Auto-Linear Phenomenon in Subsurface Imaging

Yinan Feng [1] **Yinpeng Chen** [2] **Peng Jin** [3] **Shihang Feng** [4] **Youzuo Lin** [5]

Abstract

Subsurface imaging involves solving full waveform inversion (FWI) to predict geophysical properties from measurements. This problem can be reframed as an image-to-image translation, with the usual approach being to train an encoder-decoder network using paired data from two domains: geophysical property and measurement. A recent seminal work (InvLINT) demonstrates there is only a linear mapping between the latent spaces of the two domains, and the decoder requires paired data for training.

This paper extends this direction by demonstrating that only linear mapping necessitates paired data, while both the encoder and decoder can be learned from their respective domains through self-supervised learning. This unveils an intriguing phenomenon (named Auto-Linear) where the self-learned features of two separate domains are automatically linearly correlated. Compared with existing methods, our Auto-Linear has four advantages: (a) solving both forward and inverse modeling simultaneously, (b) applicable to different subsurface imaging tasks and achieving markedly better results than previous methods, (c)enhanced performance, especially in scenarios with limited paired data and in the presence of noisy data, and (d) strong generalization ability of the trained encoder and decoder.

1. Introduction

Subsurface imaging is crucial for revealing subsurface layering and geophysical properties (such as velocity and conductivity), supporting important applications such as energy exploration, carbon capture, and earthquake early warning systems. In this field, the full waveform inversion (FWI) is a well-known method to infer subsurface velocity maps from the seismic data. Concurrently, the forward process involves computing the pressure wavefield using these velocity maps. Specifically, seismic data are obtained via seismic surveys which employ receivers to record reflected and refracted seismic waves generated by controlled sources. Each receiver records a 1D time series signal, and the signals recorded by all the receivers form the seismic data. They are mathematically connected by an acoustic wave equation as:

$$\nabla^2 p(x,z,t) - \frac{1}{c^2(x,z)}\frac{\partial^2}{\partial t^2}p(x,z,t) = s(x,z,t), \quad (1)$$

where $p(x,z,t)$ represents the seismic data and $c(x,z)$ is the velocity map. $s(x,z,t)$ is the source term. x is the hor-

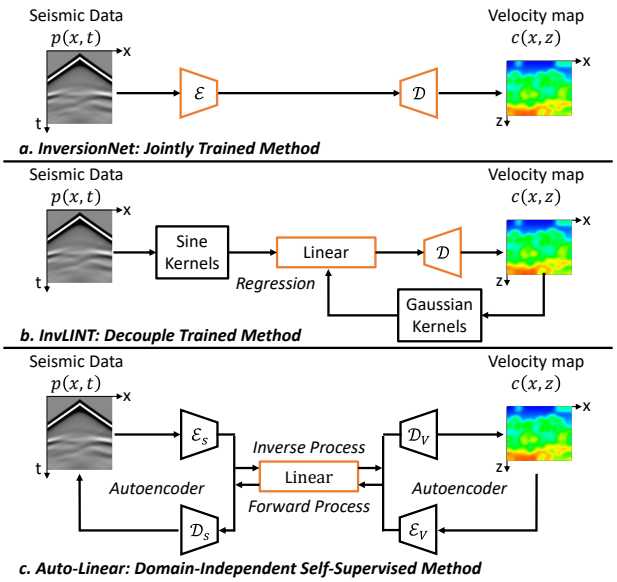

Figure 1: **Overview** of the jointly trained encoder-decoder (top), InvLINT (middle), and Auto-Linear (bottom). The orange color indicates that components need to be trained with paired data. Auto-Linear decouples both the encoder and decoder and self-supervised trains them separately in their own domains. Linear converters are learned to connect the frozen, pre-trained encoders and decoders.

[1]Department of Computer Science, The University of North Carolina at Chapel Hill,USA [2]Google Research, USA [3]College of Information Sciences and Technology, The Pennsylvania State University, USA [4]Earth and Environmental Sciences Division, Los Alamos National Laboratory,USA [5]School of Data Science and Society, The University of North Carolina at Chapel Hill,USA. Correspondence to: Youzuo Lin <yzlin@unc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

izontal direction, z is the depth, t denotes time, and ∇^2 is the Laplacian operator. In practice, seismic data is typically collected by surface sensors (i.e., $p(x, z = 0, t)$, abbreviated as $p(x, t)$). This inversion problem is "ill-posed," presenting challenges due to sensitivity to initial conditions and potential for multiple solutions. Furthermore, the substantial costs and logistical hurdles of acquiring real subsurface data often make it infeasible to gather extensive real-world datasets. Consequently, much of the current research primarily depends on full-physics simulations for experimental work, driven by the scarcity of publicly available real datasets.

Recent works (Wu & Lin, 2019; Zhang et al., 2019; Sun et al., 2021; Jin et al., 2022) consider FWI as an *image-to-image translation problem* constrained by a wave equation, and leverage deep neural networks to achieve a significant performance boost. As shown in Figure 1-a, they learn an encoder-decoder architecture to map seismic data to velocity. Note that the encoder and decoder are jointly trained from the supervision of paired seismic data and velocity maps.

In Feng et al. (2022), the authors of InvLINT laid the foundation for exploring linear relationships in latent spaces by separating encoders and the decoder into two distinct domains and retaining a linear component for joint training. Shown in Figure 1-b, their approach utilizes two predetermined integral transforms, with Sine and Gaussian kernels, to encode seismic data and velocity maps. While the encoders are decoupled, the decoder still relies on supervised training due to the absence of an explicit inversion for the Gaussian integral transform. This also restricts the decoder's ability to generalize among multiple datasets. Moreover, the kernel solution faces notable limitations: a) poor performance with datasets containing large variations and high-frequency components (e.g., OpenFWI (Deng et al., 2022)), b) a lack of noise resistance, and c) no clear rule for selecting the appropriate kernels for different situations. Thus, InvLINT may not apply to broader scenarios.

In this paper, we provide a more modular framework for data-driven subsurface imaging that can overcome all the limitations of InvLINT via self-supervised learning. This approach fully decouples the training of the encoder and decoder. Specifically, we first independently train two masked autoencoders (MAEs) (He et al., 2022) separately, one for seismic data and another for velocity maps, as depicted in Figure 1-c. After self-pretraining, the encoder and decoder are frozen, and a linear converter is then trained to connect these components using paired seismic data and velocity maps. This efficient framework uniquely enables us to address both the inverse problem and forward modeling simultaneously, with minimal additional training required. Moreover, it exhibits considerable improvement in few-shot scenarios with paired data, presenting a substantial advancement over traditional image-to-image translation-based inversion networks like InversionNet (Wu & Lin, 2019).

Moreover, we formula the above phenomenon that *the **independently** self-supervised encoder and decoder of two different domains can be **automatically** integrated into an end-to-end model through a supervised **linear** mapping* as *Auto-Linear Phenomenon*. The phenomenon offers an intriguing insight: an inherent and strong cross-domain correlation, with a simpler linear mapping, exists within the domain-independent self-consistent representations. This introduces a significant change in our perspective of the problem, moving away from the conventional, complex image-to-image translation task towards a more simplified, linear approach achieved through self-supervised learning. This modification not only streamlines the process but also enhances our comprehension of the auto-alignment of representation learned at each domain, offering a deeper insight into the interplay of data relationships across domains.

In experiments, Auto-Linear achieves solid performance on multiple datasets. Compared with joint training methods (e.g., InversionNet (Wu & Lin, 2019)), Auto-Linear has comparable results in the inverse problem, with its superiority in forward modeling and few-shot contexts with limited paired data and enhanced noise robustness. Moreover, Auto-Linear outperforms the previous decouple-trained method, InvLINT, and exhibits greater robustness to noisy data. Furthermore, it excels over both InvLINT and InversionNet in electromagnetic (EM) inversion, another subsurface imaging task. Our framework also enhances our understanding of relationships among multiple FWI datasets with distinct subsurface structures. We found that while these datasets can share both encoders and decoders, they require different linear mappings. In addition, we observe a correlation between the linear layer's singular values and the complexity of the dataset. Essentially, these suggest a consistent embedding method across datasets, with varying subsurface characteristics captured by cross-domain linear mappings. This leads to a piece-wise linearity between the two domains.

2. Related Works

Recently, data-driven methods for FWI have been developed. They consider the FWI as an image-to-image problem and jointly train the encoder-decoder network to solve it. Araya-Polo et al. (2018) use a fully connected network to invert velocity maps. Wu & Lin (2019) adopted an encoder-decoder CNN to solve. Zhang et al. (2019) employ GAN and transfer learning to improve the generalization. In Zeng et al. (2021), authors present an efficient and scalable encoder-decoder network for 3D FWI. Feng et al. (2021) develop a multi-scale framework with two convolutional neural networks to reconstruct the low- and high-frequency components of velocity maps. A thorough review of deep learning for FWI can be found in Lin et al. (2023).

Jin et al. (2022) use the finite difference to approximate the forward modeling as a differentiable operator and integrate it and a deep neural network (DNN) in a loop to construct an unsupervised learning method. Chen et al. (2021) proposed a self-supervised approach to solve the inverse problem from the perspective of image invariance. These purely self-supervised and unsupervised methods focus on how to solve problems without labels and still treat the network as a black box. Unlike them, our method uses self-supervised learning as a tool with the aim of simplifying the problem and decoupling the inverse process. We hope this can help the field better understand the problem and the relationship among different subsurface structures.

Recently, OpenFWI was released. It is the first open-source collection of large-scale multi-structural benchmark datasets for FWI (Deng et al., 2022). It includes 12 datasets (11 2D datasets and one 3D dataset) synthesized from multiple sources. The datasets cover diverse domains in geophysics, such as interfaces, faults, and CO2 reservoirs, and feature a variety of subsurface structures, including flat and curved geologies. Along with the dataset, they also report performance benchmarks by using state-of-art data-driven methods and the physics-driven method.

An alternative self-supervised approach for inverse problems involves pre-training a generative model on physical properties to capture their prior distribution. Subsequently, the model is adapted by integrating measurements and a physical model of the measurement process into the sampling process (Wang et al., 2023; Song et al., 2021). The proposed Auto-Linear and generative model-based methods offer **complementary insights** into inverse problems, each revealing the problem from different perspectives. Auto-Linear leverages self-supervised learning to streamline connecting seismic data with subsurface velocity maps, focusing on understanding the latent space properties of two physical quantities connected via PDE and enriching our understanding of their relationships. Conversely, generative models focus on capturing the data's complex distributions to the entire space of possible solutions. They excel in integrating the forward modeling into sampling, and only relying on samples from prior distributions without necessitating paired samples.

A key insight of our method is the ability of self-learned representations, generated by an autoencoder in one domain, to retain essential information for accurate reconstruction in another domain via minimum transformation. The crucial aspect here is the **automatic** alignment of representations across domains, while the linear relationship emphasizes the simplicity of this relationship and its strong correlation. Thus, our work provides a new perspective on solving FWI problems. It presents a complementary, yet distinct, approach to existing generative model-based approaches.

