This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3171694, IEEE Transactions on Geoscience and Remote Sensing

A Stronger Baseline for Seismic Facies Classification with Less Data

Xiaoyu Chen, Qi Zou*, Xixia Xu, Nan Wang

***Abstract*—With the great success of deep learning in computer vision, the application of Convolution Neural Network (CNN) in seismic facies classification is growing rapidly. However, most of the previous works based on pure state-of-the-art CNN architectures still suffer from coarse segmentation results. In this paper, we study the challenges of seismic facies classification and propose a stronger baseline. More specifically, we propose a simple yet effective unsupervised approach named Spatial Pyramid Sampling (SPS) to choose representative samples for training to reduce the labeling costs. Next, we propose a Multi-modal Fusion (M2F) module to extract and fuse the edge and frequency information from selected seismic images to build a stable multimodal representation. Finally, we propose a Local to Global (L2G) module, which improves the recognition power by capturing the local relationship between pixels and enhancing the global context representation. Experimental results demonstrate that the proposed method achieves superior performance with less labeled training data, especially for small categories.**

***Index Terms*—Seismic facies classification, semantic segmentation, multi-modal knowledge, local awareness, context aggregation.**

I. Introduction

SEISMIC facies classification refers to the interpretation of facies type from the seismic reflector information[1]. Since seismic facies classification is a labor-intensive and time-consuming task, robust automated and semi-automated classification algorithms are pursued in both industry and academy[2].

In the past, there were many methods[3–5] based on machine learning to analyze seismic facies. With the great success of deep learning in computer vision, the application of CNN in seismic facies classification is growing rapidly[6–9]. However, it still faces many challenges: (1) annotation is time-consuming and it overloads human interpreters with the increasing amount of geophysical information; (2) different from natural images, seismic images have no obvious semantic objects, but more edge and texture information; (3) poor performance in small categories due to class imbalance.

To cope with the first challenge, existing works[2, 10] resort to specific data partition strategies as shown in Fig.1(a) to reduce labeling costs. Since the training set and test set are similar, the performance on the test set will not be unsatisfactory though the model may be overfitted. However, when the special setting where training set and test set are similar are broken, such as the data split as in Fig.1(b), the performance on the test set is extremely limited. Based on the observation as shown in Fig.1(b) that adjacent samples are very similar in both sample space and feature space, we hope to select representative samples for annotation to reduce the feature redundancy. Out of this intuition, we propose SPS which adopts Harris corner detection[11] to construct spatial pyramid features and can quickly select representative samples without any supervision. Experiments testify that our method can improve the performance and greatly reduce the training data. Furthermore, our method is suitable for arbitrary data splitting form, which is more in line with the real-world applications.

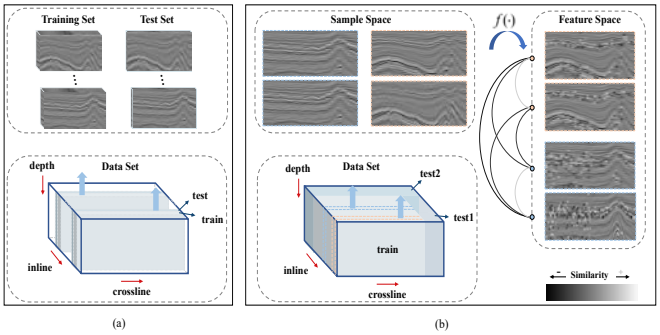


Fig. 1. Different ways of reducing labeling costs. (a) Specific data splitting to make test data highly similar to training data. 3D seismic volume is uniformly divided into different blocks. In each block, the first slices go to the training set and the adjacent rest to the test set. (b) Selecting informative samples using feature similarity. This can be used in arbitrary data splitting. Intuitively, adjacent slices are more similar. $f(\cdot)$ represents the feature extractor.

To tackle the second challenge, we fuse the edge and frequency information in seismic images to get complementary features. Seismic facies classification involves the external geometry, continuity and frequency of seismic wave reflection. Intuitively, simple grayscale images cannot efficiently describe these characteristics. Therefore, we propose EDGE module to help the network use edge and texture information as a clue for distinguishing different seismic facies. On the other hand, we utilize the FDM (Feature De-drifting Module)[12] to help the network exploit different frequency information, from high frequency areas reflecting abrupt changes of seismic events to low frequency areas reflecting the overall trends, as shown in Fig.2(a). Overall, we integrate EDGE and FDM modules and further propose the M2F module to extract different modal information such as edges and frequency, and fuse them with the original input into a stable multimodal representation.

In addition, pixel-wise relations and global context are

* Corresponding author
Xiaoyu Chen, Qi Zou, Xixia Xu and Nan Wang are with School of Computer and Information Technology, Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, 100044, China (email : 20120344@bjtu.edu.cn; qzou@bjtu.edu.cn; 19112036@bjtu.edu.cn;16112070@bjtu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3171694, IEEE Transactions on Geoscience and Remote Sensing

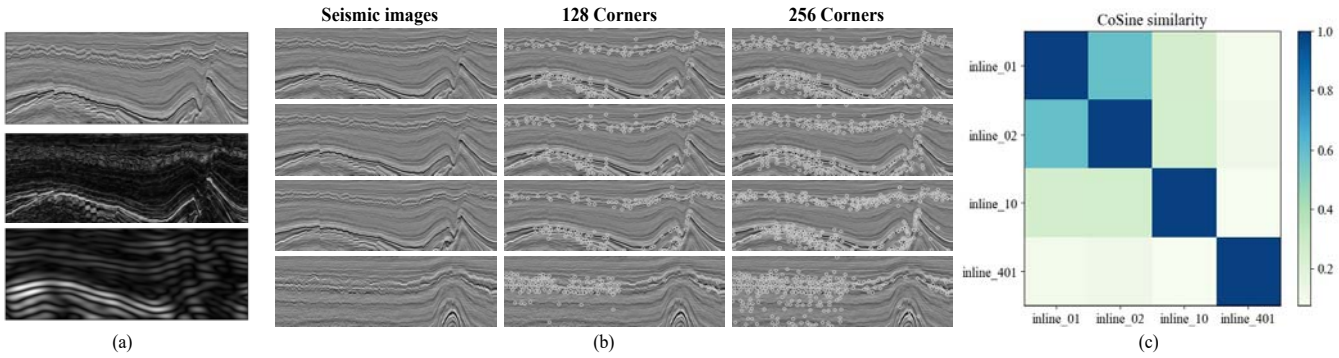


Fig. 2. (a) Results of the seismic image through the frequency filter, from top to bottom, the seismic image, the image after high-pass filtering, and the image after low-pass filtering. (b) Harris corner detection on seismic images. The 1^{st} column presents the raw seismic data. From top to bottom presents inline 01/inline 02/inline 10/inline 401, and the 2^{nd}/3^{rd} column presents the number of corners 128/256 detected on seismic data separately. These examples are selected from the F3 training set. We can see that closer lines have more similar density distributions of corners. (c) We adopt SPS ($L = 1$) to construct one-dimensional features for each seismic image, and calculate the cosine similarity between them. The results demonstrate that closer lines have higher cosine similarity of their feature vectors.

