Towards deep learning for seismic demultiple

Mario R Fernandez, Norman Ettrich, Matthias Delescluse, Alain Rabaute, Janis Keuper

Received: 26 October 2023 | Accepted: 27 January 2025

DOI: 10.1111/1365-2478.13672

Geophysical Prospecting | EAGE EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS

ORIGINAL ARTICLE

Towards deep learning for seismic demultiple

Mario R. Fernandez[1,2] | **Norman Ettrich**[1] | **Matthias Delescluse**[2] | **Alain Rabaute**[3] | **Janis Keuper**[1,4]

[1]Competence Center High Performance Computing, Fraunhofer ITWM, Kaiserslautern, Rheinland Pfalz, Germany

[2]Laboratoire de géologie - CNRS UMR 8538, École Normale Supérieure, PSL University, Paris, Île-de-France, France

[3]Institut des Sciences de la Terre de Paris, Sorbonne Université, CNRS-INSU, Paris, Île-de-France, France

[4]Institute for Machine Learning and Analytics, Offenburg University, Offenburg, Germany

Correspondence
Mario R. Fernandez, Competence Center High Performance Computing, Fraunhofer ITWM, Kaiserslautern, Rheinland Pfalz, 67663, Germany. Email: mario.ruben.fernandez@itwm.fraunhofer.de

Abstract

Multiple attenuation is an important step in seismic data processing, leading to improved imaging and interpretation. Radon-based algorithms are commonly used for discriminating primaries and multiples in common depth point seismic gathers. This process implies a large number of parameters that need to be optimized for a satisfactory result. Moreover, Radon-based approaches sometimes present challenges in discriminating primaries and multiples with similar moveouts. Deep learning, based on convolutional neural networks, has recently shown promising results in seismic processing tasks that could mitigate the challenges of conventional methods. In this work, we detail how to train convolutional neural networks with only synthetic seismic data for assessing the demultiple problem in field datasets. We compare different training strategies for multiples removal based on different loss functions. We evaluate the performance of the different strategies on 400 clean and noisy synthetic data. We found that training a convolutional neural network to predict the multiples and then subtracting them from the input image is the most effective strategy for demultiple, especially for noisy data. Finally, we test our model to predict multiples on an elastic synthetic dataset and four distinctive field datasets. Our proposed approach reports successful generalization capabilities predicting and eliminating internal and surface-related multiples before and after migration while mitigating Radon challenges and relieving the user from any manual tasks. As a result, our effectively trained models bring a new valuable tool for seismic demultiple to consider in existing processing workflows.

KEYWORDS
data processing, deep learning, demultiple, multiples, noise, seismic

INTRODUCTION

In conventional seismic processing, only primary reflections are used for imaging the Earth's interior. Multiple reflections are therefore considered coherent noise. The presence of multiple reflections can severely impact the accuracy and reliability of seismic interpretation, making demultiple an essential process in various applications, including sedimentary and tectonic interpretation, hydrocarbon exploration and reservoir characterization (Oz, 2001).

Multiple attenuation methods are traditionally divided into two main categories (Vershuur, 2013): (1) methods based on periodicity and predictability and (2) separation based on transform domains. Predictability-based methods exploit the repetitive nature of multiples and their connection to primaries. These methods typically involve a

prediction–subtraction workflow, where predicted multiples are subtracted from the data. Some widely used techniques include the wavefield extrapolation method (Wiggins, 1998, 1999) and surface-related multiple elimination (SRME) (Vershuur, 2013; Verschuur et al., 1999). These methods are known for effectively suppressing water bottom multiples. However, they are computationally demanding, require dense seismic acquisition, need quality near-offset data or must rely on data interpolation techniques (Vershuur, 2013). On the other hand, separation-based methods transform seismic data into intermediate domains that allow better discrimination of multiples and primaries than in the regular time-offset domain, enabling the selective elimination of multiples. One widely used approach is the Radon transform (RT), which maps common depth point gathers (CDP) from time-offset to the Radon domain where seismic events with different moveouts can be discriminated and then back-transformed (Beylkin, 1987; Hampson, 1986). A general challenge of this method is that the RT operator is not orthogonal; therefore, the forward and inverse mapping without data loss is not trivial (Trad et al., 2003). Parabolic RT addresses this problem by fitting the data with a combination of parabolic events, where the Radon domain is optimized to reconstruct the original data in the least square sense (Hampson, 1986). This method, however, presents challenges when discriminating events with similar moveouts due to low resolution in the Radon domain. RT discrimination also may introduce artefacts due to possible sharp cutoffs in the transform domain. Several variants of the RT have been proposed to overcome these challenges. High-resolution RT is achieved by stochastic inversion in the time domain (Thorson & Claerbout, 1985), or the frequency domain by sparse inversion (Sacchi & Ulrych, 1995) or mixed time-frequency methods (Lu, 2013; Trad et al., 2003). Nonetheless, its application is difficult due to the sensitivity of the hyperparameter search needed for optimal results (Trad et al., 2003). Additionally, parabolic RT assumes that the seismic events follow an idealized parabolic pattern and do not account for deviations from it, often present in field data. Analogous problems are encountered when linear (Abbasi & Jaiswal, 2013; Vershuur, 2013; Taner, 1980) or hyperbolic RT (Vershuur, 2013; Foster & Mosher, 1992) are used. Another practical limitation of RT methods is that the cutoff function that discriminates multiples and primaries in the Radon domain is usually calculated by one reference CDP. This function might not be optimal across all CDP gathers, requiring a mute definition in several reference CDPs with interpolation between them.

Machine learning (ML) methods are data-driven methods that learn to solve a task by extracting information from data without being explicitly programmed (Goodfellow et al., 2016a; LeCun et al., 2015). Deep learning (DL), based on convolutional neural networks (CNNs), is an ML method that uses multiple layers to extract progressively abstract features of the data (Goodfellow et al., 2016a; LeCun et al., 2015). DL has recently obtained successful results in several computer vision tasks such as image restoration (Isola et al., 2016; Ulyanov et al., 2017), image generation (Goodfellow et al., 2016b), classification (Krizhevsky et al., 2012; LeCun et al., 1998), among others. The success in this field and their similarity to some of the seismic challenges have created a new venue of research for applying DL to seismic problems (Mousavi et al., 2024; Harsuko & Alkhalifah, 2022). In particular, DL approaches have shown an alternative solution for seismic processing tasks such as interpolation (Chai et al., 2020; Fernandez et al., 2022c, 2022b; Mandelli et al., 2019), denoising (Birnie et al., 2021; Qiu et al., 2022; Zhang et al., 2019), deblending (Shirui et al., 2021), ground-roll attenuation (Guo et al., 2020; Kaur, Fomel, et al., 2020), velocity picking (Wang et al., 2021), dispersion curve estimation (Chamorro et al., 2023), among others. For the demultiple problem, supervised DL has shown alternative solutions to the conventional approaches, mitigating challenges in traditional methods. Considering shot/receiver seismic gathers, a DL model has been trained for performing the adaptive subtraction of surface-related multiples required in the SRME method (Li & Gao, 2020; Zhang et al., 2021) or to reconstruct near-offset data needed to improve SRME performance (Qu et al., 2021). Also, a DL network has been proposed to approximate the estimation of primaries by a sparse inversion algorithm to eliminate surface-related multiples (Siahkoohi et al., 2019). Alternative solutions with DL methods were also proposed for demultiple based on move-out discrimination in CDP seismic gathers. A CNN has been trained to compute the hyperbolic Radon transform for separating multiples and primaries in the CDP (Kaur, Pham, et al., 2020). Also, CNNs have been trained to predict primaries for a given input CDP with primaries and residual multiples in post-migrated seismic data (Breuer et al., 2020; Bugge et al., 2021; Durall et al., 2022; Fernandez et al., 2023a). Additionally, other methods for training CNNs were applied for demultiple in post-migrated CDP gathers, such as generative adversarial networks (Fernandez et al., 2023b), and diffusion models (Durall et al., 2023). The main advantage of DL methods for multiple attenuation is that parameters of the CNNs are trained once during the training phase, resulting in a parameter-free user-friendly approach during inference time, reducing computational time and time-consuming manual tasks, that is, picking the mute function in the Radon domain and searching for optimal parameters. However, to the best of our knowledge, previous studies of DL for seismic demultiple on CDP gathers only consider the case of post-migrated data, where some demultiple processing was already applied. Also, previous studies consider the prediction of primaries but do not investigate multiple prediction or joint multiple and primary prediction.

In this paper, we tackle the seismic demultiple problem in normal moveout corrected CDP gathers by leveraging a supervised DL model. Our goal is to obtain a DL model, allowing direct application of surface-related and internal multiple attenuation to many different field datasets. We first describe the process of constructing a synthetic dataset tailored for the seismic demultiple task of a wide range of field applications. We then perform an extensive evaluation of various training strategies, emphasizing the significant impact of different loss functions on model performance, in particular for different kinds of noisy data. Using our preferred learning strategy, we show that a single U-Net network, trained in a supervised manner with only synthetic data, can generalize across diverse field datasets, effectively predicting and eliminating surface-related and internal multiples, both pre- and post-migration. Additionally, we compare the efficiency of our U-Net model with traditional seismic demultiple-processing workflows, including SRME, parabolic Radon transform, and linear Radon transform. Lastly, we discuss the advantages and limitations of employing this DL approach for seismic demultiple, providing a comprehensive analysis of its potential in practical applications.

METHODOLOGY

In this section, we describe our seismic demultiple methodology with supervised deep learning. We first detail the process of building synthetic training data for seismic demultiple based on moveout discrimination in normal moveout (NMO)-corrected common depth point (CDP) gathers. Then, we describe different strategies for multiple attenuation based on different training loss functions. Finally, we discuss the choice of our neural network architecture.

Training data

Training data is a fundamental aspect of supervised deep learning applications. In the demultiple problem, we do not have real data labels, that is, field data with merged primaries and multiples and corresponding separated primaries and multiples. Real data labels could be obtained by conventional demultiple methods, for example, Radon. However, this approach is time-consuming, and the performance of the trained network would have the same limitations as the conventional methods used to create the training data. Therefore, we consider only synthetic training data to overcome the lack of real labels. In this case, the challenge is to generate synthetic data that are sufficiently realistic to apply the train networks on field data during inference time.

The training data consists of one hundred thousand synthetic CDP seismic gathers (C), that is, merged primaries and multiples, and the corresponding separated primaries (P), and multiples (M) true labels. We modelled the training data through a convolutional method for a large number of variations (Figure 1). We obtain those variations by scanning a high-dimensional parameter space to include possible real scenarios of mixed multiple and primary energy. The following workflow details the synthetic data creation:

- The starting point for modelling each synthetic CDP with the convolution model is defining the source time function $S(t)$. We consider the Ricker wavelets and vary peak frequency, phase and the superposition of a pair of Ricker wavelets (Figure 1a).
- Following, the zero offset ($h = 0$) one-dimensional reflectivity function $R(t, h = 0) = R_0(t)$ is constructed as a white reflectivity spectrum, biased towards the sparseness of the strong reflection coefficients (Figure 1b).
- For the extension to non-zero offset, hyperbolic moveout is calculated by defining a time-dependent velocity function $v(t)$. In addition, the reflectivity coefficients are scaled by the Shuey approximation (Shuey, 1985), including amplitude variations with offsets/angles. Moreover, kinematic inaccuracies are simulated by adding time perturbations δ_t to the reflectivity coefficient's time, providing the final two-dimensional (2D) reflectivity $R(R_0(t), v(t), h)$ (Figure 1c).
- We obtain the 2D seismic gathers convolving the source time function $S(t)$ with our 2D reflectivity series $R(t, h)$; moreover, we incorporate a frequency decay along the time axis to compute the actual wavelet that is set active at convolution time $t(t_0, h)$. The final CDP gather of the primaries is obtained by simulating far offset amplitude decay for some CDP samples (Figure 1d).
- The synthetic gathers with primaries are corrected with the NMO equation, using the velocity model of the primaries, perturbed by inaccuracies that simulate imperfections of the process of velocity picking performed at each CDP. This step mimics two important characteristics observed in real data: (1) NMO stretch in primaries and (2) residual moveout effect (RMO) in primaries, not completely flattened due to inaccurate velocity picking. Modelling RMO in primaries allows the networks to avoid eliminating primaries not completely flattened after NMO and includes over-corrected primaries as well (Figure 1e).
- Analogously, we create the NMO-uncorrected CDP gather of the multiples using velocities lower than the ones used for the primaries. The multiples are NMO corrected with the velocity used to flatten the primaries, providing multiples with stronger moveouts than primaries and presenting also stretching at far offsets. Note that we do not pay any attention to physically relating the reflectivity series of the multiples to the reflectivity series of the primaries. Therefore, the network will learn to distinguish between

