KALMAN-INSPIRED CROSS-SCALE SELECTIVE FUSION IN U-NETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard U-Net decoding is globally multi-scale, yet each decoder stage usually receives only its matching encoder feature. We study whether aligned non-matching scales contain task-relevant information beyond that matching skip and how such evidence can be fused. Specifically, we formulate cross-scale fusion as sequential inference from complementary observations with varying reliability. Within a linear-Gaussian latent observation model along the scale axis, the classical Kalman recursion motivates prediction, innovation, and adaptive correction. Guided by this insight, we develop Cross-scale Selective Fusion U-Net (CSF U-Net), which replaces same-scale skip concatenation with a recurrent prediction–innovation–correction scan over aligned encoder scales. Experiments on four medical image segmentation benchmarks show that CSF U-Net consistently improves segmentation accuracy and boundary quality over vanilla U-Net and performs competitively with representative U-Net variants. Ablation studies further support the effectiveness of the recurrent scan, innovation subtraction, and adaptive correction for cross-scale fusion.

1 INTRODUCTION

U-Net-style encoder–decoder architectures (Ronneberger et al., 2015; Çiçek et al., 2016; Milletari et al., 2016; Zhou et al., 2018; Oktay et al., 2018; Isensee et al., 2021; Cao et al., 2022) are a standard backbone for dense prediction, especially in medical image segmentation (Azad et al., 2024). A U-Net encoder produces multi-resolution feature maps with progressively larger receptive fields, while the decoder reconstructs a high-resolution output by repeatedly combining coarse context with fine localization through skip connections. In the vanilla U-Net, each decoder stage receives only the encoder feature at the matching resolution, typically through concatenation followed by a local convolution. Many variants modify this pathway through attention (Oktay et al., 2018), gating (Asadi-Aghbolaghi et al., 2020), pyramid aggregation (Asgari et al., 2019), nested skips (Zhou et al., 2018), or transformer-style interactions (Wang et al., 2022; He et al., 2023). These designs improve skip fusion, but their fusion rules are often motivated architecturally rather than derived from an explicit estimation model.

This paper asks a sharper question: once features from several encoder scales are available during decoding, how much can be lost by restricting a stage to the matching skip? We compare the best estimator that relies only on the matching skip with the best estimator that can also use the remaining aligned scales. An exact risk-gap equality shows that the benefit of additional scales is the squared change in the posterior mean beyond what is available from the decoder feature and matching skip. The equality makes cross-scale complementarity explicit, while frozen-feature probes on trained U-Nets consistently show that aligned non-matching scales reduce prediction error beyond the matching skip.

To understand how information from multiple scales might be fused, we cast decoding as an inference problem. The latent semantic content needed for prediction is inferred from multi-scale encoder features that provide partial, noisy, and complementary evidence. We make this view explicit with a linear-Gaussian latent observation model in which aligned scale features are noisy observations of shared latent semantics. Within this model, the Bayes-optimal fusion rule takes the form of a Kalman-style recursion (Kalman, 1960) along the scale axis: the hidden state is a running estimate of latent semantics, the innovation measures what the current scale adds beyond previously

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.

fused context, and the gain determines how strongly that evidence should update the state. This prediction–innovation–correction structure provides the architectural principle for our design.

This inference view also suggests an appropriate general model class. Under the surrogate model, a selective state-space scan with context-dependent transition and input matrices can represent the Bayes-optimal recursion when the oracle gain is identifiable from the available scan context. A practical realization, however, requires a structured and computationally feasible parameterization. We therefore use input-dependent channelwise modulation followed by learned channel mixing, retaining the prediction–innovation–correction structure while restricting the gain parameterization.

The conditional-expectation projection property underlying our risk-gap equality and the classical Kalman recursion used in the surrogate model are established estimation tools. Our contribution lies in formulating decoder-stage fusion along the scale axis, using these tools to derive explicit and empirically testable architectural implications, and instantiating those implications in a practical cross-scale fusion module.

These estimation arguments provide concrete design insights for a practical U-Net modification. Specifically, we retain the original encoder–decoder architecture while replacing the standard same-scale skip connection with a cross-scale selective fusion block. The block integrates encoder features across multiple scales, propagates a hidden state along the scale axis, updates the state through prediction and innovation, and uses an input-adaptive gate to regulate the correction. We call the resulting model Cross-scale Selective Fusion U-Net (CSF U-Net). In this sense, the proposed architecture is a structured implementation of the scale-axis prediction–innovation–correction principle identified by the analysis.

Empirically, frozen-feature probes show that adding paired non-matching scales consistently reduces linear-prediction error, while end-to-end experiments assess segmentation accuracy, the roles of the main CSF components, and computational cost. CSF consistently improves vanilla U-Net and achieves the strongest overall performance in mean Dice and the 95th-percentile Hausdorff distance (HD95), although it does not lead every dataset–metric pair and its sequential scan incurs additional memory and latency. Our contributions are:

1. We formulate cross-scale complementarity as a measurable reduction in prediction risk, prove an exact squared-loss risk-gap equality, and probe this phenomenon using paired and shuffled frozen features.
2. Within an explicit linear-Gaussian scale-axis model, we use the classical Kalman recursion to show why heterogeneous observation reliability calls for adaptive correction.
3. We characterize when a general selective scan can realize the Bayes-optimal recursion and instantiate a structured selective gain in bidirectional CSF U-Net, a practical drop-in replacement for U-Net skip fusion.
4. Across four medical segmentation benchmarks, CSF improves vanilla U-Net and achieves the strongest overall mean Dice and HD95 performance, with a measurable cost in latency and training memory.

2 Related Work

Skip-fusion variants in U-Net-style architectures. A substantial body of work improves U-Net-style encoder–decoder networks by modifying the skip pathway. U-Net++ (Zhou et al., 2018) introduces nested and dense skip connections to reduce the semantic gap between encoder and decoder features, while U-Net 3+ (Huang et al., 2020) performs full-scale skip aggregation by connecting each decoder stage to encoder features from multiple resolutions. Attention U-Net (Oktay et al., 2018) and context-gating variants (Asadi-Aghbolaghi et al., 2020) modulate skip features using decoder context. MultiResUNet (Ibtehaz & Rahman, 2020) and pyramid-based designs (Asgari et al., 2019) enlarge the receptive-field range inside the skip or decoder pathway. More recent methods use transformer-style or cross-scale attention for medical segmentation, including TransUNet (Chen et al., 2021), UCTransNet (Wang et al., 2022), TransCeption (Azad et al., 2023), and ScaleFusionNet (Qamar et al., 2025). These methods establish the importance of skip-pathway design, but they typically treat fusion primarily as an architectural design choice. In contrast, we ask when restricting a decoder stage to the same-scale skip loses information that is available from other aligned

scales. Our analysis formalizes this comparison and uses a stated surrogate to motivate an adaptive recursive fusion rule.

Theoretical analyses of U-Net and principled fusion. Several recent works provide mathematical explanations for U-Net-like architectures. Falck et al. (2022) formulate U-Nets in a multi-resolution function-space framework and connect average pooling to projection in a Haar wavelet basis, giving a wavelet interpretation of U-Net representations. Williams et al. (2023) develop a general U-Net design and analysis framework, characterizing encoder and decoder roles, high-resolution scaling limits, and a preconditioning interpretation. Tai et al. (2024) derive U-Net from a control problem solved by operator splitting and multigrid decomposition, showing that a one-step splitting algorithm recovers the U-Net architecture. Mei (2025) interprets U-Nets as efficient approximations to belief propagation denoising in generative hierarchical models. Closely related in spirit, FuseUNet (He et al., 2025) views skip fusion through a multistep predictor–corrector mechanism inspired by ordinary differential equations.

Our analysis differs in scope. While existing work explains global U-Net principles such as multi-resolution representation, preconditioning, operator splitting, belief propagation, or dynamical-system-inspired decoding, we study how different encoder scales can be fused at each decoder stage. We first formulate and empirically probe cross-scale complementarity, with the formal statement and proof of the squared-loss risk-gap equality provided in the appendix. We then use a stated linear-Gaussian surrogate to derive a conditional Bayes recursion and motivate adaptive correction under heterogeneous observation reliability.

Selective state-space models and Mamba-like vision networks. Selective state-space models such as Mamba (Gu & Dao, 2023) provide efficient sequence modeling through input-dependent state transitions and hardware-aware selective scans. Recent vision variants apply these ideas to image modeling by scanning spatial patches, rows, columns, windows, or flattened visual sequences (Zhu et al., 2024; Liu et al., 2024). These methods use selective scans primarily as efficient alternatives to attention for spatial or patch-wise dependency modeling. Our construction instead scans the *scale axis* at each fixed decoder location. The sequence length is the number of aligned encoder scales, and the hidden state represents a running estimate of cross-scale semantic content. This selective recurrence provides a suitable general class for adaptive cross-scale estimation.

