Data-Efficient Inference of Neural Fluid Fields via SciML Foundation Model

Yuqiu Liu[1*], Jingxuan Xu[2*], Mauricio Soroco[1], Yunchao Wei[2,3†], Wuyang Chen[1†]
[1]Simon Fraser University [2]Beijing Jiaotong University [3]Beijing Academy of Artificial Intelligence

Abstract

*Recent developments in 3D vision have enabled significant progress in inferring **neural fluid fields** and realistic rendering of fluid dynamics. However, these methods require dense captures of real-world flow, which demand specialized lab setups, making the process costly and challenging. Scientific machine learning (**SciML**) **foundation models**, pretrained on extensive simulations of partial differential equations (PDEs), encode rich multiphysics knowledge and thus provide promising sources of domain priors for inferring fluid fields. Nevertheless, the transferability of these foundation models to real-world vision problems remains largely underexplored. In this work, we demonstrate that SciML foundation models can significantly reduce the **data costs** of inferring real-world 3D fluid dynamics with improved generalization. Our method leverages strong forecasting capabilities and meaningful representations of SciML foundation models. We equip neural fluid fields with a novel collaborative training that utilizes augmented frames, and fluid features extracted by our foundation model. We demonstrate significant advancements in both quantitative metrics and visual quality over previous methods, improving 9∼36% peak signal-to-noise ratio (PSNR) in future prediction with 25∼50% reduction in the number of training frames, thereby showcasing the practical applicability of SciML foundation models in real-world fluid dynamics. We release our code at: https://github.com/delta-lab-ai/SciML-HY.*

1. Introduction

Fluid phenomena are ubiquitous in our 3D world, from the powerful ocean currents, to the turbulent jet streams in the air. One important yet open challenge in understanding fluids is to recover fluid dynamics from visual observations, also known as the inference of **3D fluid fields**. Formally stated, given visual inputs (2D images or video sequences), this task aims to infer invisible quantities like density and velocity in the spatiotemporal domain (3+1D) (Figure 1 left). This facilitates downstream rendering of realistic fluids in computer games and videos [71], and even applications of broad impact, such as weather forecasting [52] and airfoil design [65]. Unlike rigid bodies, fluids present a unique challenge due to their dynamic and complex nature, requiring advanced computational methods.

Recent advancements in 3D vision have enabled significant progress in inferring fluid fields. This includes both multi-view benchmarks [17] of high-quality flow videos with well-calibrated camera poses, and also neural fluid fields [11–13, 27, 81] jointly optimized by rendering loss and physics constraints. However, learning neural fluid fields is notorious for its **high costs of acquiring dense fluid views** [1]. Methods like HyFluid [81] require four videos with 120 continuous frames each. This requirement relies on *specialized* lab setups. For example, collecting and calibrating the ScalarFlow dataset [17] requires insulated containers with heaters, fog machines with servo-controlled valves, and multiple cameras, with an estimated total cost of around $1,100. Many fluid dynamics phenomena occur rapidly, necessitating the use of high-speed cameras to capture detailed visualizations. These specialized imaging systems can add significant expenses to experimental setups, with costs reaching thousands of dollars per camera. [1, 82]. With mobile devices or drones, capturing real-world fluid views in the wild will become even more challenging.

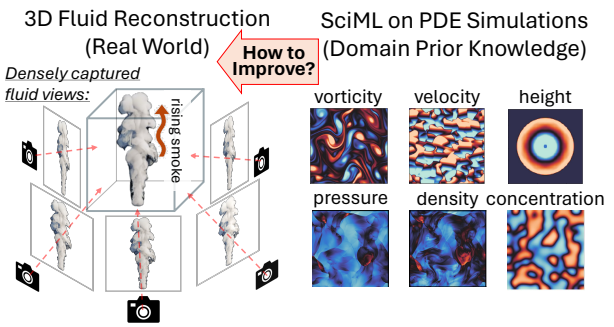

Figure 1. Inferring neural fluid fields requires densely captured views. Meanwhile, PDE simulations are important for building SciML foundation models. How to utilize this rich domain knowledge to improve 3D fluid reconstruction in the real world?

*The first two authors contributed equally.
†Co-corresponding authors.

[1]"Frame" and "view" are used interchangeably in the context of fluid field reconstruction, following [81]. We will unify them under "frame" to avoid ambiguity.

A common strategy to achieve data efficiency and improved generalization is to introduce prior knowledge. Scientific machine learning (SciML), which aims to learn physical dynamics, is a promising source of prior (Figure 1, right). Deep neural networks (DNNs) provide surrogate models for approximations of partial differential equations (PDEs) and real-world challenges like weather forecasting [41, 52] and turbulent flow [34]. **SciML foundation models** are further advanced in recent works [28, 29, 46, 54, 60, 62, 70, 80]. By scaling up extensive training datasets to incorporate multiphysical domains and PDE simulations (such as Navier-Stokes, Burgers', shallow water), SciML foundation models aim to encode common physical behaviors and improve generalization in scientific contexts.

Although promising, SciML foundation models are mainly pretrained and evaluated on synthetic PDE simulations [48, 61, 63]. These simulations, while encoding rich physical domain knowledge, still differ from real fluid captures with multiscale patterns and noisy measurements (Figure 1). This poses questions about the transferability of SciML foundation models in real-world 3D fluid problems. In contrast, foundation models in popular ML domains have been widely utilized as strong priors. Vision models such as DINO [6, 51] and CLIP [53] have been leveraged to support generalizable representations and semantic awareness [7, 45, 64, 68, 73, 75, 79]. Large language models (LLMs) [2, 16, 66, 67] are pretrained on high-quality corpora and interact with the informal spoken language of human users every day. Therefore, we ask our core question:

How to utilize SciML foundation models to advance 3D reconstruction of real-world fluids?

In this work, we provide affirmative answers (Figure 2), and demonstrate that pretrained SciML foundation models can enhance data efficiency in inferring neural fluid fields from sparse videos. We establish the foundation for incorporating pretrained physics knowledge as a prior for real-world fluid reconstruction. Our core idea is to leverage the strong forecasting and meaningful representations of SciML foundation model, and "distill" this prior into neural fluid fields. We demonstrate both improved quantitative metrics and the high-quality visual appearance of our method on real-world flow captures with significantly reduced training input frames. Specifically:

1. Given extremely sparse initial frames from short videos of flows, our foundation model forecasts future steps (temporal frames) and enables a collaborative training strategy for neural fluid fields with more augmented fluid frames (Section 3.2 and Figure 6).
2. To improve generalization, we introduce meaningful representations of flows into neural fluid fields. These representations are extracted by our foundation model and carefully aligned with the camera rays used in the fluid field (Section 3.3 and Figure 7).