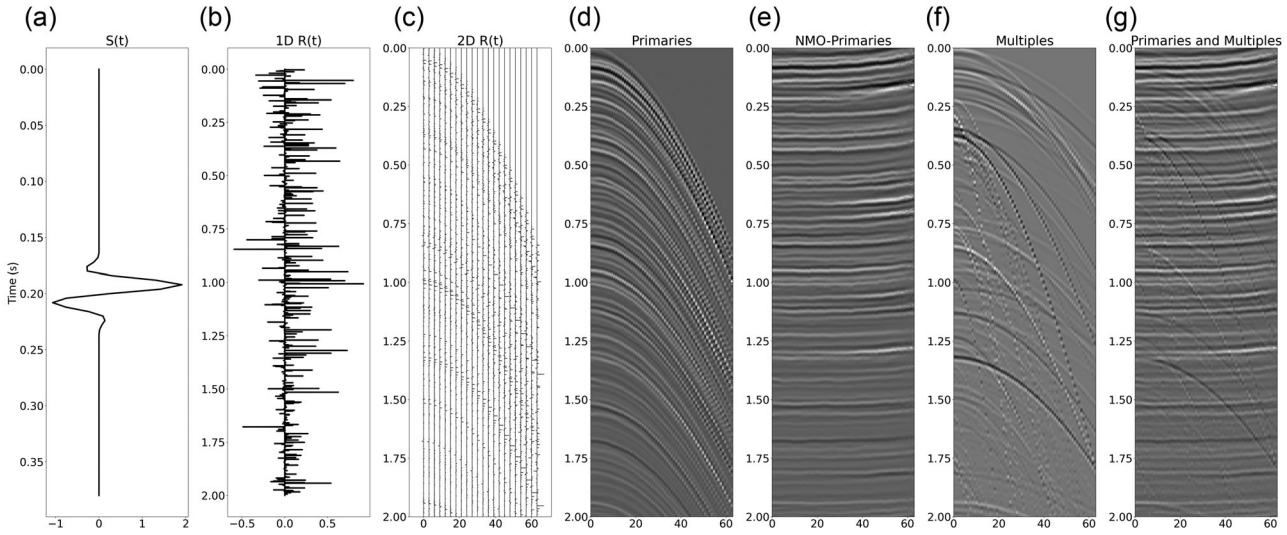

FIGURE 1 Generation of training data. (a) Source time function. (b) 1D reflectivity series. (c) 2D reflectivity series. (d) Synthetic primaries obtained by convolution of the source time function and the 2D reflectivity series. (e) Primaries (P) after NMO correction. (f) Multiples (M) generated following the same workflow as primaries. (g) Mixed primaries and multiples ($C = P + M$).

primaries with little RMO and crossing multiples with larger RMO, while it does not learn to identify periodicity nor relate multiples to primaries generating reflectors (Figure 1f).

- Realistic CDP seismic gathers with the corresponding ground-true labels are generated by mixing the primaries and multiples (Figure 1g).
- Additionally, linear seismic noise is included in the multiple images, considering steep near and far offset events, to asses challenging cases for the Radon transform (RT) method. Finally, the synthetic data are also augmented during training by adding different levels of random uniform noise to further mitigate the difference between synthetic and real data and improve the robustness of our models to this perturbation often present in field data. In particular, we consider random noise from the uniform distribution $X \sim \mathcal{U}(-a, a)$, where the value a is also randomly sampled from a distribution at each epoch to allow for different signal-to-noise rations during training, that is, $a \sim \mathcal{U}(0,0.1)$. Figure 2 shows some examples of the synthetic CDP gathers created with our workflow.

Strategies

As we want to test the significance of predicting primaries, multiples or both simultaneously, we present three supervised strategies to train the convolutional neural network (CNN)'s weights (θ) to solve the seismic demultiple task with synthetic data generated by the workflow described in the previous section. The first strategy trains the weights of a CNN to map seismic data (C) with merged primaries (P) and multiples (M) to the desired image with only primaries ($f_\theta(\boldsymbol{C}) = \boldsymbol{P}^*$) (Figure 3a). This is equivalent to predicting primaries. Multiples are calculated by subtracting from the input image the estimated primaries $\boldsymbol{M}^* = \boldsymbol{C} - \boldsymbol{P}^*$, constraining the training with the information of multiples labels in the second term of Equation (1).

$$\min_\theta\{\alpha_1 L(\boldsymbol{P} - \boldsymbol{P}^*) + \alpha_2 L(\boldsymbol{M} - (\boldsymbol{C} - \boldsymbol{P}^*))\}. \quad (1)$$

The second strategy consists of training a CNN to predict the multiples ($f_\theta(\boldsymbol{C}) = \boldsymbol{M}^*$) for a given input seismic data (C) with combined primaries and multiples (Figure 3b). Primaries are obtained by subtracting from the input image the estimated multiples $\boldsymbol{P}^* = \boldsymbol{C} - \boldsymbol{M}^*$, constraining the training with the information of primaries labels in the second term of Equation (2).

$$\min_\theta\{\alpha_1 L(\boldsymbol{M} - \boldsymbol{M}^*) + \alpha_2 L(\boldsymbol{P} - (\boldsymbol{C} - \boldsymbol{M}^*))\}. \quad (2)$$

Finally, strategy 3 learns how to map an image (C) with joint primaries and multiples to both multiples (M*) and primaries (P*) simultaneously ($f_\theta(\boldsymbol{C}) = (\boldsymbol{P}^*, \boldsymbol{M}^*)$) as a double output channel of the CNN (Equation 3) (Figure 3c). Both outputs are trained simultaneously with the corresponding labels. In this study, we consider the parameters α_i in the losses equal to 1, and the mean absolute error (L = L1).

$$\min_\theta\{\alpha_1 L(\boldsymbol{P} - \boldsymbol{P}^*) + \alpha_2 L(\boldsymbol{M} - \boldsymbol{M}^*)\}. \quad (3)$$

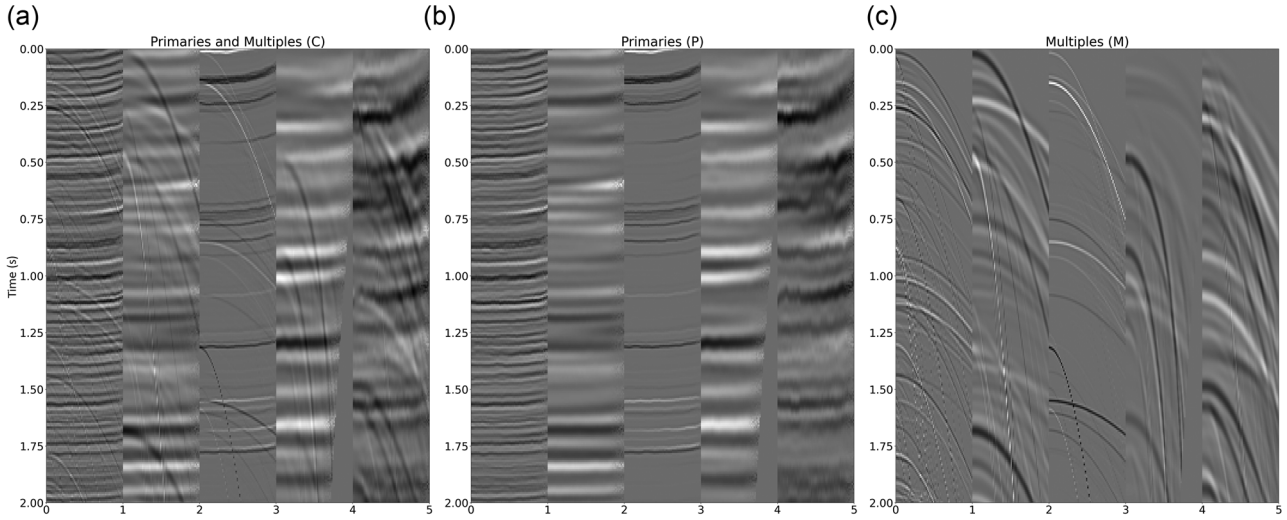

FIGURE 2 Training data examples in the form of five CDP gathers of varied frequency and reflectivity content. (a) Synthetic gathers (labelled C in the text) with mixed primaries and multiples. (b) Primaries (labelled P in the text). (c) Multiples (labelled M in the text).

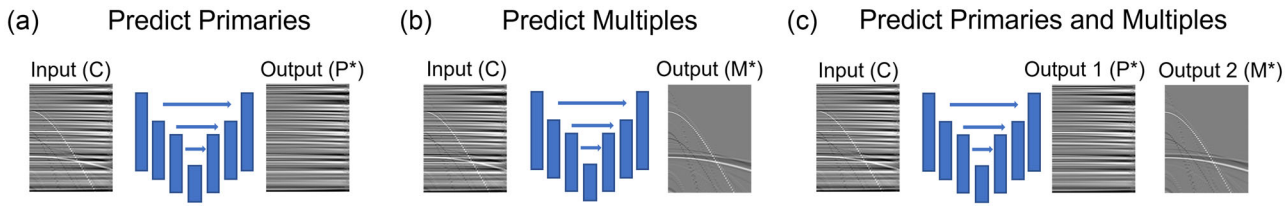

FIGURE 3 Demultiple strategies with CNNs. (a) Predicting primaries (strategies 1). (b) Predicting multiples (strategy 2). (c) Predicting simultaneously both multiples and primaries (strategy 3).

Architecture and hyperparameters

Different deep learning architectures have been proposed to extract information from data according to the different data characteristics and applications. For computer vision tasks, architectures based on CNNs have dominated the field (Long et al., 2015; Isola et al., 2016; Krizhevsky et al., 2012; LeCun et al., 1998; Ulyanov et al., 2017), while for natural language processing transformer-based architectures are the state-of-art choice (Devlin et al., 2019; Vaswani et al., 2023). Applying CNNs or transformer architectures to seismic tasks makes the assumption of handling the seismic data as images or text, with respective advantages and disadvantages.

CNNs are designed to capture local features and build hierarchical representations (Goodfellow et al., 2016a; LeCun et al., 1998; Krizhevsky et al., 2012), which align well with the structure of images. Additionally, the convolutional and pooling operations in CNNs provide a natural way to handle translations within images, making them robust to variations in object positions (Goodfellow et al., 2016a; Gu et al., 2017; Krizhevsky et al., 2012; LeCun et al., 1998), for example, in our case, multiples and primaries. Another advantage is that the built-in biases towards local feature learning and spatial hierarchies make CNNs effective with relatively smaller datasets compared to transformers (Dosovitskiy et al., 2021; Liu et al., 2021). The main disadvantage of CNNs is the difficulty in capturing global context and long-range dependencies within an image, which can be a limitation for certain tasks (Liu et al., 2021). On the other hand, self-attention mechanisms allow transformers to capture global relationships across an image, which is useful for understanding complex scenes (Vaswani et al., 2023). However, the self-attention mechanism has quadratic complexity to the input size, making transformers computationally expensive, especially for high-resolution images (Vaswani et al., 2023). Additionally, capturing global context patterns for data is a desired feature, but it comes with the cost of requiring more training data (Liu et al., 2021).