mined to enhance the representation of seismic facies. This works together with the SPS and M2F, improving seismic classification especially in rare classes. We notice that there is a strong correlation among different categories and pixels of seismic images. For example, the formation time of different seismic facies is inconsistent, and there are sedimentary sequences and interactions among them. Based on the above considerations, we propose a L2G module which adopts the Convolutional Block Attention Module (CBAM)[13] to capture the local relationship among pixels and introduces the Object-Contextual Representations (OCR)[14] scheme to further enhance the global context representation.

The contributions of this paper are summarized as follows:

• We propose a simple yet effecient unsupervised method named SPS to select representative samples in seismic images, which can improve the performance and considerably reduce the training data.

• We propose a M2F module to extract various modal information such as edges and frequency, and fuse them with the original input into a stable multimodal representation, which helps the network extract more diverse and robust features.

• We propose a L2G module to enhance the local awareness and context representation of the model. Cooprating with SPS and M2F, we can achieve higher accuracy espacially for small categories.

II. Related Work

Semantic Segmentation. Semantic segmentation aims to assign a label to each pixel in an image. With the development of deep learning, the FCN-based method [15] combined with encoder-decoder architectures [16–20] have become more and more popular for semantic segmentation. After that, a great deal of research improved the performance of semantic segmentation in different aspects. To enlarge the receptive field, several approaches proposed dilated or atrous convolutions [21–24]. To aggregate information of different scales, pyramid scene parsing network (PSPNet) [25] utilized a pyramid pooling module (PPM) to fuse features under different pyramid scales. DeepLabV2 [26] proposed an atrous spatial pyramid pooling (ASPP) module to adopt pyramid dilated convolutions with different dilation rates. High-Resolution Network (HRNet) [27] exchanges information on parallel multi-resolution subnetworks for multi-scale fusion. However, these models are not customized for seismic images.

Feature Representation and Enhancement. The most relevant feature enhancement methods for our work include attention mechanism, contextual aggregation, and multimodal knowledge fusion. Attention-based models[13, 28–31] are widely used in enhancing feature representation. Squeeze-and-Excitation Networks (SENet)[29] adaptively calibrates channel-wise feature responses by learning the interdependence between channels. CBAM[13] connects channel attention module and spatial attention module to adapt feature refinement. In recent years, numerous works on context representation have emerged[14, 32–35] to descirbe each pixel by its relationship with the surrounding pixels. For instance, OCR[14] characterizes a pixel by exploiting the representation of the corresponding object class to augment context representation. Multimodal approaches aim to extract and combine relevant information from the different modalities and hence take better decisions than using only one. Inspired by previous works[12, 36–38], we aggregate the edge information and frequency information to build multi-modal expression.

Seismic Facies Classification. In recent years, several works used deep learning to classify seismic facies. In the early stage, seismic facies classification is regarded as image classification. [6] proposed a series of patch-based models to classify seismic facies. [7] used fine-tune to benefit from pre-trained networks and evaluate their performance on seismic data. The disadvantages of these methods is the loss of context and spatial information caused by inconsistent patch scale. Afterwards, seismic facies classification is regarded as the task of semantic segmentation. [8] have introduced an encoder-decoder CNN model for seismic facies classification and compared it with the patch-based models. [9] published a fully annotated 3D geological model of the Netherlands F3 Block and provided a baseline model for seismic facies segmentation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3171694, IEEE Transactions on Geoscience and Remote Sensing

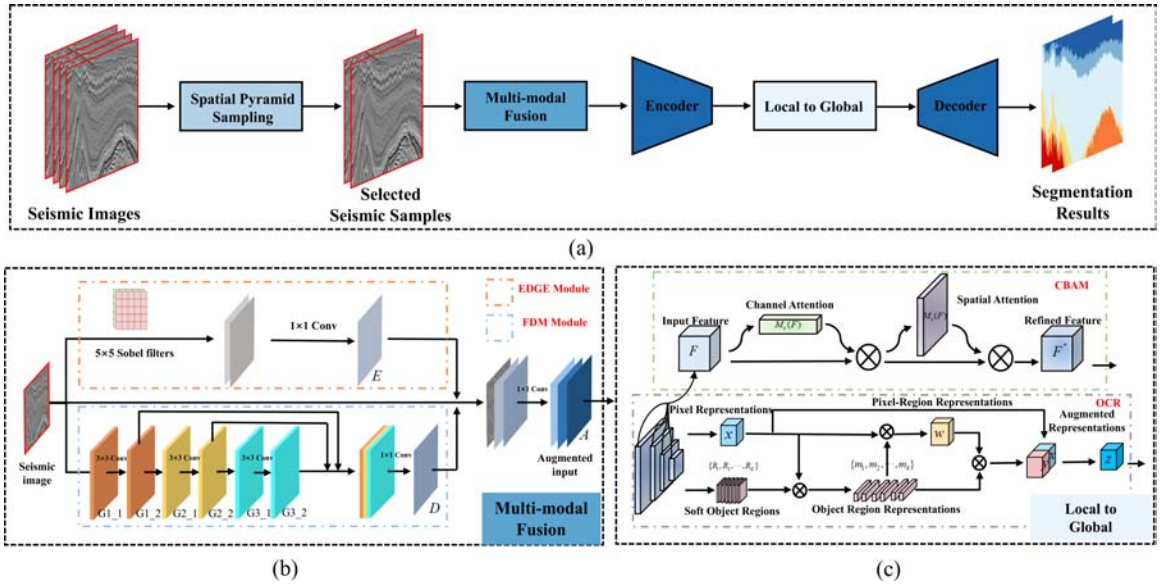


Fig. 3. (a) The pipeline of our model in training stage. The seismic images are first selected by SPS, and then send to (b) M2F module to build multimodal representations. Finally, we fed the output from M2F to the encoder-decoder network with (c) L2G module to get the segmentation results.