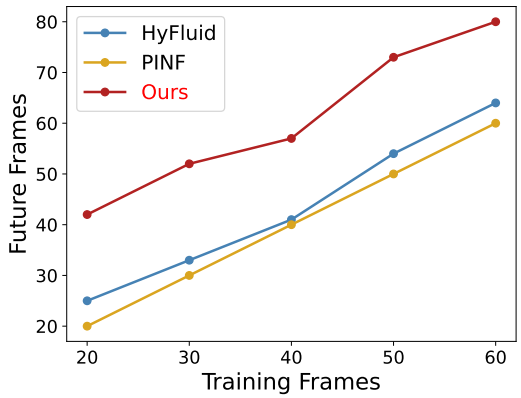

Figure 2. Our method is significantly more data-efficient than previous works (PINF [11], HyFluid [81]) on future prediction. X-axis: different numbers of training frames (`nf`) per video. Y-axis: temporal index of reliably predicted future frames with peak signal-to-noise ratio (PSNR) threshold of 25 (higher is better).

3. We provide comprehensive experiments and ablation studies (Section 4). Our method not only unlocks extreme data efficiency (25∼50% reduction in the number of training frames), but also achieves both improved reconstruction error and visual quality (10∼36% improved peak signal-to-noise ratio in future prediction).

2. Preliminary

2.1. Inferring Fluid Fields from Videos

Our method is developed to work with HyFluid [81], which infers neural fluid fields (density and velocity) from videos.

Problem Definition. Given videos of smoke rising upwards (Figure 1, left), with the number of frames (views) used in each video denoted by `nf`, neural fluid fields aim to infer the density field $\sigma(x, y, z, t)$ and the underlying velocity field $\boldsymbol{u}(x, y, z, t) = (v_x, v_y, v_z)$ of the smoke, both parameterized by deep networks.

For the density field $\sigma(x, y, z, t)$, HyFluid randomly samples camera rays (x, y, z, t) and reconstructs the density using a 4D extension of iNGP [49], which accelerates the neural rendering with multiscale hash encoding of spatiotemporal positions. During training, this density field is optimized by comparing input and rendered views via differentiable volume rendering (Figure 4 bottom). Similarly, the velocity field $\boldsymbol{u}(x, y, z, t)$ is inferred by another iNGP model, and is supervised by physics-informed losses that enforce mass conservation for incompressible flows and divergence-free velocity. We follow the assumptions of the original ScalarFlow dataset, whose reconstruction model assumes incompressible flow (see Section 3 of [17]). Under atmospheric pressure and low Mach numbers, this is a standard and reasonable approximation for smoke.

During inference, the density field is used to render the visual appearance of the smoke, and the learned velocity

field can be used to advect (evolve) the density over temporal steps for both *re-simulation* (interpolation of the temporal range seen during training) and *future prediction* (extrapolation of unseen future temporal ranges).

The ScalarFlow Dataset: Smoke Videos with Calibrated Cameras. Recent works on fluid field reconstruction focus on the ScalarFlow dataset [17]: a comprehensive collection of volumetric reconstructions of real-world smoke plumes (Figure 1 left). It encompasses a wide array of complex, buoyancy-driven flows rising upwards that transition into turbulence, capturing observable scalar transport processes. To the best of our knowledge, the ScalarFlow dataset is by far the best-calibrated benchmark on real-world fluid (smoke) dynamics.

2.2. SciML Foundation Model

For time-dependent PDEs, the solution is a mapping from the joint of a spatial and temporal domain to the dynamics of the physical system (e.g. density, velocity, vorticity of the fluid at a certain spatiotemporal location): $\mathbf{v} := \mathcal{T} \times \mathcal{S} \rightarrow \mathbb{R}^d$. In current literature [41, 42, 46, 63], the **forward modeling** operator $\mathcal{N}$ computes the PDE solution given $T_{in} \in \mathbb{Z}^+$ consecutive previous timesteps: $\mathcal{N} := \mathbf{v}(t - T_{in} \cdot \Delta t, \cdot), \ldots, \mathbf{v}(t - \Delta t, \cdot) \mapsto \mathbf{v}(t, \cdot)$, where Δt is the granularity of the temporal grid. This enables finite-difference approximations of the temporal derivatives of PDEs. See Figure 3 for an illustration.

SciML aims to find ML-based surrogate models for forward modeling by learning an approximation from data $\mathcal{N}_\phi \simeq \mathcal{N}$ (ϕ for learnable parameters). To optimize $\mathcal{N}_\phi$, we take a dataset $\mathcal{D}$ comprising N discretized PDE simulations ("samples") $\mathcal{D} := \left\{\mathbf{v}^{(i)} \mid i = 1, \ldots, N\right\}$, and minimize a loss functional L, typically the normalized root of the mean squared error ($\mathrm{nRMSE} \equiv \frac{\|\mathbf{v}_{\text{pred}} - \mathbf{v}\|_2}{\|\mathbf{v}\|_2}$ where $\mathbf{v}_{\text{pred}}$ is the prediction from $\mathcal{N}_\phi$).

Traditionally, SciML models focus on learning simulations of one PDE [41, 42, 44]. However, recent works explored and verified benefits of scaling up the pretraining data to include diverse PDE systems, thus developing **SciML foundation models** [28, 29, 46, 54, 60, 62, 70, 80]. Intuitively, although these PDEs model very different physical systems, this "multi-tasking" strategy 1) implicitly enforce the learning of the compositionality of PDEs (which describe core components like nonlinear advection or diffusion in common and also augment specialized terms like buoyancy or system constraints); 2) facilitate transfer learning and knowledge sharing across multiple PDE families.

3. Methods

In our work, we aim to reduce the number of video frames (`nf`) required by learning neural fluid fields, thereby improving data efficiency. Our method can be applied to any NeRF-based fluid models, in this paper we mainly use HyFluid as our baseline, see the results on other baselines (PINF [11]) in Section A.8 of the supplementary material. In Figure 4, we overview our proposed framework[2].