Table 1 presents a side-by-side comparison between our approach and generative model-based approaches. First, Generative model-based approaches are trained in a purely self-supervised manner do not require paired data, while our approach necessitates paired data for training the linear converter. Second, unlike diffusion models that require multi-step denoising, our setup enables efficient single-step inference. Third, we target a conditional average with minimum error, focusing on a singular outcome based on the training data's distribution, contrasting with generative models-based methods that aim to explore multiple potential solutions through distribution-to-distribution mappings. Additionally, generative model-based methods integrate forward modeling into the learning process and benefit from existing knowledge of physical processes; however, they struggle in situations where the forward model is unknown or lacks an explicit formulation. For instance, Kimberlina carbon sequestration problem (Alumbaugh et al., 2021b), while data, generated by Maxwell's Equations, has been made available, the forward modeling remains proprietary and challenging to replicate. In such scenarios, our approach presents an effective alternative, offering a different way for inverse analysis. A more detailed discussion is given in the Supplementary Material.

Proporties	Auto-Linear	Generative model-based
Paired Data	Required	Not Required
Inference	Single Step	Single/Multiple Step
Solution	Condition Average with Minimum Error	Multiple Instances
Forward Modeling	Not Required	Required
Solving Forward Problem	Can	Cannot

Table 1: Differences Between Auto-Linear and Generative Model-Based Approaches.

3. Review of previous methods

Let's begin by reviewing the previous joint training method (InversionNet (Wu & Lin, 2019)) and the decouple training method (InvLInT (Feng et al., 2022)). InversionNet considers FWI as an *image-to-image translation problem*. As shown in Figure 1-a, they train an encoder-decoder convolutional network to map seismic data $p(x,t)$ and velocity maps $c(x,z)$. This kind of jointly trained method can be formulated as:

$$\theta^*, \eta^* = \underset{\theta,\eta}{\operatorname{argmin}}\, \mathcal{L}(c, (\mathcal{D}_\eta \circ \mathcal{E}_\theta)(p)), \tag{2}$$

where $\mathcal{L}$ is the loss function. The encoder $\mathcal{E}$'s parameters θ and decoder $\mathcal{D}$'s parameters η are jointly trained from the supervision of paired seismic data and velocity maps.

In InvLINT, the authors also try to decouple the encoder and decoder, and use a linear layer to connect two latent spaces. They use two pre-determined integral transforms, with Sine and Gaussian kernels, to embed the seismic data and velocity into high-dimensional spaces. They also provide a theoretical analysis with some hypotheses to establish

Variable	Definition
$p(x,t)$	seismic data
$c(x,z)$	velocity maps
$\mathcal{E}_s$	encoder of seismic data
$\mathcal{D}_s$	decoder of seismic data
$\mathcal{E}_v$	encoder of velocity map
$\mathcal{D}_v$	decoder of velocity map
$\boldsymbol{A}, \boldsymbol{B}$	Matrix

Table 2: Table of Notation.

a near-linear relationship in latent spaces when appropriate transforms are used. This method can be formulated as:

$$\begin{aligned}
&\boldsymbol{P} = [P_1, \ldots, P_N]^T, \;\; P_n = \iint p(x,t)\Phi_n(x,t)dxdt, \\
&\boldsymbol{C} = [C_1, \ldots, C_M]^T, \;\; C_m = \iint c(x,z)\Psi_m(x,z)dxdz, \\
&\boldsymbol{A}^* = \underset{\boldsymbol{A}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{C}, \boldsymbol{A}\boldsymbol{P}), \\
&\eta^* = \underset{\eta}{\operatorname{argmin}}(c, (\mathcal{D}_\eta \circ \boldsymbol{A}^*)(\boldsymbol{P})),
\end{aligned} \tag{3}$$

where Φ_n and Ψ_m are kernels for integral transforms. Notably, InvLINT only decouples the training of encoders. Since there is no explicit inverse transformation of the Gaussian integral transform, the training of their decoder remains dependent on the seismic data c for the specific dataset. Moreover, the kernel-based solution encounters significant limitations, including poor performance on datasets containing large variations and high-frequency components (e.g., OpenFWI (Deng et al., 2022)) and a lack of noise resistance and systematic rule for kernel selection tailored to different situations. Consequently, this pre-determined kernel solution might not suit wider applications.

4. Methodology

Expanding upon InvLINT's partially decoupled approach, our framework utilizes the advanced self-supervised learning strategy to decouple the training processes of the encoder and decoder completely, significantly enhancing performance beyond InvLINT's. In this section, we present the formula of the Auto-Linear Phenomenon and describe how to apply it to the FWI problem. Table 2 lists used notations.

4.1. Auto-Linear Phenomenon

Here, we define the Auto-Linear Phenomenon as the automatic integration of independently self-trained encoder and decoder from two different domains into an end-to-end model through a linear mapping. This phenomenon highlights two key principles: network integration and feature correlation. Network integration is facilitated by a linear mapping that effectively connects the independently trained encoder and decoder into a cohesive end-to-end model of subsurface imaging. This ensures the encoder and decoder operate independently yet synergistically within the broader framework. Moreover, the Auto-Linear Phenomenon reveals that the latent representations learned independently from each domain are inherently linearly correlated.

Based on the above definition, the relationship of different components in Auto-Linear can be formulated as

$$p = (\mathcal{D}_s \circ \mathcal{E}_s)(p), \tag{4.1}$$

$$c = (\mathcal{D}_v \circ \mathcal{E}_v)(c), \tag{4.2}$$

$$\mathcal{E}_v(c) = \boldsymbol{A}\mathcal{E}_s(p), \tag{4.3}$$

$$\boldsymbol{B}\mathcal{E}_v(c) = \mathcal{E}_s(p), \tag{4.4}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the linear mappings, and not necessary to be full rank. In the above formulation, Eq. 4.1 and 4.2 describe two domain-independent autoencoders. Eq. 4.3 and 4.4 illustrate the linear correlations in the latent representations. Then, by plugging Eq. 4.3 into Eq. 4.2 (Eq. 4.4 into Eq. 4.1), we can construct an inverse (forward) model, demonstrating a seamless integration of these processes.

In contrast, InvLINT focuses on the linear correlation between two domain embeddings. This results in a model that lacks the comprehensive encoder-decoder structure necessary for subsurface imaging. Consequently, InvLINT necessitates an additional training phase for a domain-dependent decoder, distinguishing it from the Auto-Linear. Therefore, while InvLINT introduces the concept of linear correlation, its partial approach can be described as *semi*-Auto-Linear because it only achieves linearity after encoding.

4.2. Auto-Linear Framework in FWI

Leveraging the above formulations, we introduce a novel approach in subsurface imaging that fully decouples the training of the encoder and decoder, enabling a more modular approach to model development in FWI. The training process can be formulated as:

$$\theta_s^*, \eta_s^* = \underset{\theta_s, \eta_s}{\operatorname{argmin}}\, \mathcal{L}(p, (\mathcal{D}_{s\eta_s} \circ \mathcal{E}_{s\theta_s})(p)), \tag{5.1}$$

$$\theta_v^*, \eta_v^* = \underset{\theta_v, \eta_v}{\operatorname{argmin}}\, \mathcal{L}(c, (\mathcal{D}_{v\eta_v} \circ \mathcal{E}_{v\theta_v})(c)), \tag{5.2}$$

$$\boldsymbol{A}^* = \underset{\boldsymbol{A}}{\operatorname{argmin}}\, \mathcal{L}(c, (\mathcal{D}_{v\eta_v^*} \circ \boldsymbol{A} \circ \mathcal{E}_{s\theta_s^*})(p)), \tag{5.3}$$

$$\boldsymbol{B}^* = \underset{\boldsymbol{B}}{\operatorname{argmin}}\, \mathcal{L}(p, (\mathcal{D}_{s\eta_s^*} \circ \boldsymbol{B} \circ \mathcal{E}_{v\theta_v^*})(c)), \tag{5.4}$$

where $\mathcal{L}$ is the loss function. The entire model is trained in two steps: a domain-independent self-supervised learning step to train two autoencoders, and a supervised learning step to train the linear converters. In the self-supervised learning step, as outlined in Eq. 5.1 and 5.2, paired data are not needed. The encoders and decoders for seismic data, $\mathcal{E}_s$ and $\mathcal{D}_s$, and for velocity maps, $\mathcal{E}_v$ and $\mathcal{D}_v$, are trained independently using data from their respective domains. In the self-supervised learning step of the inverse process, as per 5.3, the trained parameters of $\mathcal{E}_s$ and $\mathcal{D}_v$ are frozen. Subsequently, a linear converter $\boldsymbol{A}$ is trained using paired

data to establish a connection between them. This approach also applies to forward modeling, as indicated in 5.4.

Although the seismic data is generated from the velocity, the automatic linear correlation observed between the two learned representations is not trivial. This is because the parameters of seismic autoencoder $\mathcal{E}_s, \mathcal{D}_s$ are independent of the velocity p and the parameters of velocity autoencoder $\mathcal{E}_v, \mathcal{D}_v$ are not influenced by the seismic c. Two MAEs are trained completely separately, and there is no inherent mechanism to guarantee that their respective learned representations will be naturally correlated.

Notably, due to a lack of constraint, the self-supervised learner can easily learn shortcuts for reconstruction. We chose to employ Masked Autoencoder (MAE) (He et al., 2022) as it generates better latent representations and can learn essential information about two physical quantities through the use of masks as noise, making it easier to connect the latent spaces of two modalities and transform and reconstruct the other. Furthermore, in practice, we decompose the linear converter into two linear layers with a low-dimensional bottleneck to constrain its rank. This effectively reduces redundancy in the network.

4.3. Benefits of Auto-Linear Framework

The Auto-Linear Framework introduces a range of benefits, spanning from its framework structure to the model property, and its practical performance. At the framework level, it can simultaneously acquire encoder and decoder components for both forward and inverse problems. This eliminates the previous methods' need to retrain entirely new networks for different tasks, as only the linear layers require specific training. From the perspective of model properties, the self-supervised pre-training captures essential information from both domains, granting encoders and decoders strong generalization abilities. This allows them to be effectively shared across datasets with various subsurface structures. In terms of actual performance, our model significantly outperforms InvLINT, especially in noise robustness. Compared to InversionNet, our model achieves comparable results, with only the linear layer needing paired data. Being smaller and having a simpler supervised training component, it requires less data and is less susceptible to overfitting. Thus, our model exhibits superior performance in few-shot scenarios and has improved noise robustness.

5. Experiments

We evaluate our approach on OpenFWI (Deng et al., 2022), the first and only large-scale collection of openly accessible multi-structural seismic FWI datasets with benchmarks. We compare our method with the state-of-the-art works, including InversionNet (Wu & Lin, 2019), i.e., the method that jointly trains the encoder and decoder, and InvLINT (Feng et al., 2022), i.e., the method that separates the encoder and decoder. We also evaluate Auto-Linear's generalizability for other imaging and PDE tasks. In particular, we test it on the electromagnetic (EM) inversion task controlled by Maxwell's equations. In the Supplementary Material, we compare the latent representation learned by our method and InvLINT, evaluating the generalization ability of the encoder and decoder on a newly constructed dataset, discuss different factors that affect performance, and explore applying Auto-Linear to the elastic FWI, which involves multiple-input to multiple-output maps. For interested readers, Deng et al. (2022) provides a detailed comparison of the physics-driven method.

