# PHYSICS-INSPIRED DUAL-GRAPH REASONING FOR SUBMESOSCALE OCEAN SUPER-RESOLUTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Accurately monitoring submesoscale (kilometer-scale) ocean dynamics is essential for understanding fine-grained energy transport and exchange, yet the high cost of in-situ observations results in limited data availability. Advances in visual super-resolution (SR) enable high-resolution (HR) predictions by refining abundant low-resolution (LR) remote sensing data. However, recent generative SR approaches can introduce visually rich but unsupported details, reducing reconstruction fidelity on fine-scale structures. To address this problem, we emphasize inductive biases inspired by multiscale ocean dynamics and propose a **d**ual-g**r**aph reas**o**ning **ne**twork, namely DRONE[1]. DRONE combines a heterogeneous channel graph with a spectral-aware pixel graph to improve reconstruction fidelity without imposing explicit physical equations or conservation constraints. The channel graph couples temporal observations with spatial-frequency representations, while the pixel graph allocates connectivity according to local spectral complexity. Extensive experiments on an HR MITgcm ocean simulation dataset demonstrate that DRONE surpasses Transformer-, diffusion-, and GNN-based SR methods on standard fidelity metrics, frontal-structure measurements, and dataset-wide spectral evaluations.

## 1 INTRODUCTION

Submesoscale ocean dynamics, such as kilometer-scale fronts, filaments, and eddies, play an important role in vertical energy transport, nutrient upwelling, and air–sea exchange (Thomas et al., 2008; McWilliams, 2019; Lévy et al., 2018; Taylor & Thompson, 2023). Capturing these processes at sufficiently high spatial and temporal resolution is essential for both operational oceanography and climate studies. However, high-resolution in-situ measurements are expensive to collect and remain sparse in space (McWilliams, 2019). Satellite observations, although much broader in coverage, are usually too coarse to resolve kilometer-scale structures (Chelton et al., 2019). Super-resolution (SR), which aims to reconstruct high-resolution (HR) fields from low-resolution (LR) inputs, therefore provides a promising data-driven way to narrow this observational gap (Dong et al., 2014; Ledig et al., 2017). Recent advances in natural-image SR (Dong et al., 2014; Lim et al., 2017; Zhang et al., 2018; Liang et al., 2021; Wang et al., 2018) have further encouraged its use in geophysical downscaling (Bassetti et al., 2024; Mardani et al., 2025; Wan et al., 2023).

In many existing studies, state-of-the-art SR models originally designed for natural images are directly applied to geophysical fields, often with only minor changes (Stengel et al., 2020; Leinonen et al., 2020; Harris et al., 2022). This practice implicitly assumes that a model capable of generating visually plausible image textures can also recover salient ocean structures. While this assumption may seem reasonable, our results suggest otherwise. On a high-resolution MITgcm (Menemenlis et al., 2021) ocean simulation benchmark with $16\times$ upscaling (Table 1), more than half of the evaluated state-of-the-art methods, including transformer- (Dosovitskiy et al., 2021; Liu et al., 2021; Liang et al., 2021), diffusion- (Ho et al., 2020; Saharia et al., 2022; Li et al., 2022), and GNN-based (Tian et al., 2024; Zhou et al., 2020) models, *do not outperform bicubic interpolation* on standard fidelity metrics. Some of the remaining methods only achieve limited gains. The fact that learned models with millions of parameters can fall behind a parameter-free interpolation baseline points to a clear

[1]The anonymous code repository can be found at https://anonymous.4open.science/r/DRONE-1E41.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mismatch between image-domain inductive biases and the structure of geophysical data, especially under extreme upscaling ratios.

We attribute this mismatch to two main limitations. The first is insufficient structure-aware fidelity. Diffusion-based methods (Ho et al., 2020; Song et al.; Saharia et al., 2022; Li et al., 2022) typically generate fine-scale details from an isotropic Gaussian prior. As a result, the generated textures may look spatially plausible, but they are not necessarily consistent with the resolved field or with the coupling among its observed channels (Bassetti et al., 2024; Mardani et al., 2025; Price et al., 2025; Wan et al., 2023). Transformer-based methods (Dosovitskiy et al., 2021; Liu et al., 2021) can model long-range statistical dependencies through self-attention, yet many important relationships in ocean dynamics, including advective transport, wave propagation, and frontal interactions, are directional rather than purely correlational. Several weather and climate models built on transformers (Nguyen et al., 2023; Bi et al., 2023; Pathak et al., 2022) have achieved impressive forecasting skill, but their architectural assumptions do not directly address the spectral fidelity requirements of spatial super-resolution. The second limitation is insufficient *spectral fidelity*. Common aggregation operations, whether based on attention (Dosovitskiy et al., 2021; Liu et al., 2021) or message passing (Tian et al., 2024; Zhou et al., 2020), tend to treat different regions in a relatively uniform way. This can blur high-frequency information without distinction. At a $16\times$ upscaling ratio, the spectral gap between LR inputs and HR targets becomes substantial. Uniform aggregation may either remove real fine-scale structures or turn reconstruction errors into structured artefacts that resemble physical features, which can be especially problematic for downstream scientific analysis.

To address these issues, we propose **DRONE** (Dual-gRaph reasOning NEtwork), an architecture built around inductive biases inspired by the multiscale and evolving structure of geophysical data rather than borrowed directly from natural-image SR. DRONE adopts a dual-graph design. The first component is a *spatiotemporal channel graph*, which models cross-channel and cross-timestep interactions between temporal observations and spatial-frequency representations. The second component is a *spectral-aware propagation graph*, which performs pixel-level aggregation with graph connectivity set inversely proportional to local spectral complexity. This encourages high-frequency structures to be processed with localized neighborhoods while allowing broad aggregation in smooth regions. In addition, we use a cascaded multi-stage pipeline to decompose the challenging upscaling task into several progressive steps (Lai et al., 2017). At each intermediate resolution, both graph mechanisms are re-instantiated to keep the reconstruction process well conditioned. Experiments on a high-resolution MITgcm (Menemenlis et al., 2021) ocean simulation dataset show that DRONE consistently outperforms all compared baselines in standard fidelity metrics, a newly introduced frontal-structure metric, and dataset-wide spectral-fidelity evaluations.

Our main contributions are summarized as follows:

- **Spatial-frequency decomposition with a structured kernel bank (§3.2).** We design a bank of spatial kernels that provides spectral coverage, scale separation, and structural non-redundancy. This kernel bank decomposes the raw geophysical field into multiple channels, each highlighting a specific frequency band or differential characteristic. Since physical processes at different length scales leave signatures in different spectral ranges, a single observed field often mixes several types of information together. Our decomposition makes this hidden multi-scale structure more explicit and easier to separate, providing the spectral basis for the following graph modules.
- **Spatiotemporal channel graph for structure-aware fidelity (§3.3).** We introduce a channel-level graph in which spatial-frequency channels and temporal channels are treated as heterogeneous nodes. These nodes interact through both intra-group and inter-group message passing, allowing the model to use cross-channel dynamics and temporal context rather than relying only on local texture synthesis.
- **Spectral-aware propagation graph for spectral fidelity (§3.4).** We build a pixel-level spatial graph whose node degree is set *inversely* proportional to a local spectral-complexity indicator. With this reversed allocation, smooth low-frequency regions receive broader connectivity, which supports globally coherent aggregation, while textured high-frequency regions are restricted to sparser and more localized neighborhoods. This design reduces the structured artefacts that uniform-degree aggregation often introduces at high wavenumbers. Together with our front-aware evaluation metric, FIU-MAE, this module provides both a modeling mechanism and an evaluation criterion for improving spectral fidelity in geophysical super-resolution.

## 2 Related Work

Dedicated spatiotemporal super-resolution methods for geophysical fields remain scarce; in practice, models developed for natural image or video SR are repurposed for Earth-science data with minimal modification. We review the two most prominent families through the lens of a question this transfer raises: when a model designed for perceptual quality is applied to a geophysical field, how well does it preserve the reconstructed structure and spectrum?

**Diffusion-based methods.** Score-matching diffusion models (Ho et al., 2020; Song et al.) achieve state-of-the-art perceptual quality in image SR (Saharia et al., 2022; Li et al., 2022), motivating their adoption for geophysical downscaling (Bassetti et al., 2024; Mardani et al., 2025; Price et al., 2025). In the natural-image setting, generating fine-scale texture from an isotropic Gaussian prior is reasonable: plausible hair strands or fabric weaves need only look correct. Geophysical fields, however, carry cross-field constraints. A downscaled precipitation field may exhibit realistic spatial texture yet place convective cells where concurrent thermodynamic fields do not support deep convection; a downstream impact model would inherit this inconsistency as silently plausible hazard estimates. Physics-informed losses (Wan et al., 2023) and residual formulations (Mardani et al., 2025) mitigate the most apparent artefacts, but these constraints are external to a generative pathway originating from uninformed noise.

**Transformer-based methods.** Vision transformers (Dosovitskiy et al., 2021; Liu et al., 2021) and their SR variants (Liang et al., 2021; Chen et al., 2023; Lu et al., 2022) have likewise been transferred to geophysical contexts (Nguyen et al., 2023; Bi et al., 2023; Pathak et al., 2022). Self-attention suits natural images, where long-range similarity is a useful prior. In geophysical fields, however, the relevant long-range dependencies are causal–advective transport, wave propagation, teleconnections, and attention captures statistical co-occurrence rather than directed physical causality. Under climate non-stationarity the model defaults to historical correlation patterns even when the underlying mechanism is no longer active, a failure mode especially consequential for applications where the user's interest is in how relationships change. Window-based approximations (Liu et al., 2021) further impose uniform spatial partitioning agnostic to anisotropic atmospheric propagation.

**Present work.** Rather than adapting an image-domain architecture post hoc, we design the model around the structure of geophysical data itself: a graph-based channel decomposition separating spatial scales and temporal modes, with operators that propagate resolved-scale information so that reconstructed detail is traceable to the input rather than sampled from a prior or inferred from historical co-occurrence.

## 3 Methodology

In this section, we first present the problem formulation, and then provide a detailed introduction to our proposed DRONE model, whose architectural design is illustrated in Figure 1. The design is physics-inspired: it uses temporal context, multiscale spatial representations, and local spectral behavior as inductive biases, but it does not impose governing equations or explicit conservation constraints during training.

### 3.1 Problem Formulation

We consider the task of spatiotemporal super-resolution: given a sliding window of low-resolution (LR) observations of a geophysical field, the goal is to reconstruct the corresponding high-resolution (HR) field at the central timestep. Formally, let $\{\mathbf{x}_t^{\mathrm{LR}}\}_{t=\tau-\lfloor m/2\rfloor}^{\tau+\lfloor m/2\rfloor}$ denote a window of $m$ consecutive LR frames defined on a coarse grid $\Omega^{\mathrm{LR}} \subset \mathbb{R}^2$ with spatial dimensions $H_l \times W_l$, and let $\mathbf{y}_\tau^{\mathrm{HR}}$ denote the target HR frame on a fine grid $\Omega^{\mathrm{HR}}$ with dimensions $H_h \times W_h$, where $H_h = s \cdot H_l$ and $W_h = s \cdot W_l$ for an integer upscaling factor $s$. We seek a mapping

$$f_\theta : \{\mathbf{x}_t^{\mathrm{LR}}\}_{t=\tau-\lfloor m/2\rfloor}^{\tau+\lfloor m/2\rfloor} \longrightarrow \hat{\mathbf{y}}_\tau^{\mathrm{HR}}, \tag{1}$$

parameterized by $\theta$, such that $\hat{\mathbf{y}}_\tau^{\mathrm{HR}} \approx \mathbf{y}_\tau^{\mathrm{HR}}$ under a suitable reconstruction criterion.