In this study, we consider that the CNN's advantages of successfully detecting objects in images due to local and hierarchical learning make them robust to successfully detect and predict multiples in seismic gathers. Additionally, the inductive bias towards local feature learning and spatial hierarchies makes CNNs effective with small datasets compared with transformers (Liu et al., 2021). This is indeed an advantage in many tasks, such as demultiple, where real ground-truth

labels are not available, and a synthetic training dataset, ideally small, needs to be created. In particular, we use the U-Net topology (Ronneberger et al., 2015). The U-Net network was developed for the biomedical image segmentation task (Ronneberger et al., 2015). This architecture type as well as variations of it has been broadly used due to its excellent performance in many computer vision tasks and seismic processing tasks (Birnie et al., 2021; Fernandez et al., 2022a; Mandelli et al., 2018; Yang & Ma, 2019). The architecture is notable for its U-shaped structure, which allows it to capture both local and global context in images. This architecture has shown good performance in detecting objects for biomedical images, most likely making it suitable for detecting the shape of multiples in CDP gathers. To focus our investigation on testing the different strategies for seismic demultiple and its generalization capacities for realworld field applications we consider the original U-Net. The architectural characteristics of this network could be fine-tuned for this particular application. However, the huge range of applications where this architecture has successfully been applied justifies our choice to use it as a baseline network to test our different strategies and test its generalization properties across many datasets.

For training the U-Net, we consider the mean absolute loss function (L1). Alternatively, other training methodologies could have been considered, such as generative adversarial networks (Goodfellow et al., 2014) or diffusion models (Ho et al., 2020; Sohn et al., 2015). This last method has shown successful application in eliminating residual multiplies in post-migration datasets (Durall et al., 2023; Fernandez et al., 2023b). However, these more sophisticated methodologies have not outperformed training based on mean absolute error loss function (L1) for eliminating residual multiples. Additionally, diffusion models come with high computation cost, making them less suitable for real seismic processing deployment. In our work, we attempt to further generalize the capacity of our network to consider not only post-migration multiple elimination but also pre-migrated surface-related multiples, while keeping the computational cost as minimal as possible. This makes the U-Net trained with an L1 misfit function suitable for our study.

Finally, our training workflow starts normalizing our data before training or inference with automatic gain control considering a time window of 0.3 s. This helps to mitigate the amplitude differences between real cases and training data. We train the models for 200 epochs, considering a batch size of 64 images. The optimizer is the Adam optimizer (Kingma & Ba, 2017), and the learning rate is 0.001.

RESULTS ON SYNTHETIC DATA

We train convolutional neural networks (CNNs) for each of the three strategies described above. Moreover, we repeat the training of each strategy 10 times. Repeating the training 10 times is beneficial for two main reasons. First, it can avoid local minima. Neural network training involves non-convex optimization, which means the training process can get stuck in local minima (Goodfellow et al., 2016a). Different initialization might lead to different local minima. Re-propagating experiments several times and analysing results from multiple runs can mitigate the impact of any single poor or good initialization. Second, it allows to analyse the variance of the model's predictions related to the training data. Each individual model might capture different aspects of the data, and assessing their predictions can lead to better analyses of the model robustness. We consider the same CNN architecture, training data and workflow for all the strategies. In this section, we first investigate quantitatively the generalization performance of the three CNN models on a test synthetic dataset generated also with the convolutional method, but unseen during training. Second, we further assess the generalization capabilities of our methodology to synthetic data generated with a finite difference wave equation solver, that is., synthetic data generated with a different engine.

Synthetic test dataset

For a quantitative analysis of the performance of convolutional neural networks (CNNs) trained with the three different learning strategies, we consider synthetic data generated using the convolutional method described in the Methodology section, which was not seen during the training phase. In the seismic demultiple problem, we have two labels, that is, primaries and multiples (P and M); however, the accurate estimation of one correlates with the other. The ultimate goal of the demultiple process is to obtain the best estimation of primaries. Therefore, we evaluate the performance of the demultiple process using primary ground-truth synthetic images. Note that in all the cases, the output of the different strategies, that is, primaries, multiples, or both simultaneously, are directly used to estimate the primaries without any further processing.

To quantitatively assess the results, we use the signal-to-noise ratio (SNR) metric between our final estimated primaries (P*) and the ground-truth labels (P). We evaluate the estimations of each of the three strategies on the same 400 synthetic examples not seen during training, considering synthetic data without noise and with different noise perturbations of increasing complexity as a proxy of the application to field datasets. We first evaluate the robustness of the methods by perturbing the test synthetic data with the noise distribution used during training, that is, uniform distribution $X \sim \mathcal{U}(-a, a)$ and considering $a = 0.1$ (uniform low). Then, we test the robustness of the strategies considering a higher value of $a = 0.3$ (uniform high). Note that for the case

TABLE 1 Statistical results (mean SNR) of primaries estimation considering 400 synthetic data for each of the three strategies: strategy 1 (Equation 1), strategy 2 (Equation 2) and strategy 3 (Equation 3) and different noise perturbations (uniform low, uniform high, Gaussian, Laplace and synthetic realistic noise).

	No-noise	Uniform (low)	Uniform (high)	Gaussian	Laplace	Realistic
Strategy 1	23.56	21.00	7.82	7.06	3.99	9.45
Strategy 2	26.13	23.50	20.74	16.85	15.10	12.44
Strategy 3	23.00	21.07	6.51	5.62	3.20	8.65

of uniform noise with $a = 0.3$, we consider the same distribution as the one used for training but allow three times higher noise amplitudes. To further test the generalization capacity of our strategies, we tested the trained models on different noise distributions than those used during training, specifically Gaussian noise $X \sim \mathcal{N}(\mu,\sigma)$ with a mean $\mu = 0$ standard deviation $\sigma = 0.1$, and Laplacian noise $X \sim \mathcal{L}(\mu,b)$ with mean $\mu = 0$ and scale parameter $b = 0.1$. Finally, we consider more realistic noise scenarios. In Appendix A, A we describe how to generate more realistic noise.

Table 1 shows the statistical results (mean SNR) for the 400 synthetic images. Note that higher values of SNR metric indicate better estimations of primaries. Our results show that training a CNN to predict the multiples (M*) from a given seismic dataset (C) and subtracting them is the most effective strategy for multiple removal (Equation 1). This strategy yields better metrics in noise-free examples and demonstrates greater robustness to different noise perturbations. Importantly, this strategy is more resilient to noise perturbations outside of the noise distribution considered during training, that is, uniform high, Gaussian, Laplace and realistic noise. This is a fundamental advantage of strategy 2 compared with the others, considering that in field data applications, the real noise distribution is unknown.

To provide a clearer understanding of the different strategies and their performance, we present one selected image from the 400 test dataset. We first solved the demultiple problem on a noise-free synthetic example using each of the three strategies: predicting primaries, predicting multiples and predicting both simultaneously. While results on noise-free data may not be realistic, they allow us to evaluate potential issues that are unrelated to noise. Figure 4c shows a synthetic example of mapping a seismic image (C) to the predicted primaries (P^*). The estimated multiples (M^*) are obtained by subtracting the primaries from the input image ($M^* = C - P^*$). Primaries estimations report a high SNR value. However, we observe that since the model predicts primaries for a given input, small errors occur in all pixels of the seismic primaries estimation, regardless of the presence of intersecting multiples in those pixels of the input CDP. This is more clearly visible in the multiple image estimated based on the primaries prediction, where leakage of primary energy appears throughout the image ($M^* = C - P^*$) (red arrow in Figure 4g). Figure 4d shows the results of mapping the image (C) to predicted multiples (M^*). The corresponding primaries are found by subtraction ($P^* = C - M^*$). In this case, we observed better performance than in the aforementioned case, with a slightly higher SNR value reported in the primaries estimation. Importantly, errors occur only in the pixels where multiple events are present, avoiding errors in primary estimation areas where no intersecting multiples are present, and thus better preserving the amplitude of the final primaries. Figure 4e shows the results of predicting both multiples (M^*) and primaries (P^*) simultaneously. Primaries and multiples are successfully separated, mitigating the leakage of primary energy into multiples, but the model struggles to fully predict multiples at the near offset (yellow arrow in Figure 4i).

Next, in Figure 5, we visualize the results on the same synthetic data perturbed with random noise (uniform low). Again, strategy 2 provides the best quantitative results, that is, the highest SNR of predicted primaries, as well as the best qualitative performance (Figure 5). Specifically, strategy 2 successfully predicts multiples in the near offsets, while strategies 1 and 3 show difficulties in this area (yellow arrow in Figure 5). Lastly, we show the performance of the three strategies on the same image perturbed with more realistic synthetic noise (Figure 6). Here, once again, strategy 2 provides the most accurate primary energy reconstruction. Specifically, it successfully removes multiples, including the challenging near-offset area, while preserving the amplitudes of the primaries (yellow arrow in Figure 6). It is important to note that in this case, by introducing a degree of correlation in our more realistic noise, we randomly added some coherent events with multiples-like characteristics. These new coherent noise patterns were detected by our deep learning models. Predicting primaries removes more of this coherent noise, but still at the expense of modified primary amplitudes (red arrow in Figure 6). Predicting multiples (strategy 2) still provides the best final primaries, where some coherent events were removed without affecting the primary energy.

Overall, the strategy of predicting multiples and then deriving primaries ($P^* = C - M^*$) reports to be the most effective. This strategy limits estimation errors to pixels where merged primary and multiple energy are present simultaneously, avoiding errors in areas where only primary energy is

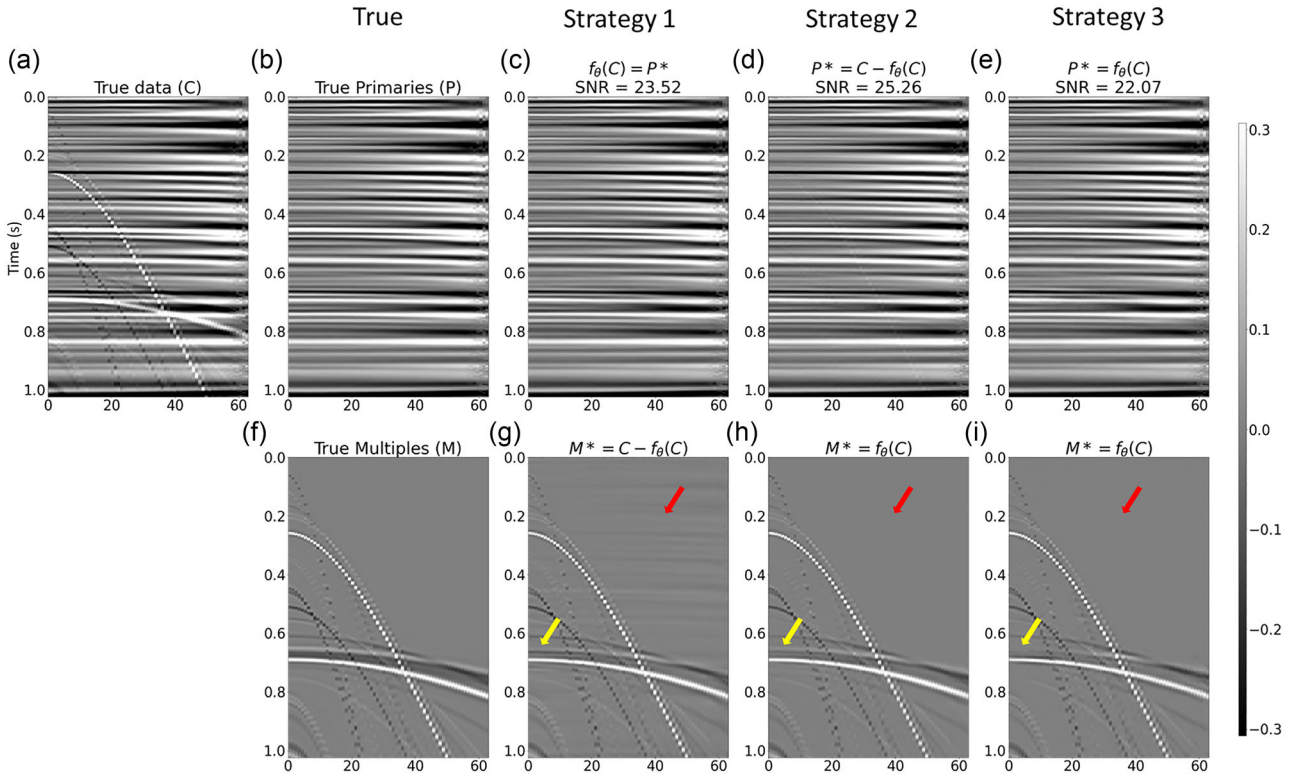

FIGURE 4 Deep learning strategies on noise-free synthetic data. (a) Synthetic gather (C). (b) True primaries (P). Predicted primaries (P*) with (c) strategy 1, (d) strategy 2, (e) strategy 3, (f) True multiples (M). (h) Estimated multiples with (g) strategy 1, (h) strategy 2 and (i) strategy 3.