3 Scale-Axis Estimation View of Adaptive Cross-Scale Fusion

Problem setup. Consider the multi-scale feature maps $F^{(1)}, \dots, F^{(S)}$ produced by the U-Net encoder at S resolutions. At decoder stage r, let $\mathcal{A}_{s\to r}$ align scale s to the stage-r grid Ω_r, and let Π_s project the aligned feature channels to a common dimension. For location $p \in \Omega_r$, define

$$\xi_{s,p} := \Pi_s\Big(\big(\mathcal{A}_{s\to r}(F^{(s)})\big)_p\Big), \qquad \Xi_p := (\xi_{1,p}, \dots, \xi_{S,p}).$$

Here $\xi_{r,p}$ is the matching skip, while Ξ_p contains all aligned scales. Let $d_{r,p}$ denote the incoming decoder feature at stage r and $z_{r,p}$ the local target to be estimated by the fusion block. Let $\mathcal{R}^\star_{\text{same}}$ and $\mathcal{R}^\star_{\text{all}}$ denote the minimum squared risks among measurable estimators based on $(d_{r,p}, \xi_{r,p})$ and $(d_{r,p}, \Xi_p)$, respectively. For any square-integrable joint distribution of these variables, the conditional-expectation projection property gives the exact risk gap

$$\mathcal{R}^\star_{\text{same}} - \mathcal{R}^\star_{\text{all}} = \mathbb{E}\,\|\mathbb{E}[z_{r,p} \mid d_{r,p}, \Xi_p] - \mathbb{E}[z_{r,p} \mid d_{r,p}, \xi_{r,p}]\|^2 \geq 0. \tag{1}$$

Thus non-matching scales can improve the optimal squared-loss estimator only when they change the conditional mean beyond the decoder feature and matching skip. Section 5 provides empirical evidence from frozen features of trained U-Nets, while Appendix A gives the formal statement and proof of Equation 1.

Scale-axis fusion as sequential estimation. Equation 1 identifies the value of additional aligned scales but leaves open how they should be combined. At a fixed decoder location, these features form a short sequence across scales. This suggests a sequential estimator that carries forward a summary of previously processed scales, compares each incoming feature with its prediction from that summary, and adjusts the update according to feature reliability. A linear-Gaussian state-space

surrogate makes this sequential update analytically explicit and provides a concrete basis for the fusion design below.

Choose any fixed scale order π and write $x_{s,p} := \xi_{\pi(s),p}$. Conditioned on the incoming decoder feature $d_{r,p}$, the surrogate is

$$\begin{aligned} u_{s,p} &= U_s u_{s-1,p} + w_{s,p}, \qquad w_{s,p} \sim \mathcal{N}(0, Q_s), \\ x_{s,p} &= H_s u_{s,p} + v_{s,p}, \qquad v_{s,p} \sim \mathcal{N}(0, R_{s,p}). \end{aligned} \tag{2}$$

Here $u_{s,p}$ is the running latent semantic state; U_s and H_s are the state-transition and observation matrices; and $w_{s,p}$ and $v_{s,p}$ are zero-mean process and observation noise with covariance matrices Q_s and $R_{s,p}$, respectively. Spatial variation in $R_{s,p}$ captures the fact that different scales may be informative at different locations. The complete probabilistic specification, including the initialization and conditional-independence assumptions, is given in Appendix B.

Within this surrogate, the classical Kalman posterior-mean update is

$$\begin{aligned} m^-_{s,p} &= U_s m_{s-1,p}, \\ \nu_{s,p} &= x_{s,p} - H_s m^-_{s,p}, \\ K_{s,p} &= P^-_{s,p} H_s^\top \left(H_s P^-_{s,p} H_s^\top + R_{s,p} \right)^{-1}, \\ m_{s,p} &= m^-_{s,p} + K_{s,p} \nu_{s,p}. \end{aligned} \tag{3}$$

Here $m^-_{s,p}$ and $P^-_{s,p}$ are the predictive mean and covariance before observing $x_{s,p}$, $\nu_{s,p}$ is the innovation, and $K_{s,p}$ is the correction gain. The recursion gives three architectural implications. First, fusion can proceed recursively along scale, with the current state summarizing the evidence processed so far. Second, the current-scale feature can be compared with the feature predicted from the accumulated state, and their residual $\nu_{s,p}$ can drive the correction. Third, because $K_{s,p}$ depends on $R_{s,p}$, heterogeneous observation reliability calls for input- and location-dependent correction rather than a single static matrix. Section 4 turns these implications into a learnable fusion block.

Appendix B derives the complete posterior recursion and proves a sharper result: when the optimal gain varies across locations, every location-independent gain has strictly larger summed estimation error.

4 Method: Cross-Scale Selective Fusion U-Net

The design rationale above suggests a direct architecture: keep the U-Net encoder–decoder backbone unchanged, and replace the same-scale skip pathway at each decoder stage by an adaptive cross-scale estimator. At stage r and location p, the Cross-scale Selective Fusion (CSF) block receives the decoder feature $d_{r,p}$ and the aligned scale sequence $(x_{1,p}, \ldots, x_{S,p})$ defined in Section 3. Its output replaces the standard same-scale skip passed to the decoder block. Figure 1 illustrates this overall framework.

For each location $p \in \Omega_r$, the block maintains a hidden state along the scale axis, initialized by $h_{0,p} = \mathcal{I}_r(d_{r,p})$. For $s = 1, \ldots, S$, the four-line update follows the same prediction–innovation–gating–correction order motivated by Equation 3:

$$\begin{aligned} \bar{h}_{s,p} &= \mathcal{P}_s(h_{s-1,p}, d_{r,p}), \qquad \bar{x}_{s,p} = \mathcal{O}_s(\bar{h}_{s,p}, d_{r,p}), \\ \Delta x_{s,p} &= x_{s,p} - \bar{x}_{s,p}, \\ \gamma_{s,p} &= \sigma(\mathcal{G}_s(d_{r,p}, x_{s,p}, \bar{x}_{s,p}, \Delta x_{s,p})), \\ h_{s,p} &= \lambda \, \mathcal{M}_s(d_{r,p}, x_{s,p}) \odot \bar{h}_{s,p} + \mathcal{V}_s(\gamma_{s,p} \odot \Delta x_{s,p}). \end{aligned} \tag{4}$$

Here $\mathcal{I}_r$ initializes the state from the decoder feature, $\mathcal{P}_s$ predicts the next state, and $\mathcal{O}_s$ maps that prediction into the aligned-feature space. The gate generator $\mathcal{G}_s$, followed by the sigmoid σ, produces the channelwise correction gate $\gamma_{s,p}$. The map $\mathcal{M}_s$ produces channelwise retention coefficients for the predicted state, $\lambda \in (0, 1]$ is a scalar transition-damping coefficient, and $\mathcal{V}_s$ maps the gated innovation back into the state space. The symbol $\odot$ denotes elementwise multiplication.

We call this fusion *selective* because the input-dependent gate $\gamma_{s,p}$ softly emphasizes or suppresses innovation channels at each location and scale. It regulates cross-scale evidence channelwise rather than choosing a single encoder scale.

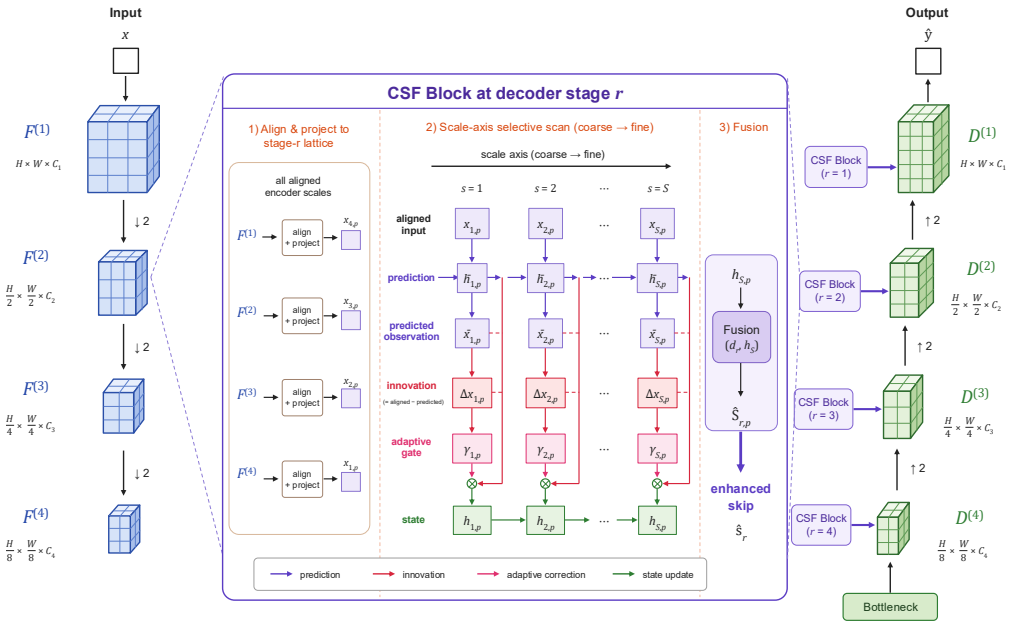

Figure 1: Overview of the proposed CSF U-Net. Multi-scale encoder features are aligned to the target decoder-stage lattice and fused along the scale axis. Input-dependent gates selectively regulate the innovation signal injected at each scale. The center depicts the coarse-to-fine branch for clarity; the default model also applies the same update form in the reverse order, and a learned fusion map combines the two terminal states into the enhanced skip representation.