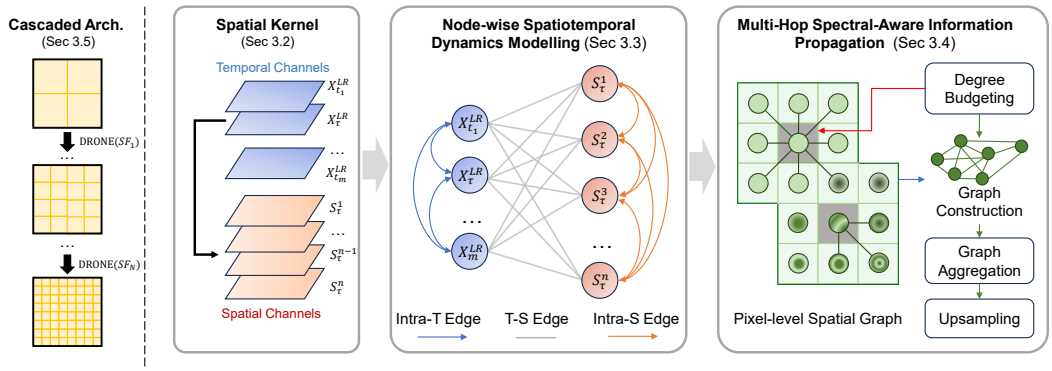


Figure 1: The overall architecture design of DRONE.

To exploit temporal context and spatial structure, we construct the input to $f_\theta$ by composing two groups of channels. The first group consists of the $m$ temporal channels given directly by the LR frames in the sliding window, capturing the local temporal neighbourhood centred at $\tau$. The second group consists of $n$ spatial-frequency channels produced by a bank of spatial kernels (§3.2) applied to the central frame $\mathbf{x}_\tau^{\mathrm{LR}}$, each exposing a different structural or spectral cue. Concatenating both groups along the channel dimension yields the composite input

$$\mathbf{z}_\tau = \left[\, \mathbf{x}^{\mathrm{LR}}_{\tau-\lfloor m/2 \rfloor},\ \ldots,\ \mathbf{x}^{\mathrm{LR}}_{\tau},\ \ldots,\ \mathbf{x}^{\mathrm{LR}}_{\tau+\lfloor m/2 \rfloor},\ S^1_\tau,\ \ldots,\ S^n_\tau \,\right] \in \mathbb{R}^{H_l \times W_l \times (m+n)}, \tag{2}$$

where $S^i_\tau = \mathcal{S}_i[\mathbf{x}^{\mathrm{LR}}_\tau]$ is the $i$-th spatial channel. The mapping $f_\theta$ then processes $\mathbf{z}_\tau$ through a sequence of graph-based modules: channel-level dynamics modelling (§3.3), spectrally adaptive pixel-level message passing (§3.4), and cascaded upscaling (§3.5), whose message-passing backbone is reviewed in Appendix A.

**Remark 3.1** (Distinction from image and video super-resolution)**.** *Our formulation maps multiple low-resolution observations from a single target variable across time, together with spatial-frequency channels, to a single high-resolution target. SISR treats each frame independently, discarding temporal context. VSR does exploit multiple frames, but assumes near-duplicate views related by sub-pixel motion, leveraging* geometric redundancy. *In geophysical data, successive time steps are distinct states of an evolving field, where features form, deform, and dissipate, so the temporal neighbourhood provides* dynamical context*, not sub-pixel sampling of a static scene.*

### 3.2 Spatial Kernels

Any observed geophysical field $y_{\mathrm{ref}}(\mathbf{x})$ contains structural information distributed across a range of spatial frequencies. Physical processes at different characteristic length scales imprint signatures in distinct spectral bands, yet a single raw observation conflates them into one composite signal. To disentangle these contributions we construct a bank of spatial kernels $\{\mathcal{S}_i\}_{i=1}^n$, each producing a channel $s_i(\mathbf{x}) = \mathcal{S}_i[y_{\mathrm{ref}}](\mathbf{x})$ whose transfer function $\hat{\mathcal{S}}_i(\boldsymbol{\omega})$ selects a distinct portion of the spatial spectrum. The bank is designed to satisfy three requirements: (i) spectral coverage: the kernels collectively span the accessible frequency support,

$$\textstyle\sum_{i=1}^n |\hat{\mathcal{S}}_i(\boldsymbol{\omega})|^2 > 0 \quad \text{for all } \|\boldsymbol{\omega}\| \in (0,\, \omega_{\mathrm{cut}}), \tag{3}$$

guaranteeing no resolvable structure is discarded; (ii) scale separation: each $\hat{\mathcal{S}}_i$ emphasizes a different frequency band or differential order; and (iii) structural non-redundancy: the kernels differ in differential order, directional selectivity, or linearity so that no channel is a linear combination of the others. The resulting spatial node set $V_S = \{s_1, \ldots, s_n\}$ encodes multi-resolution geometry and texture, providing the spatial branch of the graph with structurally diverse inputs. Specific kernel definitions and spectral justifications are given in Appendix B.

### 3.3 Node-wise Spatiotemporal Dynamics Modeling

We assemble the full channel graph $\mathcal{G} = (V, E)$ with node set $V = V_S \cup V_T$, where $V_S$ collects the spatial-frequency nodes and $V_T$ the temporal nodes. Each node $v_k$ carries a channel signal $c_k(\mathbf{x}, t)$

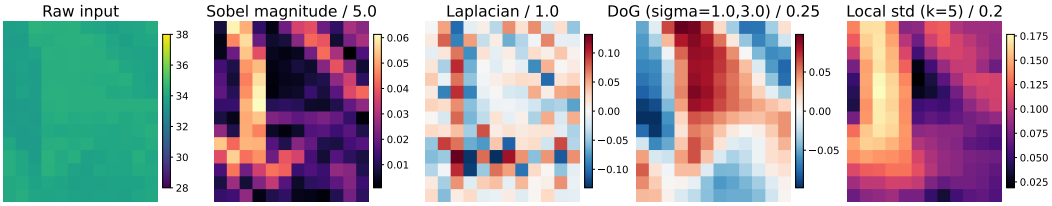


Figure 2: Spatial-kernel visualization for each input image, showing the raw input together with DRONE feature extractions (details in appendix B) to highlight edge, curvature, and texture cues used by the model.

capturing one facet, either a spatial frequency band or a temporal mode of the evolving field. The edge set is partitioned as $E = E_{SS} \cup E_{TT} \cup E_{ST}$: $E_{SS}$ and $E_{TT}$ form complete graphs within $V_S$ and $V_T$ respectively, while $E_{ST}$ forms a complete bipartite graph between the two sets.

This partition naturally induces two complementary levels of interaction:

**Intra-group modelling.** Edges in $E_{SS}$ enable inter-scale exchange among spatial-frequency nodes, allowing the network to couple coarse structures with fine-grained texture within the spatial group. Analogously, edges in $E_{TT}$ propagate information across temporal modes, capturing motion coherence and temporal correlations within the temporal group.

**Inter-group modelling.** Edges in $E_{ST}$ bridge the spatial and temporal groups, fusing frequency-decomposed spatial structure with temporal dynamics. This bipartite coupling ensures that high-frequency spatial detail can be informed by temporal context, and conversely that temporal predictions are grounded in multi-scale spatial evidence.

Together, the intra-group and inter-group message-passing rounds allow each node to integrate information both from its own modality group and from the complementary group, yielding spatiotemporally enriched node representations. The detailed formulation of the dynamics and message functions is provided in Appendix C.

### 3.4 Multi-hop Spectral-aware Information Propagation

We now define the pixel-level spatial graph that mediates local aggregation. Each of the $P = H \times W$ pixels becomes a node $p_i$ carrying the multi-channel feature $\mathbf{h}_i = [c_1(\mathbf{x}_i, t), \ldots, c_K(\mathbf{x}_i, t)]^\top \in \mathbb{R}^K$. Unlike a standard grid graph with fixed connectivity, we endow each pixel with an adaptive degree that depends on local spectral complexity.

**Complexity indicator and reversed degree budget.** Following IPG (Tian et al., 2024), we measure local high-frequency content via a downsample–upsample residual:

$$\xi_i = \left\| \mathbf{F}(\mathbf{x}_i) - (\mathcal{U}_s \circ \mathcal{D}_s)[\mathbf{F}](\mathbf{x}_i) \right\|_2, \tag{4}$$

where $\mathcal{D}_s$ and $\mathcal{U}_s$ denote downsampling and upsampling by factor $s$, and $\mathbf{F}$ is the feature map. The operation $\mathcal{U}_s \circ \mathcal{D}_s$ acts as a low-pass filter, so $\xi_i$ captures precisely the high-frequency energy at $\mathbf{x}_i$.

IPG sets the degree budget proportional to $\xi_i$, giving complex regions more neighbours. We adopt the reversed strategy where the budget is reciprocal to complexity:

$$d_i = \mathrm{clamp}\left( \left\lfloor \frac{C}{\xi_i + \epsilon} \right\rfloor,\ d_{\min},\ d_{\max} \right). \tag{5}$$

Smooth pixels ($\xi_i$ small) thus receive dense connectivity, enlarging the receptive field without spectral distortion, while textured and edge pixels ($\xi_i$ large) receive sparse connectivity, limiting the low-pass smoothing inherent to message passing and suppressing the degradation of high-frequency components into structured artifacts.

**Graph construction and message passing.** For each pixel $p_i$ we retain the $d_i$ nearest neighbours within radius $r_{\max}$, yielding a spatially adaptive edge set $E^{\mathrm{pix}}$. We then perform $L$ rounds of message

passing:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{j\in\mathcal{N}(i)} \frac{1}{\sqrt{d_i\, d_j}}\, \mathbf{W}_{\text{msg}}^{(l)}\, \mathbf{h}_j^{(l)}\right), \quad l = 0, \ldots, L-1, \tag{6}$$

where $\mathbf{W}^{(l)}, \mathbf{W}_{\text{msg}}^{(l)} \in \mathbb{R}^{K\times K}$ are learnable and $\sigma$ is a nonlinearity. The degree budget shapes both the aggregation set and the normalization, ensuring that the effective smoothing varies spatially with local spectral content. The full derivation and spectral analysis are deferred to Appendix D.

### 3.5 Cascaded Multi-Stage Upscaling

Our target regime ($\times 16$ and beyond) requires each low-resolution pixel to generate $s^2 = 256$ high-resolution pixels. A single-stage pipeline must recover the entire missing frequency band $(\pi/s,\, \pi]$ at once, with no intermediate check. At such extreme ratios the spectral gap overwhelms the channel decomposition, the complexity indicator $\xi_i$ loses discriminative power because nearly all coarse-grid pixels appear smooth, and the pixel graph contains too few nodes for meaningful spatial aggregation.

We therefore factorize $s_{\text{total}}$ into $T$ successive stages with per-stage factors satisfying $\prod_{t=1}^{T} s^{(t)} = s_{\text{total}}$. At each stage $t$ the full pipeline of Sections 3.2–3.4 is applied: spectral decomposition, adaptive-degree pixel graph construction, message passing, and upsampling by $s^{(t)}$. Because each stage only bridges a narrow spectral band, channel separation remains well-conditioned. Moreover, the complexity indicator and degree budget are recomputed at every stage; as intermediate frequencies are progressively filled in, regions that appeared uniformly smooth at coarser scales may reveal edges or textures, restoring the adaptive advantage of the degree budget.

### 3.6 Summary of the Proposed Methodology

The total upscaling factor is factorized into $T$ cascaded stages. At each stage, the spatial field is first decomposed into $n$ spectral channels via spatial kernels (§3.2), exposing frequency content as a multi-channel feature vector per pixel. A pixel-level graph is then constructed whose per-node degree is set inversely proportional to a local complexity indicator, granting smooth regions broad receptive fields and textured regions sharp, localized connectivity (§3.4). Message passing over this adaptive graph produces aggregated features that are decoded and upsampled by the stage-specific factor. The cascade repeats this procedure at progressively finer resolutions, re-instantiating the spectral decomposition and degree budget at every stage, until the full target resolution is reached.