[2] used a transposed residual unit to replaces the traditional dilated convolution for the decode block and proposed Danet-FCN2 and Danet-FCN3. [39] proposed an unsupervised deep domain adaptation network to semantically segment the seismic images. The most related work to ours is [10]. They employed the wavelet transform to implement the spectral decomposition to get vertical slices of different frequencies as inputs to the network, and proposed Spatial–Spectral Attention to enforce the model to suppress the interference from the untargeted information. While our approach utilize convolution operation to dig the potential knowledge in frequency space instead of using wavelet transform. In addition, the purpose of our attention mechanism is to enhance local awareness and context representation. Furthermore, we all use F3 data set, but the number of categories and the division of data set are different.

III. Method

In this section, we present the technical details of our methods including the SPS, the M2F module, and the L2G module. The overall pipeline of our model is illustrated in Fig.3.

A. Spatial Pyramid Sampling

Seismic facies annotation is a time-consuming task. In order to reduce the workload of labeling, we propose a simple yet effective unsupervised approach named SPS to choose representative samples for annotation.

The Harris corner detector[11] is a popular interest point detection method due to its strong invariance to rotation, translation, illumination variation, and image noise. Here we adopt it to find interesting points in seismic images, which describes the geometric properties of seismic facies (e.g., continuity, configuration, and curvature). Generally, the adjacent slices should be similar since they represent seismic characteristics of neighboring regions[2]. We illustrate some examples in Fig.2(b) and Fig.2(c) to describe this phenomenon. Out of this intuition, we are able to construct the holistic feature space by only using representative samples.

Our approach is inspired by Spatial Pyramid Matching(SPM)[40], which is an algorithm for image matching, using "spatial pyramid" image representation. Given a seismic image I of shape $H \times W$, we first divide I into L levels patches, the shape of the lth level patches is $\frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}$, for each level there are $Q_l = 4^{l-1}$ patches, in all we can obtain $Q = \sum_{l=1}^{L} Q_l = \frac{4^L-1}{3}$ patches with different levels and spatial positions.

For each patch, we construct feature map f of size $h \times w$ ($h = \lceil \frac{H}{2^L} \rceil$, $w = \lceil \frac{W}{2^L} \rceil$), then Q feature maps are flattened, and concatenate together with the weight $w_l = 2^{l-L}$ to penalize features found in larger patches because they involve increasingly dissimilar features. Finally we obtain 1D feature vector with length $\frac{4^L-1}{3} \times h \times w$ to describe I. Let's take the ith patch of lth level as an example to construct the feature map f_i^l.

The process of SPS is shown in Fig.4. First, we use Harris corner detection to get the coordinates of each corner. Then we divide the patch into $h * w$ grids evenly. Finally, we count the number of corners falling into each grid to construct the feature map f_i^l. As mentioned above, we obtain 1D feature vectors for each seismic image. And we use the KMeans clustering algorithm to select samples in different classes to form new training data. Moreover, the selection of clusters cls depends on the number of labeled samples. if there are N samples in total, and we only want half of them to be labeled, then $cls = N/2$.

Our SPS adopts "spatial pyramid" to construct features, which can capture both coarse features and fine-grained features, and also learn spatial location relations. Experiments

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3171694, IEEE Transactions on Geoscience and Remote Sensing

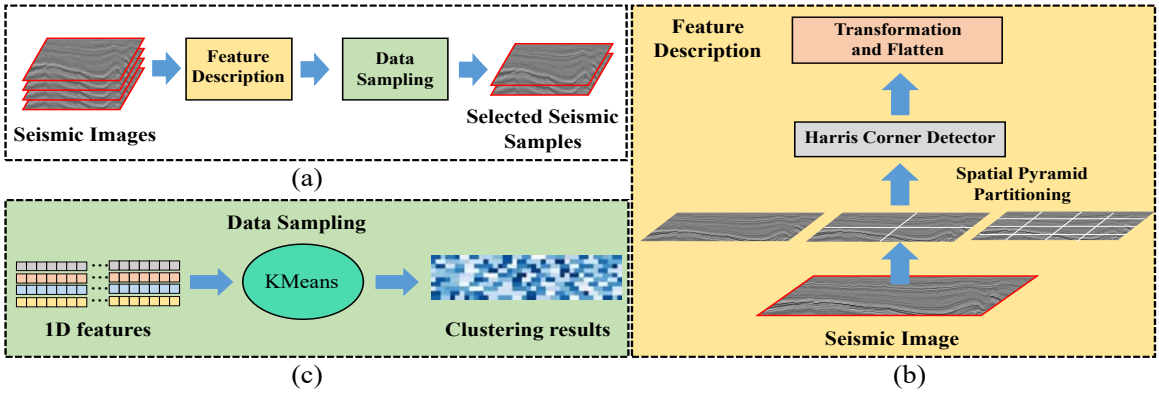


Fig. 4. The process of Spatial Pyramid Sampling. (a) The overview of SPS. (b) The details of constructing feature description for seismic image using Harris corner detection. (c) The representative samples sampling with KMeans.

testify that our method achieves better performance although with only fewer samples.

B. Multi-modal Fusion

The M2F module first extracts the multi-modal knowledge of edge information and frequency information in seismic images by the EDGE and the FDM modules, and aggregate the representations by fusing them to explore the latent geological features. As shown in Fig.3(b), the details of each component are given as follows.

EDGE Module. Considering that the edge information of seismic images can reflect the external geometry, continuity, amplitude, and other characteristics of seismic waves, so we devise the EDGE module to enhance the edge features of seismic images. Given a seismic image I of shape $1 \times H \times W$, we compute the edge map E:

$$E = s([S_x * I, S_y * I]), \tag{1}$$

where S_x and S_y are Sobel operators of shape 5×5 (other edge operators such as Roberts and Prewitt are also available), $*$ denotes the convolution operation, $s(\cdot)$ is a transformation function used to fuse the gradient features in horizontal and vertical direction, implemented by 1×1 Conv $\rightarrow$ BN $\rightarrow$ ReLU.

FDM Module. Our FDM module is based on the work [12] to dig the potential knowledge in frequency domain by simulating the linear weighting between the central, surround, and marginal frequency parts.