The relation to Equation 3 is structural rather than an exact parameter-by-parameter equivalence. The predicted state $\bar{h}_{s,p}$ plays the role of $m^-_{s,p}$, and the predicted feature $\bar{x}_{s,p}$ plays the role of $H_s m^-_{s,p}$, but the learned maps $\mathcal{P}_s$ and $\mathcal{O}_s$ are not identified with the linear matrices U_s and H_s. Because $x_{s,p}$ is the same current-scale observation in both equations, $\Delta x_{s,p} = x_{s,p} - \bar{x}_{s,p}$ is the learned analogue of the innovation $\nu_{s,p}$. The gate $\gamma_{s,p}$ together with $\mathcal{V}_s$ supplies adaptive correction, while $h_{s,p}$ plays the role of the updated mean $m_{s,p}$. The retention map $\mathcal{M}_s$ is an additional practical component with no separate counterpart in the classical recursion. Equation 4 therefore preserves the prediction–innovation–adaptive-correction pattern rather than reproducing the Kalman update parameter by parameter.

Two gaps limit this analogy. First, the oracle gain depends on local reliability, whereas the CSF gate receives local feature values rather than the spatial index p or the observation covariance $R_{s,p}$. It can emulate a location-dependent gain only when the relevant reliability is inferable from those feature values; if identical gate inputs require different oracle gains, no deterministic gate can recover both. Second, practical CSF uses channelwise state retention and the factorized correction matrix $W_s \mathrm{diag}(\gamma_{s,p})$, where W_s is the pointwise linear map implementing $\mathcal{V}_s$, rather than unrestricted input-dependent matrices. The implemented block is therefore a structured, learnable approximation motivated by the recursion, not an architecture guaranteed to contain the Bayes-optimal recursion.

The default CSF block runs this recurrence twice: $\pi_{\rightarrow}$ orders scales from coarse to fine and $\pi_{\leftarrow}$ uses the reverse order. Let $h^{\rightarrow}_{S,p}$ and $h^{\leftarrow}_{S,p}$ denote the corresponding terminal states. We concatenate them and use a learned fusion map to produce the enhanced skip tensor $\hat{s}_r$:

$$\tilde{s}_{r,p} = \mathcal{R}_r\big(d_{r,p}, \big[h^{\rightarrow}_{S,p}; h^{\leftarrow}_{S,p}\big]\big), \qquad \hat{s}_r = \mathcal{Q}_r((\tilde{s}_{r,p})_{p\in\Omega_r}). \tag{5}$$

The matching-scale encoder feature is not used as a separate bypass. It may appear inside the aligned scale sequence, but its contribution is processed by the same selective cross-scale estimator as every other scale. The feature actually injected into the decoder is therefore $\hat{s}_r$, not the original same-scale

skip. The decoder update is unchanged:

$$d_r^{\mathrm{out}} = \Phi_r(\mathrm{Up}_r(d_{r+1}), \hat{s}_r)\,.$$

Thus CSF is a drop-in replacement for skip fusion: the encoder, upsampling path, and decoder block Φ_r remain fixed, while the skip representation is produced by adaptive cross-scale estimation rather than same-scale concatenation.

The maps $\mathcal{I}_r, \mathcal{P}_s, \mathcal{O}_s, \mathcal{G}_s, \mathcal{M}_s, \mathcal{R}_r$, and $\mathcal{Q}_r$ are implemented by pointwise or small local convolutional layers, while $\mathcal{V}_s$ is implemented by the pointwise linear map W_s. The gate $\gamma_{s,p}$ controls innovation injection, while $\mathcal{M}_s$ controls retention of the predicted state. We use bounded gates or normalization in these paths to stabilize the recurrent update. Bidirectional scanning is the default practical configuration; the two directional terminal states are concatenated and fused by $\mathcal{R}_r$ as in equation 5. This configuration is selected empirically by the scan-order ablation rather than dictated by the theory.

Appendix C provides conceptual pseudocode for the complete CSF procedure at one decoder stage.

5 Experiments

We evaluate the proposed cross-scale selective fusion block within U-Net-style decoding. To isolate the skip-fusion mechanism, we keep the encoder–decoder template fixed and replace the standard same-scale skip pathway with the proposed CSF block.

Datasets and metrics. We evaluate on four two-dimensional (2D) medical image segmentation benchmarks: SegPC (Gupta et al., 2023) for cell segmentation, ISIC 2018 (ISIC) (Codella et al., 2019) for skin lesion segmentation, GLaS (Sirinukunwattana et al., 2017) for gland segmentation, and MoNuSeg (Kumar et al., 2019) for nuclei segmentation. These datasets cover different object sizes, degrees of boundary ambiguity, and appearance statistics, making them suitable for evaluating adaptive cross-scale fusion. We report the Dice score (%) and the 95th-percentile Hausdorff distance (HD95). Dice measures region overlap, while HD95 evaluates boundary quality and is less sensitive to extreme outliers than the maximum Hausdorff distance. All results are reported as the mean and standard deviation over three random seeds.

Baselines. We compare CSF U-Net with representative U-Net-style segmentation models, including vanilla U-Net (Ronneberger et al., 2015), U-Net++ (Zhou et al., 2018), U-Net 3+ (Huang et al., 2020), Attention U-Net (Oktay et al., 2018), ResUNet (Zhang et al., 2018), MultiResUNet (Ibtehaz & Rahman, 2020), FuseUNet (He et al., 2025), TransUNet (Chen et al., 2021), and UCTransNet (Wang et al., 2022). These baselines cover plain concatenation, nested dense skips, full-scale skip aggregation, attention-gated skips, residual and multi-resolution encoding, predictor–corrector fusion, and transformer-style skip interaction. U-Net 3+ is the closest full-scale aggregation baseline, while FuseUNet is closely related in its predictor–corrector motivation. This comparison tests CSF against both the vanilla skip pathway and stronger cross-scale fusion alternatives.

Implementation details. We implement all models in PyTorch and train them with Adam ($\mathrm{lr} = 10^{-4}$, weight decay 10^{-4}). The segmentation objective is an equally weighted sum of Dice loss and cross-entropy on pixel logits. We resize inputs to 224×224, use a batch size of 16, and train for up to 100 epochs on SegPC and ISIC and up to 2000 epochs on GLaS and MoNuSeg, with early stopping. All methods use the same dataset splits, input resolution, maximum epoch budget, and early-stopping rule on each dataset. No method uses ImageNet or Vision Transformer pretraining; in particular, TransUNet is trained from scratch to keep the initialization protocol comparable with those of the convolutional baselines. The long maximum schedules on GLaS and MoNuSeg are shared by all methods, and early stopping means that models are not forced to train for the full 2000 epochs. Our CSF U-Net uses a standard U-Net backbone with a base width of 48 and a lightweight selective fusion module of width 32. The default block scans once from coarse to fine and once from fine to coarse, concatenates the two terminal states, and combines them with the learned fusion map in equation 5. It uses batch normalization, rectified linear unit activations, and transition damping with $\lambda = 0.9$. Bidirectionality is selected by the component ablation rather than assumed by the theory. Experiments are run on an NVIDIA RTX 6000 with 48 GB of memory.

Main results. Table 1 shows that CSF U-Net achieves the strongest overall performance: it obtains the lowest mean HD95 on all four datasets and the highest mean Dice on SegPC, ISIC, and

Table 1: Comparison with U-Net-style baselines on four medical image segmentation datasets. Each entry reports Dice (%) ↑ and HD95 ↓ as the mean and standard deviation over three seeds. Best results are underlined.