## 4 Experiments

We first introduce the experimental settings, including the datasets, evaluation metrics, and baselines. Then, we evaluate DRONE with extensive experiments to answer the following research questions:

- **RQ1 (Temporal Generalizability)**: Does DRONE achieve state-of-the-art reconstruction quality when evaluated on temporally distant samples unseen during training?
- **RQ2 (Frontal-Structure Fidelity)**: Does DRONE accurately preserve the location and intensity of salient structures such as salinity fronts, while maintaining the spatial continuity of the reconstructed field?
- **RQ3 (Spectral Fidelity)**: Does DRONE degrade gracefully at high frequencies, distributing unrecovered energy as incoherent residuals rather than structured artifacts?
- **RQ4 (Dual-Graph Necessity)**: Is the Dual-Graph design in DRONE essential to the overall performance?

Additional experiments compare the three strongest models in the main comparison—PFT, SwinIR, and DRONE—under matched evaluation settings. They are organized as RQ5 (unseen degradation robustness), RQ6 (matched temporal inputs), RQ7 (degree-budget policy), RQ8 (FIU-MAE sensitivity), RQ9 (dataset-wide spectral fidelity), and RQ10 (transfer to sea-surface temperature) in Appendix H. Test-set metric variability is reported there as a supplementary analysis without introducing an additional research question.

### 4.1 EXPERIMENT SETUP

**Datasets**. We use the Northwest Pacific Pre-SWOT Level-4 Hourly MITgcm LLC4320 Native Grid 2 km Oceanographic Dataset Version 1.0 (Menemenlis et al., 2021) as high-resolution ground truth. The dataset provides a regional multivariate ocean state estimate from the MITgcm global simulation on a Lat-Lon-Cap (LLC) grid at $\frac{1}{48}$-degree ($\approx 2$ km at the equator) nominal horizontal resolution, covering the Northwest Pacific. We target surface salinity, a variable for which submesoscale observations are especially scarce (Klemas, 2011; Kim et al., 2023), and construct low-resolution inputs by downsampling with a factor of $16\times$. The 10 296 hourly snapshots are split chronologically (70%/15%/15% for training, validation, and testing) to prevent temporal leakage.

**Baseline**. We compare the proposed DRONE with recent SOTA single-image super-resolution models, reflecting current practice in geophysical downscaling. As discussed in Remark 3.1, neither naively channel-stacking multiple fields into an SISR model nor applying VSR alignment to geophysical fields matches the temporal-context setting studied here. We compare against three families of architectures: transformer-based methods (Liang et al., 2021; Chen et al., 2022; Zhang et al., 2023; Chen et al., 2023; 2024; Zhang et al., 2024; Long et al., 2025), graph neural network methods (Zhou et al., 2020; Tian et al., 2024), and diffusion-based methods (Yue et al., 2023; Wang et al., 2024).

**Training protocol.** All baselines are trained on the same dataset under identical input–output resolution settings ($16\times$ super-resolution), evaluated on the same chronological test split, and optimized with MSE. Learning rates and batch sizes are selected using the validation split; the concise reproducibility details are provided in Appendix G.

**Evaluation Methods**. We evaluate with metrics standard in geophysical study (MSE, MAE, RMSE), which measure field reconstruction fidelity in native units, alongside metrics standard in computer vision (PSNR, SSIM), which capture perceptual reconstruction quality. We also employ the newly designed FIU-MAE metric, introduced in Section 4.3, to evaluate frontal-structure fidelity.

### 4.2 OVERALL PERFORMANCE (RQ1)

Table 1 summarizes the quantitative comparison. Our model consistently outperforms all baselines across metrics, achieving the lowest MSE, RMSE and MAE, which indicates improved reconstruction accuracy and field fidelity. It also attains the highest PSNR (46.1344 dB) and SSIM (0.9881), reflecting better preservation of spatial structure and local details compared to recent methods such as SwinIR and PFT. Additionally, the lowest FIU-MAE (0.0167) indicates improved fidelity around salient frontal variations. Overall, these results demonstrate that our approach produces more accurate and structurally consistent reconstructions. We attribute these gains to the effective exploitation of multi-temporal context and multiscale spatial representations, which better captures geophysical-scale interactions. Selected visualizations of results are presented in Figure 3.

Table 1: Comparison of DRONE with recent SR methods.

| Method | Params | MSE↓ | MAE↓ | RMSE↓ | PSNR↑ | SSIM↑ | FIU-MAE↓ |
|---|---|---|---|---|---|---|---|
| Bicubic | - | 0.0039 | 0.0411 | 0.0628 | 44.0468 | 0.9829 | 0.0203 |
| IGNN(Zhou et al., 2020) | 2.570M | 0.0048 | 0.0473 | 0.0696 | 43.1506 | 0.9821 | 0.0290 |
| SwinIR(Liang et al., 2021) | 2.891M | <u>0.0025</u> | <u>0.0332</u> | <u>0.0502</u> | <u>45.9766</u> | <u>0.9878</u> | <u>0.0170</u> |
| CAT(Chen et al., 2022) | 3.478M | 0.0061 | 0.0498 | 0.0783 | 42.1286 | 0.9820 | 0.0351 |
| ResShift(Yue et al., 2023) | 3.357M | 0.0034 | 0.0408 | 0.0587 | 44.6327 | 0.9847 | 0.0228 |
| ART(Zhang et al., 2023) | 3.529M | 0.0086 | 0.0605 | 0.0927 | 40.6608 | 0.9789 | 0.0483 |
| HAT(Chen et al., 2023) | 3.159M | 0.0060 | 0.0499 | 0.0773 | 42.2321 | 0.9820 | 0.0348 |
| StableSR(Wang et al., 2024) | 3.099M | 0.0086 | 0.0676 | 0.0927 | 40.6585 | 0.9675 | 0.0509 |
| ATD(Zhang et al., 2024) | 4.585M | 0.0060 | 0.0490 | 0.0773 | 42.2327 | 0.9817 | 0.0333 |
| IPG(Tian et al., 2024) | 2.656M | 0.0038 | 0.0417 | 0.0619 | 44.1684 | 0.9833 | 0.0249 |
| RGT(Chen et al., 2024) | 3.567M | 0.0033 | 0.0389 | 0.0573 | 44.8433 | 0.9851 | 0.0203 |
| PFT(Long et al., 2025) | 3.541M | 0.0029 | 0.0364 | 0.0537 | 45.3972 | 0.9861 | 0.0190 |
| **DRONE (Ours)** | 3.536M | **0.0024** | **0.0320** | **0.0493** | **46.1344** | **0.9881** | **0.0167** |

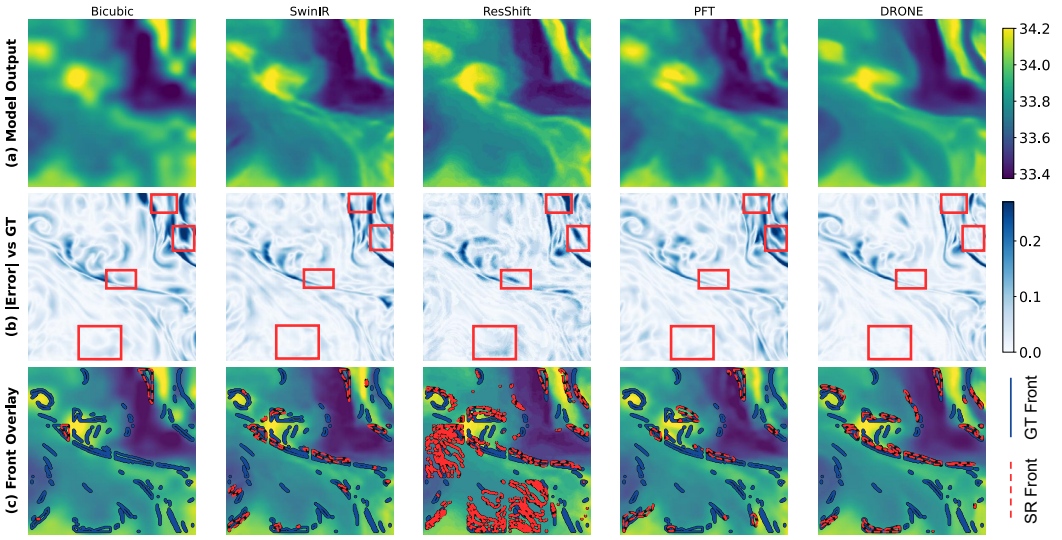


Figure 3: Row 1: reconstructed output fields for each model. Row 2: pixel-wise absolute error relative to the ground truth, plotted on a common color scale. Row 3: front detections overlaid on the reconstructed fields, where solid blue lines denote ground-truth fronts and dashed red lines denote model-detected fronts.

### 4.3 Frontal-Structure Fidelity (RQ2)

The spatiotemporal channel graph encodes cross-channel and cross-timestep dependencies, enabling the model to reconstruct fine-scale structures that are coherent with the surrounding dynamical context rather than hallucinated from local texture alone. We evaluate this property on ocean salinity fronts with the *front intensity union-weighted MAE* (FIU-MAE; algorithmic details in Appendix E, quantitative comparisons in Table 1). Ocean fronts exhibit complex topologies characterized by non-convex shapes, rings, filaments, and span a wide intensity range across basins, making single-point summaries such as centroid distance inadequate. FIU-MAE instead (i) adaptively detects front support regions per local window, (ii) evaluates gradient-magnitude differences over the *union* of predicted and true front pixels, and (iii) weights errors by true front intensity. Strong, salient fronts therefore dominate the score, while the union operation penalizes both missed detections and spurious fronts. The metric thus jointly captures front localization, spatial extent, and intensity preservation—properties that reflect the spatial continuity expected of ocean fields. As shown in the third row of Fig. 3, our model reconstructs frontal structures with notably higher fidelity.

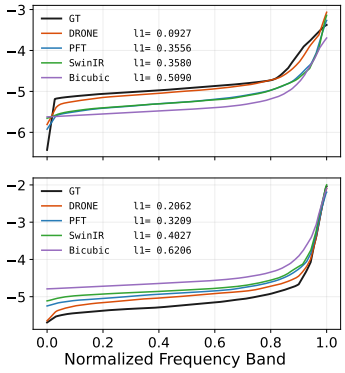


Figure 4: Selected-patch high-frequency spectral energy; inset: distance to ground truth. Computation details are in Appendix F.

### 4.4 Spectral Fidelity (RQ3)

Some high-frequency information is inevitably lost at extreme upscaling ratios; the critical question is whether the unrecovered energy manifests as gracefully attenuated amplitude or as structured artifacts that mimic physical features. Standard GNN aggregation promotes the latter: high-degree nodes accumulate disproportionate influence, creating spatially coherent distortions at the most uncertain wavenumbers. The reciprocal degree budget redistributes aggregation so that reconstruction error remains incoherent rather than organizing into spurious structures. Figure 4 shows that, for the illustrated sample, the high-frequency energy distribution of DRONE is closer to the ground truth than those of the baselines.

### 4.5 Case study: How does the dual-graph reasoning work? (RQ2, RQ3)

We visualize both graphs across consecutive timestamps for a representative DRONE sequence (Figure 5). In the bipartite graph, edge weights shift toward high-frequency channels when fronts sharpen and toward broader-scale channels during smoother transitions. Complementarily, the pixel-wise spatial graph shows neighbourhood connectivity contracting near fronts and expanding in smooth regions, consistent with spectral-aware degree assignment. Both topologies adapt dynamically to the evolving input rather than acting as static priors.

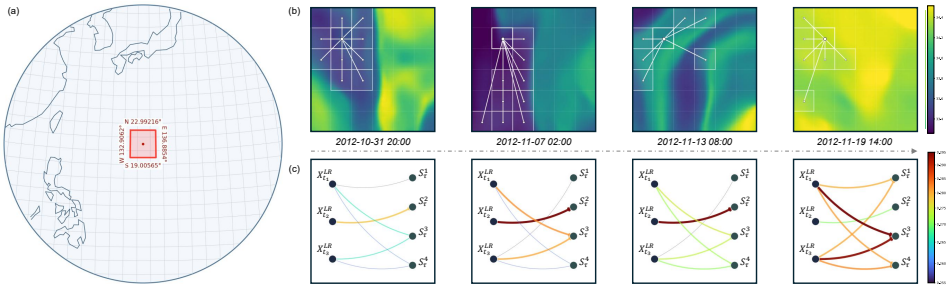


Figure 5: (a) shows the dataset source region. At each timestamp, we visualize two complementary graphs used in DRONE inference: (b) a temporal-to-spatial bipartite graph, where temporal nodes connect to spatial feature nodes and edge color/thickness indicate T→S attention strength; and (c) a pixel-level spatial graph over the SR patch, with an overlaid LR grid. A tracked LR pixel and its local neighbors are highlighted, with white edges showing spatial connectivity variations over time.