Similarly, a seismic image I as input, the output D is produced by filtering it with Gaussian filters:

$$D = d[C_1(I * G_1(\sigma_1)) + C_2(I * G_2(\sigma_2)) + C_3(I * G_3(\sigma_3))], \tag{2}$$

where G_i denote Gaussian filters with different filter bandwidths, σ_i are the scale parameters determining the filter bandwidths, $d(\cdot)$ is a transformation function implemented by 1×1 Conv $\rightarrow$ BN $\rightarrow$ ReLU,

$$G_i(\sigma_i) = \frac{1}{\sqrt{2\pi{\sigma_i}^2}} \exp(-\frac{x^2+y^2}{2\sigma_i}). \tag{3}$$

$C_1 \sim C_3$ represent the coefficients in the central, surround, and marginal frequency areas, respectively. $*$ denotes the convolution operation. We use convolution kernel G_i' to replace G_i, and use the convolution result in the first (second) term as the input of the second (third) term, thus Eq.(2) can be reformulated as:

$$\begin{aligned} D = d[C_1(I * {G_1}') + C_2((I * {G_1}') * {G_2}') \\ + C_3(((I * {G_1}') * {G_2}'){G_3}'))]. \end{aligned} \tag{4}$$

Details of the FDM module are summarized in Table I for inline data and crossline data.

TABLE I
THE DETAILS OF FDM WITH INLINE AND CROSSLINE DATA. THE NETWORK STRUCTURE OF FDM ARE SHOWN IN FIG.3(B).

Inline/Crossline	InputSize	Num	Filter	Stride	Pad
G1_1	1*255*701/401	64	3	1	1
G1_2	64*255*701/401	64	3	1	1
G2_1	64*255*701/401	32	3	1	1
G2_2	32*255*701/401	32	3	1	1
G3_1	32*255*701/401	16	3	1	1
G3_2	16*255*701/401	16	3	1	1
D	112*255*701/401	1	1	1	0

Finally, we integrate original input I, edge map E and FDM output D by a transformation function $a(\cdot)$ to obtain the augmented input A.

$$A = a([I, E, D]), \tag{5}$$

where $a(\cdot)$ is a transformation function implemented by 1×1 Conv $\rightarrow$ BN $\rightarrow$ ReLU.

C. Local to Global

L2G module imporves the feature discriminative power by exploring the category-wise and pixel-wise correlations in seismic images. As shown in Fig.3(c), L2G module contains the CBAM[13] module to capture the local relationship between each pixel, and the OCR[14] module to enhance the global context representation.

Convolutional Block Attention Module. CBAM consists of a channel attention module and a spatial attention module in sequence, focusing on 'what' and 'where' is meaningful for a given input, respectively.

Channel attention module first uses average-pooling and max-pooling operations from an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$ to generating two different spatial context descriptors: F^c_{avg} and $F^c_{\max}$, which represent average-pooled features and max-pooled features respectively. And then these two descriptors are forwarded to a shared network composed of multi-layer perceptron (MLP) with one hidden layer to produce channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$. Channel attention can be formulated as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(\text{ AvgPool }(F)) + MLP(\text{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{\text{avg}}\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right)\right), \end{aligned} \tag{6}$$

where C is the number of channels, $W_0 \in \mathbb{R}^{C/r\times C}$, $W_1 \in \mathbb{R}^{C\times C/r}$ are the weights of MLP, and σ denotes the sigmoid function.

Spatial attention module applies average-pooling and max-pooling operations along the channel axis to produce two different feature descriptions $F^s_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{\max} \in \mathbb{R}^{1\times H\times W}$, then concatenate them to generate a spatial feature map $M_s(F) \in \mathbb{R}^{H\times W}$ by a standard convolution layer. Spatial attention can be formulated as follows:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times 7}([\text{ AvgPool }(F); \text{Max Pool}(F)])\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{\text{avg}}; F^s_{\max}\right]\right)\right), \end{aligned} \tag{7}$$

where $f^{7\times 7}$ represents a convolution operation with the filter size of 7×7 and σ denotes the sigmoid function.

In general, the whole attention process can be summarized as:

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F) \otimes F', \end{aligned} \tag{8}$$

where $\otimes$ denotes element-wise multiplication.

Object-Contextual Representations. We adopt the OCR scheme to build the context representation of different seismic facies categories.

First, we compute the k soft category regions $\{R_1, R_2, \cdots, R_k\}$ from the output of the encoder, and then represents each category region as m_k:

$$m_k = \sum_{i\in I} \tilde{r}_{ki} x_i, \tag{9}$$

where x_i is the representation of pixel p_i, and $\tilde{r}_{ki}$ computed by spatial softmax is the normalized degree for pixel p_i belonging to each category region R_k.

Next, we calculate the relation between each pixel and each category region as follows:

$$w_{ik} = \frac{e^{\kappa(x_i, m_k)}}{\sum_{j=1}^{k} e^{\kappa(x_i, m_j)}}, \tag{10}$$

where $\kappa(x, m) = \phi(x)^\top \psi(m)$ denotes the unnormalized relation function, $\phi(\cdot)$ and $\psi(\cdot)$ are two transformation functions implemented by 1×1 Conv $\rightarrow$ BN $\rightarrow$ ReLU.

Then, we augment the representation for each pixel by considering its relations with all the category regions:

$$y_i = \rho(\sum_{k=1}^{K} w_{ik}\delta(m_k)), \tag{11}$$

the $\rho(\cdot)$ and $\delta(\cdot)$ are two transformation functions implemented by 1×1 Conv $\rightarrow$ BN $\rightarrow$ ReLU.

Finally, we get the augmented representation of each pixel p_i:

$$\mathbf{z}_i = g\left(\left[\mathbf{x}_i^\top \mathbf{y}_i^\top\right]^\top\right), \tag{12}$$

where x_i is the original representation, and y_i is the category contextual representation. $g(\cdot)$ is a transform function used to fuse x_i and y_i, implemented by 1×1 Conv $\rightarrow$ BN $\rightarrow$ ReLU.

IV. Experiments

A. Datasets and Metrics

Due to the confidentiality of the oil and gas industry hindering the sharing of datasets, high-quality publicly available annotated datasets for seismic interpretation are very precious. After dGB Earth Sciences made the data public, the F3 block[41] became one of the most widely known and studied seismic surveys. In 2019, Alaudah et al.[9] released a fully-annotated 3D geological model of the F3 Block and published a pixel-level annotated seismic facies classification dataset consisting of six different classes, divided into a training set and two test sets. It is worth mentioning that there is a serious data imbalance problem in this dataset, Fig.5 shows the percentage of pixels from different classes in the training set, Table II shows the details of training set and test set. We adopt the following evaluation metrics: pixel accuracy (PA), class accuracy (CA) for each individual class, mean class accuracy (MCA) for all classes, mean intersection over union (MIOU) and frequency-weighted intersection over union (FWIU).