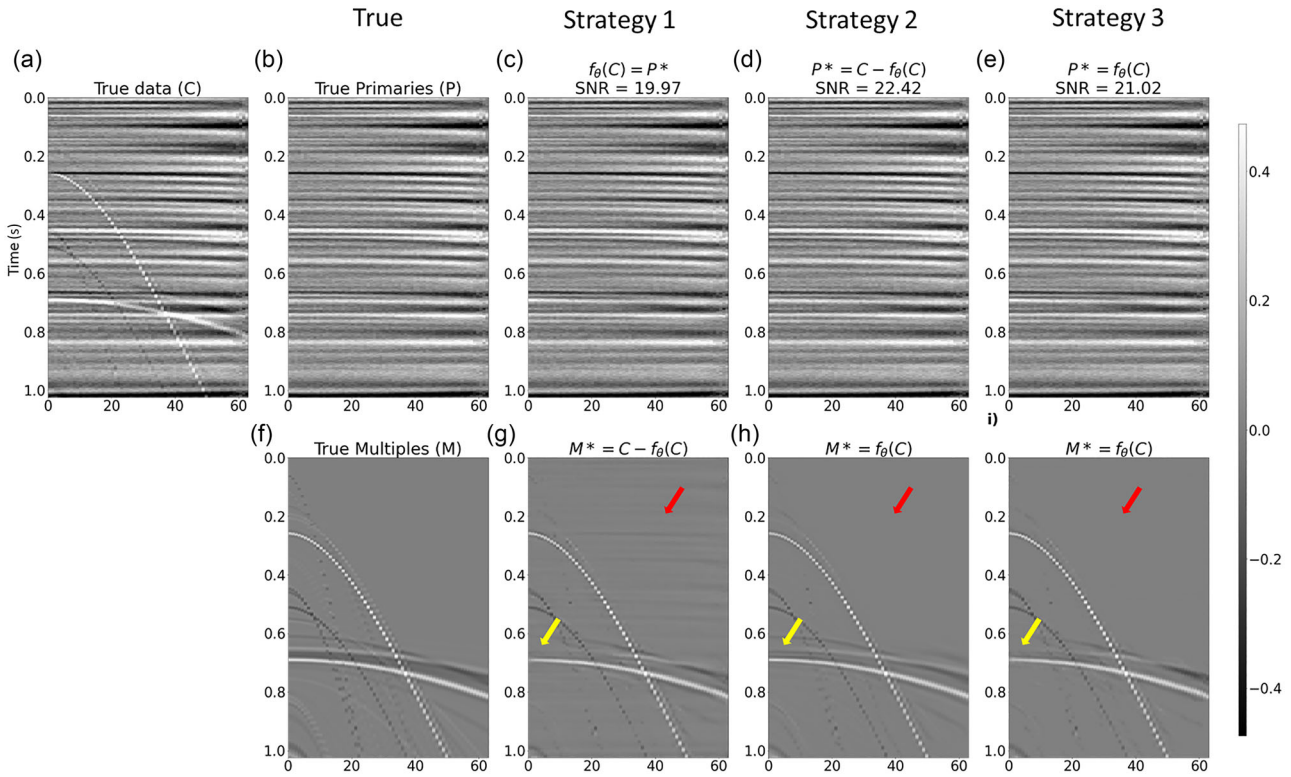

FIGURE 5 Deep learning strategies on synthetic data perturbed with uniform noise. (a) Synthetic gather (C). (b) True primaries (P). Predicted primaries (P*) with (c) strategy 1, (d) strategy 2, (e) strategy 3. (f) True multiples (M). (h) Estimated multiples with (g) strategy 1, (h) strategy 2 and (i) strategy 3.

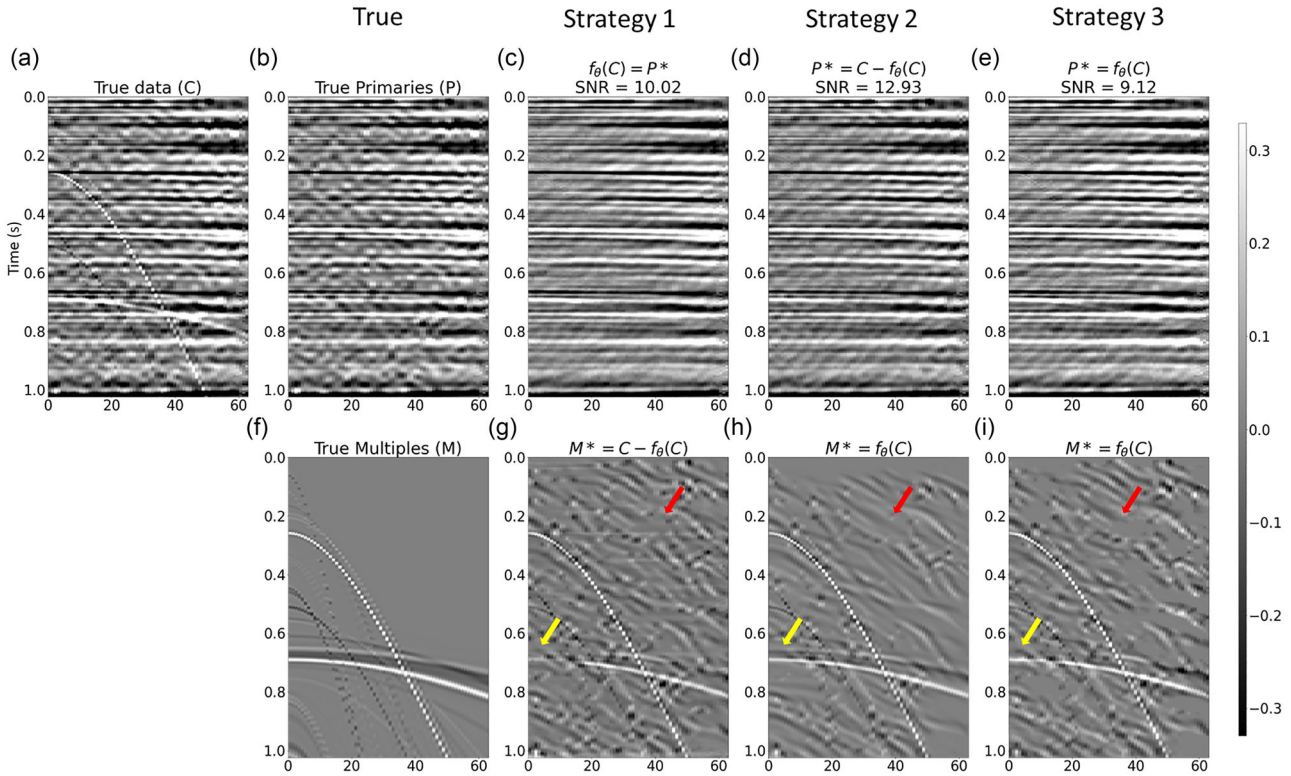

FIGURE 6 Deep learning strategies on synthetic data perturbed with realistic noise. (a) Synthetic gather (C). (b) True primaries (P). Predicted primaries (P*) with (c) strategy 1, (d) strategy 2, (e) strategy 3. (f) True multiples (M). (h) Estimated multiples with (g) strategy 1, (h) strategy 2 and (i) strategy 3.

present. This strategy is more accurate in the multiple subtractions in near offsets. Additionally, strategy 2 is more robust to noise perturbations, including out-of-distributions noise. Consequently, it results in a more precise deep learning model for separating primaries and multiples in common depth point (CDP) gathers, while preserving primary amplitudes.

Finite difference synthetic data

In order to further test the generalization capacity of our best strategy, we consider a more realistic synthetic dataset, modelled through a three-dimensional finite differences solver (courtesy of ExxonMobil). Our previous analysis showed that the most effective strategy is predicting multiples and then subtracting them from the input data while constraining the training with both, multiples and primaries (Equation 2). Therefore, we further test the deep learning network trained with this strategy. Since we do not have the true labels to quantitatively assess the performance, we qualitatively analyse the results and compare them with high-resolution Radon demultiple (Sacchi & Ulrych, 1995). Figure 7 shows the results of Radon and deep learning (DL) demultiple for a CDP of the synthetic dataset. Parabolic Radon demultiple is applied first, removing many parabolic events from the original gather (Figure 7b,c). However, near-offset steep events and far-offset linear noise could not be modelled by parabolas and, therefore, not removed from the primaries (arrows in Figure 7). To further improve the results with conventional methods, we applied linear Radon discrimination to the primaries obtained after parabolic Radon. Figure 7d,e shows the results after applying parabolic and linear Radon demultiple. Linear Radon mitigated the linear steep event in near offsets, but it could not eliminate some far-offset linear noise. The DL approach successfully predicts all multiples and coherent noise present in the CDP gather, providing a multiple-free seismic gather after subtraction. Importantly, it mitigates some of the challenges of Radon transform (RT) demultiple, that is, near offset steep events and far offset linear noise. This example shows the generalization capacity of our proposed methodology to a dataset created with a different engine (wave equation solver) than the one used for creating the training data (convolutional model). Additionally, our DL approach does not require parameterization during inference time, eliminating time-consuming optimal parameter search and picking a mute function required for successful parabolic and linear RT demultiple. Furthermore, the multiple models obtained with the neural network are directly subtracted from the input CDP without any further processing such as adaptive subtraction (Guitton & Verschuur, 2004; Verschuur

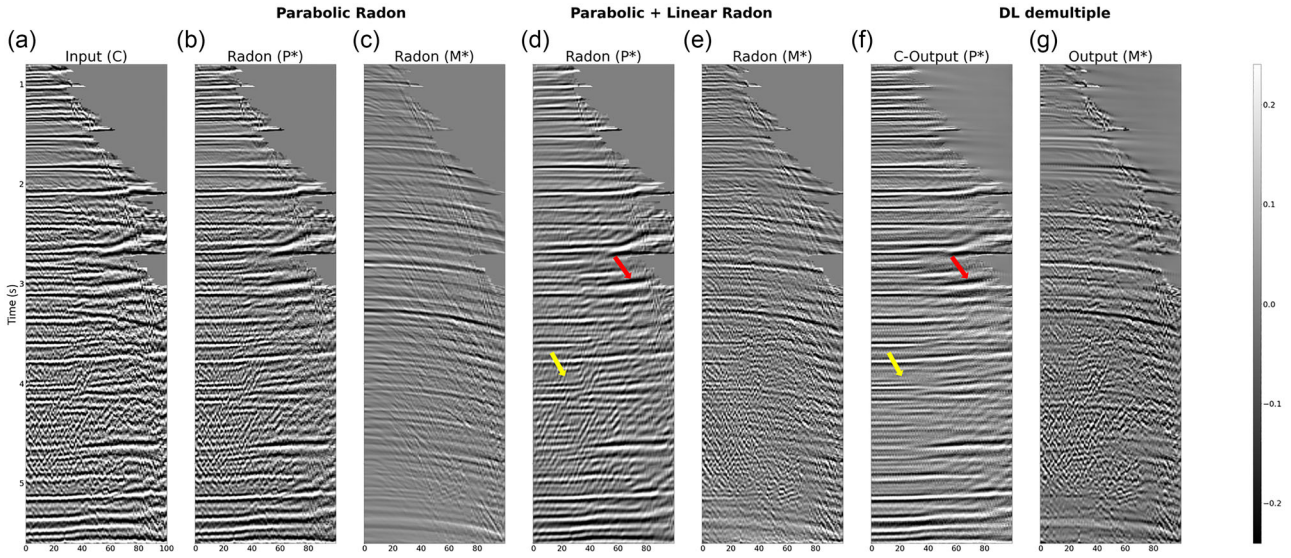

FIGURE 7 CDP gather from a synthetic dataset modelled by 3D finite difference. (a) Data with multiple and primaries. (b) Primaries estimation with parabolic Radon method. (c) Multiples estimated by parabolic Radon method. (d) Primaries estimation with parabolic and linear Radon methods. (e) Multiples estimated by parabolic and linear Radon methods. (f) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (g) Multiples estimation from the DL model trained with Equation (2) ($f_\theta(C) = M^*$).