### 4.6 Ablation Studies (RQ4)

To verify the necessity of the dual-graph design, we conduct ablation experiments by systematically degrading the graph topology of DRONE. Specifically, we derive the following variants:

- **DRONE$_{\mathbf{S}}$** retains the spatial graph, removing auxiliary temporal inputs. The spatiotemporal graph degrades to a star of depth one with $n$ leaves, whose leaf nodes form a complete spatial graph.
- **DRONE$_{\mathbf{T}}$** retains only the temporal channels, disabling all spatial aggregation kernels. The spatiotemporal graph reduces to a complete graph over $m$ temporal nodes with no intra-frame spatial reasoning.
- **DRONE$_{\mathbf{bare}}$** removes both graph components entirely, reducing the architecture to a conventional single-image super-resolution backbone.

Detailed results are presented in Table 2, confirming that each graph component yields consistent gains across all metrics and that their contributions are complementary: jointly removing both reduces the model to near-baseline performance. Matched-temporal-input and degree-policy controls are reported as RQ6 and RQ7 in Appendix H; the latter uses a common three-month subset because of computational-resource limits.

Table 2: Ablation: DRONE vs. degraded variants.

| | DRONE$_{\text{bare}}$ | DRONE$_{\text{S}}$ | DRONE$_{\text{T}}$ | DRONE |
|---|---|---|---|---|
| MSE↓ | .0030 | .0028 | .0028 | **.0024** |
| MAE↓ | .0365 | .0349 | .0354 | **.0320** |
| RMSE↓ | .0543 | .0527 | .0531 | **.0493** |
| PSNR↑ | 45.30 | 45.56 | 45.50 | **46.13** |
| SSIM↑ | .9862 | .9868 | .9867 | **.9881** |
| FIU-MAE↓ | .0186 | .0178 | .0179 | **.0167** |

## 5 Conclusion and Limitations

**Conclusion.** We present DRONE, a physics-inspired dual-graph reasoning network that uses spatiotemporal channel interactions and spectrally adaptive propagation to improve super-resolution fidelity. On 16× upscaling of HR MITgcm ocean simulations, DRONE achieves the strongest performance across the reported metrics, while several image-domain SR methods even fail to surpass bicubic interpolation. Ablation results further verify that both graph components are useful.

**Limitations.** Our method relies on paired HR–LR simulation data and harmonized spatial-temporal fields. Extending it to statistical downscaling without HR ground truth, to additional variables, and to observations with registration errors remains future work. The dual-graph design is physics-inspired but does not directly impose governing equations or conservation laws.

## REFERENCES


Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Dudley B Chelton, Michael G Schlax, Roger M Samelson, J Thomas Farrar, M Jeroen Molemaker, James C McWilliams, and Jonathan Gula. Prospects for future satellite estimation of small-scale variability of ocean surface velocity and vorticity. *Progress in Oceanography*, 173:256–350, 2019.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22367–22377, 2023.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. In *ICLR*, 2024.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pp. 184–199. Springer, 2014.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

Lucy Harris, Andrew TT McRae, Matthew Chantry, Peter D Dueben, and Tim N Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022.

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

Young Jun Kim, Daehyeon Han, Eunna Jang, Jungho Im, and Taejun Sung. Remote sensing of sea surface salinity: challenges and research directions. *GIScience & Remote Sensing*, 60(1):2166377, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Victor Klemas. Remote sensing of sea surface salinity: an overview with case studies. *Journal of Coastal Research*, 27(5):830–838, 2011.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 624–632, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2020.

Marina Lévy, Peter JS Franks, and K Shafer Smith. The role of submesoscale currents in structuring marine ecosystems. *Nature communications*, 9(1):4758, 2018.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

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*, pp. 10012–10022, 2021.

Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2279–2288, 2025.

Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 457–466, 2022.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025.

James C McWilliams. A survey of submesoscale currents. *Geoscience Letters*, 6(1):3, 2019.

D. Menemenlis, C. Hill, C. E. Henze, J. Wang, and I. Fenty. Northwest pacific pre-swot level-4 hourly mitgcm llc4320 native grid 2km oceanographic dataset version 1.0, 2021.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, pp. 25904–25938. PMLR, 2023.

Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. In *International Conference on Learning Representations*, 2020.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. In *105th Annual AMS Meeting 2025*, volume 105, pp. 449275, 2025.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Karen Stengel, Andrew Glaws, Dylan Hettinger, and Ryan N King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117(29): 16805–16815, 2020.

John R Taylor and Andrew F Thompson. Submesoscale dynamics in the upper ocean. *Annual Review of Fluid Mechanics*, 55(1):103–127, 2023.

Leif N Thomas, Amit Tandon, and Amala Mahadevan. Submesoscale processes and dynamics. 2008.

Yuchuan Tian, Hanting Chen, Chao Xu, and Yunhe Wang. Image processing gnn: Breaking rigidity in super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24117, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Zhong Yi Wan, Ricardo Baptista, Anudhyan Boral, Yi-Fan Chen, John Anderson, Fei Sha, and Leonardo Zepeda-Núñez. Debias coarsely, sample conditionally: Statistical downscaling through optimal transport and probabilistic diffusion models. *Advances in Neural Information Processing Systems*, 36:47749–47763, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems*, 36: 13294–13307, 2023.

Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023.

Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2856–2865, 2024.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018.

Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3499–3509. Curran Associates, Inc., 2020.

## A Message Passing on Graphs

Several components of the proposed method—the channel graph in §3.3 and the pixel graph in §3.4—are built on the message-passing framework for graph-structured computation, which we review here for completeness.

Let $\mathcal{G} = (V, E)$ be a graph with node set $V = \{v_1, \ldots, v_K\}$ and edge set $E \subseteq V \times V$. Each node $v_k$ carries a feature vector $\mathbf{h}_k^{(l)} \in \mathbb{R}^d$ at layer $l$, and its neighbourhood is $\mathcal{N}(k) = \{j : (v_j, v_k) \in E\}$.

A single message-passing layer updates every node representation through three stages–message construction, aggregation, and update:

$$\mathbf{m}_{j\to k}^{(l)} = \psi^{(l)}\big(\mathbf{h}_j^{(l)},\ \mathbf{h}_k^{(l)},\ \mathbf{e}_{jk}\big), \tag{7}$$

$$\mathbf{m}_k^{(l)} = \bigoplus_{j\in\mathcal{N}(k)} \mathbf{m}_{j\to k}^{(l)}, \tag{8}$$

$$\mathbf{h}_k^{(l+1)} = \gamma^{(l)}\big(\mathbf{h}_k^{(l)},\ \mathbf{m}_k^{(l)}\big), \tag{9}$$

where $\psi^{(l)}$ is a learnable message function, $\mathbf{e}_{jk}$ is an optional edge attribute encoding prior relational information, $\bigoplus$ is a permutation-invariant aggregator (e.g. summation, mean, or max), and $\gamma^{(l)}$ is a learnable update function. After $L$ successive layers the receptive field of each node spans its $L$-hop neighbourhood in $\mathcal{G}$, allowing information to propagate across the graph. This formulation is general: specific instantiations of $\psi$, $\bigoplus$, and $\gamma$ recover well-known architectures including Graph Convolutional Networks (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), and Graph Attention Networks (Veličković et al., 2018).

In the main text we instantiate this framework twice: (i) on a channel graph whose nodes correspond to spatial and temporal channels, so that each round of message passing realizes cross-channel—and hence cross-scale—information exchange (§3.3); and (ii) on a pixel graph with spectrally adaptive connectivity, where each round of message passing performs spatially varying aggregation guided by the local frequency content (§3.4).

## B Spatial Kernels

The spatial kernels comprising $V_S$ are now specified in full. Recall that each kernel $\mathcal{S}_i$ is required to select a distinct region of the spatial spectrum while satisfying the coverage, separation, and non-redundancy conditions stated in §3.2. We instantiate the bank with four classical kernels—Difference of Gaussians, Laplacian, Sobel gradient magnitude, and local standard deviation—chosen because they collectively span bandpass, highpass-isotropic, highpass-oriented, and broadband-nonlinear regimes.

**Definition B.1.** *Let $G_\sigma(\mathbf{x}) = \frac{1}{2\pi\sigma^2}\exp\big(-\|\mathbf{x}\|^2/2\sigma^2\big)$ denote the isotropic Gaussian with scale $\sigma$. We define the following spatial channel kernels applied to a reference frame $y_{\mathrm{ref}}$.*

**1. Difference of Gaussians (DoG).** A bandpass filter isolating structure between scales $\sigma_1 < \sigma_2$:

$$s_{\mathrm{DoG}}(\mathbf{x}) = (G_{\sigma_1} - G_{\sigma_2}) * y_{\mathrm{ref}}(\mathbf{x}) \tag{10}$$

with transfer function

$$\hat{s}_{\mathrm{DoG}}(\boldsymbol{\omega}) = \Big(e^{-\sigma_1^2\|\boldsymbol{\omega}\|^2/2} - e^{-\sigma_2^2\|\boldsymbol{\omega}\|^2/2}\Big)\hat{y}_{\mathrm{ref}}(\boldsymbol{\omega}), \tag{11}$$

which peaks at $\|\boldsymbol{\omega}\|^* = \sqrt{2\ln(\sigma_2/\sigma_1)/(\sigma_2^2-\sigma_1^2)}$. This selects medium-frequency texture and blob-like features.

**2. Laplacian (Lap).** The isotropic second-order derivative kernel:

$$s_{\mathrm{Lap}}(\mathbf{x}) = \nabla^2 y_{\mathrm{ref}}(\mathbf{x}) = \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}\right) y_{\mathrm{ref}}(\mathbf{x}) \tag{12}$$

with transfer function

$$\hat{s}_{\mathrm{Lap}}(\boldsymbol{\omega}) = -\|\boldsymbol{\omega}\|^2\,\hat{y}_{\mathrm{ref}}(\boldsymbol{\omega}). \tag{13}$$

This is a pure highpass kernel that amplifies fine-scale discontinuities—edges and corners—while annihilating any polynomial of degree $\leq 1$, including the DC atmospheric component $\ell_{\mathrm{ref}}$.

**3. Sobel (directional gradient).** First-order directional derivatives:

$$s_{\mathrm{Sob}}(\mathbf{x}) = \sqrt{(g_x * y_{\mathrm{ref}})^2 + (g_y * y_{\mathrm{ref}})^2} \tag{14}$$

where $g_x, g_y$ are the horizontal and vertical Sobel kernels with transfer functions $\hat{g}_x(\boldsymbol{\omega}) \approx i\omega_1\,\mathrm{sinc}(\omega_2)$. This captures oriented edge magnitude—the local gradient norm $\|\nabla f(\mathbf{x})\|$—providing geometric boundary information invariant to constant illumination offset.

**4. Local Standard Deviation (Std).** A nonlinear local contrast descriptor:

$$s_{\mathrm{Std}}(\mathbf{x}) = \sqrt{G_\sigma * y_{\mathrm{ref}}^2(\mathbf{x}) - \big(G_\sigma * y_{\mathrm{ref}}(\mathbf{x})\big)^2} \tag{15}$$

This measures the root-mean-square deviation within the Gaussian window of scale $\sigma$. In the Fourier domain it does not admit a linear transfer function, but it serves as a pointwise estimator of the local spectral energy above DC within the window, i.e.,

$$s_{\mathrm{Std}}(\mathbf{x}) \approx \left( \int_{\|\boldsymbol{\omega}\|>0} |\hat{r}_{\mathrm{loc}}(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \right)^{1/2}. \tag{16}$$

This encodes local texture richness.