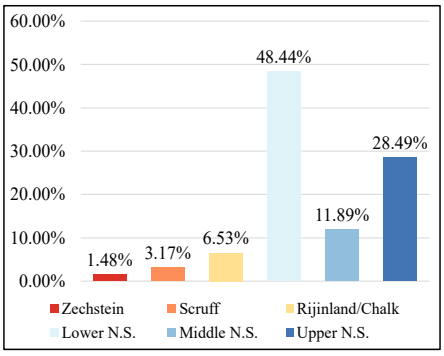


Fig. 5. The percentage of pixels from different classes in the training set.

TABLE II
THE DETAILS OF TRAINING SET AND TEST SET.

Datasets	Inlines	Quantity	Crosslines	Quantity	Total quantity
Train	[300, 700]	401	[300, 1000]	701	1102
Test1	[100, 299]	200	[300, 1000]	701	901
Test2	[100, 700]	601	[1001, 1200]	200	801

B. Implementation Details

In the process of data sampling, we detect 256 corners and we adopt $L = 3$. We use Hrnetv2-W32 proposed in [42] as our default backbone. The model parameters pre-trained on the ImageNet[43] dataset are used to initialize our model. We train our model with the Adam[44] optimizer for 61 epochs optimizing the cross-entropy loss with the learning rate of 0.001. We use standard data augmentation techniques including randomly rotating the patch or the section by up to $\pm 10°$, adding random Gaussian noise, and randomly flipping the sections horizontally. In the training stage, the input image includes two sizes of 255×701 and 255×401, which represent the inline image and crossline image. During the test, we have 255×701, 255×601 and 255×200 input images. We implement all the experiments in PyTorch[45] on a single NVIDIA TITAN XP GPU with 12 GB memory.

C. Data Sampling

We sample $1/2$, $1/3$, and $1/4$ of the training data using different methods, including naive selection such as random sampling and interval sampling, domain expert selection and feature based methods. Among the feature based methods, we compare the Principal Component Analysis (PCA) based sampling with our proposed SPS method. The experimental results are shown in Table III.The random sampling has a certain probability to select representative samples to enhance the performance(e.g. Random(1/2) have 0.8%, 1.2%, and 2.1% improvements on PA, MIOU, and MCA). However, with the decrease of sampling data, the probability is getting lower, so the performance of the model is worse than the baseline. The interval sampling is broadly used in previous works[2, 10]. Similar to random sampling, with the decrease of training data the performance of the model drops sharply, especially in small categories(e.g. the accuracy of Interval(1/3) and Interval(1/4) in Scruff group is only 55.1% and 36.7%). In the feature based methods, even if only $1/4$ samples are used, the PA, MIOU, and MCA of SPS method is 1.2%, 1.7%, and 3.6% higher than the baseline. We consider the reason for the improvement of model performance is the reduction of feature redundancy. Specifically, SPS averages the contributions of discriminative patterns from all categories instead of being inclined to some dominating samples.

On the other hand PCA sampling is inferior to PCA+ sampling, while SPS is slightly superior to PCA+. This demonstrates that spatial pyramid method is a more important step than Harris corner detection, but Harris corner detecter has better visual interpretability than PCA. The expert selection is slightly better than SPS method, but the performance gap between them is relatively small. The biggest disadvantage of expert selection is that domain experts are required to scan all the data to select the most representative samples, which is time-consuming and labor-consuming.

Generally, compared with other sampling methods, the training set with SPS outperforms the baseline significantly, demonstrating that the SPS is competent to select representative samples effectively and is more suitable for real-world scenarios.

D. Quantitative Results

Table IV summarizes the results under different settings of our proposed methods. Compared with the baseline model, the performance of DanetFCN3 on the test set is extremely limited especially on Zechsteins and Scruff groups, with accuracies of only 36.1% and 50.4%. The baseline with M2F and SPS($1/2$) shows a clear advantage compared to the pure baseline. M2F helps the baseline to gain a better feature representation, with the PA and MIOU improving 2% and 4.6% respectively, and the class accuracy score for the Zechstein and Scruff groups increasing by 17% and 6.4%. Furthermore, the SPS method is used to reduce the training data, which obtains outstanding performance. The pure HRNetV2-W2 maintains high-resolution feature expression in the whole process, which helps the model improve the overall performance compared with baseline. Based on HRNetV2-W32, our M2F module and L2G module can improve the performance of the model respectively. We achieve higher performance by combining all the components, which achieves 2.9%, 6.8%, and 4.9% improvements on PA, MIOU and MCA compared to the baseline. For smaller classes such as the Zechstein, Scruf, and Rijinland/Chalk groups, our model shows strong representation ability, which improves the class accuracy scores by 12.4%, 6.0%, and 7.3%, respectively.

E. Visualization and Analysis

We illustrate some examples in Fig.6 to show the advancement of our method visually. It can be clearly seen that the prediction results of the HRNet-based models have fewer isolated regions, which are obviously reflected in 1^{st} , 3^{rd} and 6^{th} rows, due to the high-resolution feature expression in the whole process. On the other hand, we find that the model cooperated with the M2F module can keep more details(e.g. the 3^{rd} , 5^{th} and 6^{th} rows) and achieve better performence in small categories such as Zechstein and Scruff groups, which is due to the fusion of multi-modal knowledge by M2F module. Indeed, our model achieves the best predicted results, which is substantially improved compared with baseline. However, due to the large domain deviation, there are still some errors in the segmentation results. For instance, in the 1^{st} row there are some chaotic regions, in the 4^{th} row, the Rijnland/Chalk group in the lower-right corner is recognized, and in the 5^{th} row, the Lower N.S. and Rijnland/Chalk groups are confused in the lower-left corner.

F. Ablation Studies

In this section, we present a series of ablation studies on the F3 test set to demonstrate the influence of each component.

Hrnet vs. Others. We evaluate four segmentation networks including Unet[16], PSPNet[25], FPN[47], and HRNet[27]. And we select VGG16, VGG19, ResNet-34, ResNet-50, HRNetV2-W18, HRNetV2-W32, and HRNetV2-W48 as the backbone to strike a good balance between performance and memory consumption.