& Berkhout, 1997), reducing further the processing effort for seismic demultiple.

RESULTS ON FIELD DATA

The goal of any seismic processing method is to be applied to real seismic data. For deep learning (DL) models, the possibility of providing good results in datasets not seen during training, that is, generalization, is one of the most desired properties. In order to test the generalization ability of our single network trained to predict multiples, we evaluate the results on four distinctive field datasets. We consider two post-migration datasets where residual multiples are present, and two pre-migration datasets where no previous demultiple workflow has been done. For real data cases, we also do not have ground-true labels, that is, real true multiples and primaries. Therefore, we qualitatively analyse the results and compare them with conventional demultiple processing, that is, surface-related multiple elimination (SRME), and high-resolution Radon demultiple. Note that in the field data examples of this section, we directly subtract the predicted multiples with a neural network trained with strategy 2 (Equation 2), without any further processing of the multiple models obtained (e.g., adaptive subtraction; Guitton & Verschuur, 2004; Verschuur & Berkhout, 1997), usually carried out in conventional processing demultiple (e.g., SRME).

Dataset 1 consists of common depth points (CDPs) after migration with remaining multiples. Each CDP has 64 traces on the horizontal axis and 5 s on the vertical axis. It is characterized by weak amplitude, high-frequency, parabolic multiples, and linear noise. Figure 8 shows the primaries and multiples estimated for dataset 1 with Radon and DL methods. Here again, we applied the first parabolic Radon transform (RT), eliminating several multiples, but unable to predict steep near and far offset events. We further apply linear RT to the image after parabolic RT to better eliminate coherent noise from the input gather. The primaries resulting after applying both parabolic and linear RT eliminate a big part of the coherent noise. However, we observed that near offsets are not completely free of the coherent noise. Figure 8f,g shows the results of our DL model. Our proposed DL approach successfully predicts all multiples and coherent noise present in the input gather. More importantly, the DL method is able to mitigate some of the challenges of RT. In particular, DL multiples estimation can predict and subsequently eliminate the near offset steep events and multiples with strong moveout non-predicted or not fully removed by Radon.

We consider The North Sea Volve Dataset as a second case (dataset 2). It also consists of CDPs after migration with remaining coherent noise, that is, multiples and linear noise. Each CDP has 40 traces in the horizontal direction and 4505 m depth in the vertical direction. It has stronger multiples with low- and high-frequency content. Moreover, the data present strong linear events intersecting primary energy. Figure 9 shows the demultiple results. In this case, parabolic Radon successfully eliminates a big part of the multiples, but it cannot eliminate steep linear events. Coherent linear noise is further mitigated with linear RT, but it does not successfully eliminate all the linear noise. Our DL approach shows successful multiple estimations and

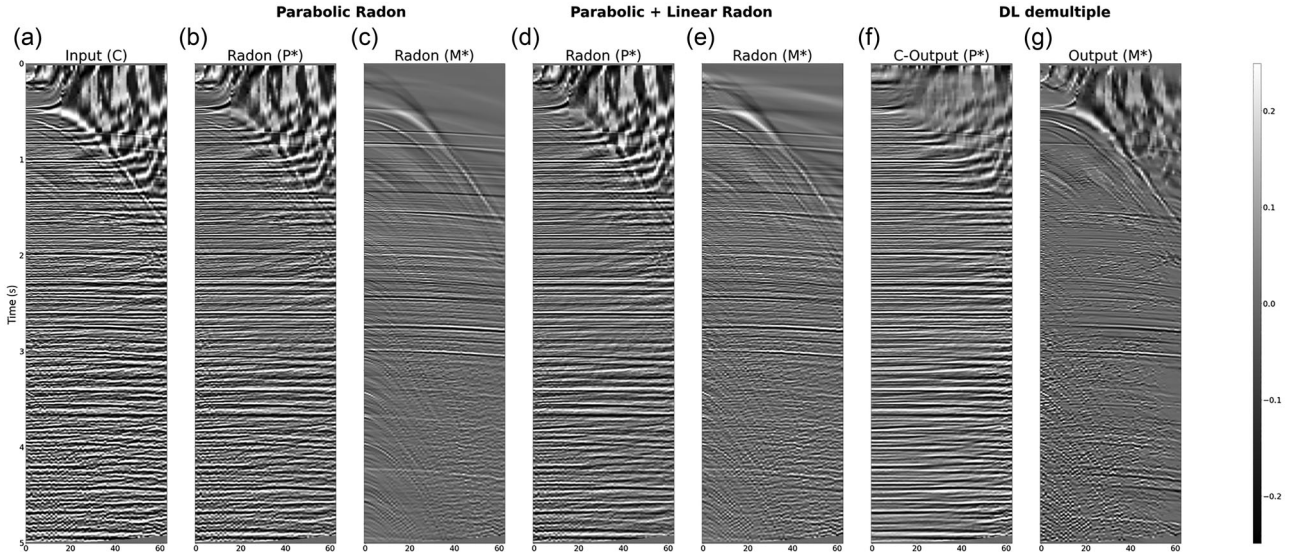

FIGURE 8 Field dataset 1. (a) Data with multiple and primaries. (b) Primaries estimation with parabolic Radon method. (c) Multiples estimated by parabolic Radon method. (d) Primaries estimation with parabolic and linear Radon methods. (e) Multiples estimated by parabolic and linear Radon methods. (f) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (g) Multiples estimation from the DL model trained with Equation (2) ($f_\theta(C) = M^*$).

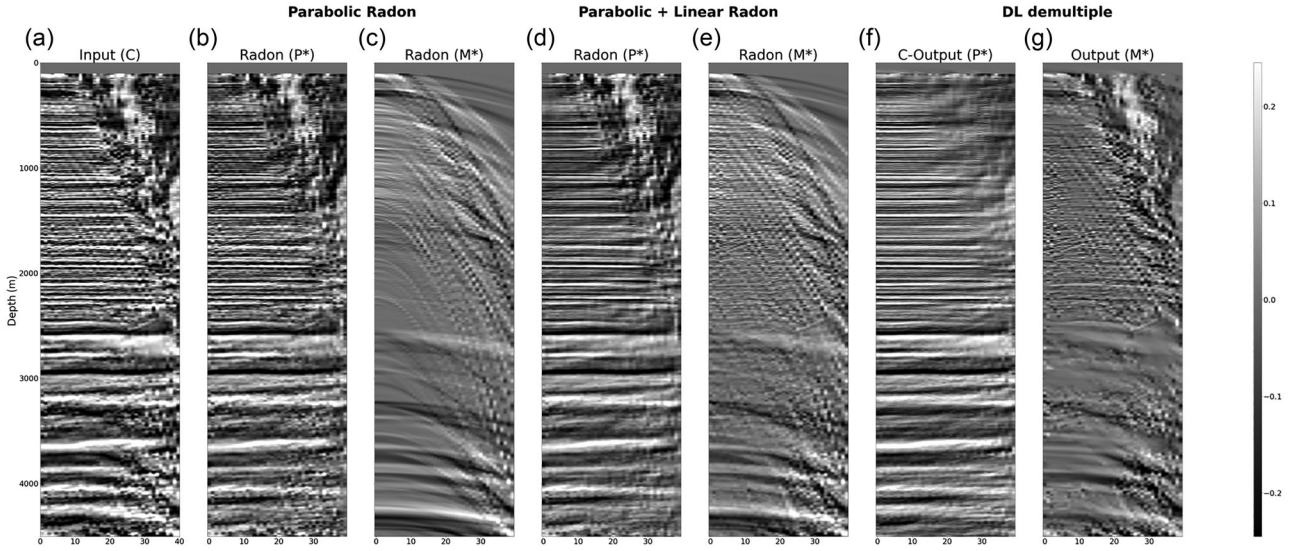

FIGURE 9 Field dataset 2. (a) Data with multiple and primaries. (b) Primaries estimation with parabolic Radon method. (c) Multiples estimated by parabolic Radon method. (d) Primaries estimation with parabolic and linear Radon methods. (e) Multiples estimated by parabolic and linear Radon methods. f) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (g) Multiples estimation from the DL model trained with Equation (2) ($f_\theta(C) = M^*$).

elimination of both, parabolic and linear events, providing a cleaner seismic gather with only primaries.

Dataset 3 consists of normal moveout (NMO) corrected CDPs before migration from the Mississippi Canyon region of the Gulf of Mexico (Dragoset, 1999). This region is well-known for multiples originated by the free-surface, and by a shallow salt body intrusion with high reflectivity contrast (Sabbione & Sacchi, 2017). Each CDP has 92 traces in the horizontal direction and 7 s in the vertical axis. In this case, no previous demultiple processing has been carried out. Therefore, it has strong multiples events (internal and surface-related multiples). In particular, around 3.5–6.5 s, the presence of a strong multiple package is present. Figure 10 shows the primaries and multiples discrimination results with Radon and DL methods. Parabolic Radon shows overall good multiple eliminations in this

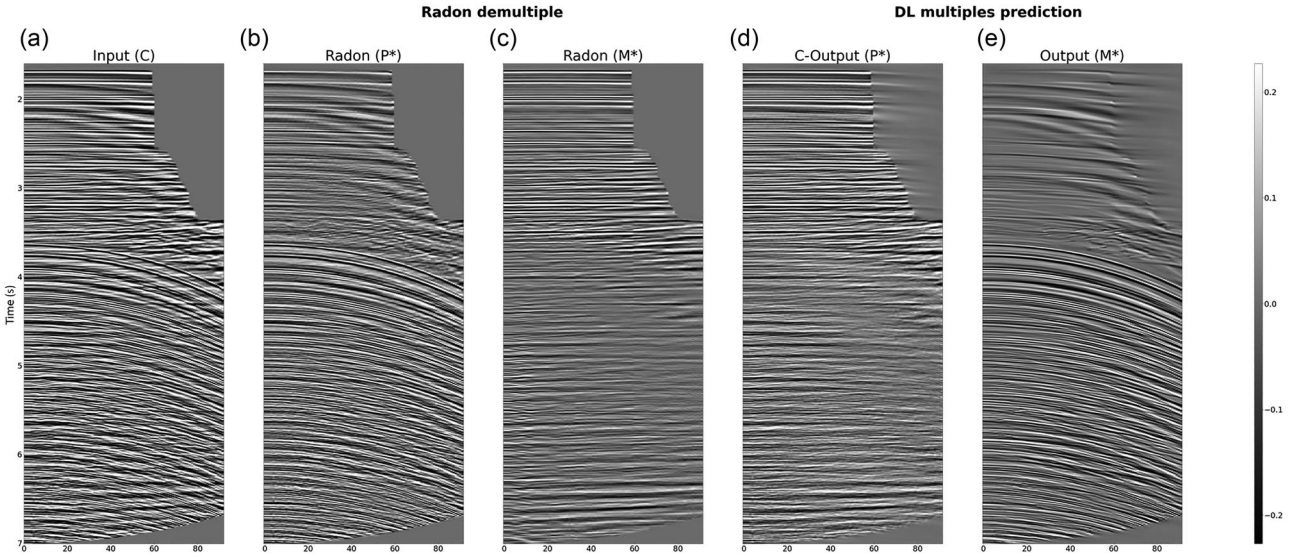

FIGURE 10 Field dataset 3. (a) Data with multiple and primaries. (b) Primaries estimation by parabolic Radon method. (c) Multiples estimated by parabolic Radon method. (d) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (e) Multiples estimation from the DL model trained with Equation (2) ($f_\theta(C) = M^*$).