5.1. Implementation Details

Datasets. While real data are extremely expensive and difficult to obtain, subsurface imaging research often relies on full-physics simulations, driven by the lack of publicly available real datasets. We verify our method on OpenFWI (Deng et al., 2022), the first open-source collection of large-scale, multi-structural benchmark datasets for data-driven seismic FWI. It contains 11 2D datasets with baseline, which can be divided into four groups: four datasets in the "Vel Family", four datasets in the "Fault Family", two datasets in the "Style Family", and one dataset in the "Kimberlina Family". Four datasets in the "Vel Family" are FlateVel-A/B, and CurveVel-A/B; four datasets in the "Fault Family" are FlateFault-A/B, and CurveFault-A/B; two datasets in"Style Family" are Style-A/B; and one dataset in "Kimberlina Family" is Kimberlina-CO_2. The first three families cover two versions: easy (A) and hard (B), in terms of the complexity of subsurface structures. We will use the abbreviations (e.g. FVA for FlatVel-A and CO_2 for Kimberlina-CO_2). More details can be found in (Deng et al., 2022).

Training Details. The input seismic data are normalized to the range [-1, 1] with a log scale. We employ AdamW (Loshchilov & Hutter, 2018) optimizer with momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 0.05 for both self-supervision and supervision steps. In the self-supervision step, we use the same hyper-parameters and the training schedule with the original MAE paper (He et al., 2022), except we change the batch size to 512 and remove the pixel normalization. We use each family together to train the MAE. Thus, in total, we trained four different models. In the supervision step, the initial learning rate is set to be 1×10^{-3}, and decayed with a cosine annealing (Loshchilov & Hutter, 2016). The batch size is set to 256. To make a fair comparison with the previous work, we use l_1 loss to train the linear layer. The exact network architectures are shown in Supplementary Material. We implement our models in Pytorch and train them on 1 NVIDIA Tesla V100 GPU.

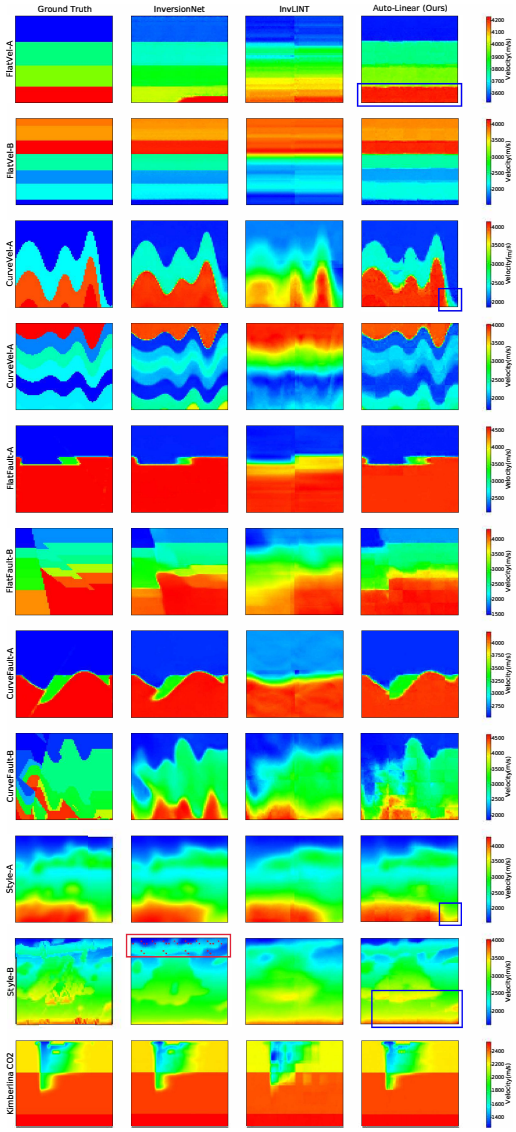

Figure 2: Illustration of results evaluated on OpenFWI, compared with InversionNet and InvLINT.

Evaluation Metrics. We apply three metrics to evaluate the generated geophysical properties: MAE, MSE, and Structural Similarity (SSIM). Following the existing literature (Wu & Lin, 2019; Feng et al., 2022; Deng et al., 2022), MAE and MSE are employed to measure the pixel-wise error, and SSIM is to measure the perceptual similarity since velocity has highly structured information, and degradation or distortion can be easily perceived by a human. We calculate them on normalized velocity maps, i.e., MAE and MSE in the scale $[-1, 1]$, and SSIM in the scale $[0, 1]$.

5.2. Auto-Linear is Simple and Effective in both Inverse and Forward Problem

Comparisons with the Joint Training Method. Table 3 shows the comparison results with InversionNet (Wu & Lin, 2019). The results of InversionNet are the reported benchmark in (Deng et al., 2022). Compared to the InversionNet, our Auto-Linear achieves comparable results on multiple datasets with only half the model size (12.3M vs. 24.4M), and only needs to supervised train the linear layer. In FlatVel-A/B, Style-B, and Kimberlina-CO_2, Auto-Linear even outperforms InversionNet in some metrics. The velocity maps inverted by different methods are shown in Figure 2. We can find InversionNet has a clearer boundary, while Auto-Linear is better at capturing the structure details in deep position (e.g., as boxed out on FaltVel-A, CurveVel-A, Style-B, and Style-A). The corresponding error map and more visualizations are provided in the Supplementary Material. Note that, InversionNet in Style-B always outputs a strange pattern in results as boxed out in red.

Comparisons with the Separate Training Method. We compare Auto-Linear with InvLINT (Feng et al., 2022), which also separates the encoder and decoder, and has a linear converter. Results are shown in Table 3. Compared to InvLINT, Auto-Linear outperforms it in terms of all three metrics. The velocity maps inverted by different methods are shown in Figure 2. The corresponding error map and more visualization results are provided in the Supplementary Material. We can clearly observe that InvLINT performs poorly for data with high-frequency layering locations and faults (i.e., “Vel Family" and “Fault Family"), but yields good results in smoother structures like “Style Family" and Kimberlina CO_2. This phenomenon may come from: 1) InvLINT model is very small and has limited expressive power. “Vel Family" and “Fault Family" are very diverse. It does not have enough capacity to learn all cases. 2) The Gaussian kernel cannot capture the small fault structure well, such as the interface and fault structures. 3) Their encoder uses frequency domain features. However, the high-frequency signal is mainly present in the reflected wave, which has a small amplitude. It is not easy to be captured by a frequency-domain encoder. A comparison of the seismic and velocity latent representations obtained by our method and InvLINT is presented in the Supplementary Material.

Auto-Linear for Forward Process. We further evaluate the efficacy of Auto-Linear for the forward process on “Fault Family". Utilizing the pre-trained velocity encoder and seismic decoder, we trained a linear converter to map the velocity latent vector to the seismic latent vector. We combine l_1 and l_2 loss as the loss function, and compare our method with a forward version InversionNet (Gupta, 2023). The results, detailed in Table 4, show a promising improvement in seismic data construction compared to InversionNet. We illustrate our output in Figure 3. While the reconstruction of reflected waves, particularly in more complex datasets, shows potential for further improvement, the overall capability to construct seismic data is evident. Additionally, we conducted experiments using the trained inversion network above to inverse the generated seismic data by the forward

Metrics	Model	FVA	FVB	CVA	CVB	FFA	FFB	CFA	CFB	SA	SB	CO2
MAE↓	Auto-Linear	**0.0081**	0.0467	0.0738	0.1820	**0.0164**	0.1208	0.0277	0.1791	0.0719	**0.0638**	**0.0060**
	InversionNet	0.0131	**0.0351**	**0.0685**	**0.1497**	0.0172	**0.1055**	**0.0260**	**0.1646**	**0.0625**	0.0689	0.0061
	InvLINT	0.0532	0.1621	0.0981	0.2462	0.0729	0.1522	0.0853	0.1955	0.1002	0.0835	0.0150
MSE↓	Auto-Linear	0.0005	0.0151	0.0188	0.1051	0.0026	0.0362	0.0061	0.0697	0.0139	**0.0097**	0.0017
	InversionNet	**0.0004**	**0.0077**	**0.0162**	**0.0836**	**0.0018**	**0.0303**	**0.0042**	**0.0614**	**0.0105**	0.0260	**0.0014**
	InvLINT	0.0085	0.0650	0.0238	0.1312	0.0190	0.0467	0.0229	0.0754	0.0209	0.0132	0.0039
SSIM↑	Auto-Linear	0.9888	0.9044	0.8057	0.6169	0.9701	0.6868	0.9426	0.5672	0.8423	**0.7275**	**0.9908**
	InversionNet	**0.9895**	**0.9461**	**0.8074**	**0.6727**	**0.9766**	**0.7208**	**0.9566**	**0.6136**	**0.8859**	0.6314	0.9872
	InvLINT	0.8457	0.6465	0.7355	0.4946	0.8506	0.6445	0.8204	0.5471	0.7916	0.6557	0.9760

Table 3: Quantitative results evaluated on OpenFWI, compared with InversionNet and InvLINT, in terms of MAE, MSE, and SSIM. Auto-Linear achieves comparable accuracy with InversionNet and outperforms InvLINT in terms of all three metrics. For each dataset, we use bold to highlight the best results, and underlined for the second best results.

model on CurveFault-A, and the visualizations are shown in Figure 4. Although the performance of this "forward-inverse" model is diminished due to the cascade amplification of errors, the visualizations affirm that the recovered velocity maps successfully capture the main structures.

Dataset	Model	MAE↓	MSE↓	SSIM↑
FlatFault-A	Auto-Linear	**0.0099**	**0.0006**	**0.9853**
	InversionNet	0.0332	0.0145	0.9659
FlatFault-B	Auto-Linear	**0.0193**	**0.0016**	**0.9604**
	InversionNet	0.0397	0.0078	0.9283
CurveFault-A	Auto-Linear	**0.0132**	**0.0009**	**0.9784**
	InversionNet	0.0373	0.0186	0.9571
CurveFault-B	Auto-Linear	**0.0253**	**0.0022**	**0.9404**
	InversionNet	0.0594	0.0132	0.8636

Table 4: Quantitative results for the forward process.

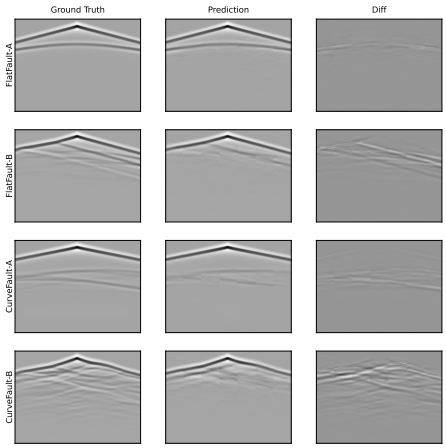

Figure 3: Illustration of forward results on "Fault Family".