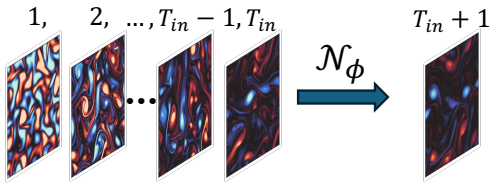

Figure 3. Forecasting by SciML foundation models [29, 46]. Given T_{in} previous steps, the model predicts the next step of the fluid dynamics (here, each frame shows the vorticity of the fluid).

3.1. How to Utilize SciML Foundation Models for Inferring Real-World Fluid Fields?

Inspired by recent works [29, 46], we first develop our SciML foundation model as follows:

1. **Architecture.** We adopt a 3D version of the Swin Transformer [43, 78] (6.5M parameters), a popular vision transformer architecture, as our foundation model[3]. It tokenizes input temporal 2D frames ($\mathbf{v}([t - T_{in} \cdot \Delta t : t - \Delta t], \cdot)$) with a 3D convolution layer, forwards through efficient windowed attentions, and predicts the next temporal step $\mathbf{v}(t, \cdot)$. Without loss of generality and following previous works [29, 46, 63], we choose $T_{in} = 10$. Tuning T_{in} may lead to better performance, but is not the focus of our method.
2. **Multiphysics Pretraining.** We utilize the PDEBench dataset [63] for pretraining. Specifically, we pretrain our foundation model on the joint of diverse simulations of the following PDEs: both compressible and incompressible Navier-Stokes, shallow water, and reaction-diffusion. See Section B.4 in the supplement for details.We sample each equation uniformly, zero-pad channels of PDEs with fewer variables, and match different PDE simulations to the same spatial resolution via interpolation. We will experimentally verify the benefits of this multiphysics pretraining in Section 4.4.
3. **Fine-tuning.** After pretraining, we fine-tune on ScalarFlow[4]. Inspired by recent works [39, 52], we em-

[2] Our method directly interacts with only the density field of HyFluid, the velocity field is implicitly improved via the density field.

[3] (1) Why not use larger model sizes? We will show that, even with a small model, we can already achieve strong improvements. Using larger models may further boost performance, but this is not the focus of our work. Recent SciML foundation models also consider sizes smaller than 10M parameters [29, 62, 69, 80]. (2) Architecture choices: We adhere to the original design of the Swin Transformer and avoid introducing ad hoc modifications. Although recent works on SciML foundation models adopt different architectures [29, 46], the commonly shared aspect of these works is their joint multiphysics pretraining, not their deep network architectures.

[4] The same set of sparse video frames that will be used to train HyFluid.

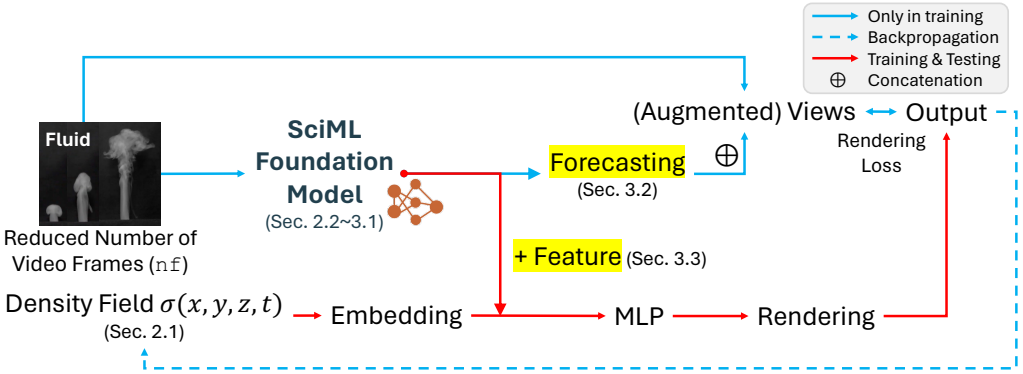

Figure 4. Overview: We improve the data efficiency (i.e., reduce the number of input fluid frames "nf") of learning neural fluid fields via the pretrained SciML foundation model. Given sparse input videos, we utilize our foundation model to: 1) forecast future steps to augment denser frames for training (Section 3.2); 2) extract flow representations and aggregate into embeddings of fluid density fields (Section 3.3).

ploy a curriculum schedule to encourage forecasting further temporal steps, gradually increasing autoregressive steps from 3 to 8 by 1 every 20 training epochs. Both pretraining and fine-tuning use the nRMSE loss.

We expect **two core benefits** of our SciML foundation model that can be utilized in the real world (highlighted with yellow in Figure 4):

1. **Strong Forecasting.** As our foundation model is pretrained with the next-frame prediction, it can natively forecast precise future steps as augmented frames of fluids to complement sparse videos (Section 3.2).
2. **Representation Learning.** As a data-driven approach similar to DINO [6, 51], the feature space constructed by our SciML foundation model can extract meaningful features of fluids to facilitate better generalization of 3D neural fluid fields (Section 3.3).

3.2. Co-Training via Foundation Model Forecasting

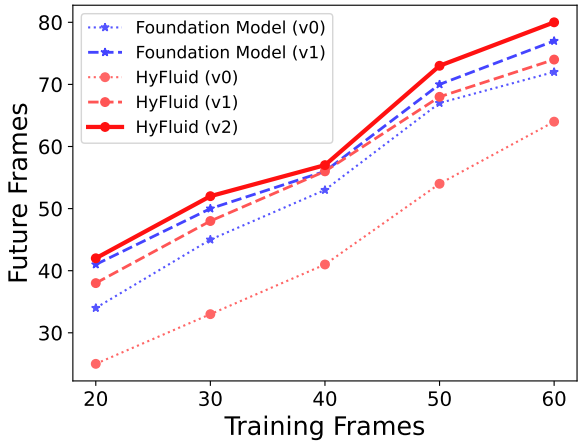

Figure 5. Collaborative training between HyFluid and the foundation model improves future predictions. "v0, v1, v2" match models annotated in Figure 6. HyFluid can be progressively improved (v0→v1→v2) with more augmented frames. Y-axis: temporal index of reliably predicted future frames (thresholded by PSNR=25). X-axis: number of training frames (nf) per video.

Given sparse smoke videos, one way to address data scarcity is to augment more frames. We first study the forecasting performance of both our SciML foundation model and the neural fluid fields. As shown by two dotted curves in Figure 5 ("Foundation Model v0" vs. "HyFluid v0"), the forecasting quality of the foundation model is much better than the neural fluid fields.