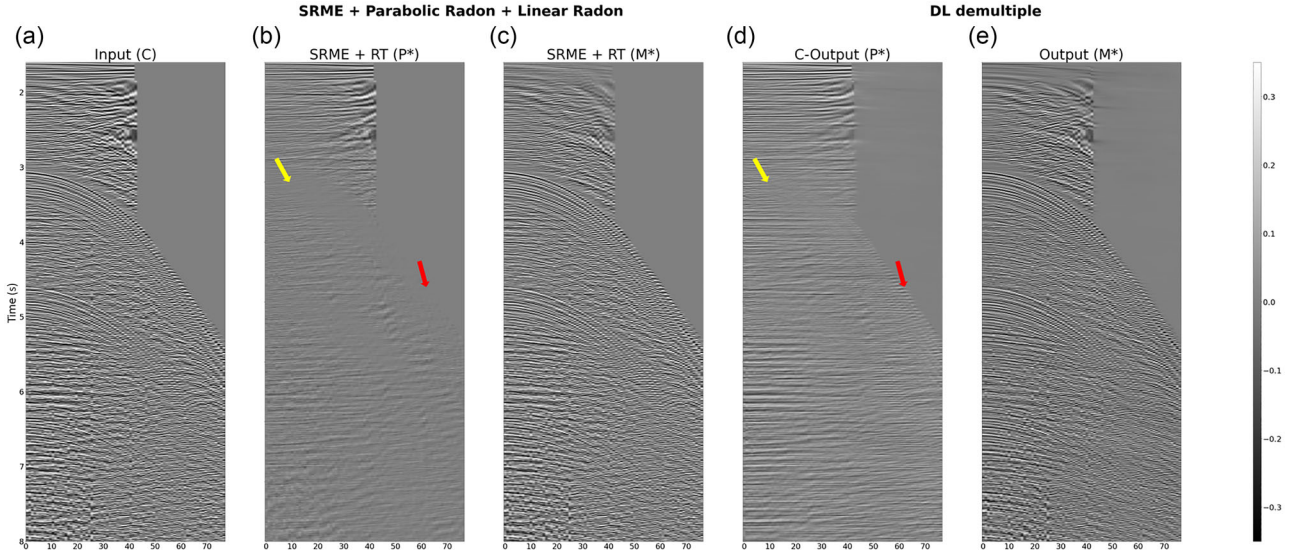

FIGURE 11 Field dataset 4. (a) Data with multiple and primaries. (b) Primaries estimation by SRME and Radon methods. (c) Multiples estimated by SRME and Radon methods. (d) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (e) Multiples estimation from the DL model trained with Equation (2) ($f_\theta(C) = M^*$).

case. However, multiples are not fully removed in near offset. Our DL approach overcomes this difficulty by being able to successfully eliminate the multiples and preserve primary energy.

Dataset 4 consists of NMO-corrected CDPs before migration from the South China Sea. The dataset is from a two-dimensional exploration profile, targeting sedimentary and crustal structures and represents deeper environments with more than of 1 km water depth (Chang et al., 2022). Each CDP has 77 traces in the horizontal direction and 16 s in time. Figure 11a shows a subsection of a CDP from 1.6 to 8 s. No previous demultiple processing has been carried out, and therefore we observe strong coherent noise. A strong package of first-order surface-related multiples is present around 3 s. Second- and third-order surface-related multiples are observed around 5 and 7 s, respectively. Internal multiples as well as linear noise seismic events with strong moveout are also observed in the CDP, making difficult the visualization of primaries. In order to discriminate primaries from coherent noise, we performed SRME, parabolic Radon and linear

Radon. Figure 11 shows the primaries and multiples discrimination results of our conventional demultiple workflow. The traditional workflow allows us to retrieve primaries; however, it presents challenges where strong surface-related multiples were present, eliminating primary energy as well for example, around 3 s (yellow arrow in Figure 11). Our DL approach overcomes this difficulty by being able to successfully eliminate the multiples and preserve primary energy. However, in this case, we also observed some multiple residuals in our primaries at far offset, for example, around 5–6 s (red arrow in Figure 11).

Overall, we show that our DL model generalizes to very distinctive field datasets, before and after migration, while being able to mitigate the challenges of the conventional high-resolution Radon demultiple. Additionally, the DL solutions do not require parameterization during inference time, avoiding time-consuming parameter search for optimal results. Importantly, it also does not require any manual tasks, that is, mute picking, reducing considerably the time of the demultiple process in the seismic processing workflow.

DISCUSSION

Generalization

The results presented on synthetic and field data show the potential of our supervised deep learning (DL) approach to overcome the limitations of conventional methods and alleviate time-consuming manual tasks, that is, mute function picking, and optimal parameter search. Furthermore, our model shows successful generalization properties, providing a demultiple tool for distinctive datasets in pre- and post-migration gathers with surface-related and/or internal multiples. In general, supervised DL algorithms report good performance when the input data are represented by the training data distribution. In the case of input data outside of the training data distribution, the performance of DL methods decays (Goodfellow et al., 2016a). Hence, it is important to create a feature-rich training dataset by considering many possible scenarios encountered in field data, which is never trivial when synthetic data are generated to use as training data. In Appendix B,B we further investigate the importance of the training data for more accurate primary estimations. Additionally, several steps are performed to further improve the generalization. First, automatic gain control normalization of the data mitigates the difference in amplitude between our training and field data distributions. Moreover, we have shown from synthetic tests that predicting the multiple leads to better multiple attenuation in noisy data (Table 1). Importantly, this strategy (Equation 2) is more robust to out-of-distribution noise (Table 1), which is an important property of field data where the noise distribution is always unknown.

Additionally, our multiple model prediction with DL shows accurate amplitude prediction in field datasets alleviating for adaptive subtraction usually carried out in conventional demultiple such us SRME. However, we think that in challenging cases of strong amplitude variation with offset, adaptive subtraction could further mitigate the generalization challenge for the seismic demultiple task.

Importantly, the synthetic multiples generated in the training dataset do not follow perfectly the moveout of multiples predicted by physical assumptions, for example, parabolic and hyperbolic shapes, and this may be an advantage. Indeed, the moveout of multiples is most often predicted following simple assumptions, for example, one-dimensional tabular medium, and ray theory, among others. Field data may include more complex shapes of seismic events, as a result of more complex physics than considered in conventional multiple elimination methods such as Radon. As a result, some of the improvements of the DL method when compared to the Radon method (Figures 7–10) may come from these properties of the training dataset. The same explanation can be given to explain why the DL method is as efficient after migration as before: it does not matter what the multiple events become after the migration operator is applied, there is no need for the physically constrained shape of the multiple for it to be removed. Our results on synthetics using "realistic noise" with some degree of coherence and field datasets show that the DL method can also be used to remove coherent events that are not parabolic multiples. The training dataset here does not include negative moveouts, but this could be implemented to fully remove coherent noises with our preferred strategy 2.

Out-of-distribution case

Figure 12 shows one of the limitations of the DL method for cases out of the training distribution, and a practical solution to mitigate this challenge. In this case, we consider the synthetic dataset generated with a three-dimensional finite difference solver, with a fold of 100 traces in the common depth point (CDP) gather (Figure 12a) instead of a fold of 64 used in the training dataset. When we input the seismic gather with 100 traces in the offset direction, we observed that the multiple around 1–3 s are not completely predicted and therefore not completely removed (yellow arrow in Figure 12c). In particular, the near offset part (flat) of that multiple is not predicted, while the far offset part with stronger moveout is correctly estimated. Most likely, the training data represents multiple geometries with stronger offset moveouts than the one presented in this case. The moveout threshold choice between primaries and multiples is learned by the network on a fold of 64 traces. Therefore, on the 100 traces dataset, some seismic events are only considered a multiple at far offsets, while the near-offset part falls into the primaries category. One practical solution would be to modify the offset

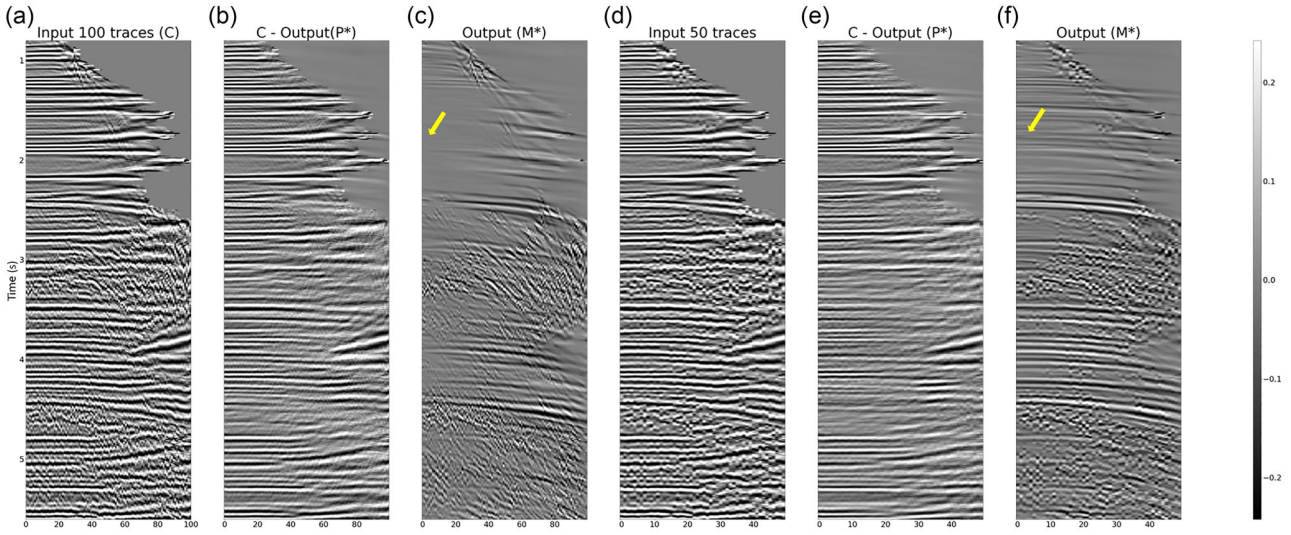

FIGURE 12 Mitigating out of distribution issue by changing the amount of input traces. (a) Data with multiple and primaries with 100 traces. (b) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (c) Multiples estimation form the DL model train with Equation (2) ($f_\theta(C) = M^*$). (d) Data with multiple and primaries with 50 traces (taking every second trace from a) to bring the CDP gather close to the geometry we use for training. (e) Primaries estimated by subtracting the DL-multiples estimation from the input image ($P^* = C - M^*$). (f) Multiples estimation from the DL model trained with Equation (2) ($f_\theta(C) = M^*$).

moveout in the input image by eliminating every second trace. In this case, the input image would have 50 traces in the offset direction (Figure 12c). If we applied our DL model to predict multiples on the decimated gather, we observed that the multiples are fully estimated and subtracted (Figure 12e,f). The data manipulation of eliminating every second trace might bring the seismic image close to the training data samples where 64 offset traces were considered, and, therefore, multiples can be fully predicted, including the near offset part (yellow arrow in Figure 12f). This shows both the complex problem of generalization and possible solutions coming from data manipulation to bring the application case close to our training distribution. For this case, a practical solution would be dividing our original 100 traces into two gathers of 50 traces each by taking every second trace, applying the DL model to eliminate the multiples on both complementary CDPs and finally recombining the multiple predictions, as presented in Figure 7.

Deep learning limitations and challenges

Deep learning models can sometimes be challenging to interpret, making it difficult to understand the underlying reasoning behind their predictions. There is no recipe or detailed workflow to build the best possible model for a given task. Therefore, training a useful model requires a considerable trial and error experimentation phase, that is, creating training data, testing convolutional neural network architectures, and optimal hyperparameters search (learning rate, batch size, etc.). On the other hand, this effort needs to be done only once during the training phase, speeding up the inference time, where no further parameter search needs to be done. It is however complicated to understand in detail the limitations of the DL model. An empirical approach to assess the limitations of a trained model is testing it on several synthetic and field datasets, as we carried out in the present work. This challenge could be further assessed by training Bayesian neural networks and being able to estimate uncertainties in the multiple predictions and elimination.