**Full Spatial Channel Set.** The spatial node set

$$V_S = \big\{ s_{\mathrm{DoG}},\ s_{\mathrm{Lap}},\ s_{\mathrm{Sob}},\ s_{\mathrm{Std}} \big\} \tag{17}$$

constitutes a collection to comprehend the spatial structure of the scene reflectance $r$ within this framework. Each channel occupies a distinct role: DoG provides selective bandpass sensitivity at intermediate scales; Lap supplies isotropic second-order curvature response at fine scales; Sob encodes oriented first-order boundary geometry; and Std delivers a nonlinear, phase-invariant measure of local texture energy across all frequencies. No channel can be derived from, or substituted by, any combination of the others—they differ simultaneously in differential order (zeroth, first, second, shaped bandpass), isotropy versus orientation, and linearity versus nonlinearity. Removing any single element leaves an irrecoverable gap in either scale, directional, or statistical coverage. All four channels share the property of suppressing the slowly varying atmospheric field $\ell_{\mathrm{ref}}$, ensuring that $V_S$ encodes intrinsic scene geometry and texture free from illumination nuisance. The four channels integrate into a full quartet for the spatiotemporal analysis that follows.

# C Node-wise Spatiotemporal Dynamics Modelling

This appendix provides the full specification of the channel graph structure and the node-wise feature transformation. We first formalize the graph topology, then describe the node-level processing and the cross-channel coupling mechanism.

## C.1 Graph Topology

Let the channel graph be $\mathcal{G} = (V, E)$ with $V = V_S \cup V_T$, where $|V_S| = K_S$ spatial nodes and $|V_T| = K_T$ temporal nodes yield $|V| = K_S + K_T$ channels in total. The edge set is composed of three structurally distinct subsets:

$$E = E_{SS} \ \cup \ E_{TT} \ \cup \ E_{ST}. \tag{18}$$

**Intra-spatial complete graph.** Within the spatial node set, every pair of nodes is connected:

$$E_{SS} = \{(v_i, v_j) : v_i, v_j \in V_S,\ i \neq j\}, \quad |E_{SS}| = K_S(K_S - 1). \tag{19}$$

Since each spatial node encodes a different frequency band or geometric feature of the observed field, full connectivity ensures that every scale can directly inform every other scale. This is motivated by domain knowledge that geophysical phenomena exhibit strong cross-scale coupling: energy cascades, frontogenesis, and instability mechanisms all involve simultaneous interaction across multiple spatial scales.

**Intra-temporal complete graph.** Analogously, the temporal nodes form a complete graph:

$$E_{TT} = \{(v_i, v_j) : v_i, v_j \in V_T,\ i \neq j\}, \quad |E_{TT}| = K_T(K_T - 1). \tag{20}$$

Different temporal modes—trends, periodic components, residual variability—interact dynamically (e.g. a seasonal cycle modulates the amplitude of sub-monthly fluctuations), and full connectivity provides the capacity to learn such inter-mode dependencies.

**Inter-set bipartite graph.** The spatial and temporal node sets are coupled through a complete bipartite graph:

$$E_{ST} = \{(v_i, v_j) : v_i \in V_S,\ v_j \in V_T\} \cup \{(v_i, v_j) : v_i \in V_T,\ v_j \in V_S\}, \quad |E_{ST}| = 2\,K_S\,K_T. \tag{21}$$

These bidirectional edges serve as the sole pathway through which spatial structure and temporal dynamics are fused. Every spatial channel can send information to every temporal channel and vice versa, enabling the model to discover which spatial scales are most predictive for which temporal modes. This design avoids imposing a priori assumptions about spatial–temporal correspondence and instead allows the attention mechanism (see §C.4) to learn the coupling structure from data.

**Overall graph properties.** The total number of directed edges is

$$|E| = K_S(K_S - 1) + K_T(K_T - 1) + 2\,K_S K_T. \tag{22}$$

By construction, $\mathcal{G}$ is not a single complete graph on $V$—the intra-set and inter-set edges play qualitatively different roles and are equipped with distinct learnable parameters (see §C.4). However, the graph diameter is at most 2: any two nodes in the same set are directly connected, and any spatial–temporal pair is connected by a single bipartite edge. After just one round of message passing, every node therefore has access to information from the entire graph.

### C.2 Node State Initialization

Each node $v_k$ carries a spatial feature map $\mathbf{h}_k \in \mathbb{R}^{H_l \times W_l \times d}$, where $d$ is the internal feature dimension. For spatial nodes, the initial state is obtained by projecting the corresponding spatial-kernel output:

$$\mathbf{h}_k^{(0)} = \phi_k^{(\text{init})}\big(S_\tau^k\big), \quad v_k \in V_S, \tag{23}$$

where $S_\tau^k = \mathcal{S}_k[\mathbf{x}_\tau^{\text{LR}}]$ is the $k$-th spatial channel from Eq. (2) and $\phi_k^{(\text{init})}$ is a learnable embedding (e.g. a $1\times1$ convolution). For temporal nodes, the initial state is similarly obtained from the corresponding frame in the sliding window:

$$\mathbf{h}_k^{(0)} = \phi_k^{(\text{init})}\big(\mathbf{x}^{\text{LR}}_{\tau-\lfloor m/2\rfloor+k-1}\big), \quad v_k \in V_T. \tag{24}$$

### C.3 Node-wise Feature Transformation

Between message-passing rounds, each node's feature map is updated by a learnable spatial operator $\Phi_k$ applied independently to each node:

$$\tilde{\mathbf{h}}_k^{(l)} = \Phi_k^{(l)}\big(\mathbf{h}_k^{(l)}\big), \tag{25}$$

where $\Phi_k^{(l)}$ is implemented as a small convolutional block (e.g. two $3\times3$ convolutions with nonlinear activations and residual connections). Each node maintains its own parameters, reflecting the inductive bias that different spatial scales and temporal modes benefit from distinct local processing.

**Remark C.1** (Connection to spatiotemporal PDEs)**.** *Although the node-wise operators $\Phi_k$ are not constrained to any particular PDE form, their structure—local convolutions acting on spatial feature maps, interleaved with cross-channel exchange via graph edges—can be interpreted as an implicit discretization of a coupled advection–diffusion–reaction system. In this view, the convolutional kernels subsume the roles of advection (directional transport) and diffusion (isotropic smoothing), while the cross-channel messages (§C.4) play the role of reaction terms coupling different fields. This connection to classical PDE dynamics is not enforced but may partly explain the model's effectiveness on geophysical data, where such processes are common in geophysical dynamics.*

### C.4 Cross-Channel Coupling via Graph Edges

The external input $\mathbf{u}_k$ to node $v_k$ is assembled from messages sent by its graph neighbours $\mathcal{N}(k) = \{j : (v_j, v_k) \in E\}$. Due to the tripartite edge structure, the neighbourhood decomposes as

$$\mathcal{N}(k) = \mathcal{N}_{SS}(k) \cup \mathcal{N}_{TT}(k) \cup \mathcal{N}_{ST}(k), \tag{26}$$

where $\mathcal{N}_{SS}(k)$ contains the intra-spatial neighbours (non-empty only if $v_k \in V_S$), $\mathcal{N}_{TT}(k)$ the intra-temporal neighbours (non-empty only if $v_k \in V_T$), and $\mathcal{N}_{ST}(k)$ the cross-set neighbours from

the bipartite edges. To allow the model to distinguish the qualitatively different roles of intra-set and inter-set messages, we employ separate projection matrices for each edge type. The aggregated input is

$$\mathbf{u}_k = \sum_{j \in \mathcal{N}_{\text{intra}}(k)} \alpha_{jk}^{\text{intra}}\, \mathbf{W}_{jk}^{(\text{intra})}\, \mathbf{h}_j \;+\; \sum_{j \in \mathcal{N}_{ST}(k)} \alpha_{jk}^{\text{inter}}\, \mathbf{W}_{jk}^{(\text{inter})}\, \mathbf{h}_j, \tag{27}$$

where $\mathcal{N}_{\text{intra}}(k) = \mathcal{N}_{SS}(k) \cup \mathcal{N}_{TT}(k)$ collects the within-set neighbours. The attention coefficients for each edge type are computed independently:

$$\alpha_{jk}^{(\cdot)} = \frac{\exp\big(\text{LeakyReLU}(\mathbf{a}_{(\cdot)}^{\top}[\mathbf{W}_{(\cdot)}^{(q)}\mathbf{h}_k,\ \mathbf{W}_{(\cdot)}^{(m)}\mathbf{h}_j])\big)}{\sum_{j' \in \mathcal{N}_{(\cdot)}(k)} \exp\big(\text{LeakyReLU}(\mathbf{a}_{(\cdot)}^{\top}[\mathbf{W}_{(\cdot)}^{(q)}\mathbf{h}_k,\ \mathbf{W}_{(\cdot)}^{(m)}\mathbf{h}_{j'}])\big)}, \tag{28}$$

where the superscript $(\cdot)$ denotes either intra or inter, and each type maintains its own attention vector $\mathbf{a}_{(\cdot)}$ and projection matrices. This separation ensures that the model does not conflate within-modality scale interactions (e.g. energy cascades across spatial frequency bands) with cross-modality fusion (e.g. a fine-scale spatial feature informing a high-frequency temporal mode).

### C.5 Full Update Rule

Combining the node-wise transformation (Eq. (25)) and the cross-channel aggregation, the complete layer-$l$ update for node $v_k$ is:

$$\mathbf{h}_k^{(l+1)} = \gamma^{(l)}\Big(\tilde{\mathbf{h}}_k^{(l)},\ \mathbf{u}_k^{(l)}\Big), \tag{29}$$

where $\gamma^{(l)}$ is a learnable update function (e.g. a gated residual block). After $L$ such layers, the final node representations $\{\mathbf{h}_k^{(L)}\}_{k=1}^{|V|}$ are forwarded to the subsequent pixel-level processing stage (§3.4).

## D Multi-hop Spectral-aware Information Propagation

This appendix provides the complete formulation of the per-pixel degree budget designed for multi-hop spectral aware information propogation, its relationship to the complexity indicator used in IPG, the spectral justification for reversing the budget, and implementation details.

### D.1 Pixel Graph Formulation

The spatial graph operates at the pixel level. Each node is a pixel, and each pixel carries a multi-channel feature vector rather than a scalar. Let $P = H \times W$ be the number of pixels. The pixel graph $\mathcal{G}^{\text{pix}} = (V^{\text{pix}}, E^{\text{pix}})$ is defined with

$$V^{\text{pix}} = \{p_1, p_2, \ldots, p_P\}, \qquad \mathbf{h}_i = \big[c_1(\mathbf{x}_i, t),\ \ldots,\ c_K(\mathbf{x}_i, t)\big]^{\top} \in \mathbb{R}^K, \tag{30}$$

assembling all $K$ spectral channel values at location $\mathbf{x}_i$ into a single feature vector. Crucially, the edge set $E^{\text{pix}}$ is not a fixed grid connectivity; instead, each pixel's neighbourhood size (degree) is governed by an adaptive budget derived from the local spectral complexity of the feature map.

### D.2 Complexity Indicator via Downsample–Upsample Residual

**Definition.** Let $\mathbf{F} \in \mathbb{R}^{H \times W \times K}$ denote the spatial feature map whose $(i)$-th row is $\mathbf{h}_i$. We define the downsample–upsample operator at scale $s$ as

$$\mathcal{R}_s[\mathbf{F}] = (\mathcal{U}_s \circ \mathcal{D}_s)[\mathbf{F}], \tag{31}$$

where $\mathcal{D}_s \colon \mathbb{R}^{H \times W \times K} \to \mathbb{R}^{\lfloor H/s \rfloor \times \lfloor W/s \rfloor \times K}$ is spatial downsampling by factor $s$ (e.g. bilinear or area averaging) and $\mathcal{U}_s$ upsamples back to $H \times W$ (e.g. bilinear interpolation). The composition $\mathcal{R}_s$ acts as a low-pass filter: downsampling removes spatial frequencies above the Nyquist limit $\pi/s$ of the coarser grid, and upsampling merely interpolates without recovering the discarded high-frequency content.