To utilize the strong forecasting of our foundation model, we propose a collaborative training strategy for neural fluid fields. The core idea is to train the foundation model and neural fluid field with augmented frames (Figure 6). We alternately concatenate the reliably predicted frames (thresholded by PSNR=25) from the foundation model or neural fluid field into the current training set, and fine-tune each other. This collaborative training can also be viewed as a knowledge distillation from the foundation model to the neural fluid field in the output space. As shown by two dashed curves in Figure 5 ("Foundation Model v1", "HyFluid v1"), the collaborative training enhances the forecasting of both models, and the final version

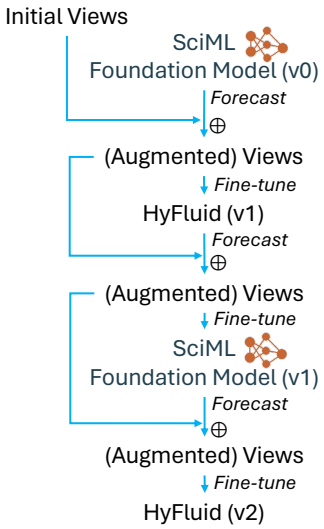

Figure 6. Collaborative training between HyFluid and our SciML foundation model via forecasting with augmented frames. "v0, v1, v2" match the corresponding curves in Figure 5.

of the neural fluid field (solid curve "HyFluid v2") achieves much stronger future predictions. By achieving comparable PSNR with fewer input frames, we demonstrate that our collaborative training can significantly improve the data efficiency of neural fluid fields. Notably, while the foundation model itself is data-driven and does not explicitly encode the Navier–Stokes equation, the training of HyFluid (v2) over extended temporal frames still introduces new physical knowledge through regularization from fluid simulation.

3.3. Feature Aggregation from Foundation Model

In addition to leveraging the augmented frames via the foundation model's forecasting, we further aggregate the learned representation from the foundation model into the neural fluid field. This can be viewed as a knowledge distillation from the foundation model to the neural fluid field in the feature space. We show our design of feature aggregation in Figure 7. This includes three steps:

1. For each camera ray (x_p, y_p, z_p), we use the camera's extrinsics and intrinsics to project the ray onto the position in image coordinates (h_{img}, w_{img}).
2. We reshape the sequence of tokens in our foundation model into 2D feature maps, and extract the feature vector corresponding to the camera ray via interpolating over the neighboring four feature coordinates. This feature vector is shared by all points sampled along the ray.
3. We use a two-layer MLP (with ReLU activation) to map the feature vector to the same feature dimensionality as the embedded features of the spatiotemporal coordinates of the density field, and sum them for aggregation.

During training, features are extracted from fluid frames from videos. During testing, since videos are not accessible, the SciML foundation model extracts features based on frames rendered by the density field from prior temporal steps. To extract features of frames before the temporal step at T_{in}, we use temporal-wise interpolation to supplement necessary frames as inputs to the foundation model.

4. Experiments

4.1. Settings

Datasets. We use real captures from the ScalarFlow dataset [17], released in the repository of HyFluid. For each scene, there are five videos from five cameras fixed at positions evenly distributed across a 120° arc centered at the rising smoke. In each video, we consider the first `nf` frames (where the smoke plumes upwards from the bottom), and adjust `nf` to study our data efficiency. Each video has a resolution of 1920×1080. These videos have been post-processed to remove backgrounds. Following HyFluid, for each scene, we use four videos for training and hold one out for testing (i.e., as the ground-truth novel view).

Tasks. We compare with two previous works on neural fluid fields: PINF [11] and HyFluid [81]. Due to the lack of true 3D volume in ScalarFlow, we evaluate the reconstruction quality using view rendering. Following [81], we consider three tasks: novel view synthesis, re-simulation, and future prediction. In novel view synthesis, the density field is used to render smoke views from unseen camera parameters; thus, its quality is evaluated based on rendering accuracy. For re-simulation and future prediction, the learned velocity field is utilized to advect the density across temporal coordinates. Thus, the quality of the learned velocity field is assessed based on its effect on the density field. In our future prediction experiments, no model is ever trained with ground-truth future frames from videos. We refer the reader to [81] for more details about these tasks.

Evaluation Metrics. We report the peak signal-noise ratio (PSNR) averaged over frames. We leave the structural similarity index measure (SSIM) and the perceptual metric LPIPS [83] in Section A.6 in Section A.6 in the supplementary material. These metrics are also widely used in previous deblurring works [21, 22].

4.2. Data-Efficiency Inference of Fluid Fields

We first report the inference of fluid density fields. By default, HyFluid [81] and PINF [11] used 120 frames (i.e., `nf`=120) from each video during training. We consider using a much fewer numbers of sparse training frames than HyFluid and PINF. During collaborative training, we use 20 augmented frames (from `nf`+1 to `nf`+20) in each round. These predicted frames are refreshed rather than accumulated. Our method improves both data efficiency and per-

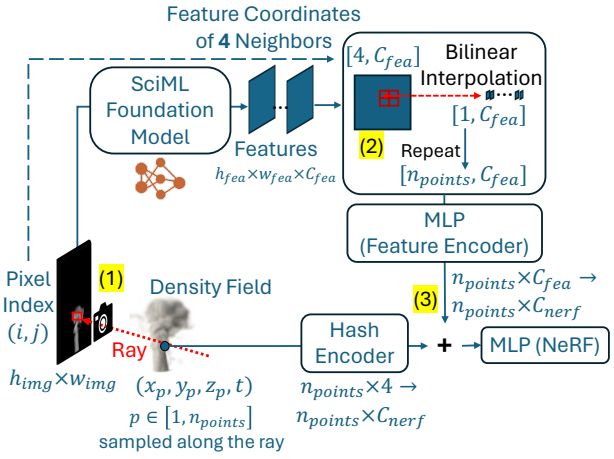

Figure 7. We aggregate representations learned by our SciML foundation model into HyFluid with three steps: (1) project from spatiotemporal location to the camera plane; (2) extract and interpolate neighboring features; (3) aggregate features from the foundation model into neural fields. "C_{fea}": feature dimension of SciML foundation model. "C_{nerf}": hidden dimension of neural density field (NeRF).

Table 1. Comparing PSNR (higher the better) of fluid field reconstruction by different methods. We report mean values over 3 random runs (see Table 5 in the supplement for standard deviations). "nf": number of input training frames (views). For future prediction, we report the PSNR averaged over 20 future frames (i.e., frames with indices from nf+1 $\rightarrow$ nf+20).