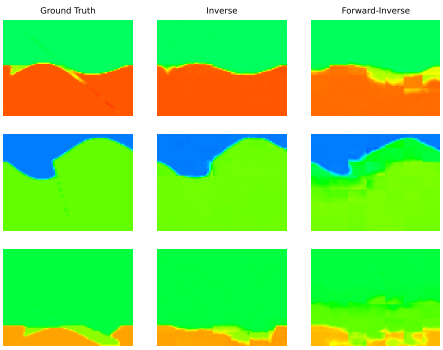

Figure 4: Illustration of results from Forward-Inverse process on CurveFault-A.

5.3. Auto-Linear is applicable to another subsurface imaging task.

We experimented on another subsurface imaging task, recovering subsurface conductivity from surface-acquired electromagnetic (EM) measurements, on the Kimberlina-Reservoir dataset (Alumbaugh et al., 2021a; Feng et al., 2022). Let $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic fields. $\mathbf{J}$ and $\mathbf{M}$ are the electric and magnetic sources. σ is the electrical conductivity and $\mu_0 = 4\pi \times 10^{-7}\Omega \cdot s/m$ is the magnetic permeability of free space. The governing equations here are Maxwell's Equations

$$\begin{aligned} \sigma\mathbf{E} - \nabla \times \mathbf{H} &= -\mathbf{J}, \\ \nabla \times \mathbf{E} + i\omega\mu_0\mathbf{H} &= -\mathbf{M}. \end{aligned} \tag{5}$$

We compared the results of our Auto-Linear model with InversionNet and results reported in InvLINT (Feng et al., 2022), presented in Table 5. Note that, to maintain consistency with InvLINT, the MAE and MSE reported below were calculated after denormalizing to the original range of $[0, 0.65]$. For all other results presented in our paper, the MAE and MSE were calculated in the normalized range of $[-1, 1]$. We observe that our proposed Auto-Linear yields significantly better performance than those obtained using InvLINT and InversionNet.

Dataset	Model	MAE↓	MSE↓	SSIM↑
Kimberlina-Reservoir	Auto-Linear	**0.00438**	**0.000192**	**0.9700**
	InversionNet	0.01330	0.000855	0.9175
	InvLINT	0.00703	0.000537	0.9370

Table 5: Quantitative results for EM inversion. MAE and MSE are calculated after denormalizing to their original range ($[0, 0.65]$). Highlighting the best results with bold and the second best results with underline.

5.4. Auto-Linear has Nice Properties

In this part, we demonstrate that our Auto-Linear has some nice properties, including the strong generalization ability of the pre-trained encoder/decoder, good noise handling, solid performance on few-shot learning, and a correlation between linear layers and the datasets' complexity.

Generalization Ability of Encoder and Decoder. We study the generalization ability of the pre-trained encoder and decoder. In particular, we choose the seismic encoder and

Metrics	Model	FVA	FVB	CVA	CVB	SA	SB
MAE↓	Auto-Linear	**0.0073**	0.0570	**0.0653**	**0.1804**	0.0725	0.0646
	InversionNet	0.0131	0.0351	0.0685	0.1497	0.0625	0.0689
MSE↓	Auto-Linear	0.0005	0.0198	**0.0159**	**0.1030**	0.0144	0.0099
	InversionNet	0.0004	0.0077	0.0162	0.0836	0.0105	0.0260
SSIM↑	Auto-Linear	**0.9895**	0.8752	**0.8192**	0.6044	0.8351	0.7222
	InversionNet	0.9895	0.9461	0.8074	0.6727	0.8423	0.7275

Table 6: Generalizability of pre-trained encoder and decoder, using InversionNet as a baseline.

velocity decoder that self-supervised trained on "Fault Family", fix it, and train the linear converter on other datasets (except the Kimberlina-CO_2, since it has different dimensions). The results are shown in Table 6.

Results show that encoders and decoders trained on the "Fault Family" excel across datasets, including Style-A and Style-B, despite their distinct subsurface structures. Notably, when applied to simpler datasets like the "Vel Family," they outperform the results trained exclusively on these datasets, as highlighted in Table 3. This demonstrates that the latent representations from self-supervision capture essential information transferable across datasets, and strategic selection of self-supervision data can enhance our method's performance. The current choice of using each family together in the self-supervision is not to lose generality

To further show the generalization ability of the pre-trained encoder and decoder and the performance improvement by picking self-supervision data, we conduct another experiment that trains MAE on cross-family datasets. In particular, CurveVel-A, FlatFault-A, and CurveFault-A are used. We test this pair of encoder and decoder in all datasets. Moreover, we constructed a new dataset from Marmousi, which is a completely different dataset from OpenFWI, to support the claim of strong generalization. Both results are in Supplementary Material.

Good handling of noise. We provide the quantitative results of the robustness test. In particular, we add Gaussian Noise with different variances to the input seismic data during testing. The noise level is chosen in accordance with the previous work (Jin et al., 2022). Table 7 shows the performance on CurveFault-A. We also include the noise's variance (σ^2) and average peak-to-noise ratio (PSNR) in the table. PSNR of a sample is defined as

$$\text{PSNR} = 10 \log_{10} \frac{(p_{max} - p_{min})^2}{\ell_2(p - p')}, \tag{4}$$

where p_{max} and p_{min} denote the maximum and minimum possible values of the seismic data in a dataset, p is the clean seismic data, and p' is the noisy data.

Compared to other models, our Auto-Linear is the most robust one to noise. The robustness of Auto-Linear shows in two aspects. First, its performance degradation on noisy data is smaller than others. Second, when the noise's variance is large ($\sigma^2 \geq$ 5e-5), our method outperforms InversionNet. This enhanced robustness can be attributed to the smaller size of our model and its simpler supervised training component, which together increase its robustness. As expected, InvLINT is extremely sensitive to the noise, as it only uses a Fourier transform as its encoder.

Strong Performance on Few-Shot Learning. One of the most important benefits of our method is it does not need paired data to train its encoder and decoder. Thus, we test Auto-Linear on the few-shot learning situation, where only a limited number of paired data exists, and compare it InversionNet. We chose five datasets as examples and tested the situation that only 1/10 or 1/20 paired data can be used in supervised learning. MAE results are reported in Table 8.

Across all datasets, our method consistently surpasses InversionNet's performance as the amount of paired data decreases, regardless of whether our model performs better (e.g., FlatVel-A and FlatFault-A) or not as well (e.g., CurveVel-A, CurveFault-A, and Style-A) compared to InversionNet with the full dataset. Additionally, the EM inversion on the Kimberlina-Reservoir dataset, detailed in 5, is another example of few-shot learning. Unlike the "Fault Family" or "Style Family," which includes 48k training data, the EM dataset comprises only 750 training samples. In this scenario, our model significantly outperforms InversionNet. Collectively, these outcomes highlight a robust advantage of our Auto-Linear approach in few-shot learning scenarios.

Correlation between the linear layer and datasets' complexity. By simplifying the image-to-image translation problem to a linear problem, our model is easy to analyze. With only a linear converter trained in a supervised manner, we can conduct a singular value decomposition analysis. The results are shown in Figure 5. We normalize it by dividing it by its maximum value and trunk it at 128 dim, which is the bottleneck dimension. We can clearly observe a strong correlation between the singular values and the dataset complexity. Generally speaking, CurveFault-B, FlatFault-B, Style-A, Style-B, and CurveFault-B have the most complex velocity map among all datasets. Their singular values are much slower to fall. On the other hand, FlatVel-A and Kimberlina-CO_2 are the simplest datasets, which are also reflected in their singular values. The original singular value is in the Supplementary Material for reference. Results prove that our linear converter correlates to the datasets' complexity. For readers who might be interested, Deng et al. (2022) provide a more detailed analysis of datasets' complexity.

5.5. Ablation Test

In this part, we test the performance of several different non-linear converters, showing that they can only provide limited improvement. We also show the comparison among different self-supervised training methods, demonstrating our framework is adaptable beyond MAE to include other

Model	σ^2 =0			σ^2 =1e-5 PSNR=70.49dB			σ^2 =5e-5 PSNR=63.48dB			σ^2 =1e-4 PSNR=60.45dB			σ^2 =5e-4 PSNR=53.39dB		
	MAE↓	MSE↓	SSIM↑	MAE↓	MSE↓	SSIM↑	MAE↓	MSE↓	SSIM↑	MAE↓	MSE↓	SSIM↑	MAE↓	MSE↓	SSIM↑
Auto-Linear	0.0277	0.0061	0.9426	0.0354	0.0070	0.9387	0.0508	0.0102	0.9255	0.0630	0.0139	0.9113	0.1093	0.0339	0.8308
Degradation (%)	\	\	\	-27.80	-14.75	-0.41	-83.39	-67.21	-1.81	-127.44	-127.87	-3.32	-294.58	-455.74	-11.86
InversionNet	0.0260	0.0042	0.9566	0.0332	0.0050	0.9539	0.0696	0.0133	0.9290	0.1439	0.0479	0.8830	0.4496	0.3948	0.6407
Degradation (%)	\	\	\	-27.69	-19.05	-0.28	-167.69	-216.67	-2.89	-453.46	-4.57	-7.69	-1629.23	-9300.00	-33.02
InvLINT	0.0853	0.0229	0.8204	3.1849	19.3293	0.0449	7.4442	103.2302	0.0172	10.1643	185.8730	0.0084	23.8050	1033.9167	0.0025
Degradation (%)	\	\	\	-3633.76	-84307.42	-94.53	-8627.08	-450686.90	-97.90	-11815.94	-22653.60	-98.98	-27807.39	-4514820.09	-99.70

Table 7: Quantitative results on CurveFault-A with Gaussian noise of varying variance σ^2 added during testing.

Dataset	Model	Ratio=1	Ratio=1/10	Ratio=1/20
FlatVel-A	Auto-Linear	**0.0081**	**0.0361**	**0.0570**
	InversionNet	0.0131	0.0590	0.0795
CurveVel-A	Auto-Linear	0.0738	**0.1245**	**0.1444**
	InversionNet	**0.0685**	0.1295	0.1450
FlatFault-A	Auto-Linear	**0.0164**	**0.0423**	**0.0592**
	InversionNet	0.0172	0.0544	0.0775
CurveFault-A	Auto-Linear	0.0277	**0.0634**	**0.0836**
	InversionNet	**0.0260**	0.0781	0.1032
Style-A	Auto-Linear	0.0719	**0.0989**	**0.1101**
	InversionNet	**0.0625**	0.1046	0.1292

Table 8: MAE quantitative results using partial data sets. Ration indicates the proportion of data sets used.

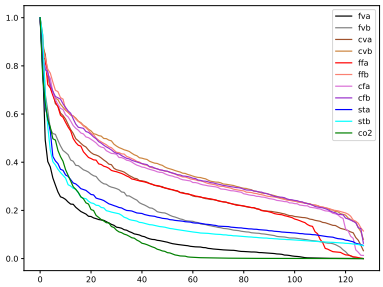

Figure 5: Normalized Singular Value of the linear layers.

advanced self-supervised learners, as they impose sufficient constraints. We then demonstrate the influence of the local linear relationship between two latent spaces and test how the model's hyper-parameters (e.g., the rank of the linear converter) will influence the performance. The detailed results are shown in the Supplementary Material.