The results are shown in Table V. It can be seen that HRNet-based models obtain notably higher MIOU and MCA than others. Due to the class imbalance problem led to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3171694, IEEE Transactions on Geoscience and Remote Sensing

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT SAMPLING METHODS ON TOP OF THE BASELINE. THE PCA REPRESENTS USING ONLY THE PCA TECHNIQUE, AND PCA+ STANDS FOR USING PCA COMBINED WITH SPATIAL PYRAMID METHOD.

Method	PA	MIOU	FWIOU	MCA	CA					
					Zechstein	Scruff	Rijinland/Chalk	Lower N.S.	Middle N.S.	Upper N.S.
Baseline[9]	0.905	0.707	0.832	0.817	0.602	0.674	0.772	0.941	0.938	0.974
Naive Selection										
Random(1/2)	0.913	0.719	0.849	0.838	0.727	0.618	0.816	0.959	0.935	0.973
Random(1/3)	0.910	0.694	0.847	0.809	0.595	0.585	0.791	0.969	**0.949**	0.965
Random(1/4)	0.901	0.681	0.834	0.816	0.752	0.509	0.792	0.963	0.928	0.973
Interval(1/2)	0.909	0.715	0.838	0.823	0.621	0.654	0.804	0.949	0.939	0.973
Interval(1/3)	0.912	0.698	0.849	0.796	0.527	0.551	0.820	**0.982**	0.930	0.968
Interval(1/4)	0.883	0.635	0.816	0.784	0.716	0.367	0.777	0.957	0.924	0.966
Feature Based Selection										
PCA(1/2)	0.906	0.713	0.835	0.821	0.649	0.653	0.775	0.948	0.929	0.974
PCA(1/3)	0.901	0.698	0.830	0.823	0.678	0.611	0.815	0.941	0.914	0.977
PCA(1/4)	0.897	0.658	0.833	0.793	0.623	0.437	0.832	0.967	0.934	0.966
PCA+(1/2)	0.923	0.738	0.865	0.850	0.742	0.688	0.798	0.968	0.940	0.966
PCA+(1/3)	0.915	0.706	0.854	0.817	0.628	0.568	0.826	0.979	0.933	0.968
PCA+(1/4)	0.913	0.721	0.849	0.834	0.676	0.676	0.784	0.953	0.944	0.970
SPS(1/2)	0.923	0.738	0.866	0.851	0.735	**0.721**	0.783	0.963	0.928	0.973
SPS(1/3)	0.918	0.733	0.857	0.850	0.732	0.693	0.811	0.954	0.934	0.975
SPS(1/4)	0.917	0.724	0.860	0.853	**0.797**	0.680	0.794	0.957	0.916	0.974
Expert Selection										
Expert(1/2)	**0.925**	**0.744**	**0.867**	0.843	0.697	0.646	0.830	0.978	0.937	0.968
Expert(1/3)	0.918	0.739	0.857	**0.856**	0.756	0.715	0.805	0.949	0.929	**0.979**
Expert(1/4)	0.918	0.720	0.862	0.838	0.708	0.624	**0.835**	0.971	0.929	0.962

TABLE IV
QUANTITATIVE RESULTS OF OUR PROPOSED METHOD.

Backbone	Method	PA	MIOU	FWIOU	MCA	CA					
						Zechstein	Scruff	Rijinland/Chalk	Lower N.S.	Middle N.S.	Upper N.S.
Danet[2](Our impl.)	DanetFCN3[2]	0.882	0.628	0.799	0.730	0.361	0.504	0.800	0.967	0.765	**0.983**
Baseline[9]	-	0.905	0.707	0.832	0.817	0.602	0.674	0.772	0.941	0.938	0.974
	M2F	0.925	0.753	0.867	0.859	**0.772**	0.738	0.775	0.965	0.932	0.970
	M2F+SPS(1/2)	0.929	0.755	0.874	0.853	0.715	**0.766**	0.768	**0.972**	0.924	0.971
HRNetV2-W32[46]	-	0.924	0.736	0.867	0.833	0.638	0.691	0.789	**0.972**	0.927	0.982
	M2F	0.929	0.754	0.875	0.852	0.714	0.734	0.781	**0.972**	0.937	0.973
	L2G	0.929	0.760	0.875	0.852	0.688	0.757	0.802	0.968	0.924	0.975
	M2F+L2G	0.932	0.768	0.878	0.858	0.711	0.747	0.815	0.971	0.926	0.979
	M2F+L2G+SPS(1/2)	**0.934**	**0.775**	**0.882**	**0.866**	0.726	0.734	**0.845**	0.971	**0.939**	0.979

poor performance in small categories, we are committed to improving the performance among them. It is worth noting that the class accuracy of the Zechstein and Scruff groups of HRNet-based models is also significantly higher than others. Generally, HRNet-based models achieve the best score relying on its high-resolution representation in the whole process to keep more spatial and contextual information. Therefore, we finally choose HRNetV2-W32 as the backbone network.

Effect of the M2F Module. M2F module is used to aggregate edge and frequency information to generate multimodal representations. We compare with the baseline[9] to illustrate the effectiveness of the proposed M2F module and the results as shown in Table VI. Compared with the baseline, our EDGE module and FDM module have 1.2%, and 4.2% improvements on MIOU, which demonstrates the edge and frequency information are beneficial for our model. We further combine them together, which achieves 4.6% improvements on MIOU compared to the baseline, showing that these two modules are complementary. The class accuracy scores for the Zechstein and Scruff groups increase by 17% and 6.4%. As well, we evaluated the effectiveness of the M2F Module on HRNetV2-W32. We can see that our M2F module can also improve the overall performance while balance the accuracy among all the categories.