Methods	Novel View Synthesis			Re-Simulation			Future Prediction		
	nf=20	nf=40	nf=60	nf=20	nf=40	nf=60	nf=20	nf=40	nf=60
PINF [11]	33.45	31.05	30.90	24.28	24.86	24.08	21.71	20.85	20.67
HyFluid [81]	33.83 (+0.38)	33.32 (+2.27)	**32.84** (+1.94)	33.89 (+9.61)	33.27 (+8.41)	32.02 (+7.94)	25.22 (+3.51)	23.98 (+3.13)	23.66 (+2.99)
Ours	**34.50** (+1.05)	**33.48** (+2.43)	**32.84** (+1.94)	**34.34** (+10.06)	**33.36** (+8.50)	**32.42** (+8.34)	**27.59** (+5.88)	**28.36** (+7.51)	**27.76** (+7.09)

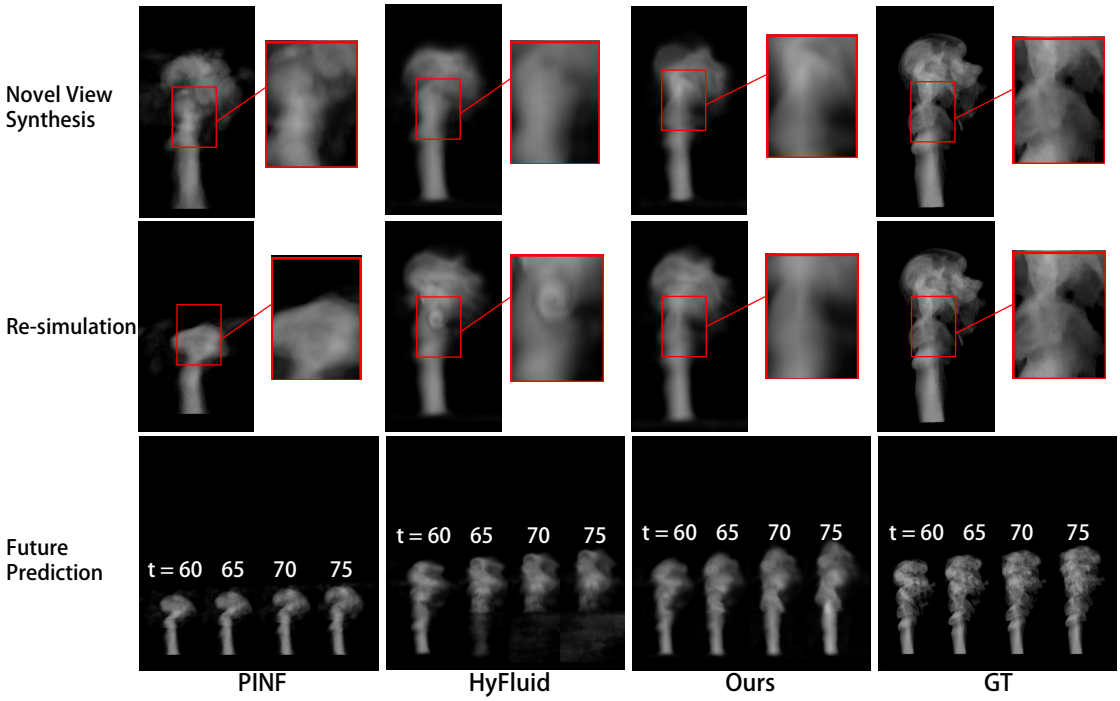

Figure 8. Visualization of novel view synthesis (top), re-simulation (middle), future prediction (bottom) on ScalarFlow [17] when 60 frames (per video) are used for training (i.e., nf=60). "GT": ground truth.

formance. As shown in Table 1, our PSNR consistently outperforms HyFluid and PINF under different sparse training frames (nf), across all three tasks[5]. Most importantly, in future prediction, our method can improve PSNR by 9% (27.59 vs. 25.22) and up to 36% (28.36 vs. 20.85) compared to HyFluid and PINF, respectively. This strong and reliable future prediction further contributes to a 25$\sim$50% reduction in the number of training frames (Figure 2), achieving significant data efficiency.

4.3. More Realistic Visual Quality

Besides measuring PSNR, it is crucial to visually assess the rendering quality of different methods to ensure realistic and artifact-free reconstructions. We present qualitative comparisons in Figure 8. For both novel view synthesis and re-simulation, our method successfully recovers fine-grained details while mitigating artifacts in HyFluid and PINF. Our ability to accurately reconstruct density fields further unlocks high-fidelity future predictions. Even when provided with sparse input frames, our method is significantly more stable and robust than HyFluid and PINF, which suffer from degraded reconstructions and weak forecasting capabilities. Our approach preserves the original structure of the fluid while maintaining a natural and physically consistent upward flow. Both the PSNR measurements and qualitative visualizations strongly indicate that our reconstruction is quantitatively superior and visually more realistic than those produced by HyFluid and PINF.

4.4. Benefits of SciML Pretraining

As the core of our SciML foundation model is the joint pretraining on diverse PDE simulations (Section 2), it is critical to study and verify the true benefits of the domain knowledge from multiphysics pretraining.

To compare different PDE sources, we also pretrain another SciML model on the Maxwell equations, which govern electromagnetic waves and largely differ from the Navier-Stokes equations of ScalarFlow[6].

[5] PSNRs across different nfs are not comparable, since the numbers of testing frames used to calculate PSNR are also adjusted to be equal to the numbers of training frames.

[6] Simulation settings for Maxwell are in Section B.4.1.

Table 2. Benefit of multiphysics pretraining on the PSNR (higher the better) of fluid field reconstruction. "nf": number of input training frames.For future prediction, we report the PSNR averaged over 20 future frames (i.e., frames with indices from nf+1 → nf+20).

Methods	Novel View Synthesis			Re-Simulation			Future Prediction		
	nf=20	nf=40	nf=60	nf=20	nf=40	nf=60	nf=20	nf=40	nf=60
No Pretraining	34.77	32.83	32.29	34.12	32.97	**32.51**	26.58	25.92	26.61
+Multiphysics Pretraining	**34.50**	**33.48**	**32.84**	**34.34**	**33.36**	32.42	**27.59**	**28.36**	**27.76**

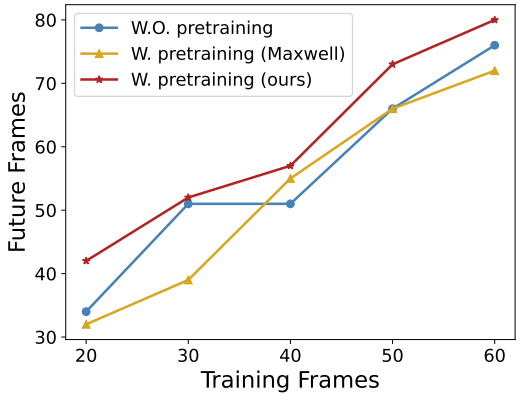

Figure 9. Benefit of multiphysics pretraining on future prediction over different numbers of initial training frames per input video (x-axis). We show the temporal index of reliably predicted future frames (thresholded by PSNR=25) on the y-axis (higher is better).