Another challenge of our proposed DL method for seismic multiple attenuation is the lack of user parameters to influence the demultiple process during inference. Having many user parameters such as in the high-resolution Radon-based approach represents an issue, but having no user parameter at all during inference could also be a problem. In particular, with our current approach, we give an input image with merged primaries and multiples and obtain as output the corresponding multiple estimations, although the classification of one event as being multiple or primary can be very subjective to the interpreter. The current method implies a fixed threshold between primaries and multiples. This discrimination is implicit in the training dataset where a residual moveout (RMO) threshold is decided. This can be challenging in field data applications where the normal moveout velocity is not accurate enough, leading to misaligned primaries. For example in Figures 8–11, there are seismic events with small moveouts considered as multiples in our DL solution. However, some of these events could be misaligned primaries due to inaccurate velocity estimation. Therefore, there is a clear

need for a more flexible model that gives the interpreter more room to decide whether to keep or eliminate a seismic event based on moveout discrimination. By incorporating a varying minimum RMO threshold into our synthetic modelling, we see a chance for a more flexible solution of a network learning to generate different versions of the multiples from which the user can select the best results in a potentially target-oriented manner. In Fernandez et al. (2024), we proposed to output simultaneously four different multiple models based on moveout discrimination, allowing selective demultiple. This control mechanism allows the processing expert to safely remove seismic events with large moveouts while being conservative with the events of small moveouts that may be unaligned primaries. There is, however, room to further improve the flexibility in the demultiple process, predicting more than four multiple models as well as discriminating seismic events based on other characteristics such as amplitude and frequency.

Our future research focuses on estimating uncertainties in our train models, providing more flexible options during inference, and better understanding the DL models' strengths and limitations. Additionally, we would like to investigate how the different building blocks of the U-Net architecture affect its performance, for example, the influence of the number of layers, number of filters, kernel size, among others. This investigation would allow us to find the optimal U-Net type architecture for the demultiple task.

CONCLUSIONS

We systematically evaluate the performance of different supervised deep learning strategies for multiple removals. Our statistical analysis of synthetic label data generated by convolutional modelling reports that training a convolutional neural network to predict multiples and then subtracting them from the input image is the most effective strategy to demultiple. Additionally, it is the most robust strategy to noise. Our qualitative analysis of synthetic data generated with finite difference solver shows that our training methodology can generalize to more physically meaningful examples, outperforming limitations of conventional high-resolution Radon demultiple. We further test the generalization capacity of the proposed deep learning approach on four distinctive field datasets before and after migration. Our proposed approach shows the ability to correctly estimate surface-related and internal multiples removing them from field datasets, overcoming some of the limitations of conventional methods and showing successful generalization capabilities. Furthermore, the proposed methods work in a parameter-free manner during inference time, relieving the user from any manual task or optimal parameter search. As a result, it can provide a new powerful tool for attenuating multiples in existing workflows.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the members of the Fraunhofer ITWM DLSeis Consortium (http://dlseis.org) for their financial support. We thank ExxonMobil for the three-dimensional synthetic dataset. We appreciate Equinor ASA, Vår Energy ASA, Petoro AS and ConocoPhillips Skandinavia AS for granting us permission to utilize their field data (dataset 1). We are also grateful to Equinor and Volve Licence partners for releasing The North Sea Volve Dataset under an Equinor Open Data Licence (dataset 2). We thank Gareth O'Brien and two anonymous reviewers for their very helpful reviews.

DATA AVAILABILITY STATEMENT

Part of the data used in this research can be obtained by contacting the authors upon reasonable request. The North Sea Volve dataset is an open seismic dataset and can be obtained by visiting https://www.equinor.com/energy/volve-data-sharing.''

ORCID

Mario R. Fernandez https://orcid.org/0009-0006-3945-1863

REFERENCES

Abbasi, S. & Jaiswal, P. (2013) Attenuating long-period multiples in short-offset 2D streamer data: Gulf of California. In: *SEG technical program expanded abstracts 2013*. Houston, TX: Society of Exploration Geophysicists, pp. 4201–4205.

Beylkin, G. (1987) Discrete Radon transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35, 162–172.

Birnie, C., Ravasi, M., Liu, S. & Alkhalifah, T. (2021) The potential of self-supervised networks for random noise suppression in seismic data. *Artificial Intelligence in Geosciences*, 2, 47–59.

Breuer, A., Ettrich, N. & Habelitz, P. (2020) Deep learning in seismic processing: trim statics and demultiple. In: *SEG technical program expanded abstracts*. Houston, TX: Society of Exploration Geophysicists, pp. 3199–3203.

Bugge, A.J., Evensen, A.K., Lie, J.E. & Nilsen, E.H. (2021) Demonstrating multiple attenuation with model-driven processing using neural networks. *The Leading Edge*, 40(11), 831–836.

Chai, X., Gu, H., Li, F., Duan, H., Hu, X. & Lin, K. (2020) Deep learning for irregularly and regularly missing data reconstruction. *Scientific Reports*, 10, 3302.

Chamorro, D., Zhao, J., Birnie, C., Staring, M., Moritz, F. & Ravasi, M. (2023) Deep learning-based extraction of surface wave dispersion curves from seismic shot gathers. arXiv preprint, arXiv:2305.13990.

Chang, S.-P., Pubellier, M., Delescluse, M., Qiu, Y., Nirrengarten, M., Mohn, G., Chamot-Rooke, N. & Liang, Y. (2022) Crustal architecture and evolution of the southwestern south china sea: Implications to continental breakup. *Marine and Petroleum Geology*, 136, 105450.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint, arXiv:2010.11929.

Dragoset, B. (1999) A practical approach to surface multiple attenuation. *The Leading Edge*, 18, 104–108.

Durall, R., Ghanim, A., Ettrich, N. & Keuper, J. (2022) Dissecting u-net for seismic application: An in-depth study on deep learning multiple removal. arXiv preprint, arXiv: 2206.12112.

Durall, R., Ghanim, A., Fernandez, M.R., Ettrich, N. & Keuper, J. (2023) Deep diffusion models for seismic processing. *Computers and Geosciences*, 177, 105377.

Fernandez, M., Durall, R., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2022a) A benchmark study of deep learning methods for seismic interpolation. In: *83rd EAGE Annual Conference & Exhibition*. European Association of Geoscientists & Engineers, pp. 1–5.

Fernandez, M., Durall, R., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2022b) A comparison of deep learning paradigms for seismic data interpolation. In *Second EAGE Digitalization Conference and Exhibition*. European Association of Geoscientists & Engineers, pp. 1–5.

Fernandez, M., Durall, R., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2022c) Image-to-image seismic interpolation. In 83rd EAGE Annual Conference and Exhibition Workshop Program. European Association of Geoscientists & Engineers, pp. 1–5.

Fernandez, M., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2023a) Deep learning strategies for seismic demultiple. In: *Third EAGE Digitalization Conference and Exhibition.* European Association of Geoscientists & Engineers, pp. 1–5.

Fernandez, M., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2023b) Seismic demultiple with deep learning. In: *84th EAGE Annual Conference and Exhibition.* European Association of Geoscientists & Engineers, pp. 1–5.

Fernandez, M., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2024) Towards flexible demultiple with deep learning. In *Fourth international meeting for applied geoscience & energy expanded abstracts*. Houston, TX: Society of Exploration Geophysicists and American Association of Petroleum Geologists.

Foster, D.J. & Mosher, C.C. (1992) Suppression of multiple reflections using the radon transform. *Geophysics*, 57, 386–395.

Goodfellow, I., Bengio, Y. & Courville, A. (2016a) *Deep learning*. Cambridge, MA: MIT Press. http://www.deeplearningbook.org.

Goodfellow, I., Bengio, Y. & Courville, A. (2016b) *Deep learning*. Cambridge, MA: MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 1–9.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J. & Chen, T. (2017) Recent advances in convolutional neural networks. arXiv preprint, arXiv:1512.07108v6.

Guitton, A. & Verschuur, D. (2004) Adaptive subtraction of multiples using the l1-norm. *Geophysical Prospecting*, 52(1), 27–38.

Guo, R., Maniar, H., Di, H., Moldoveanu, N., Abubakar, A. & Li, M. (2020) Ground roll attenuation with an unsupervised deep learning approach. In: *SEG technical program expanded abstracts 2020*. Houston, TX: Society of Exploration Geophysicists, pp. 3164–3168.

Hampson, D. (1986) Inverse velocity stacking for multiple elimination. In: *1986 SEG annual meeting, SEG 1986*, Houston, TX: Society of Exploration Geophysicists, pp, 422–424.

Harsuko, R. & Alkhalifah, T.A. (2022) Storseismic: a new paradigm in deep learning for seismic processing. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

Ho, J., Jain, A. & Abbeel, P. (2020) Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A.A. (2016) Image-to-image translation with conditional adversarial networks. arXiv preprint, arXiv:1611.07004.

Kaur, H., Fomel, S. & Pham, N. (2020) Seismic ground-roll noise attenuation using deep learning. *Geophysical Prospecting*, 68(7), 2064–2077.

Kaur, H., Pham, N. & Fomel, S. (2020) Separating primaries and multiples using hyperbolic radon transform with deep learning. In: *SEG technical program expanded abstracts 2020*. Houston, TX: Society of Exploration Geophysicists, pp. 1496–1500.

Kingma, D.P. & Ba, J. (2017) Adam: a method for stochastic optimization. arXiv preprint, arXiv:1412.6980v9.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 84–90.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, 521(7553), 436–444.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Li, Z. & Gao, H. (2020) Feature extraction based on the convolutional neural network for adaptive multiple subtraction. *Marine Geophysical Research*, 41, 10.

Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B. & Nadai, M.D. (2021) Efficient training of visual transformers with small datasets. arXiv preprint, arXiv:2106.03746v2.

Long, J., Shelhamer, E. & Darrell, T. (2015) Fully convolutional networks for semantic segmentation. arXiv preprint, arXiv:1411.4038v2.

Lu, W. (2013) An accelerated sparse time-invariant Radon transform in the mixed frequency-time domain based on iterative 2D model shrinkage. *Geophysics*, 78, V147–V155.

Mandelli, S., Borra, F., Lipari, V., Bestagini, P., Sarti, A. & Tubaro, S. (2018) Seismic data interpolation through convolutional autoencoder. In *SEG technical program expanded abstracts 2018*. Houston, TX: Society of Exploration Geophysicists, pp. 4101–4105.

Mandelli, S., Lipari, V., Bestagini, P. & Tubaro, S. (2019) Interpolation and denoising of seismic data using convolutional neural networks. arXiv preprint, arXiv:1901.07927.

Mousavi, S.M., Beroza, G.C., Mukerji, T. & Rasht-Behesht, M. (2024) Applications of deep neural networks in exploration seismology: a technical survey. *Geophysics*, 89(1), WA95-WA115.

Oz, Y. (2001) *Seismic data analysis*. Tulsa, OK: Society of Exploration Geophysicists.

Qiu, C., Wu, B., Liu, N., Zhu, X. & Ren, H. (2022) Deep learning prior model for unsupervised seismic data random noise attenuation. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.

Qu, S., Verschuur, E., Zhang, D. & Chen, Y. (2021) Training deep networks with only synthetic data: deep-learning-based near-offset reconstruction for (closed-loop) surface-related multiple estimation on shallow-water field data. *Geophysics*, 86, A39–A43.

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-Net: convolutional networks for biomedical image segmentation. In *International Conference on medical image computing and computer-assisted intervention*, Berlin: Springer, pp. 234–241.

Sabbione, J.I. & Sacchi, M.D. (2017) Attenuating multiples with the restricted domain hyperbolic Radon transform. In: *15th International Congress of the Brazilian Geophysical Society & EXPOGEF, Rio de Janeiro, Brazil, 31 July-3 August 2017*. Rio de Janeiro: Brazilian Geophysical Society, pp. 603–608.