Method	SegPC		ISIC		GLaS		MoNuSeg	
	Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
U-Net	88.92±0.19	11.82±0.79	85.84±0.58	26.52±1.86	86.25±2.13	24.90±4.26	68.88±2.71	5.22±0.62
U-Net++	91.42±0.23	7.37±0.04	86.98±0.29	19.30±0.41	89.44±0.44	18.09±1.25	73.34±0.49	3.35±0.13
U-Net 3+	91.91±0.08	7.25±0.46	86.80±0.58	19.06±0.41	90.71±0.43	17.58±0.88	68.20±0.49	4.77±0.32
Attention U-Net	91.67±0.10	7.41±0.14	87.59±0.15	19.31±0.89	89.85±0.59	17.86±1.44	72.71±0.20	3.29±0.16
ResUNet	87.68±0.15	16.76±0.48	83.74±1.15	32.30±2.28	88.67±0.94	21.95±1.22	67.68±2.46	5.12±0.35
MultiResUNet	84.64±2.63	14.64±2.54	84.92±0.84	33.51±3.92	83.09±1.52	26.80±3.60	65.93±3.31	4.94±0.39
FuseUNet	89.18±0.07	14.72±0.44	84.63±0.43	30.35±1.23	89.27±0.84	21.15±2.70	69.07±3.53	4.82±0.75
TransUNet	81.71±0.89	14.16±0.31	81.96±0.48	31.64±1.02	86.98±0.38	23.48±0.39	72.98±0.59	4.07±0.14
UCTransNet	91.53±0.15	7.17±0.18	88.01±0.49	18.93±0.61	87.70±0.49	22.11±0.49	73.53±1.62	3.52±0.49
CSF U-Net	92.31±0.26	6.72±0.24	88.16±0.72	17.24±0.79	90.52±0.64	16.27±0.86	74.90±0.89	3.00±0.13

MoNuSeg. For GLaS Dice, U-Net 3+ slightly outperforms CSF U-Net (90.71 versus 90.52). Compared with vanilla U-Net, CSF U-Net yields relative Dice gains of 3.8% on SegPC, 2.7% on ISIC, 5.0% on GLaS, and 8.7% on MoNuSeg. The boundary metric shows the same trend: relative HD95 reductions are 43.1% on SegPC, 35.0% on ISIC, 34.7% on GLaS, and 42.5% on MoNuSeg. Comparisons with U-Net 3+ and FuseUNet further indicate that the improvements are not explained solely by access to full-scale features or by a generic predictor–corrector skip pathway. Overall, adaptive cross-scale estimation improves both region overlap and boundary localization, with U-Net 3+ retaining the highest GLaS Dice. The 0.15-point ISIC Dice difference between CSF and UCTransNet lies within the reported seed variation, so we treat their performance on this metric as comparable.

Component ablations. Figure 3(b) summarizes the component ablations on GLaS and MoNuSeg under fixed backbones, splits, and seed sets, while Table 4 in Appendix D.2 reports the exact values. The all-scale concatenation control receives the same aligned scales as CSF but replaces the recurrent scan by concatenation, with nearly matched capacity (19.77M versus 19.85M parameters). The static-gain variant replaces the adaptive gate by a learned per-scale, per-channel vector broadcast across locations and inputs, while leaving prediction, innovation, and recurrence unchanged. Full CSF improves Dice over all-scale concatenation by 0.84 points on GLaS and 4.06 points on MoNuSeg, showing that cross-scale access and parameter count alone do not reproduce its performance. Removing adaptive gain, innovation subtraction, or recurrence consistently reduces Dice. The preferred unidirectional order is dataset-dependent, but neither direction matches the bidirectional model on both datasets. We therefore use bidirectional scanning as the default empirical configuration; the analysis itself remains order-agnostic. Removing transition damping also degrades the full configuration. These ablations test the practical value of the three architectural ingredients motivated above; they do not validate the Gaussian or Markov assumptions or show that CSF recovers an oracle Kalman gain.

Figure 2 visualizes the internal behavior of CSF on a representative test image. The effective correction maps $C_s(p) = \|\gamma_{s,p} \odot \Delta x_{s,p}\|_2$ show, for this example, that different scales update different spatial regions: coarse corrections appear as low-resolution semantic updates, whereas finer corrections concentrate on more localized structures. The scale-dominance map assigns each high-response pixel to the scale with the largest effective correction, illustrating spatially varying scale usage rather than a fixed global weighting. The removal maps $D_s(p) = |p_{\text{full}}(p) - p_{\text{drop}(s)}(p)|$ further show, for this example, that ablating individual scales changes the prediction mainly near boundaries and ambiguous structures, indicating that CSF uses non-matching scales during decoding.

Frozen-feature probe of cross-scale complementarity. To probe whether non-matching scales carry predictive information beyond the matching skip, we freeze vanilla U-Nets and extract, at each decoder stage r, the decoder feature d_r, matching skip ξ_r, and aligned multi-scale collection Ξ. We fit pointwise linear predictors of the one-hot segmentation labels under squared loss and compare evaluation-set squared errors using (d_r, ξ_r) (R_{same}) and (d_r, Ξ) (R_{all}). We also shuffle the non-matching features across images before fitting an otherwise identical predictor (R_{shuf}), preserving the feature dimension while breaking image-specific cross-scale correspondence. Figure 3(a) summarizes the coarsest-stage results, and Table 3 in Appendix D.1 reports all three stages. Adding

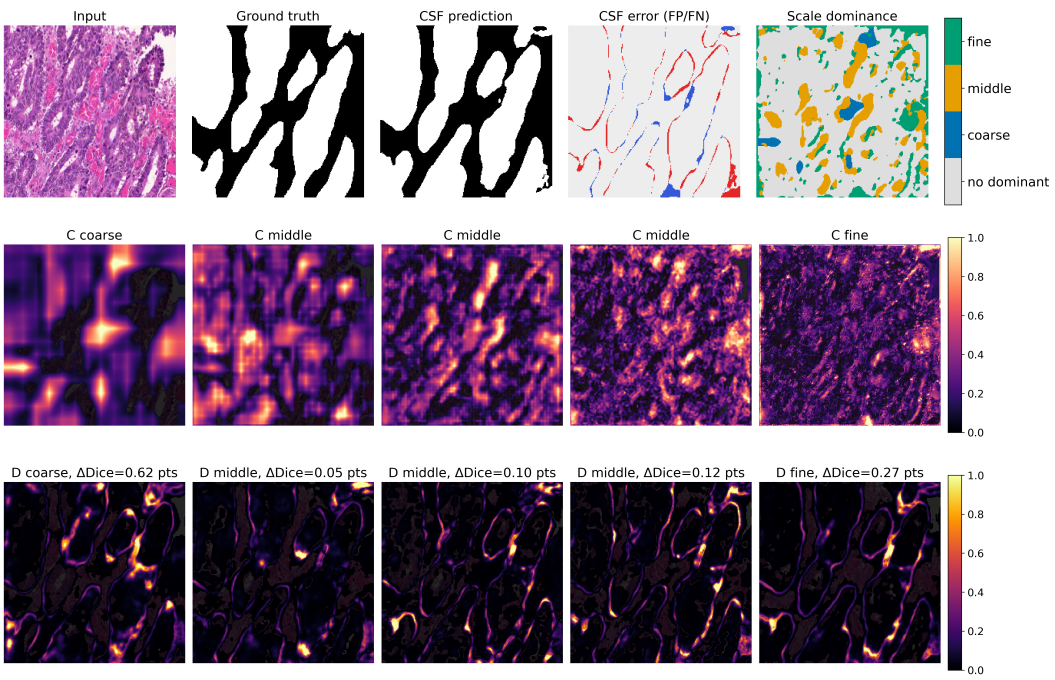

Figure 2: Internal behavior of CSF. Top: input, ground truth, prediction, foreground error, and scale dominance. Middle: effective corrections $C_s(p) = \|\gamma_{s,p} \odot \Delta x_{s,p}\|_2$ for the displayed coarse-to-fine branch. Bottom: scale-removal influence $D_s(p) = |p_{\text{full}}(p) - p_{\text{drop}(s)}(p)|$, with Dice decreases reported in the titles. The maps show spatially varying use of aligned scales.

Table 2: Computational cost at input resolution 224×224. Peak training memory is measured with batch size 16, and inference latency with batch size 1.

Model	**Params**	**GFLOPs**	**Train mem.**	**Latency**
U-Net	14.80M	25.2	3.6 GB	1.7 ms
All-scale concat	19.77M	31.2	7.4 GB	4.4 ms
U-Net 3+	26.97M	153.1	20.3 GB	8.6 ms
FuseUNet	18.75M	42.2	4.8 GB	1.8 ms
CSF U-Net	19.85M	48.8	28.0 GB	17.7 ms

paired non-matching scales lowers the error on every dataset, while shuffling increases it relative to paired all-scale features. This provides evidence that image-specific non-matching features contain information useful to these linear predictors. It does not validate the exact Bayes risk-gap equality, directly estimate the unrestricted Bayes risks, or establish that the linear-Gaussian surrogate is the true feature distribution.

Computational cost. Table 2 reports complexity under the same input size and hardware. CSF is not designed to be cheaper than concatenation. Relative to the parameter-matched all-scale concatenation baseline, CSF uses approximately $1.6\times$ the giga floating-point operations (GFLOPs), $4.0\times$ the batch-1 inference latency, and $3.8\times$ the peak training memory, despite nearly identical parameter counts. The overhead therefore comes primarily from the current sequential bidirectional scan rather than additional parameters. The component ablations show an accuracy benefit from the scan, especially on MoNuSeg, but the resulting accuracy–efficiency trade-off remains an important limitation of the present implementation.

Overall, the results support the central design principle motivated by the analysis: decoder-stage fusion benefits from explicitly aligning multiple encoder scales, updating recursively along the scale axis, and using input-adaptive correction rather than static same-scale concatenation.