We analyze and identify two benefits below.

Improved generalization of neural fluid fields. We compare the performance of neural fluid fields equipped with our foundation model, with and without multiphysics pretraining. As shown in Figure 9, our multiphysics pretraining can largely improve the data efficiency of neural fluid fields during future prediction. In contrast, both SciML models—without pretraining or pretrained on irrelevant PDE simulations (Maxwell)—lead to worse future predictions. Moreover, over all three fluid reconstruction tasks, the utilization of multiphysics pretraining leads to much improved PSNR, as shown in Table 2. We also evaluate DPOT [29] with pretrained weights, see results in Section A.7 in the supplementary material. These results validate the necessity of high-quality pretraining of our SciML foundation model, and the lack of a strong prior is the key to the worse performance of HyFluid and PINF.

Faster convergence during fine-tuning. Multiphysics pretraining also enables fast convergence during fine-tuning on real-world fluid data. As shown in Figure 10, despite gaps between PDE simulations and Scalarflow, our pretrained weights can still be quickly adapted to achieve accurate predictions and forecasting. In comparison, SciML models without pretraining or pretrained on Maxwell converge much more slowly during fine-tuning.

4.5. Ablation Study

We further provide ablation studies on our decoupled framework to demonstrate the benefits of each individual component. As shown in Table 3 and Figure 11, both our methods outperform the HyFluid baseline, with the combined approach achieving the best performance. For more ablation studies and comparison with other SciML foundation models, please read Section A in the supplementary material.

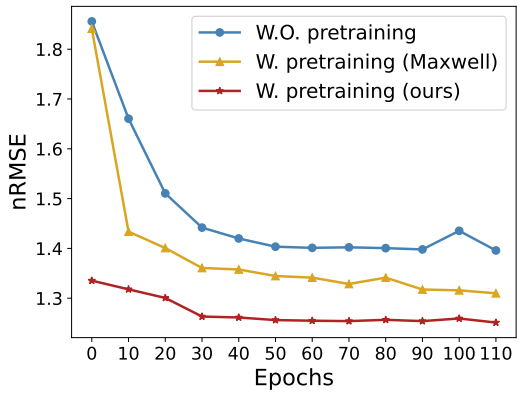

Figure 10. Multiphysics pretraining accelerates the convergence during fine-tuning of our SciML foundation model (on 40 initial frames from each of the four training videos in ScalarFlow).

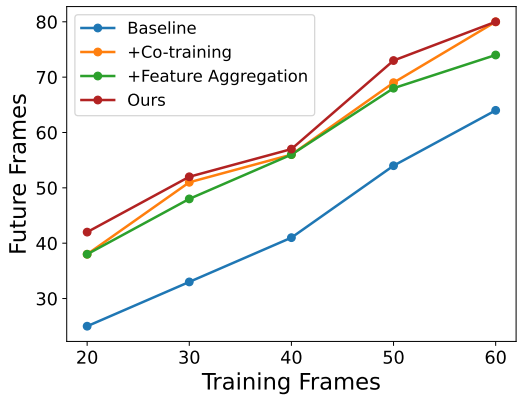

Figure 11. Ablation study of our decomposed methods on future prediction. X-axis: different numbers of initial training frames per video. Y-axis: temporal index of reliably predicted future frames (thresholded by PSNR=25) (higher is better).

5. Related works

5.1. 3D Reconstruction of Fluid

To reconstruct 3D fluid from visual measurements, traditional approaches utilized active sensing [26, 31, 33] or particle imaging velocimetry (PIV) [3, 19]. While effective, they necessitated sophisticated and controlled lab environments. Supervised view synthesis was recently proposed.

Table 3. Ablation study of our methods on the PSNR (higher the better) of fluid field reconstruction. "nf": number of input training frames. For future prediction, we report the PSNR averaged over 20 future frames (i.e., frames with indices from nf+1 → nf+20).

Methods	Novel View Synthesis			Re-Simulation			Future Prediction		
	nf=20	nf=40	nf=60	nf=20	nf=40	nf=60	nf=20	nf=40	nf=60
Baseline (HyFluid [81])	33.52	32.12	31.64	33.27	32.98	31.56	23.91	23.98	23.84
+ Co-training	**34.56**	33.19	32.31	34.03	33.13	**32.61**	**28.02**	25.20	26.87
+ Feature Aggregation	33.88	33.18	32.76	33.88	33.29	32.09	26.58	27.13	25.61
Ours	34.50	**33.48**	**32.84**	**34.34**	**33.36**	32.42	27.59	**28.36**	**27.76**

In [82], regularizers on view interpolation and projection consistency were designed for reconstruction from light tomography views. NeRFlow [15] learned 4D spatiotemporal representations of dynamic scenes by capturing 3D occupancy, radiance, and dynamics while enforcing consistency across different modalities. PINF [11] proposed to reconstruct fluid dynamics by leveraging PDEs (Navier-Stokes) to train a continuous spatiotemporal scene representation with a neural radiance field. NeuroFluid [27] proposed a particle-driven neural renderer that integrates fluid physical properties into volume rendering and includes a particle transition model to minimize differences between rendered and observed fluid views. HyFluid [81] proposed hybrid neural fluid fields to jointly infer fluid density and velocity fields, using a set of physics-based losses to enforce physically plausible density and velocity fields. However, no previous works explored the introduction of prior knowledge for data efficiency and improve generalization.