Effect of the L2G Moduel. We adopte the L2G module to guide the network to find out 'where' and 'what' local respondence is significant, and then enhance the global context representation of each pixel. We conducted ablation studies based on HRNetV2-W32, the results are shown in Table VII. Firstly, based on M2F, we added the OCR scheme, compared with pure HRNetV2-W32, PA improves from 92.4% to 93.0%, MIOU improves from 73.6% to 75.4% and MCA improves from 83.3% to 84.6%. Moreover, we add CBAM and OCR together, we get 93.2% PA, and we have 3.2%, 1.1%, and 2.5% improvements on MIOU, FWIOU, and MCA, respectively. Generally, CBAM and OCR are compatible and mutually beneficial. For small classes such as the Scruf, and Rijinland/Chalk groups, L2G moduel brings stronger representation ability compare to M2F, which improves the class accuracy scores by 1.3% and 3.4%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3171694, IEEE Transactions on Geoscience and Remote Sensing

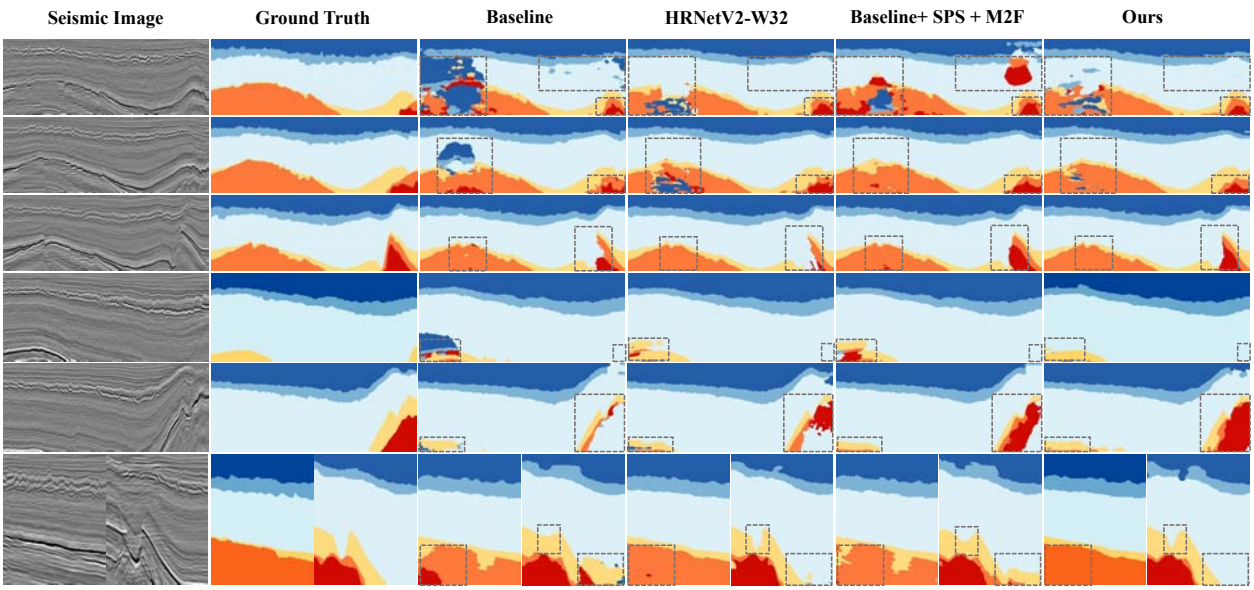


Fig. 6. Qualitative results comparison of different models. The example images are random selected from F3 test set. The first four rows are inline samples and the last row are two samples from crossline. Our method alleviates the inner blur problem and fixes missing details as shown in gray boxes.

TABLE V
COMPARISON WITH DIFFERENT MODELS.

Method	Backbone	PA	MIOU	FWIOU	MCA	CA					
						Zechstein	Scruff	Rijnland/Chalk	Lower N.S.	Middle N.S.	Upper N.S.
Unet[16]	VGG16[48]	0.880	0.649	0.788	0.746	0.404	0.493	0.819	0.944	0.832	0.985
	VGG19[48]	0.890	0.648	0.806	0.747	0.397	0.445	0.822	0.967	0.873	0.980
	ResNet-34[49]	0.914	0.694	0.852	0.795	0.472	0.591	0.840	**0.977**	0.920	0.967
	ResNet-50[49]	0.889	0.665	0.807	0.775	0.509	0.511	**0.842**	0.948	0.853	0.984
PSPNet[25]	VGG16[48]	0.876	0.608	0.783	0.703	0.171	0.466	0.774	0.952	0.879	0.975
	VGG19[48]	0.881	0.615	0.788	0.704	0.165	0.452	0.796	0.961	0.871	0.980
	ResNet-34[49]	0.825	0.555	0.711	0.667	0.164	0.482	0.774	0.874	0.719	**0.990**
	ResNet-50[49]	0.850	0.554	0.736	0.649	0.130	0.36	0.776	0.96	0.705	0.965
FPN[47]	VGG16[48]	0.896	0.669	0.814	0.756	0.346	0.530	0.821	0.963	0.903	0.971
	VGG19[48]	0.858	0.608	0.752	0.711	0.300	0.430	0.768	0.921	0.870	0.975
	ResNet-34[49]	0.914	0.702	0.849	0.787	0.393	0.687	0.807	0.973	0.890	0.974
	ResNet-50[49]	0.882	0.648	0.794	0.753	0.384	0.492	0.838	0.941	0.890	0.974
HRNet[27]	HRNetV2-W18[46]	0.918	0.725	0.856	0.832	0.605	0.686	0.780	0.964	0.926	0.979
	HRNetV2-W32 [46]	**0.924**	0.736	**0.867**	**0.833**	**0.638**	0.691	0.789	0.972	0.927	0.982
	HRNetV2-W48 [46]	0.921	**0.737**	0.860	0.832	0.625	**0.714**	0.783	0.964	**0.931**	0.978

V. CONCLUSION

In this paper, we study the characteristics of seismic images and point out a series of challenges. To cope with these challenges, we first propose a simple yet effecient unsupervised method named SPS to select representative samples to reduce training data. Furthermore, we explore the multi-modal representation of seismic images and propose the M2F module to fuse the knowledge of edges and frequencies. Finally, we propose the L2G module which integrates the local information and global context information to get the salient features. The results show that our model outperforms the baseline and improves the accuracy of small categories substantially with less training data. In the future, we hope to discover more effective weakly supervised or unsupervised methods to further reduce the labeling costs while maintaining high performance.

ACKNOWLEDGEMENT

This research is supported by National Natural Science Foundation of China (No.62106017).

REFERENCES

[1] K. O. Olowoyo, *Structural and Seismic Facies Interpretation of Fabi Field, Onshore Niger Delta, Nigeria*. Universal-Publishers, 2010.

[2] D. Civitarese, D. Szwarcman, E. V. Brazil, and B. Zadrozny, "Semantic segmentation of seismic images," *arXiv preprint arXiv:1905.04307*, 2019.

[3] T. Zhao, V. Jayaram, A. Roy, and K. Marfurt, "A comparison of classification techniques for seismic facies recognition," *Interpretation*, vol. 3, pp. SAE29–SAE58, 11 2015.