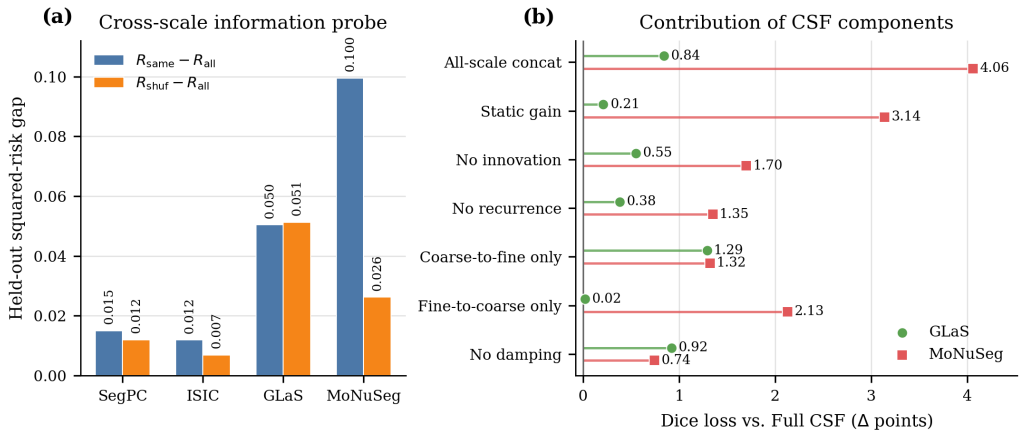

Figure 3: Summary of the cross-scale information probes and CSF component ablations. **(a)** Coarsest-stage frozen-feature probe. Positive $R_{\text{same}} - R_{\text{all}}$ shows the evaluation-error reduction from adding paired non-matching scales, while positive $R_{\text{shuf}} - R_{\text{all}}$ shows the penalty from breaking image-specific cross-scale correspondence. **(b)** Dice decrease relative to Full CSF after replacing or removing one component; larger positive values indicate a larger benefit from the corresponding component in the full model. Exact values are reported in Tables 3 and 4 of Appendix D.

6 Conclusion

This paper formulates decoder-stage cross-scale fusion as an estimation problem using classical projection and filtering tools. Cross-scale complementarity is posed as a prediction-risk question, and frozen-feature probes show that paired non-matching scales are useful to the evaluated linear predictors. The appendix provides a formal statement and proof of the corresponding exact squared-loss risk-gap equality. A linear-Gaussian scale-axis model then yields a conditional prediction–innovation–correction rule and explains why heterogeneous observation reliability favors adaptive correction. General scan realizability requires an identifiable oracle gain, while the practical network uses the structured parameterization described in Section 4.

Motivated by these structural implications, CSF U-Net implements a bidirectional scale scan with a structured selective gain. The ablations indicate that its gains depend on the combination of recurrence, innovation, adaptive gating, and bidirectional scanning rather than all-scale access alone. CSF improves vanilla U-Net on all four benchmarks, achieves the lowest mean HD95 on all four and the highest mean Dice on three datasets, and remains competitive with the evaluated U-Net variants, although U-Net 3+ achieves a slightly higher GLaS Dice. This improvement is accompanied by substantially higher latency and training memory than the parameter-matched all-scale concatenation baseline.

Limitations. The linear-Gaussian model is an analytical idealization rather than a validated generative model of trained features, so its optimality and realizability results remain conditional on the model, squared loss, and gain-identifiability assumption. They do not establish oracle optimality for real U-Nets. Moreover, the general scan class permits arbitrary matrix-valued updates, whereas CSF uses the structured gain $W_s\mathrm{diag}(\gamma_{s,p})$; exact representation of every oracle update is therefore not guaranteed. The sequential bidirectional implementation also costs approximately $4.0\times$ the batch-1 latency and $3.8\times$ the peak training memory of the parameter-matched concatenation baseline, motivating more efficient implementations.

Experiments cover four 2D medical datasets and a U-Net backbone; generalization to three-dimensional (3D) U-Nets, nnU-Net pipelines, larger datasets, or non-medical tasks is not established. Scan direction, bidirectionality, alignment, and damping are empirical choices rather than unique consequences of the theory. Results use three seeds without formal hypothesis testing, so small differences, including the 0.15-point ISIC Dice margin over UCTransNet, should not be interpreted as statistically significant.

AI USE STATEMENT

Generative AI tools were used to edit and reorganize manuscript text, improve clarity and notation consistency, assist with LaTeX formatting and cross-reference checks, and provide critical feedback on the formulation and scope of the theoretical exposition. They were not used to generate datasets, produce experimental measurements, or select the reported results. The authors reviewed all AI-assisted text and mathematical edits against the derivations, implementation, cited literature, and experimental records. The authors take full responsibility for the final content of the paper, including all text, claims, equations, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. The experiments use existing publicly released medical-image segmentation benchmarks cited in Section 5; no participants were recruited and no clinical data were collected for this study. The proposed method is a research prototype, and benchmark performance should not be interpreted as evidence of clinical safety, fairness across patient groups, or readiness for deployment. Dataset sampling, acquisition, and annotation biases may limit generalization across institutions and populations. Any clinical use would require appropriate institutional approval, privacy and security review, subgroup evaluation, external validation, and prospective assessment. The computational costs and efficiency limitations of the current implementation are reported explicitly in Section 5 and the limitations discussion.

REPRODUCIBILITY STATEMENT

Section 4 specifies the CSF architecture and update equations, while Section 5 reports the datasets, preprocessing, optimization settings, training budgets, evaluation metrics, hardware, seed aggregation, model capacity, and computational cost. The main results and computational costs are reported in the main-paper tables, while the full component ablations and frozen-feature probes are reported in the appendix. The appendix also provides the complete surrogate assumptions and proofs, implementation-oriented pseudocode, and decoder-stability analysis. These materials document the mathematical and experimental procedures used to obtain the reported findings.

REFERENCES

Maryam Asadi-Aghbolaghi, Reza Azad, Mahmood Fathy, and Sergio Escalera. Multi-level context gating of embedded collective knowledge for medical image segmentation. *arXiv preprint arXiv:2003.05056*, 2020.

Rhona Asgari, Sebastian Waldstein, Ferdinand Schlanitz, Magdalena Baratsits, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. U-net with spatial pyramid pooling for drusen segmentation in optical coherence tomography. In *International Workshop on Ophthalmic Medical Image Analysis*, pp. 77–85. Springer, 2019.

Reza Azad, Yiwei Jia, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Enhancing medical image segmentation with transception: A multi-scale feature fusion approach. *arXiv preprint arXiv:2301.10847*, 2023.

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pp. 205–218. Springer, 2022.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted intervention*, pp. 424–432. Springer, 2016.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Fabian Falck, Christopher Williams, Dominic Danks, George Deligiannidis, Christopher Yau, Chris C Holmes, Arnaud Doucet, and Matthew Willetts. A multi-resolution framework for u-nets with applications to hierarchical vaes. *Advances in Neural Information Processing Systems*, 35:15529–15544, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Anubha Gupta, Shiv Gehlot, Shubham Goswami, Sachin Motwani, Ritu Gupta, Álvaro García Faura, Dejan Štepec, Tomaž Martinčič, Reza Azad, Dorit Merhof, et al. Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Medical Image Analysis*, 83:102677, 2023.

Quansong He, Xiangde Min, Kaishen Wang, and Tao He. FuseUNet: A multi-scale feature fusion method for u-like networks. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KYiynifZUk.

Sheng He, Rina Bao, P Ellen Grant, and Yangming Ou. U-netmer: U-net meets transformer for medical image segmentation. *arXiv preprint arXiv:2304.01401*, 2023.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1055–1059. IEEE, 2020.

Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

RE Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging*, 39(5):1380–1391, 2019.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in Neural Information Processing Systems*, 37:103031–103063, 2024.

Song Mei. U-nets as belief propagation: Efficient classification, denoising, and diffusion in generative hierarchical models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sy1lbQxj9J.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pp. 565–571. IEEE, 2016.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018. URL https://openreview.net/forum?id=Skft7cijM.

Saqib Qamar, Syed Furqan Qadri, Roobaea Alroobaea, Goram Mufarah Alshmrani, Mohd Fazil, and Richard Jiang. Scalefusionnet: transformer-guided multi-scale feature fusion for skin lesion segmentation. *Scientific Reports*, 15(1):34393, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35:489–502, 2017.

Xue-Cheng Tai, Hao Liu, Raymond H Chan, and Lingfeng Li. A mathematical explanation of unet. *arXiv preprint arXiv:2410.04434*, 2024.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2441–2449, 2022.

Christopher Williams, Fabian Falck, George Deligiannidis, Chris C Holmes, Arnaud Doucet, and Saifuddin Syed. A unified framework for u-net design and analysis. *Advances in Neural Information Processing Systems*, 36:27745–27782, 2023.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, pp. 3–11. Springer, 2018.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, pp. 62429–62442. PMLR, 2024.