5.2. Scientific Machine Learning

SciML, fueled by advancements in deep learning, models physical phenomena and differential equations [8, 9, 37, 38]. Physics-informed neural networks (PINNs)[24, 25, 55, 58, 87] aim to incorporate physics into neural networks by including the differential form of the PDE as an additional physics loss regularization term. However, this paradigm has been confined to specific PDE scenarios (e.g., fixed PDE coefficients). Moreover, recent work has highlighted several fundamental issues with PINN-based methods[18, 36]. In contrast, operator learning methods, including Fourier Neural Operators [35, 40, 41] and the Deep Operator Network [44], have made progress in approximating the solution operators of PDEs. Although these data-driven approaches show promise in learning PDE solutions, they rely on vast quantities of high-fidelity labeled data. Researchers have also explored generating synthetic PDE solutions to train SciML models [30]. More recently, SciML foundation models have been developed [28, 29, 46, 54, 60, 62, 70, 80] by scaling up training datasets to incorporate multiple PDE simulations. SciML foundation models aim to encode common physical behaviors and enhance the generalization and scalability of SciML.

5.3. Foundation Models for 3D Reconstruction

Foundation models for vision are large-scale models pretrained on vast amounts of images or videos, designed to generalize across downstream vision tasks [6, 53, 56, 57, 59, 74, 77, 86]. CLIP [53] employed contrastive learning with extensive image-text data and achieved zero-shot performance. DINO [6] exemplified self-supervised learning and achieved impressive segmentation with minimal supervision, proving useful for visual correspondence and recognition [4, 10, 47, 72]. Diffusion models [14, 32, 50, 59] demonstrated exceptional image generation capabilities, with their learned feature spaces also serving recognition purposes, such as in semantic segmentation [5, 76]. To leverage these 2D vision foundation models in 3D reconstruction, researchers increasingly explored the potential of distilling 2D features into 3D space, exemplified by generalizable neural radiance fields (NeRFs) proposed to bridge this gap [7, 45, 64, 68, 73, 75, 79, 84]. Fine-tuning pretrained source models while training 3D reconstruction is also a common strategy. For example, Condense [85] enhances downstream task performance by jointly pretraining 2D and 3D features through multi-view images, creating a unified 2D-3D feature embedding space. From a broader perspective, when applying DINO to downstream domains with significantly different data distributions, such as medical imaging [20] or image matting [23], further fine-tuning is often necessary.

6. Conclusions

In this work, we demonstrate that integrating SciML foundation models with neural fluid fields provides a substantial improvement in data efficiency and generalization for inferring 3D fluid fields. Through a collaborative training approach, our method leverages the foundation model's forecasting capabilities to augment data, thereby reducing the reliance on extensive training frames. Additionally, the aggregation of pretrained representations enables more accurate reconstructions of fluid dynamics from sparse video frames. The results indicate that this strategy not only enhances reconstruction quality but also achieves robust performance in novel view synthesis and future prediction. Our work highlights the practical applicability of SciML foundation models in real-world fluid dynamics.

References

[1] Chronos 1.4 high-speed camera. https://www.4kshooters.net/2018/03/15/chronos-1-4-high-speed-camera-now-in-stock-raw-samples-available-for-download/, 2018. 1

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[3] Ronald J Adrian and Jerry Westerweel. *Particle image velocimetry*. Number 30. Cambridge university press, 2011. 7

[4] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 8

[5] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 8

[7] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2, 8

[8] Tianping Chen and Hong Chen. Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks*, 6(4):904–910, 1995. 8

[9] Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE transactions on neural networks*, 6(4):911–917, 1995. 8

[10] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021. 8

[11] Mengyu Chu, Lingjie Liu, Quan Zheng, Erik Franz, Hans-Peter Seidel, Christian Theobalt, and Rhaleb Zayer. Physics informed neural fields for smoke reconstruction with sparse data. *ACM Transactions on Graphics (ToG)*, 41(4):1–14, 2022. 1, 2, 3, 5, 6, 8, 4

[12] Yitong Deng, Hong-Xing Yu, Jiajun Wu, and Bo Zhu. Learning vortex dynamics for fluid inference and prediction. *arXiv preprint arXiv:2301.11494*, 2023.

[13] Yitong Deng, Hong-Xing Yu, Diyang Zhang, Jiajun Wu, and Bo Zhu. Fluid simulation on neural flow maps. *ACM Transactions on Graphics (TOG)*, 42(6):1–21, 2023. 1

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 8

[15] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021. 8

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[17] Marie-Lena Eckert, Kiwon Um, and Nils Thuerey. Scalarflow: a large-scale volumetric data set of real-world scalar transport flows for computer animation and machine learning. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019. 1, 2, 3, 5, 6

[18] C. Edwards. Neural networks learn to speed up simulations. *Communications of the ACM*, 65(5):27–29, 2022. 8

[19] Gerrit E Elsinga, Fulvio Scarano, Bernhard Wieneke, and Bas W van Oudheusden. Tomographic particle image velocimetry. *Experiments in fluids*, 41(6):933–947, 2006. 7

[20] Cui et al. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *IJCARS*, 2024. 8

[21] Kupyn et al. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 5

[22] Nah et al. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 5

[23] Yao et al. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 2024. 8

[24] Han Gao, Luning Sun, and Jian-Xun Wang. Phygeonet: Physics-informed geometry-adaptive convolutional neural networks for solving parameterized steady-state pdes on irregular domain. *Journal of Computational Physics*, 428: 110079, 2021. 8

[25] Nicholas Geneva and Nicholas Zabaras. Modeling the dynamics of pde systems with physics-constrained deep auto-regressive networks. *Journal of Computational Physics*, 403: 109056, 2020. 8

[26] Jinwei Gu, Shree K Nayar, Eitan Grinspun, Peter N Belhumeur, and Ravi Ramamoorthi. Compressive structured light for recovering inhomogeneous participating media. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):1–1, 2012. 7

[27] Shanyan Guan, Huayu Deng, Yunbo Wang, and Xiaokang Yang. Neurofluid: Fluid dynamics grounding with particle-driven neural radiance fields. In *International Conference on Machine Learning*, pages 7919–7929. PMLR, 2022. 1, 8

[28] Zhou Hang, Yuezhou Ma, Haixu Wu, Haowen Wang, and Mingsheng Long. Unisolver: Pde-conditional transformers are universal pde solvers. *arXiv preprint arXiv:2405.17527*, 2024. 2, 3, 8

[29] Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. Dpot: Auto-regressive denoising operator transformer for large-scale pde pre-training. *arXiv preprint arXiv:2403.03542*, 2024. 2, 3, 7, 8, 4, 6

[30] Erisa Hasani and Rachel A Ward. Generating synthetic data for neural operators. *arXiv preprint arXiv:2401.02398*, 2024. 8

[31] Tim Hawkins, Per Einarsson, and Paul Debevec. Acquisition of time-varying participating media. *ACM Transactions on Graphics (ToG)*, 24(3):812–815, 2005. 7