6. Discussion

Limitation about the collection of real data. Seismic surveys for collecting real data for FWI typically involve deploying an array of sensors, such as geophones or hydrophones, to record reflected and refracted seismic waves generated by controlled sources. However, the financial and computational demands, alongside the need for expert analysis, make acquiring extensive real datasets for supervised learning extremely difficult and expensive. To our knowledge, no open-source real-data dataset is available in seismic imaging. Due to the lack of labeled real datasets in subsurface geophysics, it is hard to train and evaluate our model on real data.

Simulated data plays a crucial role in this area's research, offering large-scale, clear datasets without noise. This is vital for analyzing complex relationships between seismic data and velocity maps. Despite challenges in training models solely on real data due to the lack of labeled real data in subsurface geophysics, this is a common limitation in the seismic inversion community. The "Sim2Real" technique is well-received to transfer knowledge learned in simulation to real data (James et al., 2019). To mitigate the gap between simulation and real scenarios, we also have tested our model in velocity maps that yield physically realistic subsurface structures, i.e., Style-A and Style-B (Feng et al., 2021). Additionally, we have imposed noise to simulate more realistic measurement procedures. Our method demonstrated promising performance in both scenarios. We will explore how to train the converter with purely unpaired data and mitigate the knowledge gap of real data in our future work.

Fixed grid solutions. Our approach, along with many other data-driven methods using image-to-image translation or generative methods, duel the problem with a fixed grid. This contrasts with some physical-driven methods, which treat physical quantities as continuous functions. While physics-driven methods theoretically offer infinite resolution, they often produce overly smooth, low-resolution results in practice and struggle with large data sets and inter-dataset relationships. Conversely, although the computer vision method fixes the grid, at this fixed resolution, they can always achieve high-resolution results that surpass the physical method. The operator learning method presents a grid-free machine learning method; however, it still faces limitations (e.g., the requirements for the coverage of the input space, the smoothness of the output space, etc.), and the grid-based method is still the most adopted one of many scientific problems in both data-driven methods and traditional numerical solutions for PDEs.

7. Conclusion

We introduce a novel Auto-Linear Phenomenon in subsurface imaging, which describes the automatic emergence of a linear correlation between the latent spaces of two self-supervised autoencoders trained independently in different domains. Based on the phenomenon, we present a new framework that decouples the encoder's and decoder's training and simplifies the problem from image-to-image translation into a linear problem. In experiments, Auto-Linear achieved comparable performance and showed solid performance in a few-shot situation and robustness test.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alumbaugh, D., Commer, M., Crandall, D., Gasperikova, E., Feng, S., Harbert, W., Li, Y., Lin, Y., Manthila Samarasinghe, S., and Yang, X. Development of a multi-scale synthetic data set for the testing of subsurface CO_2 storage monitoring strategies. In American Geophysical Union (AGU), 2021a.

Alumbaugh, D., Commer, M., Crandall, D., Gasperikova, E., Feng, S., Harbert, W., Li, Y., Lin, Y., Samarasinghe, S., and Yang, X. Development of a multi-scale synthetic data set for the testing of subsurface co2 storage monitoring strategies. In AGU Fall Meeting Abstracts, volume 2021, pp. S25A–0212, 2021b.

Araya-Polo, M., Jennings, J., Adler, A., and Dahlke, T. Deep-learning tomography. The Leading Edge, 37(1): 58–66, 2018.

Chen, D., Tachella, J., and Davies, M. E. Equivariant imaging: Learning beyond the range space. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4379–4388, 2021.

Chen, Y., Dai, X., Chen, D., Liu, M., Yuan, L., Liu, Z., and Lin, Y. Image is first-order norm+ linear autoregressive. arXiv preprint arXiv:2305.16319, 2023.

Deng, C., Feng, S., Wang, H., Zhang, X., Jin, P., Feng, Y., Zeng, Q., Chen, Y., and Lin, Y. Openfwi: Large-scale multi-structural benchmark datasets for full waveform inversion. 2022.

Feng, S., Fu, L., Feng, Z., and Schuster, G. T. Multiscale phase inversion for vertical transverse isotropic media. Geophysical Prospecting, 69(8-9):1634–1649, 2021.

Feng, S., Wang, H., Deng, C., Feng, Y., Liu, Y., Zhu, M., Jin, P., Chen, Y., and Lin, Y. $\mathbf{E^{FWI}}$: Multi-parameter benchmark datasets for elastic full waveform inversion of geophysical properties. CoRR, 2023.

Feng, Y., Chen, Y., Feng, S., Jin, P., Liu, Z., and Lin, Y. An intriguing property of geophysics inversion. The Thirty-ninth International Conference on Machine Learning, 2022.

Gupta, N. Solving Forward and Inverse Problems for Seismic Imaging using Invertible Neural Networks. PhD thesis, Virginia Tech, 2023.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009, 2022.

James, S., Wohlhart, P., Kalakrishnan, M., Kalashnikov, D., Irpan, A., Ibarz, J., Levine, S., Hadsell, R., and Bousmalis, K. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12627–12637, 2019.

Jin, P., Zhang, X., Chen, Y., Huang, S. X., Liu, Z., and Lin, Y. Unsupervised learning of full-waveform inversion: Connecting CNN and partial differential equation in a loop. In Proceedings of the Tenth International Conference on Learning Representations (ICLR), 2022.

Lin, Y., Theiler, J., and Wohlberg, B. Physics-guided data-driven seismic inversion: Recent progress and future opportunities in full waveform inversion. IEEE Signal Processing Magazine, 40:115–133, 2023.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In Sixth International Conference on Learning Representations (ICLR), 2018.

Schuster, G. T. Seismic inversion. Society of Exploration Geophysicists, 2017.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. arXiv preprint arXiv:2111.08005, 2021.

Sun, J., Innanen, K. A., and Huang, C. Physics-guided deep learning for seismic inversion with hybrid training and uncertainty analysis. Geophysics, 86(3):R303–R317, 2021.

Wang, F., Huang, X., and Alkhalifah, T. A. A prior regularized full waveform inversion using generative diffusion models. IEEE Transactions on Geoscience and Remote Sensing, 61:1–11, 2023.

Wu, Y. and Lin, Y. InversionNet: An efficient and accurate data-driven full waveform inversion. IEEE Transactions on Computational Imaging, 6:419–433, 2019.

Zeng, Q., Feng, S., Wohlberg, B., and Lin, Y. Inversionnet3d: Efficient and scalable learning for 3-d full-waveform inversion. IEEE Transactions on Geoscience and Remote Sensing, 60:1–16, 2021.

Zhang, Z., Wu, Y., Zhou, Z., and Lin, Y. Velocitygan: Subsurface velocity image estimation using conditional adversarial networks. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 705–714. IEEE, 2019.

A. Appendix

A.1. Comparing with generative model-based methods

An alternative self-supervised approach for inverse problems involves pre-training a generative model on physical properties (e.g., velocity model and medical image) to capture their prior distribution. Subsequently, the model is adapted by integrating measurements and a physical model of the measurement process into the sampling process (Wang et al., 2023; Song et al., 2021). Table 1 presents a side-by-side comparison, highlighting differences between our approach and generative model-based approaches. We compared the Auto-liner and generative model-based methods in detail.

Firstly generative-model-based approaches can be trained in a purely self-supervised manner and eliminate the need for paired data. In contrast, our approach needs the paired data to train the linear converter. Generally, real data only contains one or a limited number of unlabeled data, which makes it challenging to train our approach but the generative model-based approaches can still work on that. For example, in Wang et al. (2023), a diffusion model-based method, trained with OpenFWI, was successfully applied to real marine data.

Secondly, our encoder-decoder architecture enables single-step inference, unlike generative model-based methods that usually require multiple-step denoising when the diffusion model is used.

Thirdly, in addressing the solution of FWI problems, our method aims for a maximum likelihood solution, also referred to as conditional average with minimum error, based on the distribution of training data, which can only generate one of the potential solutions. This approach is distinct from generative models, which seek distribution-to-distribution mappings and inherently can generate multiple solutions.

Additionally, generative models-based methods incorporate the forward modeling process into learning, which requires a thorough understanding of the governing physics. This approach can greatly benefit from the existing knowledge of the forward operator, particularly when forward modeling is accessible, as in the cases of FWI, CT, and MRI. However, challenges arise in scenarios where the forward modeling is partially unknown, completely unknown, or lacks an explicit formulation. For instance, in the Kimberlina carbon sequestration problem (Alumbaugh et al., 2021b), while data, generated by Maxwell's Equations, has been made available, the forward modeling remains proprietary and challenging to replicate. In such scenarios, our approach presents an effective alternative, offering a different way for inverse analysis.

Moreover, for problems like FWI, forward modeling is usually computationally expensive and time-consuming, leading to inefficient inference. For real-world data challenges, generative model-based methods can create paired data, which can be further utilized to train our Auto-Linear model, thus enhancing inference speed. This collaboration allows the Auto-Linear model and generative approaches to complement each other effectively in practical applications.

Finally, it is also important to note that generative models-based methods, while relying on forward modeling, are currently focused on inverse problems. In contrast, Auto-Linear has the capability to solve both forward modeling and inverse problem, simultaneously.

In summary, Auto-Linear and generative model-based methods each contribute uniquely to our understanding of FWI, offering insights from different perspectives. Both approaches shed light on distinct aspects of the problem and suggest diverse strategies for tackling the inherent challenges of the inverse problem.

A.2. Architecture.

The exact transformer architectures and layer dimensions of the seismic and velocity autoencoders are provided in Table 9. For all the datasets except Kimberlina, the size of seismic data is 1000×70 and the size of velocity maps is 70×70. We choose the patch size 100×10 for seismic data and 10×10 for velocity maps. Thus, the latent dimension of seismic data is 132×70, and the latent dimension of velocity maps is 516×49. For Kimberlina, the patch size of seismic data is 250×10, and of velocity maps is 20×40. The latent dimension of seismic data is 132×50, and the latent dimension of velocity maps is 516×70. The rank of the linear converter is set to 128. The mask ratio for training MAE is set to 0.75.

A.3. Generalizability.

In the following experiment, CurveVel-A, FlatFault-A, and CurveFault-A are used to pre-train the MAEs. We test the generalizability of this pair of encoder and decoder in all datasets. Results are shown in Table 10.

Model	#Layers	Embedded Dim	MLP Dim	#Heads
Seismic Encoder	2	132	528	12
Seismic Decoder	2	512	144	16
Velocity Encoder	3	516	2064	12
Velocity Decoder	2	512	2064	16

Table 9: Details of seismic and velocity autoencoders

Dataset	MAE↓	MSE↓	SSIM↑
CurveVel-A*	0.0634	0.0155	0.8267
FlatFault-A*	0.0166	0.0026	0.9698
CurveFault-A*	0.0271	0.006	0.9434
FlatVel-A	0.0072	0.0004	0.9912
FlatVel-B	0.0552	0.0179	0.8783
CurveVel-B	0.1754	0.0981	0.6157
FlatFault-B	0.1260	0.0381	0.6734
CurveFault-B	0.1837	0.0711	0.5590
Style-A	0.0744	0.0146	0.8311
Style-B	0.0653	0.0102	0.7175

Table 10: Quantitative results of the generalization ability of pre-trained encoder and decoder. The encoder and decoder are trained across datasets' families. (*) indicates the datasets used to train the encoder and decoder.