The complexity indicator at pixel $p_i$ is then the per-pixel residual norm:

$$\xi_i = \big\|\mathbf{F}(\mathbf{x}_i) - \mathcal{R}_s[\mathbf{F}](\mathbf{x}_i)\big\|_2 = \left(\sum_{k=1}^{K} \big|c_k(\mathbf{x}_i, t) - [\mathcal{R}_s[\mathbf{F}]]_k(\mathbf{x}_i)\big|^2\right)^{1/2}. \tag{32}$$

**Interpretation.** The residual $\mathbf{F} - \mathcal{R}_s[\mathbf{F}]$ isolates exactly the spatial frequencies in the band $(\pi/s,\ \pi]$. A large $\xi_i$ signifies that pixel $p_i$ lies in a region rich in high-frequency structure (edges, textures, fine patterns), whereas a small $\xi_i$ indicates a locally smooth region. This is the same indicator used in the IPG framework (Tian et al., 2024) to characterize per-node information complexity.

### D.3 Degree Budget: IPG vs. Our Reversed Strategy

**IPG budget (proportional).** In the IPG framework, the adaptive propagation budget for each node is set proportional to the complexity indicator. When applied to a pixel graph, this gives each pixel $p_i$ a degree budget

$$d_i^{\mathrm{IPG}} = d_{\min} + \left\lfloor (d_{\max} - d_{\min}) \cdot \frac{\xi_i - \xi_{\min}}{\xi_{\max} - \xi_{\min}} \right\rfloor, \tag{33}$$

where $\xi_{\min} = \min_i \xi_i$ and $\xi_{\max} = \max_i \xi_i$. Under this policy, pixels in complex regions receive more neighbours, under the reasoning that information-rich nodes benefit from broader aggregation.

**Our budget (reciprocal).** We argue that this reasoning is spectrally backwards for the pixel graph after spectral decomposition. Since the multi-channel feature $\mathbf{h}_i$ has been explicitly constructed to encode a spectrum of spatial frequencies, the goal of the pixel graph is not to gather more information from complex regions but to preserve the high-frequency content that was carefully separated. Broader aggregation at a high-frequency pixel subjects its features to additional low-pass smoothing, defeating the purpose of the spectral decomposition.

We therefore reverse the relationship and define the degree budget as the reciprocal of the complexity indicator:

$$d_i = \mathrm{clamp}\left( \left\lfloor \frac{C}{\xi_i + \epsilon} \right\rfloor,\ d_{\min},\ d_{\max} \right), \tag{34}$$

where $C > 0$ is a global scaling constant chosen so that the median budget falls near $(d_{\min} + d_{\max})/2$, and $\epsilon > 0$ is a small constant for numerical stability. Equivalently, using the normalized indicator:

$$d_i = d_{\max} - \left\lfloor (d_{\max} - d_{\min}) \cdot \frac{\xi_i - \xi_{\min}}{\xi_{\max} - \xi_{\min}} \right\rfloor. \tag{35}$$

Table 3: Comparison of degree-budget strategies.

| | **IPG** | **Ours (Reversed)** |
|---|---|---|
| Budget relationship | $d_i \propto \xi_i$ | $d_i \propto 1/\xi_i$ |
| Complex/textured pixels | High degree (many neighbours) | Low degree (few neighbours) |
| Smooth pixels | Low degree (few neighbours) | High degree (many neighbours) |
| Underlying rationale | Information gathering | Spectral preservation |
| Treats high-freq. content as | Signal to aggregate | Signal to protect |

**Comparison.**

### D.4 Spectral Justification

**Low-pass filtering and degree.** Let $\hat{\mathbf{A}}$ denote the symmetrically normalized adjacency of $\mathcal{G}^{\mathrm{pix}}$. Each round of message passing applies the operator $\hat{\mathbf{A}}$ to the feature map, which acts as a low-pass filter on the graph spectrum (Nt & Maehara, 2020; Oono & Suzuki, 2020). The strength of this smoothing at a given node $p_i$ is controlled by two factors: (i) the number of hops $L$, and (ii) the degree $d_i$, which determines how many neighbours contribute to the local average.

At a node with large degree, each hop replaces the node's feature with an average over a wider set of neighbours, resulting in stronger smoothing per hop. Conversely, a node with small degree averages over fewer neighbours, and the per-hop smoothing effect is milder.

**Alignment with spectral content.** Pixels in textured regions carry significant energy in the high-frequency channels of $\mathbf{h}_i$. By assigning these pixels a small degree budget, we limit the per-hop smoothing strength, ensuring that the high-frequency content encoded in channels $c_k$ for large $\omega_k$ is not unduly attenuated. Conversely, pixels in smooth regions carry energy predominantly in the low-frequency channels. These channels are robust to smoothing (their energy sits near the DC component of the graph spectrum), so a large degree budget broadens the receptive field without distorting the signal.

**Quantitative bound.** Consider a pixel $p_i$ with degree $d_i$ and its local neighbourhood subgraph. The effective low-pass suppression at graph eigenvalue $\lambda$ after $L$ hops on the local star subgraph is approximately

$$g(\lambda, d_i, L) \approx \left(1 - \frac{\lambda}{d_i}\right)^L. \tag{36}$$

For fixed $\lambda$ and $L$, this is a monotonically increasing function of $d_i$: larger degree means the eigenvalue $\lambda$ is more "diluted," yielding weaker suppression per hop. However, the aggregate effect over the neighbourhood also broadens the spatial averaging kernel, which is appropriate only when the signal is spatially smooth. Our reciprocal budget ensures:

- High $\xi_i$ (textured pixel, high-frequency energy): $d_i$ is small $\Rightarrow$ narrow aggregation kernel $\Rightarrow$ high-frequency detail preserved.
- Low $\xi_i$ (smooth pixel, low-frequency energy): $d_i$ is large $\Rightarrow$ wide aggregation kernel $\Rightarrow$ large receptive field, no spectral distortion.

### D.5 Adaptive Graph Construction

Given the per-pixel degree budgets $\{d_i\}_{i=1}^P$, we construct $E^{\text{pix}}$ as follows:

1. **Candidate generation.** For each pixel $p_i$, identify all pixels within a maximum search radius: $\mathcal{C}(i) = \{p_j : \|\mathbf{x}_i - \mathbf{x}_j\|_2 \le r_{\max},\ j \ne i\}$.
2. **Proximity ranking.** Sort $\mathcal{C}(i)$ by Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ in ascending order.
3. **Degree-budgeted selection.** Retain the closest $d_i$ neighbours:
$$\mathcal{N}(i) = \text{top-d}_\text{i}\big(\mathcal{C}(i)\big). \tag{37}$$
4. **Symmetrization (optional).** To ensure an undirected graph, add the reverse edge for every selected edge: if $p_j \in \mathcal{N}(i)$, then $p_i \in \mathcal{N}(j)$ (this may cause some nodes to slightly exceed their budget; the budget is treated as a soft upper bound).

On a regular grid with $r_{\max}$ set to include at least the $d_{\max}$-nearest pixels (e.g. $r_{\max} = \sqrt{2}\lceil\sqrt{d_{\max}}\rceil$ for a square grid), this procedure smoothly interpolates between sparse 4-connectivity (when $d_i \approx 4$) and dense block connectivity (when $d_i$ is large).

### D.6 Message Passing on the Adaptive-Degree Graph

With $\mathcal{G}^{\text{pix}}$ constructed, we perform $L$ rounds of message passing on the multi-channel pixel features:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{j\in\mathcal{N}(i)} \frac{1}{\sqrt{|\mathcal{N}(i)|\,|\mathcal{N}(j)|}}\,\mathbf{W}_{\text{msg}}^{(l)}\,\mathbf{h}_j^{(l)}\right), \tag{38}$$

where $\mathbf{W}^{(l)}, \mathbf{W}_{\text{msg}}^{(l)} \in \mathbb{R}^{K\times K}$ are learnable weight matrices, and the normalization $1/\sqrt{|\mathcal{N}(i)|\,|\mathcal{N}(j)|}$ is the standard symmetric normalization adapted to the variable-degree graph. The output after $L$ hops is $\tilde{\mathbf{h}}_i = \mathbf{h}_i^{(L)} \in \mathbb{R}^K$.

Because $|\mathcal{N}(i)| = d_i$ varies across pixels, the effective smoothing operator is spatially heterogeneous:

- At smooth pixels with large $d_i$, each hop averages over many neighbours, producing broad spatial smoothing that is spectrally appropriate for the low-frequency content present.

- At textured pixels with small $d_i$, each hop averages over very few neighbours, applying only mild smoothing that preserves the high-frequency detail encoded in the multi-channel feature vector.

### D.7 Choice of Scale Factor $s$

The downsample factor $s$ in the complexity indicator (32) controls which spatial frequencies contribute to $\xi_i$. A small $s$ (e.g. $s = 2$) isolates only the very highest frequencies (near-Nyquist), while a large $s$ captures a broader range. In practice, we tie $s$ to the spectral decomposition: if the decomposition separates the signal at cutoff frequencies $\{\omega_k\}$, we set $s$ such that $\pi/s$ aligns with the median cutoff, ensuring that $\xi_i$ discriminates between pixels that are "spectrally simple" (energy below $\pi/s$) and "spectrally complex" (significant energy above $\pi/s$).

### D.8 Computational Complexity

Computing the complexity indicator requires one downsample and one upsample pass over the feature map, costing $\mathcal{O}(PK)$. The $k$-nearest-neighbour graph construction with maximum degree $d_{\max}$ within radius $r_{\max}$ can be performed in $\mathcal{O}(P \cdot |\mathcal{C}| \log |\mathcal{C}|)$ where $|\mathcal{C}|$ is the average candidate set size determined by $r_{\max}$; on a regular grid this is $\mathcal{O}(P \cdot d_{\max} \log d_{\max})$. Message passing over $L$ hops costs $\mathcal{O}(L \cdot |E^{\text{pix}}| \cdot K)$, where $|E^{\text{pix}}| = \sum_i d_i = \bar{d} \cdot P$ with $\bar{d}$ being the average degree. Since high-frequency pixels (which are typically more numerous in natural images) receive small $d_i$, the average degree $\bar{d}$ is substantially below $d_{\max}$, yielding computational savings over a uniform $d_{\max}$-regular graph.

## E Front Evaluation Metrics

FIU-MAE is an evaluation metric only. It is not used for training, hyperparameter selection, early stopping, or checkpoint selection; all models are trained and selected with MSE. Its purpose is to report fidelity around salient frontal structures in addition to global field metrics.

The proposed FIU-MAE 1 is given as below.