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 8

[33] Yu Ji, Jinwei Ye, and Jingyi Yu. Reconstructing gas flows using light-path approximation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2514, 2013. 7

[34] Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021. 2

[35] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. 8

[36] Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34: 26548–26560, 2021. 8

[37] Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on neural networks*, 9 (5):987–1000, 1998. 8

[38] Isaac E Lagaris, Aristidis C Likas, and Dimitris G Papageorgiou. Neural-network methods for boundary value problems with irregular boundaries. *IEEE Transactions on Neural Networks*, 11(5):1041–1049, 2000. 8

[39] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022. 3

[40] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew Stuart, Kaushik Bhattacharya, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. *Advances in Neural Information Processing Systems*, 33:6755–6766, 2020. 8

[41] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. 2, 3, 8

[42] Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *arXiv preprint arXiv:2111.03794*, 2021. 3

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[44] Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019. 3, 8

[45] Mana Masuda, Jinhyung Park, Shun Iwase, Rawal Khirodkar, and Kris Kitani. Generalizable neural human renderer. *arXiv preprint arXiv:2404.14199*, 2024. 2, 8

[46] Michael McCabe, Bruno Regaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023. 2, 3, 8, 4, 6

[47] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 8

[48] Gregoire Mialon, Quentin Garrido, Hannah Lawrence, Danyal Rehman, Yann LeCun, and Bobak Kiani. Self-supervised learning with lie symmetries for partial differential equations. *Advances in Neural Information Processing Systems*, 36:28973–29004, 2023. 2

[49] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[50] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 8

[51] Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4

[52] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 1, 2, 3

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 8

[54] Md Ashiqur Rahman, Robert Joseph George, Mogab Elleithy, Daniel Leibovici, Zongyi Li, Boris Bonev, Colin White, Julius Berner, Raymond A Yeh, Jean Kossaifi, et al. Pretraining codomain attention neural operators for solving multiphysics pdes. *arXiv preprint arXiv:2403.12553*, 2024. 2, 3, 8

[55] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019. 8

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 8

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 8

[58] Pu Ren, Chengping Rao, Yang Liu, Jian-Xun Wang, and Hao Sun. Phycrnet: Physics-informed convolutional-recurrent network for solving spatiotemporal pdes. *Computer Methods in Applied Mechanics and Engineering*, 389:114399, 2022. 8

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 8

[60] Junhong Shen, Tanya Marwah, and Ameet Talwalkar. Ups: Efficiently building foundation models for pde solving via cross-modal adaptation. In *ICML 2024 AI for Science Workshop*. 2, 3, 8

[61] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. *arXiv preprint arXiv:2306.00258*, 2023. 2

[62] Jingmin Sun, Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Towards a foundation model for partial differential equation: Multi-operator learning and extrapolation. *arXiv preprint arXiv:2404.12355*, 2024. 2, 3, 8

[63] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022. 2, 3, 5

[64] Songlin Tang, Wenjie Pei, Xin Tao, Tanghui Jia, Guangming Lu, and Yu-Wing Tai. Scene-generalizable interactive segmentation of radiance fields. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6744–6755, 2023. 2, 8

[65] Nils Thuerey, Konstantin Weißenow, Lukas Prantl, and Xiangyu Hu. Deep learning methods for reynolds-averaged navier–stokes simulations of airfoil flows. *AIAA Journal*, 58 (1):25–36, 2020. 1

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[67] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[68] Peihao Wang, Zhiwen Fan, Zhangyang Wang, Hao Su, Ravi Ramamoorthi, et al. Lift3d: Zero-shot lifting of any 2d vision model to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21367–21377, 2024. 2, 8

[69] Sifan Wang, Jacob H Seidman, Shyam Sankaran, Hanwen Wang, George J Pappas, and Paris Perdikaris. Cvit: Continuous vision transformer for operator learning. *arXiv preprint arXiv:2405.13998*, 2024. 3

[70] Tian Wang and Chuang Wang. Latent neural operator pretraining for solving time-dependent pdes. *arXiv preprint arXiv:2410.20100*, 2024. 2, 3, 8

[71] Xiaokun Wang, Yanrui Xu, Sinuo Liu, Bo Ren, Jiri Kosinka, Alexandru C Telea, Jiamin Wang, Chongming Song, Jian Chang, Chenfeng Li, et al. Physics-based fluid simulation in computer graphics: Survey, research trends, and challenges. *Computational Visual Media*, pages 1–56, 2024. 1

[72] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 8

[73] Yixuan Wang, Mingtong Zhang, Zhuoran Li, Tarik Kelestemur, Katherine Rose Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. Dynamic 3d descriptor fields for zero-shot generalizable rearrangement. In *8th Annual Conference on Robot Learning*, 2024. 2, 8

[74] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 8

[75] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. 2, 8

[76] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022. 8

[77] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021. 8

[78] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 3

[79] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8962–8973, 2023. 2, 8

[80] Zhanhong Ye, Xiang Huang, Leheng Chen, Hongsheng Liu, Zidong Wang, and Bin Dong. Pdeformer: Towards a foundation model for one-dimensional partial differential equations. *arXiv preprint arXiv:2402.12652*, 2024. 2, 3, 8

[81] Hong-Xing Yu, Yang Zheng, Yuan Gao, Yitong Deng, Bo Zhu, and Jiajun Wu. Inferring hybrid neural fluid fields from videos. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5, 6, 8, 3

[82] Guangming Zang, Ramzi Idoughi, Congli Wang, Anthony Bennett, Jianguo Du, Scott Skeen, William L Roberts, Peter Wonka, and Wolfgang Heidrich. Tomofluid: Reconstructing dynamic fluid from sparse view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1870–1879, 2020. 1, 8

[83] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 2, 3

[84] Xiaoshuai Zhang, Zhicheng Wang, Howard Zhou, Soham Ghosh, Danushen Gnanapragasam, Varun Jampani, Hao Su, and Leonidas Guibas. Condense: Consistent 2d/3d pretraining for dense and sparse features from multi-view images. *arXiv preprint arXiv:2408.17027*, 2024. 8

[85] Xiaoshuai Zhang, Zhicheng Wang, Howard Zhou, Soham Ghosh, Danushen Gnanapragasam, Varun Jampani, Hao Su, and Leonidas Guibas. Condense: Consistent 2d/3d pretraining for dense and sparse features from multi-view images. In *European Conference on Computer Vision*, pages 19–38. Springer, 2025. 8

[86] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 8

[87] Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, 394:56–81, 2019. 8