawal, and A. A. Efros, ‘‘What makes ImageNet good for transfer learning?’’ 2016, *arXiv:1608.08614*.

[22] A. Paszke, ‘‘PyTorch: An imperative style, high-performance deep learning library,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[23] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ 2015, *arXiv:1512.03385*.

[24] O. Ronneberger, P. Fischer, and T. Brox, ‘‘U-Net: Convolutional networks for biomedical image segmentation,’’ 2015, *arXiv:1505.04597*.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ‘‘DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,’’ 2016, *arXiv:1606.00915*.

[26] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ‘‘Rethinking atrous convolution for semantic image segmentation,’’ 2017, *arXiv:1706.05587*.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, ‘‘Encoder–decoder with atrous separable convolution for semantic image segmentation,’’ 2018, *arXiv:1802.02611*.

[28] S. Xie and Z. Tu, ‘‘Holistically-nested edge detection,’’ 2015, *arXiv:1504.06375*.

[29] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, ‘‘Richer convolutional features for edge detection,’’ 2016, *arXiv:1612.02103*.

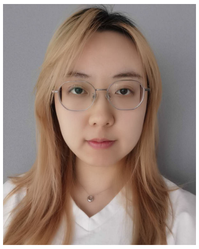
[30] D. P. Kingma and J. Ba, ‘‘Adam: A method for stochastic optimization,’’ 2014, *arXiv:1412.6980*.

[31] S. Walfish, ‘‘A review of statistical outlier methods,’’ *Pharmaceutical Technol.*, vol. 33, no. 11, p. 82, 2006.

[32] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, ‘‘Visualizing the loss landscape of neural nets,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[33] M. Mosbach, M. Andriushchenko, and D. Klakow, ‘‘On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines,’’ 2020, *arXiv:2006.04884*.

[34] Y. Hao, L. Dong, F. Wei, and K. Xu, ‘‘Visualizing and understanding the effectiveness of BERT,’’ 2019, *arXiv:1908.05620*.



YU AN received the B.S. degree in Internet of Things from University College Dublin, Dublin, Ireland, in 2018, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include deep learning, computer vision, and seismic data interpretation.

RUIHAI DONG received the Ph.D. degree in computer science from University College Dublin, Dublin, Ireland, in 2015. He is currently an Assistant Professor with the School of Computer Science, University College Dublin. He has published in top peer-reviewed journals and conferences, such as WWW, RECSYS, IUI, and IJCAI. His research interests include machine learning, deep learning, and their applications in recommender systems, finance, health, and geoscience.