Sacchi, M.D. & Ulrych, T.J. (1995) High-resolution velocity gathers and offset space reconstruction. *Geophysics*, 60, 1169–1177.

Shirui, W., Wenyi, H., Pengyu, Y., Xuqing, W., Qunshan, Z., Prashanth, N., German, O.B. & Jiefu, C. (2021) Seismic deblending by self-supervised deep learning with a blind-trace network. In: *First International Meeting for Applied Geoscience & Energy Expanded Abstracts*, 40(11), 831–836.

Shuey, R.T. (1985) A simplification of the Zoeppritz equations. *Geophysics*, 50(4), 609–614.

Siahkoohi, A., Verschuur, D.J. & Herrmann, F.J. (2019) Surface-related multiple elimination with deep learning. In: *SEG technical program expanded abstracts 2019*. Houston, TX: Society of Exploration Geophysicists, pp. 4629–4634.

Sohn, K., Lee, H. & Yan, X. (2015) Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28, 3483–3491.

Taner, M. (1980) Long period sea-floor multiples and their suppression. *Geophysical Prospecting*, 28, 30–48.

Thorson, J.R. & Claerbout, J.F. (1985) Velocity-stack and slant-stack stochastic inversion. *Geophysics*, 50, 2727–2741.

Trad, D., Ulrych, T. & Sacchi, M. (2003) Latest views of the sparse radon transform. *Geophysics*, 68, 386–399.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2017) Deep image prior. arXiv preprint, arXiv:1711.1092.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2023) Attention is all you need. arXiv preprint, arXiv:1706.03762v7.

Verschuur, D. & Berkhout, A. (1997) Estimation of multiple scattering by iterative inversion, part ii: Practical aspects and examples. *Geophysics*, 62(5), 1596–1611. ISSN: 0016-8033.

Verschuur, D. J., Berkhout, A. J. & Wapenaar, C. P. A. (1999) Adaptive surface-related multiple elimination. *Geophysics*, 57, 1166–1177.

Vershuur, D. J. (2013) *Seismic multiple removal techniques: past, present and future*. Houten, the Netherlands: European Association of Geoscientists & Engineers.

Wang, W., McMechan, G.A., Ma, J. & Xie, F. (2021) Automatic velocity picking from semblances with a new deep-learning regression strategy: Comparison with a classification approach. *Geophysics*, 86(2), U1–U13.

Wiggins, W. (1998) Attenuation of complex water-bottom multiples by wave equation prediction and subtraction. *Geophysics*, 53(12), 1527–1539.

Wiggins, W. (1999) Multiple attenuation by explicit wave extrapolation to an interpreted horizon. *The Leading Edge*, 18, 46–54.

Yang, F. & Ma, J. (2019) Deep-learning inversion: A next-generation seismic velocity model building method. *Geophysics*, 84, R583–R599.

Zhang, D., de Leeuw, M. & Verschuur, E. (2021) Deep learning-based seismic surface-related multiple adaptive subtraction with synthetic primary labels. In *First International Meeting for Applied Geoscience & Energy Expanded Abstracts*. Houston, TX: Society of Exploration Geophysicists, pp. 2844–2848.

Zhang, M., Liu, Y. & Chen, Y. (2019) Unsupervised seismic random noise attenuation based on deep convolutional neural network. *IEEE Access*, 7, 179810–179822.

How to cite this article: Fernandez, M.R., Ettrich, N., Delescluse, M., Rabaute, A. & Keuper, J. (2025) Towards deep learning for seismic demultiple. *Geophysical Prospecting*, 73, 1185–1203. https://doi.org/10.1111/1365-2478.13672

APPENDIX A: REALISTIC SYNTHETIC NOISE

Generating synthetically realistic noise is challenging. However, here we attempt to create more realistic synthetic noise for a qualitative analysis in our synthetic test dataset. One of the characteristics of real noise is that it is usually not completely uncorrelated. In fact, there is a level of correlation in time, space or both (Birnie et al., 2021). To synthetically generate a noise matrix with more realistic characteristics, we propose to create an initial matrix based on a random noise distribution, for example, Gaussian, $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, we interpolated the noise matrix, introducing some degree of coherence and, finally, we applied a band-pass filter between 10 and 50 Hz to further eliminate the random unrealistic texture, resulting in a more realistic noise. Figure A1 shows the initial Gaussian noise, the interpolated noise, and the final frequency-filtered noise.

APPENDIX B: TEST ON DIFFERENT PERCENTAGES OF TRAINING DATA

Training data plays a crucial role in supervised deep learning methods. Our training data consists of 100,000 synthetically generated common depth points (CDPs) and corresponding multiples and primaries ground-truth labels. Here we aim to further investigate the influence of our training data on the final trained models. We analyse the performance of the preferred strategy 2 for different percentages of the total training dataset considering our test dataset of 400 synthetic CDPs, and different noise perturbations. In particular, we consider training with 75%, 50%, 25%, 10%, and 5% of the initial training dataset. Table B1 reports the overall decrease in performance when the training data is smaller. To further understand the impact of the training data, we visualize one synthetic sample. Figure B1 shows the decrease in the accuracy of the predicted multiples when the model is trained with a smaller dataset, highlighting the importance of a feature-rich training dataset for optimal demultiple. Specially, we observed that the first area affected by the reduction of training data is near-offset, where multiple predictions start to be less accurate.

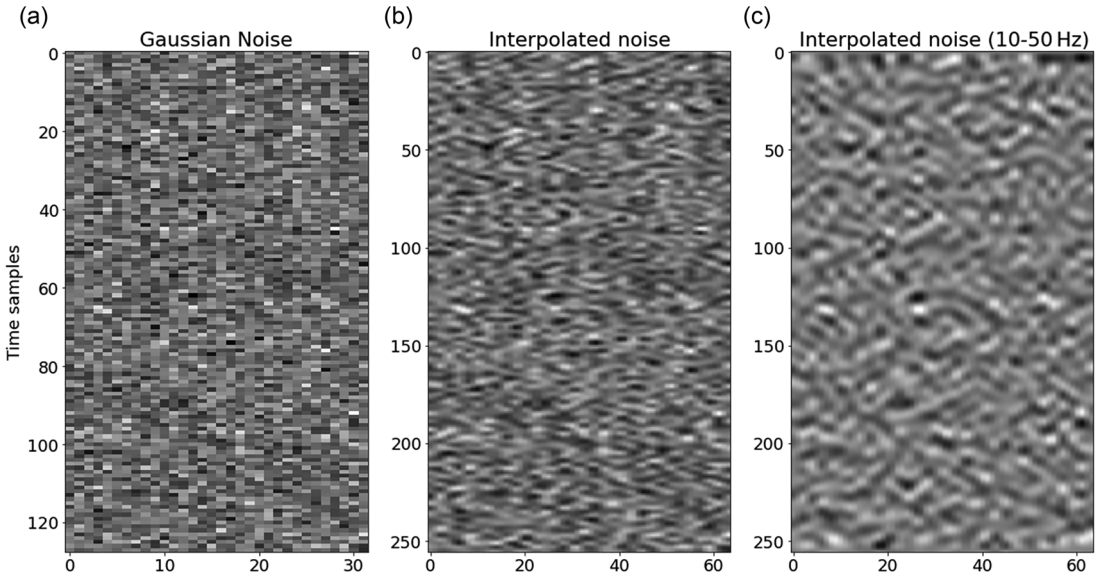

FIGURE A1 Generating realistic seismic noise from a Gaussian distribution. (a) Gaussian random noise. (b) Interpolated Gaussian noise. (c) Filtered interpolated noise (10–50 Hz).

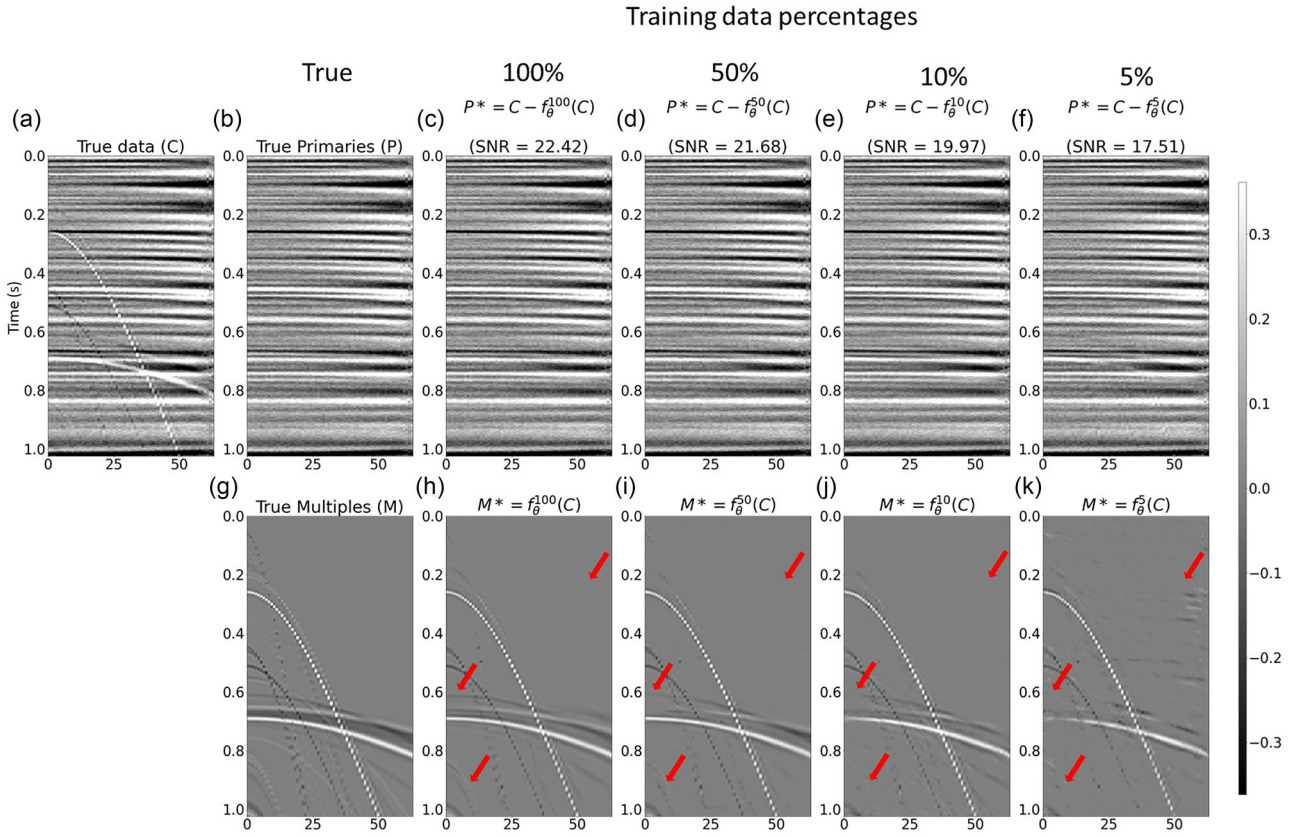

FIGURE B1 Effect of training data in multiple predictions. (a) Input synthetic CDP. (b) Primary label. Primaries obtained training the network with (c) 100% of training data, (d) 50% of training data, (e) 10% of training data, and (f) 5% of training data. (g) Multiple labels. Multiples obtained training with (h) 100% of training data, (i) 50% of training data, (j) 10% of training data and (k) 5% of training data.

TABLE B1 Statistical results on test dataset (mean SNR) of primaries estimations with the preferred strategy 2, considering different percentages of the initial training dataset (100%, 75%, 50%, 25%, 10%, 5%) and different noise perturbations (uniform low, uniform high, Gaussian, and Laplace).

	100%	**75%**	**50%**	**25%**	**10%**	**5%**
Uniform 1	23.50	22.37	21.92	20.74	20.37	18.05
Uniform 2	20.74	19.47	19.32	18.42	18.22	16.80
Gaussian	16.85	15.76	15.53	15.27	15.20	14.23
Laplace	15.10	14.50	14.15	14.13	13.60	13.54