To support our claim of strong generalization, we conducted another experiment by evaluating our encoder and decoder on a completely new dataset. Since there is no suitable dataset for training the model, we created a new dataset derived from the original Marmousi velocity map. The Marmousi velocity map is a standard and complex benchmark widely recognized in the exploration geophysics community. The size of the original Marmousi is [13601, 2801]. After removing the water layer at the top, we obtained a velocity map of size [13601, 2381]. Utilizing a 70×70 sliding window, we generated a dataset with 25,000 samples, referred to as the "Marmousi slice". Additionally, we constructed another downsampled version by first downsampling the raw velocity to [2197, 731] and then applying the sliding window to produce 1,000 samples, referred to as "Marmousi downsamples" in the following. This downsampled version contains fewer samples and encapsulates more complex structures within each sample. For both datasets, we allocated 80% of the data for training the linear converter, with the remaining 20% used for validation. We tested the encoder and decoder, pre-trained on the "Fault Family" dataset, and compared the performance against supervised InversionNet trained on these two datasets as baselines. The results in Table 11 show that our pre-trained models achieve comparable performance to the supervised baselines, indicating a strong generalization ability.

Dataset	Model	MAE↓	MSE↓	SSIM↑
Marmousi slice	Auto-Linear	0.0116	0.0013	0.9770
	InversionNet	**0.0101**	**0.0012**	**0.9832**
Marmousi downsample	Auto-Linear	**0.0971**	**0.0296**	0.8020
	InversionNet	0.1094	**0.0296**	**0.8311**

Table 11: Quantitative results for generalization ability on completely different datasets.

A.4. Ablation Test.

In this part, we test the performance of several different non-linear converters and demonstrate the influence of the local linear relationship between two latent spaces. We also show the comparison between using masked autoencoder and autoencoder as self-supervised learners; and test the performance of several different non-linear converters and how the rank of the linear converter will influence the performance.

Non-Linear Converter. We evaluate networks with a more complicated nonlinear converter on CurveFault-A. We tested four different settings: 1) a two-layer MLP; 2) a two-piece Maxout layer; 3) a two-layer U-Net; and 4) a four-layer U-Net. The results are provided in Table 12. From the results, we can see that 1) a simple nonlinear mapping (e.g., two-layer MLP or U-Net) has no positive effect on final performance; and 2) a piece-wise linear mapping (Maxout) or a much more complex nonlinear mapping (four-layer U-Net) can only provide limited improvement. These results are consistent with our conclusion of a near-linear relationship.

Model	MAE↓	MSE↓	SSIM↑
Linear	0.0277	0.0061	0.9426
Two-Layer MLP	0.0280	0.0064	0.9433
Two-Pieces Maxout	0.0260	0.0057	0.9472
2-Layer U-Net	0.0285	0.0062	0.9414
4-Layer U-Net	0.0259	0.0056	0.9465

Table 12: Quantitative results on CurveFault-A with different nonlinear converters.

Metrics	Model	FlatVel-A	FlatVel-B	CurveVel-A	CurveVel-B	FlatFault-A	FlatFault-B	CurveFault-A	CurveFault-B	Style-A	Style-B
MAE↓	Original	0.0081	0.0467	0.0738	0.1820	0.0164	0.1208	0.0277	0.1791	0.0719	0.0638
	Sharing Linear	0.0191	0.0545	0.0761	**0.1709**	0.0306	**0.1198**	0.0421	**0.1697**	**0.0699**	**0.0636**
MSE↓	Original	0.0005	0.0151	0.0188	0.1051	0.0026	0.0362	0.0061	0.0697	0.0139	0.0097
	Sharing Linear	0.0015	0.0161	0.0193	**0.0963**	0.0054	**0.0339**	0.0093	**0.063**	**0.013**	**0.0094**
SSIM↑	Original	0.9888	0.9044	0.8057	0.6169	0.9701	0.6868	0.9426	0.5672	0.8423	0.7275
	Sharing Linear	0.9633	0.8827	0.8007	**0.6326**	0.9411	**0.6913**	0.9096	**0.5873**	**0.8512**	0.7248

Table 13: Quantitative results of sharing linear converter over multiple datasets, compared with original results. Both the encoder/decoder and the linear layer are shared across each dataset family. We highlight the improvement of the results after sharing the linear converter.

Local Linear Relationship. We demonstrate the property we found that each dataset shows linear relation locally, and there is a piece-wise linear relation globally over multiple datasets. In particular, we let datasets in each family share not only the encoder and decoder but also the linear converter. In other words, we use all datasets in each family to train the linear converter. We report the results and performance change in Table 13. In the table, we highlight the improvement of the results after sharing the linear converter. It is quite interesting that, generally, the datasets with a more complex subsurface structure show a performance improvement. In contrast, simpler datasets' performance drops a lot. The results come from the fact that a complex dataset covers a larger range in the latent space. The scope of simple datasets is covered by those complex ones in the same family. Thus, with more data to use, Auto-Linear achieves better results on complex datasets. But, for simple datasets, out-of-distribution data make the learning results deviate substantially from their local linear relationship.

MAE v.s. Other Self-Supervised Learning Method. In a further experiment, we utilized vanilla autoencoders (i.e., mask ratio equals zero.) and another self-supervised algorithm, Masked FINOLA-B (Chen et al., 2023), with the same architecture as the self-supervised training models. These models were pre-trained on the "Fault Family" dataset and then applied to train and validate a linear converter on CurveFault-A as an illustrative case. The reconstruction and inversion results are shown in Table 14. As demonstrated, our framework is adaptable beyond MAE to include other advanced self-supervised learners, as they impose sufficient constraints. However, a simple vanilla autoencoder cannot capture the crucial information that is necessary for both reconstructing and connecting to another domain. If we simply consider both seismic data and velocity maps as pure images and ignore the physical meaning behind them, the vanilla autoencoder would learn too many shortcuts that are only useful to reconstruct the image but lose the essential information reflecting its physical properties. This is because seismic data and velocity maps are not as diverse as natural images. On the other hand, if a model can embed the essential underlying physics information of these two quantities, it will naturally enhance the generalization ability.

Model	MAE↓	MSE↓	SSIM↑	Seismic Pre-training MAE↓	Velocity Pre-training MAE↓
Masked Autoencoder	**0.0277**	**0.0061**	0.9426	0.1703	0.0410
Autoencoder	0.0614	0.0174	0.8302	**0.0008**	**0.0005**
Masked FINOLA-B	0.0287	0.0062	**0.9430**	0.0083	0.0579

Table 14: Comparison between different pre-training strategies on CurveFault-A. In addition to the quantitative results of inversion, the mean absolute reconstruction errors (with masks) of the pre-trained models (Columns 5 & 6) are also reported.

Rank of Linear. We evaluate performances over five different numbers of ranks of the linear converter, varying from 32 to 128. The quantitative results are shown in Table 15. Results indicate that increasing the rank makes the model much larger, but the growth of the results is limited. On the other hand, decreasing the model's rank also does not reduce its capacity a lot but results in a smaller number of parameters. This allows for the balance of performance and computational cost based on specific requirements and available resources, highlighting the flexibility of our model.

Model Complexity and Mask Ratio. We test the influence of model complexity and self-supervised training hyperparameters (i.e., mask ratio). We test a simpler encoder and smaller latent dimension (namely Auto-Linear-SE), a more complex encoder and larger latent dimension (namely Auto-Linear-LE and Auto-Linear-LD), and corresponding settings for the

Dataset	Dim	#Param	MAE↓	MSE↓	SSIM↑
CurveFault-A	128*	12.3M	0.0277	0.0061	0.9426
	512	26.0M	0.0271	0.0058	0.9441
	256	16.8M	0.0274	0.0059	0.9434
	64	10.0M	0.0280	0.0064	0.9392
	32	8.9M	0.0304	0.0075	0.9300

Table 15: Quantitative results of different dimensions of bottleneck in the linear converter, and the corresponding number of parameters. As a reference, the number of parameters of InversionNet is 24.4M. (*) indicates the default decoder option.

decoder (namely Auto-Linear-SD and Auto-Linear-LD). For Auto-Linear-SE, we use a one-layer transformer as the encoder and reduce the latent dimension of seismic data to 72. For Auto-Linear-LE, we use a three-layer transformer as the encoder and increase the latent dimension of seismic data to 264. For Auto-Linear-SD, we use a one-layer transformer as the decoder and reduce the latent dimension of velocity to 252. For Auto-Linear-LD, we also use a three-layer transformer as the decoder and increase its width to 768. We also test a different masking ratio (MR) of 0.5 for the encoder and decoder separately. The results are shown in Table 16.

Model	MAE↓	MSE↓	SSIM↑
Auto-Linear	0.0738	0.0188	0.8057
Auto-Linear-SE	0.0786	0.0206	0.7928
Auto-Linear-LE	0.0685	0.0170	0.8191
Auto-Linear-SD	0.0760	0.0188	0.7880
Auto-Linear-LD	0.0729	0.0194	0.8033
Auto-Linear (encoder MR 0.5)	0.0785	0.0205	0.7941
Auto-Linear (decoder MR 0.5)	0.0744	0.0196	0.7922

Table 16: Quantitative results on CurveVel-A with different hyperparameters.

The results highlight a few key observations: 1) our approach exhibits robustness across various hyperparameter selections; 2) a small model and latent space will relatively influence the model capacity, while a more complex model can achieve even better results. This suggests that our new paradigm has a substantial potential to achieve even better results. Our primary focus is to provide novel insights and introduce a new paradigm. Thus, our choices of hyperparameters are designed to balance results and model complexity, maintaining generality rather than being specifically tailored for a specific dataset. Regarding the masking ratio, for now, we can see that a very high masking ratio (i.e., 0.75) benefits the downstream task, which is consistent with the conclusion in the original MAE paper. Further discussion of the impact of masks in representation learning and whether masking is the best pre-task for image-like data is out of the scope of this paper. We would love to explore more advanced methods for learning better representations that make the latent-space representations closer to the physical nature in our future work.

Patching size. Our choice for the current patch size ensures a balanced and reasonable patch size and the number of patches, avoiding extremes in either direction. The selected sizes—$(100, 10)$ for seismic and $(10, 10)$ for velocity—uphold broad applicability. We conducted additional experiments on CurveVel-A using two different patch sizes for seismic data: $(250, 10)$ and $(500, 14)$. The results, presented in Table 17, indicate that the patch size has a small influence on the final performance. Our model demonstrates robustness across varying patch sizes, reinforcing the logic behind our current selection.