[4] A. Amendola, G. Gabbriellini, P. Dell'Aversana, and A. Marini, "Seismic facies analysis through musical attributes," *Geophysical Prospecting*, vol. 65, 03 2017.

TABLE VI
INFLUENCE OF M2F. BOTH EDGE MODULE AND FDM MODULE IMPROVE THE MODEL PERFORMANCE AND COMPLEMENT EACH OTHER.

Backbone	Method	PA	MIOU	FWIOU	MCA	CA					
						Zechstein	Scruff	Rijinland/Chalk	Lower N.S.	Middle N.S.	Upper N.S.
Baseline[9]	-	0.905	0.707	0.832	0.817	0.602	0.674	0.772	0.941	**0.938**	0.974
	EDGE	0.916	0.719	0.854	0.833	0.641	0.697	0.797	0.957	**0.938**	0.967
	FDM	0.923	0.749	0.862	0.851	0.710	0.710	**0.820**	0.962	0.930	0.973
	M2F	0.925	0.753	0.867	**0.859**	**0.772**	0.738	0.775	0.965	0.932	0.970
HRNetV2-W32[46]	-	0.924	0.736	0.867	0.833	0.638	0.691	0.789	**0.972**	0.927	**0.982**
	EDGE	0.927	0.750	0.871	0.837	0.603	0.727	0.812	0.971	0.933	0.975
	FDM	0.926	0.750	0.870	0.848	0.685	**0.742**	0.781	0.966	**0.938**	0.975
	M2F	**0.929**	**0.754**	**0.875**	0.852	0.714	0.734	0.781	**0.972**	0.937	0.973

TABLE VII
INFLUENCE OF L2G.

Backbone	Method	PA	MIOU	FWIOU	MCA	CA					
						Zechstein	Scruff	Rijinland/Chalk	Lower N.S.	Middle N.S.	Upper N.S.
HRNetV2-W32[46]	-	0.924	0.736	0.867	0.833	0.638	0.691	0.789	**0.972**	0.927	**0.982**
	M2F	0.929	0.754	0.875	0.852	**0.714**	0.734	0.781	**0.972**	**0.937**	0.973
	M2F+OCR	0.930	0.754	0.877	0.846	0.635	**0.759**	0.805	**0.972**	0.929	0.973
	M2F+L2G	**0.932**	**0.768**	**0.878**	**0.858**	0.711	0.747	**0.815**	0.971	0.926	0.979

[5] T. Wrona, I. Pan, R. Gawthorpe, and H. Fossen, "Seismic facies analysis using machine learning," *GEOPHYSICS*, vol. 83, pp. 1–34, 06 2018.

[6] D. S. Chevitarese, D. Szwarcman, E. V. Brazil, and B. Zadrozny, "Efficient classification of seismic textures," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[7] J. S. Dramsch and M. Lüthje, "Deep-learning seismic facies on state-of-the-art cnn architectures," in *Seg technical program expanded abstracts 2018*. Society of Exploration Geophysicists, 2018, pp. 2036–2040.

[8] T. Zhao, "Seismic facies classification using different deep convolutional neural networks," in *SEG Technical Program Expanded Abstracts 2018*. Society of Exploration Geophysicists, 2018, pp. 2046–2050.

[9] Y. Alaudah, P. Michałowicz, M. Alfarraj, and G. AlRegib, "A machine-learning benchmark for facies classification," *Interpretation*, vol. 7, no. 3, pp. SE175–SE187, 2019.

[10] F. Li, H. Zhou, Z. Wang, and X. Wu, "Addcnn: An attention-based deep dilated convolutional neural network for seismic facies analysis with interpretable spatial-spectral maps," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[11] K. G. Derpanis, "The harris corner detector," *York University*, vol. 2, 2004.

[12] Y. Wang, Y. Cao, Z. J. Zha, J. Zhang, and Z. Xiong, "Deep degradation prior for low-quality image classification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[14] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[18] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[19] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

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

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[26] 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," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[27] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[30] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[31] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3024–3033.

[32] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding, "Acfnet: Attentional class feature network for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6798–6807.

[33] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 548–557.

[34] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[35] K. Yang, J. Zhang, S. Reiß, X. Hu, and R. Stiefelhagen, "Capturing omni-range context for omnidirectional segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1376–1386.

[36] C. Wang, Y. Zhang, M. Cui, J. Liu, P. Ren, Y. Yang, X. Xie, X. Hua, H. Bao, and W. Xu, "Active boundary loss for semantic segmentation," *arXiv preprint arXiv:2102.02696*, 2021.

[37] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.

[38] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 435–452.

[39] M. Q. Nasim, T. Maiti, A. Shrivastava, T. Singh, and J. Mei, "Seismic facies analysis: A deep domain adaptation approach," *arXiv preprint arXiv:2011.10510*, 2020.

[40] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006.

[41] R. M. Silva, L. Baroni, R. S. Ferreira, D. Civitarese, D. Szwarcman, and E. V. Brazil, "Netherlands dataset: A new public dataset for machine learning in seismic interpretation," *arXiv preprint arXiv:1904.00770*, 2019.

[42] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

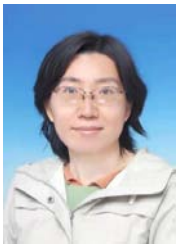
[46] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019.

[47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Xiaoyu Chen received the B.S. degree in Geophysics from China University of Mining and Technology, Beijing, China, in 2020. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her research interests include computer vision, image processing, and machine learning with the applications on seismic data analysis.



Qi Zou received the Ph.D. degree in computer science from Beijing Jiao Tong University, Beijing, China, in 2006. In 2014, she was a Visiting Researcher with the Department of Computer and Information Science, Temple university, Philadelphia, USA. She is currently a Professor and a Doctoral Supervisor in School of Computer and Information Technology, Beijing Jiao Tong University. Her research interests include computer vision and intelligent transportation systems. She has published over 30 papers in peer-reviewed journals including IEEE Trans. PAMI, TMM, TITS, TCSVT and conferences like CVPR, AAAI, IJCAI, ECCV, ACM MM.

Xixia Xu received the B.S. degree in software engineering from the Lanzhou Jiao tong University, Lan Zhou, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiao tong University, Beijing, China, in 2019. Her current research interests include computer vision, image processing, and machine learning with the applications on 2d multi-person pose estimation analysis and human centric behavior analysis

Nan Wang was born in 1989. He is currently pursuing the Ph.D. degree in computer science and technology with the School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include machine learning, computer vision and object tracking.