**Algorithm 1** Front Intensity Union-Weighted MAE (FIU-MAE)

**Require:** Predicted salinity field $\hat{S}$, ground-truth salinity field $S$, local window size $w$, percentile $p$
**Ensure:** Front intensity union-weighted MAE $\mathcal{E}$

Compute gradient magnitudes: $G = \|\nabla S\|$, $\hat{G} = \|\nabla \hat{S}\|$
*// Adaptive front detection*
**for** each spatial block $B_k$ of size $w \times w$ **do**
    Compute local threshold: $\tau_k = \text{Percentile}(G|_{B_k},\ p)$
    Generate local masks: $M_k = \mathbf{1}[G|_{B_k} > \tau_k]$, $\hat{M}_k = \mathbf{1}[\hat{G}|_{B_k} > \tau_k]$
**end for**
Assemble full masks: $M = \bigcup_k M_k$, $\hat{M} = \bigcup_k \hat{M}_k$
*// Union of front support*
$U = M \cup \hat{M}$
*// Intensity-weighted evaluation*
Compute weights: $w_{ij} = G(i, j)$ for $(i, j) \in U$
**return** $\mathcal{E} = \frac{\sum_{(i,j)\in U} w_{ij} \left|\hat{G}(i,j) - G(i,j)\right|}{\sum_{(i,j)\in U} w_{ij}}$

## F Spectral Fidelity Computation

For each patch we compute the mean-removed 2D FFT, form $\text{PSD} = |F|^2$, and retain coefficients in the high-frequency annulus $r \in [0.7, 1.0]$. Curves show sorted $\log_{10}(\text{PSD} + \epsilon)$ values ($\epsilon = 10^{-12}$) versus normalized rank. The spectral distance to ground truth is measured by the high-band log-PSD

Table 4: PSNR under unseen observation degradations. Higher is better.

| Model | D1: blur | D2: noise | D3: masking |
|---|---|---|---|
| SwinIR | 45.92 | 45.59 | 45.85 |
| PFT | 44.80 | 44.56 | 44.75 |
| **DRONE** | **46.04** | **45.83** | **46.03** |

$L_1$ distance:

$$d_{\text{logPSD-}L_1} = N^{-1}\sum_{k=1}^{N}|\tilde{p}_k - \tilde{g}_k|,$$

where $\tilde{p}_k, \tilde{g}_k$ are sorted prediction and ground-truth high-band log-PSD samples over $N$ points.

# G Training and Evaluation Protocol (Ocean SR $\times 16$)

This appendix records the training and evaluation details needed to reproduce the reported comparisons. All primary experiments use the same chronological train/validation/test partition and the same $\times 16$ single-channel super-resolution task.

## G.1 Data and optimization

The target variable is normalized with fixed min–max bounds $[v_{\min}, v_{\max}] = [28.0, 38.0]$; predictions are de-normalized before reporting metrics. Training uses MSE loss, batch size 16, a 10,000-iteration warm-up, validation every 500 iterations, and early stopping on validation MSE with patience 10,000 iterations. Checkpoints are selected by the lowest validation MSE. Learning rates and batch sizes are selected from the validation split, and no test-set information is used for selection. Models are trained from scratch on a single NVIDIA A800 (80 GB) GPU.

## G.2 Evaluation and capacity

All methods use the same no-quantization evaluation path. We report MSE, MAE, RMSE, PSNR, SSIM, and FIU-MAE. All reported variants use the same spatial scale, target variable, and chronological test partition unless an experiment explicitly states otherwise.

# H Additional Experiments

The additional experiments focus on the three strongest models in the main comparison: PFT, SwinIR, and DRONE. Unless an experiment isolates an internal DRONE mechanism, these models are trained and evaluated under the settings in Appendix G. The degree policy study compares DRONE variants by design. For the SST experiment, the same training settings and evaluation procedure are used; PFT did not converge under these settings and repeatedly stopped early, so no PFT SST result is reported.

## H.1 Robustness to unseen observation degradations (RQ5)

**Research question.** Does DRONE retain its advantage when the test-time observation process differs from the bicubic degradation used for training?

**Method.** PFT, SwinIR, and DRONE are trained only with bicubic inputs. At test time, D1 applies isotropic Gaussian blur ($\sigma = 1.5$) before bicubic downsampling, D2 adds zero-mean Gaussian noise ($\sigma = 0.02$ in the normalized $[0, 1]$ scale) after downsampling, and D3 occludes 15% of the HR field with irregular 3–15-pixel masks followed by inpainting and downsampling. No degradation-specific fine-tuning or model selection is used.

The results are reported in Table 4.

DRONE has the highest PSNR under all three shifts, indicating that its improvement is not restricted to the exact bicubic observation operator.

Table 5: Matched-input temporal comparison. Lower is better for MSE, MAE, RMSE, and FIU-MAE; higher is better for PSNR and SSIM.

| Model | MSE | MAE | RMSE | PSNR | SSIM | FIU-MAE |
|---|---|---|---|---|---|---|
| SwinIR-Stack | .0040 | .0424 | .0634 | 43.96 | .9828 | .0465 |
| PFT-Stack | .0049 | .0494 | .0701 | 43.09 | .9802 | .0539 |
| DRONE$_T$ | .0028 | .0354 | .0531 | 45.50 | .9867 | .0179 |
| **DRONE** | **.0024** | **.0320** | **.0493** | **46.13** | **.9881** | **.0167** |

### H.2 Matched temporal inputs (RQ6)

**Research question.** Does the gain come from explicit graph reasoning rather than simply from giving DRONE more temporal frames?

**Method.** SwinIR-Stack and PFT-Stack receive the same $m$ raw salinity frames as DRONE, concatenated along the input-channel dimension. DRONE$_T$ receives those same frames but removes the spatial-kernel channels. All variants are trained from scratch with the same split and optimization settings.

The matched-input results are summarized in Table 5.

With identical temporal inputs, DRONE$_T$ outperforms both stacked baselines, and the spatial-kernel channels provide a further improvement in the complete DRONE model.

### H.3 Degree-budget policies (RQ7)

**Research question.** Does the direction of the complexity-to-degree mapping matter beyond using adaptive connectivity?

**Method.** Only the degree mapping is changed inside DRONE. We explore four variants: reciprocal (the proposed inverse-complexity policy of DRONE), proportional (the original IPG-style policy), constant (the same degree budget for every node), random (uniformly sampled degree budgets). The comparison uses a common three-month subset because of computational-resource limits; all degree mapping variants use the same subset, training settings, and evaluation procedure.

The comparison is reported in Table 6.

Table 6: Degree-budget policies on the common three-month subset.

| Policy | MSE↓ | MAE↓ | RMSE↓ | PSNR↑ | SSIM↑ | FIU-MAE↓ |
|---|---|---|---|---|---|---|
| Reciprocal | **.0041** | **.0429** | **.0612** | **44.62** | **.9917** | **.0395** |
| Proportional | .0106 | .0748 | .0946 | 41.06 | .9905 | .0407 |
| Constant | .0044 | .0457 | .0637 | 44.24 | .9915 | .0414 |
| Random | .0117 | .0813 | .1082 | 40.54 | .9892 | .0456 |

Reciprocal allocation is best across all six metrics. Because this is a controlled subset experiment, its values should not be compared directly with the full-test-set values in Table 1.

### H.4 FIU-MAE sensitivity (RQ8)

**Research question.** Does the model ranking depend on one particular local-window size or percentile threshold?

**Method.** We evaluate the same three models for $w \in \{3, 6\}$ and $p \in \{0.85, 0.90, 0.95\}$.

The sensitivity results are shown in Table 7.

DRONE retains the best ranking in every tested configuration.

Table 7: FIU-MAE sensitivity to local window $w$ and percentile $p$. Lower is better.

| $w$ | $p$ | SwinIR | PFT | DRONE |
|---|---|---|---|---|
| 3 | .85 | .0146 | .0163 | **.0136** |
| 3 | .90 | .0171 | .0192 | **.0156** |
| 3 | .95 | .0196 | .0220 | **.0174** |
| 6 | .85 | .0139 | .0150 | **.0131** |
| 6 | .90 | .0161 | .0180 | **.0147** |
| 6 | .95 | .0170 | .0190 | **.0167** |

### H.5 Dataset-wide spectral fidelity (RQ9)

**Research question.** Does the spectral advantage observed in selected patches persist over the complete held-out test set?

**Method.** We compute FFT Magnitude L1 (overall discrepancy between predicted and ground-truth Fourier magnitudes), high-band radial log-PSD L1 (discrepancy in radially averaged high-frequency log-power spectra), and radial PSD Wasserstein-1 over the complete test set (distributional shift between normalized radial spectral-energy profiles).

The complete-test-set spectral results are reported in Table 8.

Table 8: Dataset-wide spectral distances. Lower is better.

| Model | FFT Magnitude L1 | High-band log-PSD L1 | PSD Wasserstein-1 |
|---|---|---|---|
| SwinIR | .00995 | .84931 | .00085 |
| PFT | .01110 | .97071 | .00098 |
| **DRONE** | **.00931** | **.80929** | **.00057** |

DRONE has the smallest distance for all three dataset-wide spectral measures, complementing the selected-patch visualization in Figure 4.

### H.6 Complete-test-set metric variability

This supplementary analysis reports mean $\pm$ standard deviation over the complete held-out test set, describing sample-level metric dispersion.

The resulting means and dispersions are reported in Table 9.

Table 9: Complete-test-set metric variability.

| Model | MSE | MAE | RMSE | PSNR | SSIM | FIU-MAE |
|---|---|---|---|---|---|---|
| **DRONE** | .00240±.00183 | .03198±.00979 | .04927±.01612 | 46.1344±2.6785 | .98808±.00729 | .01671±.00060 |
| SwinIR | .00251±.00175 | .03324±.00984 | .05018±.01504 | 45.9766±2.4848 | .98780±.00759 | .01702±.00061 |
| PFT | .00291±.00197 | .03638±.01012 | .05372±.01509 | 45.3972±2.2970 | .98609±.00807 | .01902±.00060 |

DRONE retains the best mean performance across the complete test set, while its metric dispersion is comparable to that of the strongest baselines.

### H.7 Transfer to sea-surface temperature (RQ10)

**Research question.** Does the same architecture transfer to another ocean scalar variable without changing the dual-graph design?

**Method.** DRONE, SwinIR, and PFT use the same $\times 16$ training settings, data partitioning, normalization procedure, and evaluation procedure as the salinity experiment. PFT did not converge under these settings and repeatedly stopped early, so its SST result is omitted.

The SST results are shown in Table 10.

Table 10: SST super-resolution results. Lower is better for MSE, MAE, RMSE, and FIU-MAE; higher is better for PSNR and SSIM.

| Model | MSE | MAE | RMSE | PSNR | SSIM | FIU-MAE |
|---|---|---|---|---|---|---|
| SwinIR | .0193 | .0888 | .1390 | 38.73 | .9622 | .0758 |
| **DRONE** | **.0159** | **.0792** | **.1261** | **39.57** | **.9652** | **.0680** |

DRONE improves all reported SST metrics over SwinIR under the same training settings.

### H.8 Additional qualitative cases

We add four extra cases (equidistant sampling) to visualize the performance of the models. In each of figs. 6 to 9: Row 1 stands for reconstructed output fields for each model; Row 2 stands for pixel-wise absolute error relative to the ground truth, plotted on a common color scale. Row 3 stands for front detections overlaid on the reconstructed fields, where solid blue lines denote ground-truth fronts and dashed red lines denote model-detected fronts.