Seismic Patch Size	MAE↓	MSE↓	SSIM↑
(100, 10)*	0.0738	0.0188	0.8057
(250, 10)	0.0779	0.0203	0.7924
(500, 14)	0.0791	0.0210	0.7891

Table 17: Quantitative results on CurveVel-A with different patch sizes. * is the default configuration.

Model's Sensitivity to the frequency of the wavelet used in seismic data generation. We conducted an additional test on CurveFault-A. With pretraining the seismic masked autoencoder on seismic data generated by a 15Hz source wavelet, we train the linear converter with input seismic data generated by a 25Hz source wavelet. The results are displayed in Table 18. It can be seen that our model, working on the spatial-temporal domain (an image-to-image translation) rather than the frequency domain, is compatible with varying frequencies. Additionally, our model can also benefit from higher frequencies, which typically offer higher resolution and sensitivity within a certain depth.

Source Frequency	MAE↓	MSE↓	SSIM↑
15Hz*	0.0277	0.0061	0.9426
25Hz	0.0181	0.0036	0.9626

Table 18: Quantitative results on CurveFault-A with different source frequencies. * is the default configuration.

A.5. Comparing the latent representations of Auto-Linear and InvLINT.

To further analyze the relationship between our Auto-Linear and InvLINT, in this part, we compare the latent representations of seismic data and velocity maps obtained by our method to those obtained by InvLINT. First, We conducted experiments on CurveFault-A that use a sine kernel from InvLINT as the encoder and use our pre-trained decoder to construct the inversion network, respectively. The converter is still linear. The results are shown in Table 19. These results show that using the latent seismic representation from the sine kernel is difficult to regress the latent velocity representation from our method.

Model	MAE↓	MSE↓	SSIM↑
Auto-Linear	0.0277	0.0061	0.9426
Sine Kernel Encoder	0.0426	0.0093	0.9233

Table 19: Comparison between latent representations of seismic data obtained by Auto-Linear and InvLINT on CurveFault-A.

To further compare the latent representations, we use one latent representation to predict another with linear regression, for seismic data and velocity maps respectively. We report the coefficient of determination (R^2 score) in Table 20.

Variable	Source	Target	R^2
Seismic	Auto-Linear	InvLINT	0.9869
	InvLINT	Auto-Linear	0.6700
Velocity	Auto-Linear	InvLINT	0.9996
	InvLINT	Auto-Linear	0.4871

Table 20: Predicting the target latent representations from the source latent representations with linear regression.

These show that our latent space with a higher dimension contains more information. As a preliminary comparison, we can roughly conclude that their latent space is a linear subspace of our latent space.

A.6. Step forward for more complicated seismic models

Currently, we focus on the acoustic FWI, which is based on one-to-one mapping. For more complicated seismic forward models like the elastic wave equation that includes both P- and S-wave, the scenario shifts to a multiple-input to multiple-output problem.

In light of this, we conducted an additional experiment using our Auto-Linear framework on the elastic FWI dataset, $\mathbb{E}^{\mathbf{FWI}}$ (Feng et al., 2023). We trained four Masked Autoencoders (MAEs): two for the different wave types and two for the velocity maps, using the “Vel Family" dataset. For training the linear converters, we implemented two linear layers to map the embedding of ux and uz into the 128-dimensional bottlenecks, individually. These embeddings were then concatenated and fed through two additional linear layers to produce the latent representations with appropriate dimensions of the respective decoders. The results, shown in Table 21, from the FlatVel-A and CurevVel-A datasets are promising, especially when compared to the ElasticNet benchmark mentioned in the original paper. It is important to note that these are preliminary results, based on our existing one-to-one mapping framework without fine-tuning any hyperparameters. There is still room to explore more effective ways to integrate the four latent spaces, such as how to combine the latent representations from different wave types. However, even at this early stage, our method already achieves a similar level of performance with the supervised benchmarks. This gives us confidence in the potential of the Auto-Linear’s applicability for complex seismic models. We will keep exploring these latents’ properties across a broader spectrum of tasks in our future work.

A.7. Evaluation on the Recovery of Faint Seep Reflectors

To further evaluate the performance of our method, we show extra evaluation results of the conventional physics-driven seismic imaging methods. To evaluate the detection of subtle subsurface reflectors at considerable depths, we apply the reverse time migration (RTM) technique. RTM is a computational approach utilized in seismic imaging, enabling the creation of high-resolution visualizations of subsurface structures beneath the Earth’s surface. Besides, we perform a

Dataset	Model	V_P			V_S			Pr		
		MAE↓	RMSE↓	SSIM↑	MAE↓	RMSE↓	SSIM↑	MAE↓	RMSE↓	SSIM↑
$\mathbb{E}^{\mathbf{FVA}}$	Auto-Linear	0.0330	0.0672	0.9403	0.0241	0.0520	0.9425	0.0366	0.0752	0.7719
	ElasticNet	0.0308	0.0559	0.9615	0.0259	0.0500	0.9596	0.0329	0.0664	0.8455
$\mathbb{E}^{\mathbf{CVA}}$	Auto-Linear	0.0781	0.1332	0.8052	0.0607	0.1039	0.8081	0.0771	0.1280	0.4721
	ElasticNet	0.0745	0.1345	0.8055	0.0600	0.1080	0.8051	0.0574	0.1156	0.5766

Table 21: Quantitative results of Eelastic FWI.

zero-offset least square reverse-time migration (LSRTM) of the predicted velocity maps under the Born approximation scenario in order to show a evaluation of the recovery of deep reflectors. The LSRTM is calculated for one predicted sample from each dataset used in the paper, with 20 optimization iterations.

For comparison purposes, we also conduct physics-driven FWI (Schuster, 2017) alongside its corresponding RTM. As Auto-Linear obviates the need for an initial model and avoids data preprocessing in this study, to ensure an equitable comparison, the physical FWI is executed with a uniform background and data generated using a 15 Hz Ricker wavelet without applying any bandpass filtering. The outcomes are presented in Figure 6. It is evident that Auto-Linear outperforms Physical FWI in restoring the velocity maps. However, due to the survey's restricted aperture, certain deep reflectors exhibit suboptimal recovery. Nonetheless, the RTM images produced through Auto-Linear still offer improved accuracy and finer details regarding the reflector positions compared to those obtained via physical FWI.

A.8. Visualizations

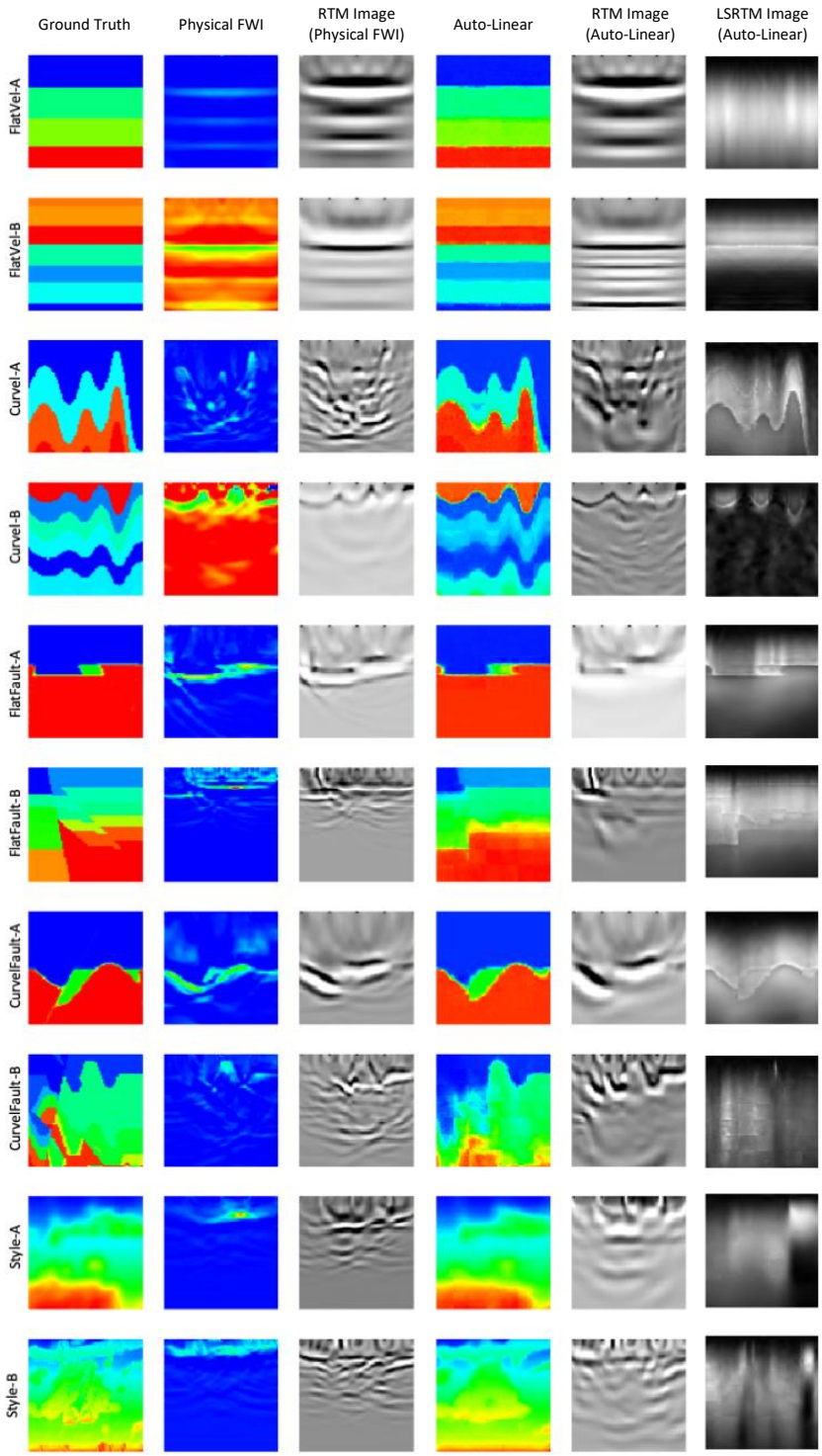

Figure 6: Ground truth, velocity maps and RTM images obtained with physics-driven FWI versus velocity maps and RTM images obtained with Auto-Linear

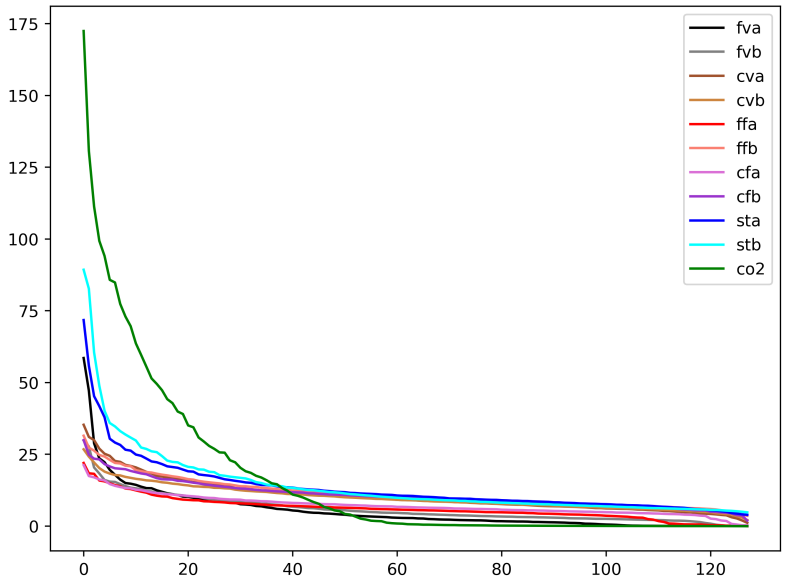

Figure 7: Original Results of Singular Value Decomposition on different datasets.