A Exact Risk Gap from Additional Scales

Consider a U-Net-style encoder–decoder with S resolution scales, indexed by $s \in \{1, \dots, S\}$, where $s = 1$ denotes the finest spatial resolution and $s = S$ the coarsest. Let $F^{(s)} \in \mathbb{R}^{H_s \times W_s \times C_s}$ be the encoder feature map at scale s. Fix decoder stage r with spatial grid $\Omega_r = \{1, \dots, H_r\} \times \{1, \dots, W_r\}$ and location $p \in \Omega_r$. For each encoder scale s, let

$$\widetilde{F}^{(s \to r)} := \mathcal{A}_{s \to r}\big(F^{(s)}\big) \in \mathbb{R}^{H_r \times W_r \times \widetilde{C}_s}$$

be the feature aligned to the stage-r grid. Let $\Pi_s : \mathbb{R}^{\widetilde{C}_s} \to \mathbb{R}^d$ be a pointwise projection to a common fusion dimension and define

$$\xi_{s,p} := \Pi_s\Big(\big(\widetilde{F}^{(s \to r)}\big)_p\Big), \qquad \Xi_p := (\xi_{1,p}, \dots, \xi_{S,p}),$$
$$\Xi_{-r,p} := (\xi_{1,p}, \dots, \xi_{r-1,p}, \xi_{r+1,p}, \dots, \xi_{S,p}).$$

Thus $\xi_{r,p}$ is the matching encoder skip, Ξ_p is the collection of all aligned scales, and $\Xi_{-r,p}$ contains the non-matching scales.

Let $d_{r,p} \in \mathbb{R}^{d_d}$ denote the incoming decoder feature and let $z_{r,p} \in \mathbb{R}^q$ denote the local semantic target to be estimated before the downstream decoder block. We compare the estimator classes

$$\mathfrak{U}^{\text{same}}_{r,p} := \Big\{\phi(d_{r,p}, \xi_{r,p}) \ : \ \phi \text{ is measurable and } \mathbb{E}\,\|\phi(d_{r,p}, \xi_{r,p})\|^2 < \infty\Big\},$$
$$\mathfrak{U}^{\text{all}}_{r,p} := \Big\{\psi(d_{r,p}, \Xi_p) \ : \ \psi \text{ is measurable and } \mathbb{E}\,\|\psi(d_{r,p}, \Xi_p)\|^2 < \infty\Big\}.$$

The first class uses the decoder feature and matching skip, whereas the second additionally accesses the aligned non-matching scales. Define their optimal local squared risks by

$$\mathcal{R}^\star_{\text{same}} := \inf_{\hat{z} \in \mathfrak{U}^{\text{same}}_{r,p}} \mathbb{E}\,\|z_{r,p} - \hat{z}\|^2, \qquad \mathcal{R}^\star_{\text{all}} := \inf_{\hat{z} \in \mathfrak{U}^{\text{all}}_{r,p}} \mathbb{E}\,\|z_{r,p} - \hat{z}\|^2.$$

Proposition 1 (Risk reduction from additional scales)**.** *Fix a decoder stage r and location p, and assume $\mathbb{E}\,\|z_{r,p}\|^2 < \infty$. Let*

$$m_{\text{same}} := \mathbb{E}[z_{r,p} \mid d_{r,p}, \xi_{r,p}], \qquad m_{\text{all}} := \mathbb{E}[z_{r,p} \mid d_{r,p}, \Xi_p].$$

Then

$$\mathcal{R}^\star_{\text{same}} - \mathcal{R}^\star_{\text{all}} = \mathbb{E}\,\|m_{\text{all}} - m_{\text{same}}\|^2 \geq 0. \tag{6}$$

Consequently, access to the aligned non-matching scales yields a strict improvement if and only if $m_{\text{all}} \neq m_{\text{same}}$ on a set of positive probability.

The proposition follows from the standard Pythagorean equality for two nested L^2 projections. It holds for any square-integrable joint distribution of $(z_{r,p}, d_{r,p}, \Xi_p)$ and does not require Gaussianity or any other parametric family. It states the estimator comparison precisely, while the frozen-feature probes provide empirical evidence of strict improvement within the evaluated linear predictor class. All statements are pointwise in p; summing equation 6 over $p \in \Omega_r$ gives the corresponding stage-level risk-gap equality.

Proof of Proposition 1

Proof. Write $z = z_{r,p}$, $d = d_{r,p}$, $\xi = \xi_{r,p}$, and $\Xi = \Xi_p$. Let $m_{\text{same}} := \mathbb{E}[z \mid d, \xi]$ and $m_{\text{all}} := \mathbb{E}[z \mid d, \Xi]$. By the definition of the estimator classes and the L^2 projection property of conditional expectation,

$$\inf_{\hat{z} \in \mathfrak{U}^{\text{same}}_{r,p}} \mathbb{E}\,\|z - \hat{z}\|^2 = \mathbb{E}\,\|z - m_{\text{same}}\|^2, \qquad \inf_{\hat{z} \in \mathfrak{U}^{\text{all}}_{r,p}} \mathbb{E}\,\|z - \hat{z}\|^2 = \mathbb{E}\,\|z - m_{\text{all}}\|^2.$$

Since (d, ξ) is measurable with respect to (d, Ξ), the difference $m_{\text{all}} - m_{\text{same}}$ is a measurable function of (d, Ξ), while $z - m_{\text{all}}$ is orthogonal to every such function. Therefore

$$\mathbb{E}\,\|z - m_{\text{same}}\|^2 = \mathbb{E}\,\|z - m_{\text{all}}\|^2 + \mathbb{E}\,\|m_{\text{all}} - m_{\text{same}}\|^2.$$

Hence the full aligned multi-scale input yields a strict risk improvement if and only if

$$\mathbb{E}\,\|m_{\text{all}} - m_{\text{same}}\|^2 > 0,$$

which is equivalent to

$$\mathbb{E}[z_{r,p} \mid d_{r,p}, \Xi_p] \neq \mathbb{E}[z_{r,p} \mid d_{r,p}, \xi_{r,p}]$$

with positive probability. This proves Proposition 1. □

Gaussian specialization. Now assume additionally that $(z_{r,p}, d_{r,p}, \Xi_p)$ is jointly Gaussian and that $\mathrm{Cov}(\Xi_{-r,p} \mid d_{r,p}, \xi_{r,p})$ is positive definite almost surely. Conditioning on $(d_{r,p}, \xi_{r,p})$, the Gaussian regression formula gives

$$\begin{aligned}\mathbb{E}[z_{r,p} \mid d_{r,p}, \Xi_p] = \mathbb{E}[z_{r,p} \mid d_{r,p}, \xi_{r,p}] \\ + \mathrm{Cov}(z_{r,p}, \Xi_{-r,p} \mid d_{r,p}, \xi_{r,p})\mathrm{Cov}(\Xi_{-r,p} \mid d_{r,p}, \xi_{r,p})^{-1} \\ \cdot \Big(\Xi_{-r,p} - \mathbb{E}[\Xi_{-r,p} \mid d_{r,p}, \xi_{r,p}]\Big). \end{aligned} \tag{7}$$

Because the residual $\Xi_{-r,p} - \mathbb{E}[\Xi_{-r,p} \mid d_{r,p}, \xi_{r,p}]$ is conditionally nondegenerate, the correction term in equation 7 vanishes almost surely if and only if

$$\mathrm{Cov}(z_{r,p}, \Xi_{-r,p} \mid d_{r,p}, \xi_{r,p}) = 0.$$

Thus strict risk improvement in equation 6 is equivalent to nonzero conditional cross-covariance. Finally, for jointly Gaussian variables, positive conditional mutual information is equivalent to nonzero conditional cross-covariance:

$$\mathrm{MI}(z_{r,p}; \Xi_{-r,p} \mid d_{r,p}, \xi_{r,p}) > 0 \iff \mathrm{Cov}(z_{r,p}, \Xi_{-r,p} \mid d_{r,p}, \xi_{r,p}) \neq 0.$$

Here MI denotes mutual information.

B Formal Linear-Gaussian Surrogate and Supporting Results

This section records the complete surrogate assumptions and the formal results summarized in Section 3. They are conditional statements about the model below, not assertions that trained U-Net features follow a linear-Gaussian generative process.