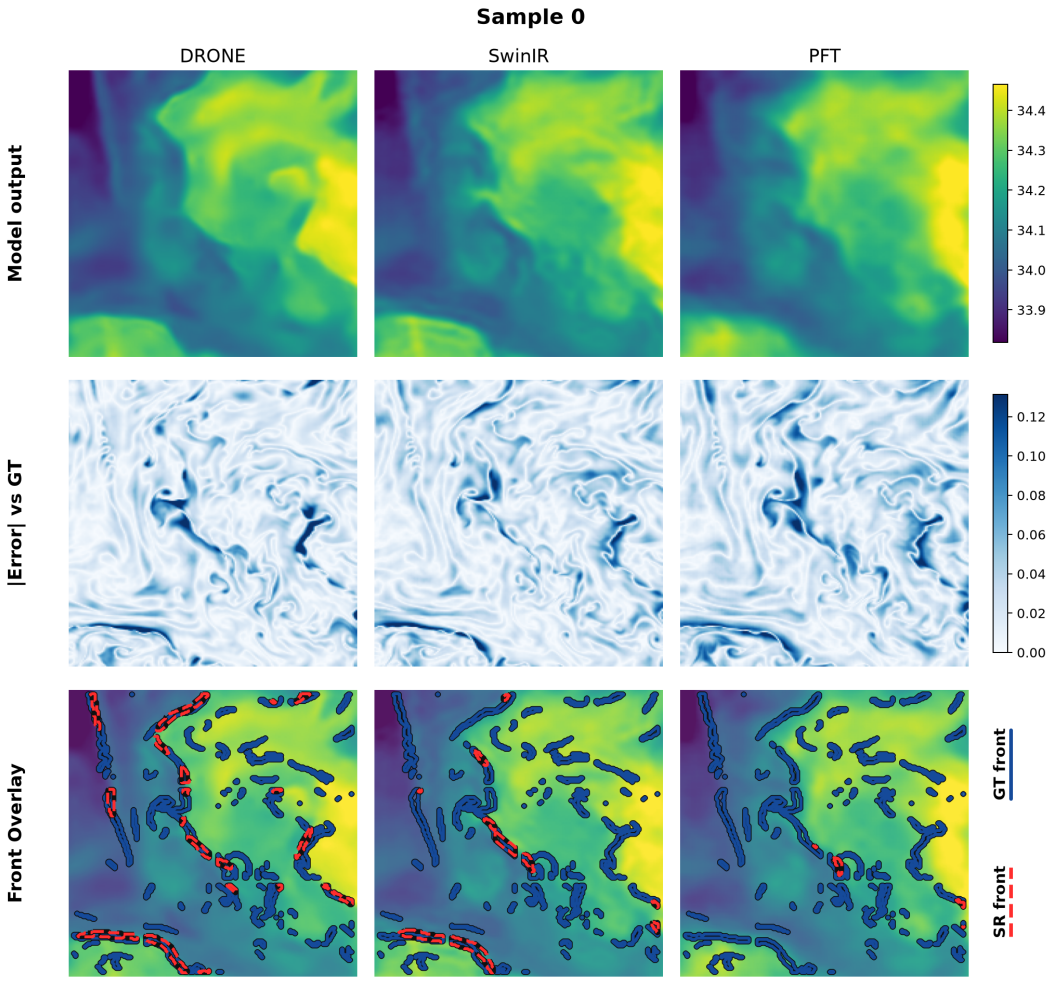


Figure 6: Additional Visualization-1.

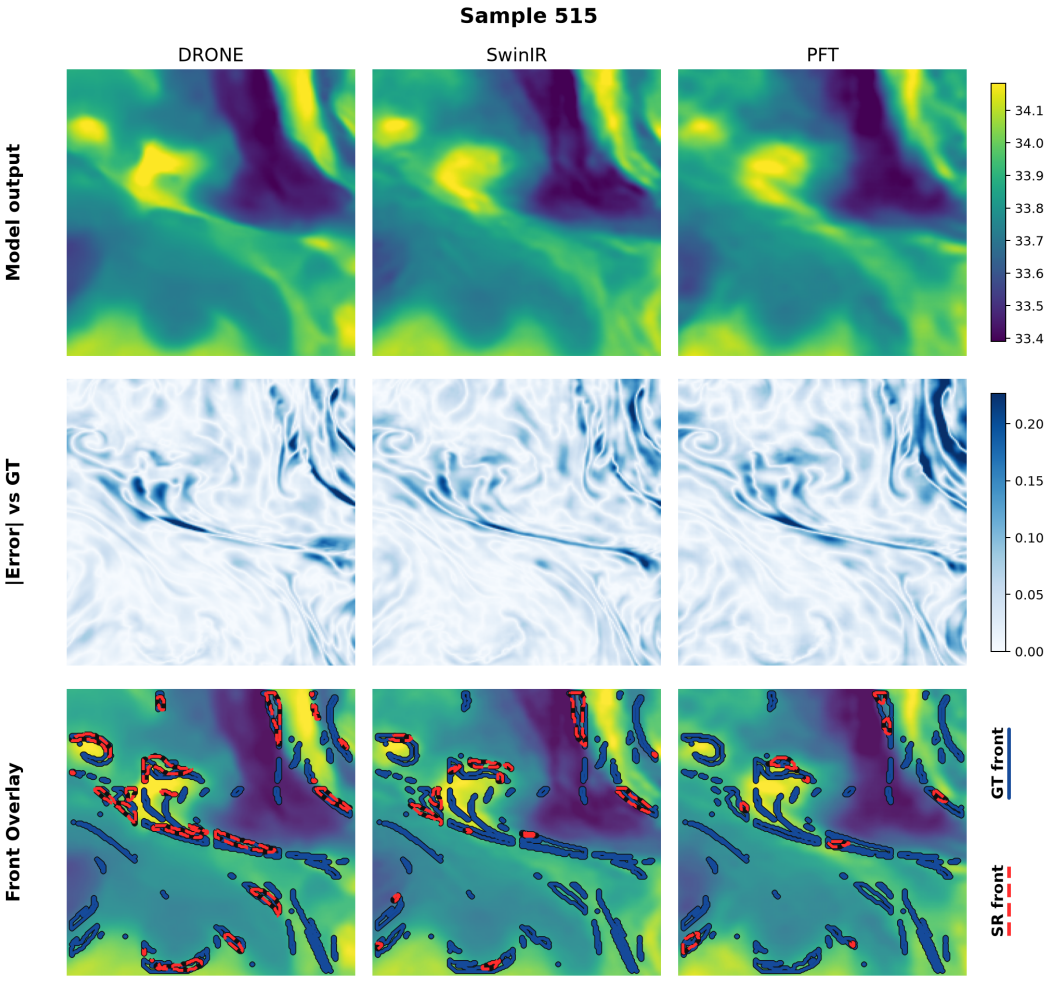


Figure 7: Additional Visualization-2.

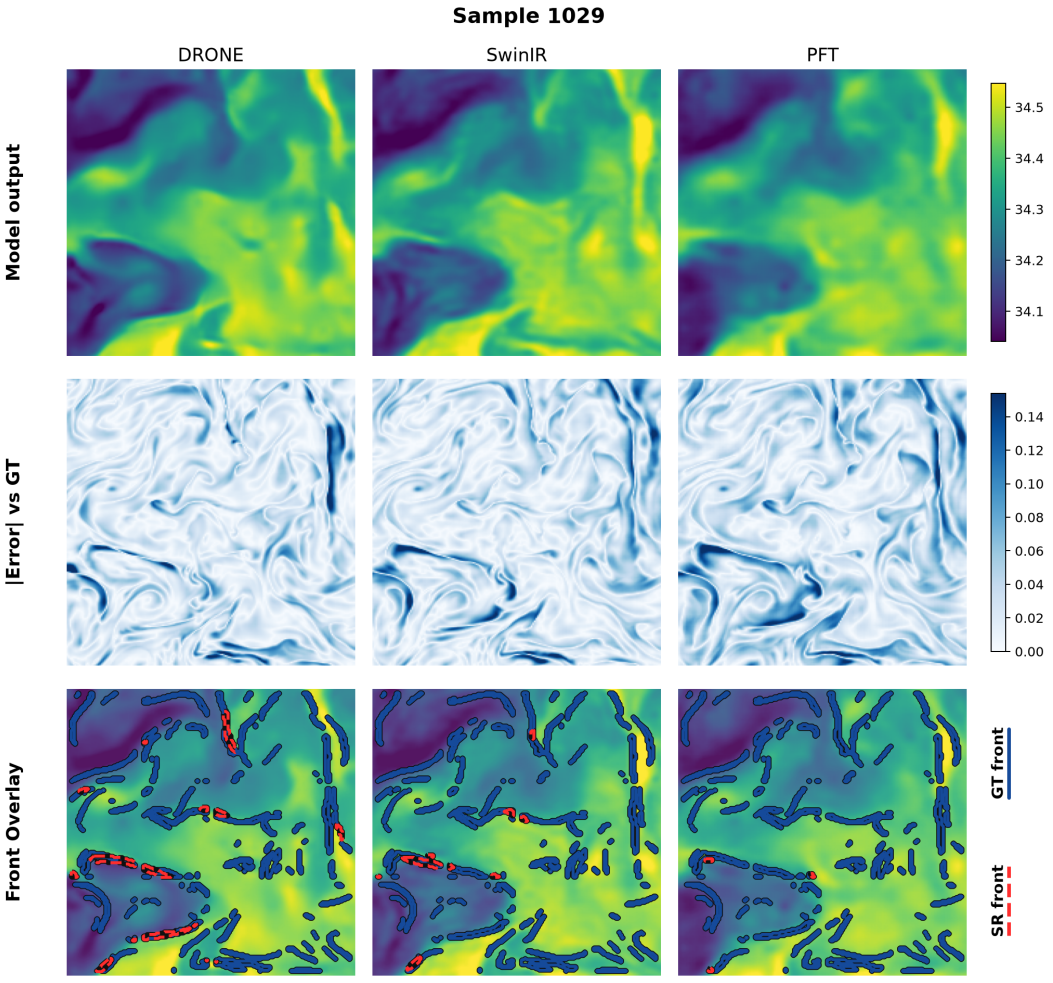


Figure 8: Additional Visualization-3.

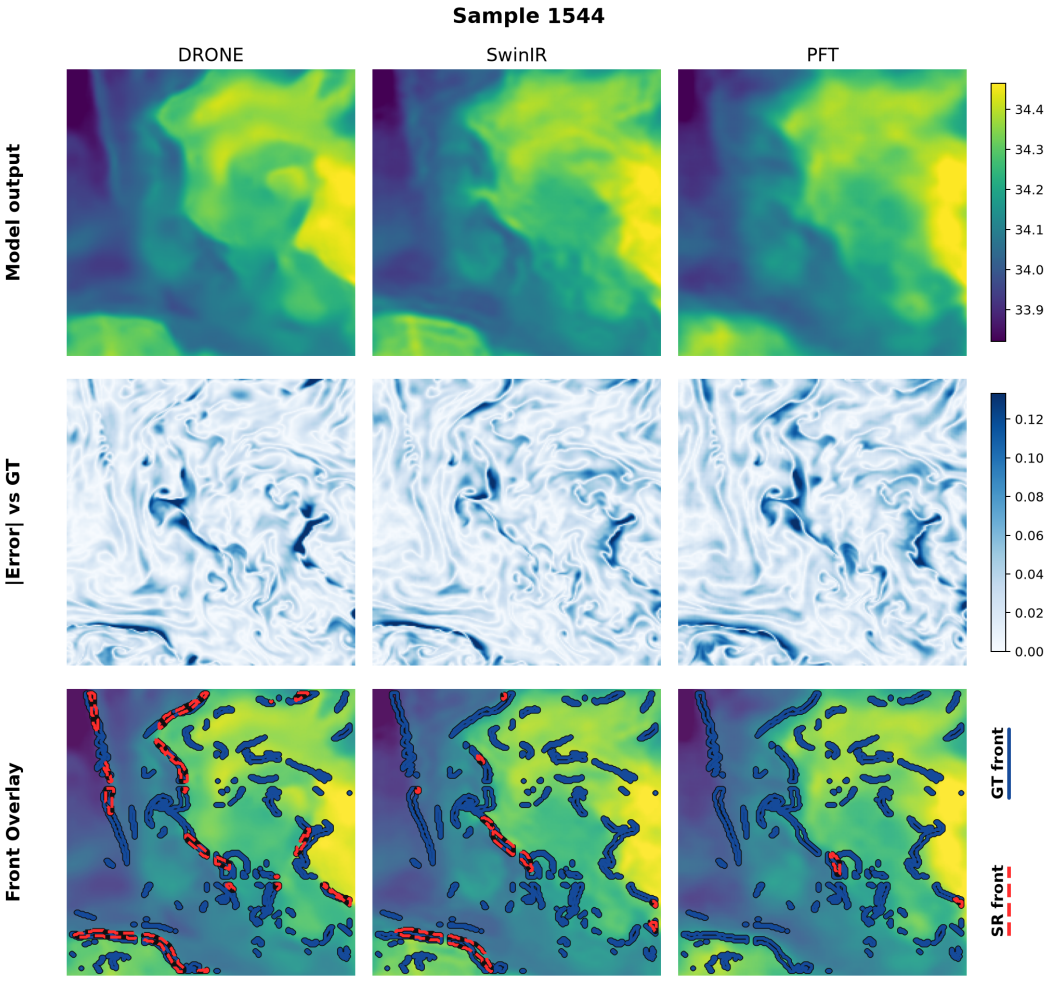


Figure 9: Additional Visualization-4.