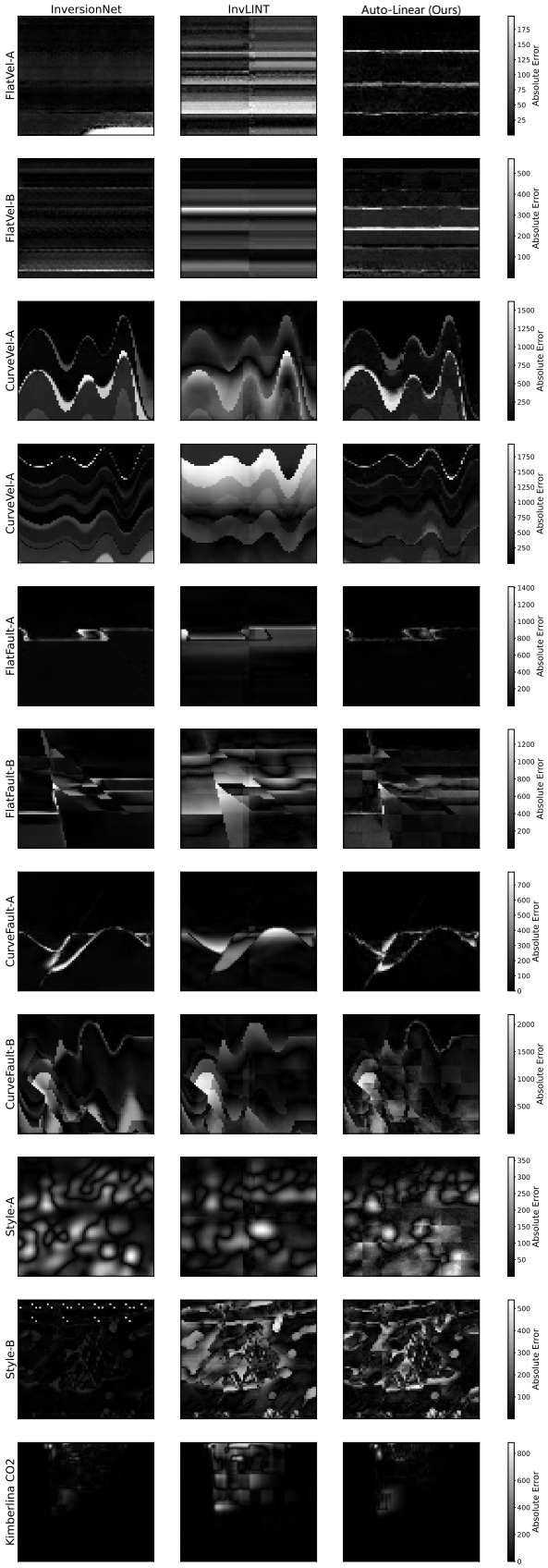

Figure 8: Illustration of absolute error map on OpenFWI, compared, Auto-Linear, InversionNet (Wu & Lin, 2019) and InvLINT (Feng et al., 2022) to the ground truth.

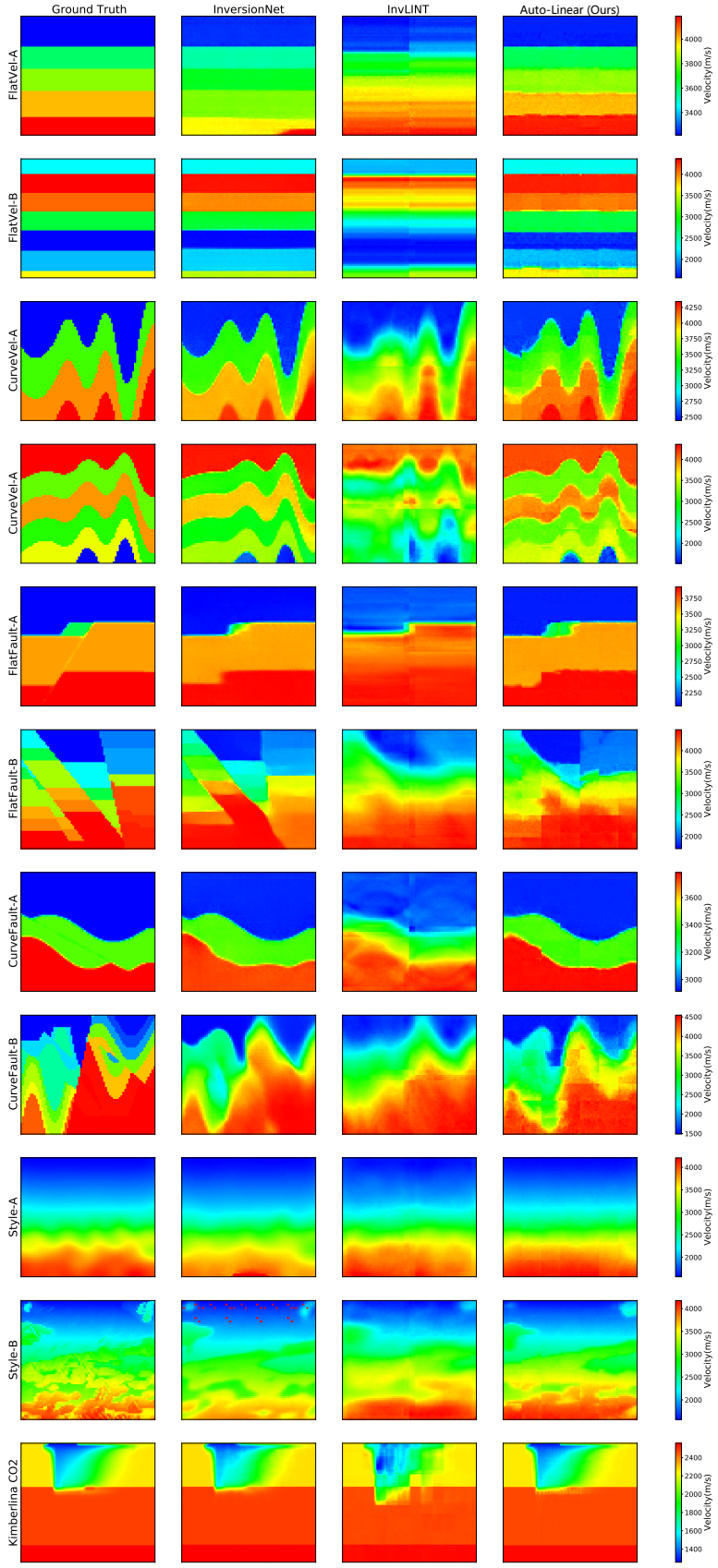

Figure 9: Illustration of results evaluated on OpenFWI, compared with InversionNet (Wu & Lin, 2019) and InvLINT (Feng et al., 2022).

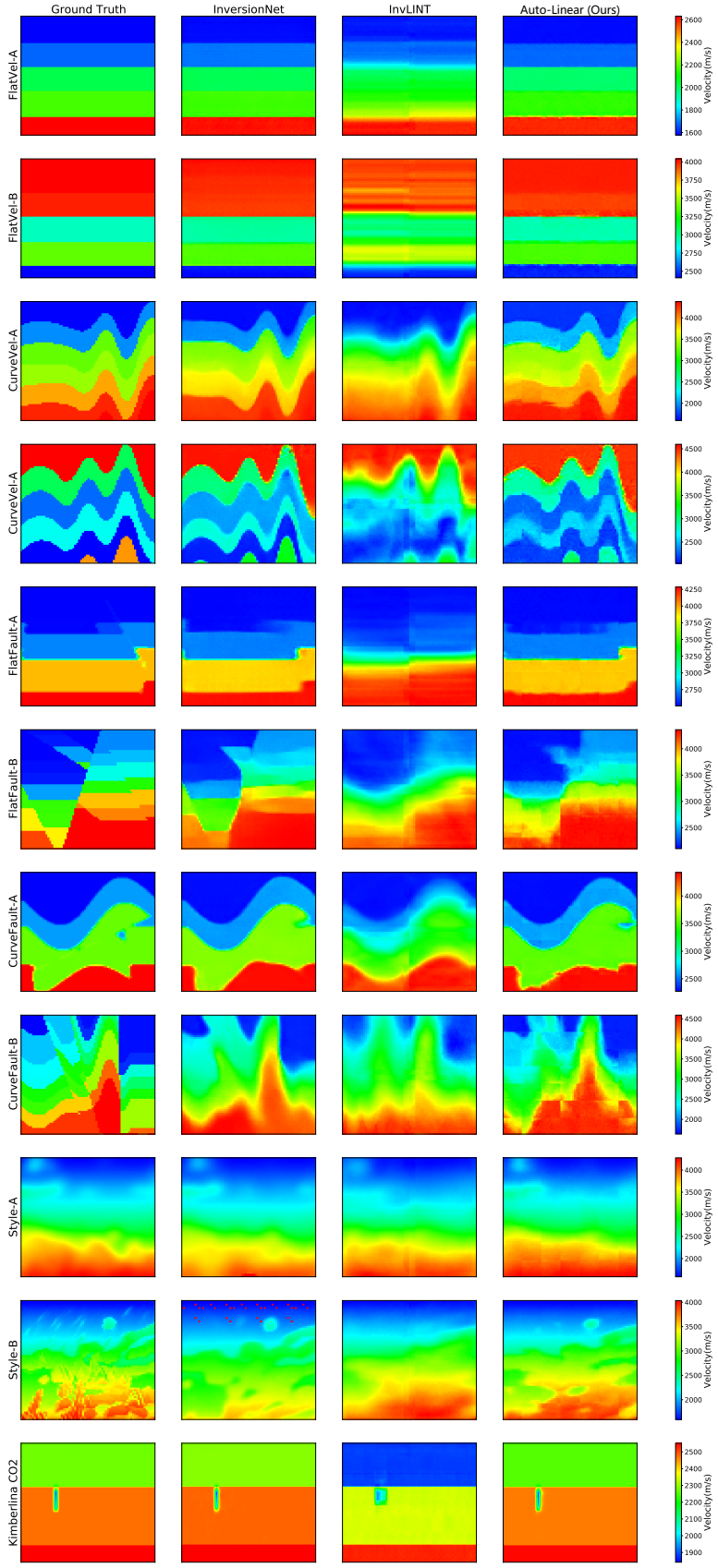

Figure 10: Illustration of results evaluated on OpenFWI, compared with InversionNet (Wu & Lin, 2019) and InvLINT (Feng et al., 2022).