Assumption 2 (Latent multi-scale observation model)**.** *Fix a decoder stage r and location p. Consider the aligned multi-scale input $\Xi_p = (\xi_{1,p}, \ldots, \xi_{S,p})$. Let $x_{s,p} := \xi_{\pi(s),p}$ denote the aligned observation processed at recursion step s, where π is any fixed order on the scales. Conditioned on the decoder prior $d_{r,p}$, there exists a latent state sequence $u_{0,p}, u_{1,p}, \ldots, u_{S,p}$ with $u_{s,p} \in \mathbb{R}^q$ such that*

$$u_{0,p} \mid d_{r,p} \sim \mathcal{N}(m_{0,p}, P_{0,p}), \tag{8}$$

$$u_{s,p} = U_s u_{s-1,p} + w_{s,p}, \qquad w_{s,p} \sim \mathcal{N}(0, Q_s), \qquad s = 1, \ldots, S, \tag{9}$$

$$x_{s,p} = H_s u_{s,p} + v_{s,p}, \qquad v_{s,p} \sim \mathcal{N}(0, R_{s,p}), \qquad s = 1, \ldots, S. \tag{10}$$

Here $U_s \in \mathbb{R}^{q \times q}$ is a deterministic transition matrix and $H_s \in \mathbb{R}^{d \times q}$ is an observation matrix. The covariance matrices satisfy $P_{0,p} \succeq 0$, $Q_s \succeq 0$, and $R_{s,p} \succ 0$, ensuring that the innovation covariance is invertible. The noises $w_{s,p}$ and $v_{s,p}$ are mutually independent across scales and independent of $u_{0,p}$ conditioned on $d_{r,p}$. The decoder-stage target is $z_{r,p} = L_r u_{S,p}$ for some matrix L_r.

Theorem 3 (Posterior recursion under the linear-Gaussian surrogate)**.** *For each scale $s = 1, \ldots, S$, define the predictive moments*

$$m^-_{s,p} := \mathbb{E}[u_{s,p} \mid d_{r,p}, x_{1,p}, \ldots, x_{s-1,p}], \qquad P^-_{s,p} := \mathrm{Cov}(u_{s,p} \mid d_{r,p}, x_{1,p}, \ldots, x_{s-1,p}),$$

and the posterior moments

$$m_{s,p} := \mathbb{E}[u_{s,p} \mid d_{r,p}, x_{1,p}, \ldots, x_{s,p}], \qquad P_{s,p} := \mathrm{Cov}(u_{s,p} \mid d_{r,p}, x_{1,p}, \ldots, x_{s,p}).$$

Under Assumption 2, the corresponding conditional distributions are Gaussian and their moments satisfy

$$m^-_{s,p} = U_s m_{s-1,p}, \qquad P^-_{s,p} = U_s P_{s-1,p} U_s^\top + Q_s, \tag{11}$$

$$\nu_{s,p} = x_{s,p} - H_s m^-_{s,p}, \qquad S_{s,p} = H_s P^-_{s,p} H_s^\top + R_{s,p}, \tag{12}$$

$$K_{s,p} = P^-_{s,p} H_s^\top S^{-1}_{s,p}, \tag{13}$$

$$m_{s,p} = m^-_{s,p} + K_{s,p}\nu_{s,p}, \qquad P_{s,p} = (\mathrm{I} - K_{s,p}H_s)P^-_{s,p}. \tag{14}$$

The recursion is initialized with $(m_{0,p}, P_{0,p})$ from equation 8. Since $z_{r,p} = L_r u_{S,p}$, the Bayes-optimal decoder-stage estimator under squared loss is

$$\hat{z}^{\mathrm{Bayes}}_{r,p} = \mathbb{E}[z_{r,p} \mid d_{r,p}, x_{1,p}, \ldots, x_{S,p}] = L_r m_{S,p}.$$

Proof. Condition on $d_{r,p}$ throughout. The model in Assumption 2 is a linear-Gaussian state-space model indexed by scale. Assume inductively that

$$u_{s-1,p} \mid x_{1,p}, \ldots, x_{s-1,p} \sim \mathcal{N}(m_{s-1,p}, P_{s-1,p}).$$

By the linear transition model,

$$u_{s,p} \mid x_{1,p}, \ldots, x_{s-1,p} \sim \mathcal{N}(m^-_{s,p}, P^-_{s,p}),$$

with equation 11. Combining this predicted state with the observation model equation 10 yields the jointly Gaussian vector

$$\begin{bmatrix} u_{s,p} \\ x_{s,p} \end{bmatrix} \Bigg| (x_{1,p}, \ldots, x_{s-1,p}) \sim \mathcal{N}\left(\begin{bmatrix} m^-_{s,p} \\ H_s m^-_{s,p} \end{bmatrix}, \begin{bmatrix} P^-_{s,p} & P^-_{s,p} H_s^\top \\ H_s P^-_{s,p} & H_s P^-_{s,p} H_s^\top + R_{s,p} \end{bmatrix} \right).$$

Conditioning a jointly Gaussian vector gives equation 12–equation 14. Because $z_{r,p} = L_r u_{S,p}$ is a linear map,

$$\mathbb{E}[z_{r,p} \mid d_{r,p}, x_{1,p}, \ldots, x_{S,p}] = L_r m_{S,p}.$$

Under squared loss, the posterior mean is Bayes-optimal. □

Theorem 4 (Adaptive correction under heterogeneous surrogate reliability)**.** *Under Assumption 2, fix a scan step s, and let $m^-_{s,p}$, $P^-_{s,p}$, $P_{s,p}$, and $K^\star_{s,p}$ denote the predictive moments, posterior covariance, and Bayes-optimal gain from Theorem 3. Consider the location-independent affine correction*

$$\hat{u}_{s,p}(K) := m^-_{s,p} + K\big(x_{s,p} - H_s m^-_{s,p}\big), \qquad K \in \mathbb{R}^{q \times d}. \tag{15}$$

Then, for every location p,

$$\mathrm{Cov}\big(u_{s,p} - \hat{u}_{s,p}(K) \mid d_{r,p}, x_{1,p}, \ldots, x_{s-1,p}\big) = P_{s,p} + (K - K^\star_{s,p})\big(H_s P^-_{s,p} H_s^\top + R_{s,p}\big)(K - K^\star_{s,p})^\top. \tag{16}$$

Consequently, $K^\star_{s,p}$ is the unique Bayes-optimal local correction. If the Bayes-optimal gain varies across locations in Ω_r, every location-independent matrix K is strictly suboptimal for the sum of local trace risks and cannot be Bayes-optimal at all locations simultaneously. In particular, if $P^-_{s,p} = P^-_s$ is constant across locations, H_s is fixed, and $P^-_s H_s^\top$ is left-invertible, then spatial heterogeneity in $R_{s,p}$ forces adaptive rather than static correction.

Proof. Fix (s, p) and condition on $(d_{r,p}, x_{1,p}, \ldots, x_{s-1,p})$. Let

$$e^-_{s,p} := u_{s,p} - m^-_{s,p}.$$

Then $\mathbb{E}[e^-_{s,p}] = 0$, $\mathrm{Cov}(e^-_{s,p}) = P^-_{s,p}$, and by the observation model

$$x_{s,p} - H_s m^-_{s,p} = H_s e^-_{s,p} + v_{s,p}.$$

The post-update error under equation 15 is therefore

$$\begin{aligned} u_{s,p} - \hat{u}_{s,p}(K) &= u_{s,p} - m^-_{s,p} - K\big(x_{s,p} - H_s m^-_{s,p}\big) \\ &= (\mathrm{I} - K H_s) e^-_{s,p} - K v_{s,p}. \end{aligned}$$

Using independence of $e^-_{s,p}$ and $v_{s,p}$, we obtain

$$\mathrm{Cov}\big(u_{s,p} - \hat{u}_{s,p}(K)\big) = (\mathrm{I} - K H_s) P^-_{s,p} (\mathrm{I} - K H_s)^\top + K R_{s,p} K^\top.$$

Let

$$S_{s,p} := H_s P^-_{s,p} H_s^\top + R_{s,p}, \qquad K^\star_{s,p} := P^-_{s,p} H_s^\top S^{-1}_{s,p}.$$

Expanding the covariance gives

$$\begin{aligned} &(\mathrm{I} - K H_s) P^-_{s,p} (\mathrm{I} - K H_s)^\top + K R_{s,p} K^\top \\ &= P^-_{s,p} - K H_s P^-_{s,p} - P^-_{s,p} H_s^\top K^\top + K S_{s,p} K^\top. \end{aligned}$$

Since

$$K^\star_{s,p} S_{s,p} = P^-_{s,p} H_s^\top, \qquad S_{s,p} (K^\star_{s,p})^\top = H_s P^-_{s,p},$$

completing the square gives

$$\begin{aligned}&(\mathrm{I}-KH_s)P_{s,p}^{-}(\mathrm{I}-KH_s)^\top + KR_{s,p}K^\top \\ &= (\mathrm{I}-K_{s,p}^\star H_s)P_{s,p}^{-}(\mathrm{I}-K_{s,p}^\star H_s)^\top + K_{s,p}^\star R_{s,p}(K_{s,p}^\star)^\top \\ &\quad + (K-K_{s,p}^\star)S_{s,p}(K-K_{s,p}^\star)^\top.\end{aligned}$$

The first two terms on the right equal $P_{s,p}$, and the final term is positive semidefinite with equality if and only if $K = K_{s,p}^\star$. This proves equation 16 and the uniqueness of the local optimum. If $K_{s,p}^\star \neq K_{s,p'}^\star$, no single K can equal both minimizers.

For the final statement, write $A := P_s^- H_s^\top$ and $B := H_s P_s^- H_s^\top$. If A is left-invertible and $K_{s,p}^\star = K_{s,p'}^\star$, then

$$A(B+R_{s,p})^{-1} = A(B+R_{s,p'})^{-1}.$$

Applying a left inverse of A gives $(B+R_{s,p})^{-1} = (B+R_{s,p'})^{-1}$, hence $R_{s,p} = R_{s,p'}$. Therefore $R_{s,p} \neq R_{s,p'}$ implies $K_{s,p}^\star \neq K_{s,p'}^\star$. □

Definition 5 (Selective cross-scale scan)**.** *A selective cross-scale scan at decoder stage r is any recurrence of the form*

$$h_{0,p} = h_0(d_{r,p}), \tag{17}$$
$$h_{s,p} = A_s(\eta_{s,p})h_{s-1,p} + B_s(\eta_{s,p})x_{s,p}, \qquad s = 1,\dots,S, \tag{18}$$
$$\hat{z}_{r,p} = C_r(d_{r,p}, h_{S,p}). \tag{19}$$

Here $\eta_{s,p}$ is any measurable function of the information available before the s-th update, namely $(d_{r,p}, x_{1,p}, \dots, x_{s,p}, h_{s-1,p})$.

Theorem 6 (Conditional scan realizability of the surrogate recursion)**.** *Under Assumption 2, let $m_{s,p}$ and $K_{s,p}$ be the posterior mean and Kalman gain from Theorem 3, and define $h_{s,p} := m_{s,p}$. Then*

$$h_{s,p} = (\mathrm{I}-K_{s,p}H_s)U_s h_{s-1,p} + K_{s,p}x_{s,p}, \qquad s = 1,\dots,S. \tag{20}$$

Moreover, if for each s there exists a measurable map κ_s such that $K_{s,p} = \kappa_s(\eta_{s,p})$ for a context $\eta_{s,p}$ from Definition 5, then the Bayes-optimal estimator belongs to the selective cross-scale scan class by choosing

$$A_s(\eta) := (\mathrm{I}-\kappa_s(\eta)H_s)U_s, \qquad B_s(\eta) := \kappa_s(\eta), \qquad C_r(d,h) := L_r h.$$

Proof. By Theorem 3,

$$m_{s,p} = m_{s,p}^- + K_{s,p}(x_{s,p} - H_s m_{s,p}^-), \qquad m_{s,p}^- = U_s m_{s-1,p}.$$

Substituting the prediction equation into the update gives

$$m_{s,p} = U_s m_{s-1,p} + K_{s,p}(x_{s,p} - H_s U_s m_{s-1,p}).$$

Rearranging,

$$m_{s,p} = (\mathrm{I}-K_{s,p}H_s)U_s m_{s-1,p} + K_{s,p}x_{s,p}.$$

Since $h_{s,p} := m_{s,p}$, this is exactly equation 20. The output map follows from

$$\hat{z}_{r,p}^{\mathrm{Bayes}} = \mathbb{E}[z_{r,p} \mid d_{r,p}, x_{1,p}, \dots, x_{S,p}] = L_r m_{S,p} = L_r h_{S,p}.$$

Finally, if $K_{s,p} = \kappa_s(\eta_{s,p})$, then

$$h_{s,p} = (\mathrm{I}-\kappa_s(\eta_{s,p})H_s)U_s h_{s-1,p} + \kappa_s(\eta_{s,p})x_{s,p},$$

which is the selective scan form in Definition 5 with the stated choices of A_s, B_s, and C_r. □

C Implementation-Oriented Pseudocode

Algorithm 1 Kalman-inspired cross-scale selective fusion for one reference stage

Require: Encoder features $\{E_1, \dots, E_S\}$, decoder feature $(d_{r,p})_{p\in\Omega_r}$, reference stage r, damping coefficient λ

Align each E_s to the reference lattice Ω_r
Project each aligned feature to fusion dimension d
for each $p \in \Omega_r$ **do**
    Form $X_p = (x_p^{(1)}, \dots, x_p^{(S)})$ and its coarse-to-fine and fine-to-coarse orderings
    for $\delta \in \{\rightarrow, \leftarrow\}$ **do**
        Initialize directional state $h_0^\delta = \mathcal{I}_r(d_{r,p})$
        for $s = 1$ to S **do**
            Let x_s^δ be the s-th feature in ordering δ
            Predict state: $\bar{h}_s^\delta = \mathcal{P}_s^\delta(h_{s-1}^\delta, d_{r,p})$
            Predict observation: $\bar{x}_s^\delta = \mathcal{O}_s^\delta(\bar{h}_s^\delta, d_{r,p})$
            Innovation: $\Delta x_s^\delta = x_s^\delta - \bar{x}_s^\delta$
            Correction gate: $\gamma_s^\delta = \sigma(\mathcal{G}_s^\delta(d_{r,p}, x_s^\delta, \bar{x}_s^\delta, \Delta x_s^\delta))$
            Update: $h_s^\delta = \lambda\, \mathcal{M}_s^\delta(d_{r,p}, x_s^\delta) \odot \bar{h}_s^\delta + \mathcal{V}_s^\delta(\gamma_s^\delta \odot \Delta x_s^\delta)$
    $\tilde{s}_{r,p} = \mathcal{R}_r(d_{r,p}, [h_S^{\rightarrow}; h_S^{\leftarrow}])$
Return $\hat{s}_r = \mathcal{Q}_r((\tilde{s}_{r,p})_{p\in\Omega_r})$

D Additional Experimental Results

This section collects the complete frozen-feature probe, full component-ablation results, and qualitative comparisons.

D.1 Full Frozen-Feature Probe Results

Table 3 reports the frozen-feature probe at all three evaluated decoder stages, ordered from the coarsest stage $r = 3$ to the finest stage $r = 1$. The paired all-scale predictor has lower evaluation-set squared error than the same-scale predictor at every evaluated stage and dataset. The error reduction is generally largest at coarser stages, while the shuffled control indicates that feature dimension alone does not explain the improvement.

Table 3: Full frozen-feature linear-probe results. Lower evaluation-set squared error is better.

Dataset	Stage	R_{same}	R_{all}	R_{shuf}	$R_{\text{same}} - R_{\text{all}}$	$R_{\text{shuf}} - R_{\text{all}}$
SegPC	$r = 3$ (coarse)	0.0716	0.0566	0.0686	+0.0150	+0.0120
	$r = 2$	0.0444	0.0395	0.0444	+0.0050	+0.0049
	$r = 1$ (fine)	0.0285	0.0276	0.0283	+0.0009	+0.0007
ISIC	$r = 3$ (coarse)	0.1541	0.1423	0.1492	+0.0119	+0.0069
	$r = 2$	0.1233	0.1195	0.1207	+0.0038	+0.0012
	$r = 1$ (fine)	0.1147	0.1136	0.1138	+0.0011	+0.0002
GLaS	$r = 3$ (coarse)	0.3609	0.3105	0.3618	+0.0504	+0.0513
	$r = 2$	0.3002	0.2660	0.2974	+0.0342	+0.0314
	$r = 1$ (fine)	0.2646	0.2363	0.2591	+0.0283	+0.0229
MoNuSeg	$r = 3$ (coarse)	0.4218	0.3223	0.3485	+0.0995	+0.0262
	$r = 2$	0.3795	0.3060	0.3532	+0.0735	+0.0472
	$r = 1$ (fine)	0.2937	0.2885	0.2924	+0.0052	+0.0038

D.2 Full Component Ablation Results

Table 4 reports the complete component-ablation results summarized in Figure 3(b).

Table 4: Component ablations on GLaS and MoNuSeg. Each entry reports the mean and standard deviation over three seeds. The static-gain variant is location- and input-independent; the full model uses bidirectional scanning and transition damping $\lambda = 0.9$. Best results are bold.

Variant	GLaS		MoNuSeg	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
All-scale concat	89.68±0.21	17.70±0.31	70.84±3.21	3.91±1.00
Static gain	90.31±0.48	17.47±2.19	71.76±1.39	3.74±0.69
No innovation	89.97±0.73	17.62±0.43	73.20±0.78	3.27±0.30
No recurrence	90.14±0.37	16.87±0.79	73.55±0.47	3.16±0.20
Coarse → fine only	89.23±0.88	18.56±1.86	73.58±2.00	3.32±1.17
Fine → coarse only	90.50±0.34	17.18±1.27	72.77±1.32	3.11±0.69
No damping ($\lambda = 1.0$)	89.60±0.92	19.15±2.43	74.16±1.48	3.14±0.25
Full CSF	**90.52±0.64**	**16.27±0.86**	**74.90±0.89**	**3.00±0.13**

D.3 Qualitative Comparisons

Figures 4–7 present representative qualitative comparisons on SegPC, ISIC, GLaS, and MoNuSeg. These visual trends are consistent with the quantitative results: CSF U-Net produces more complete object delineation and cleaner boundaries, especially in challenging regions where vanilla same-scale fusion and stronger skip-fusion baselines tend to under-segment, over-segment, or leak across boundaries.

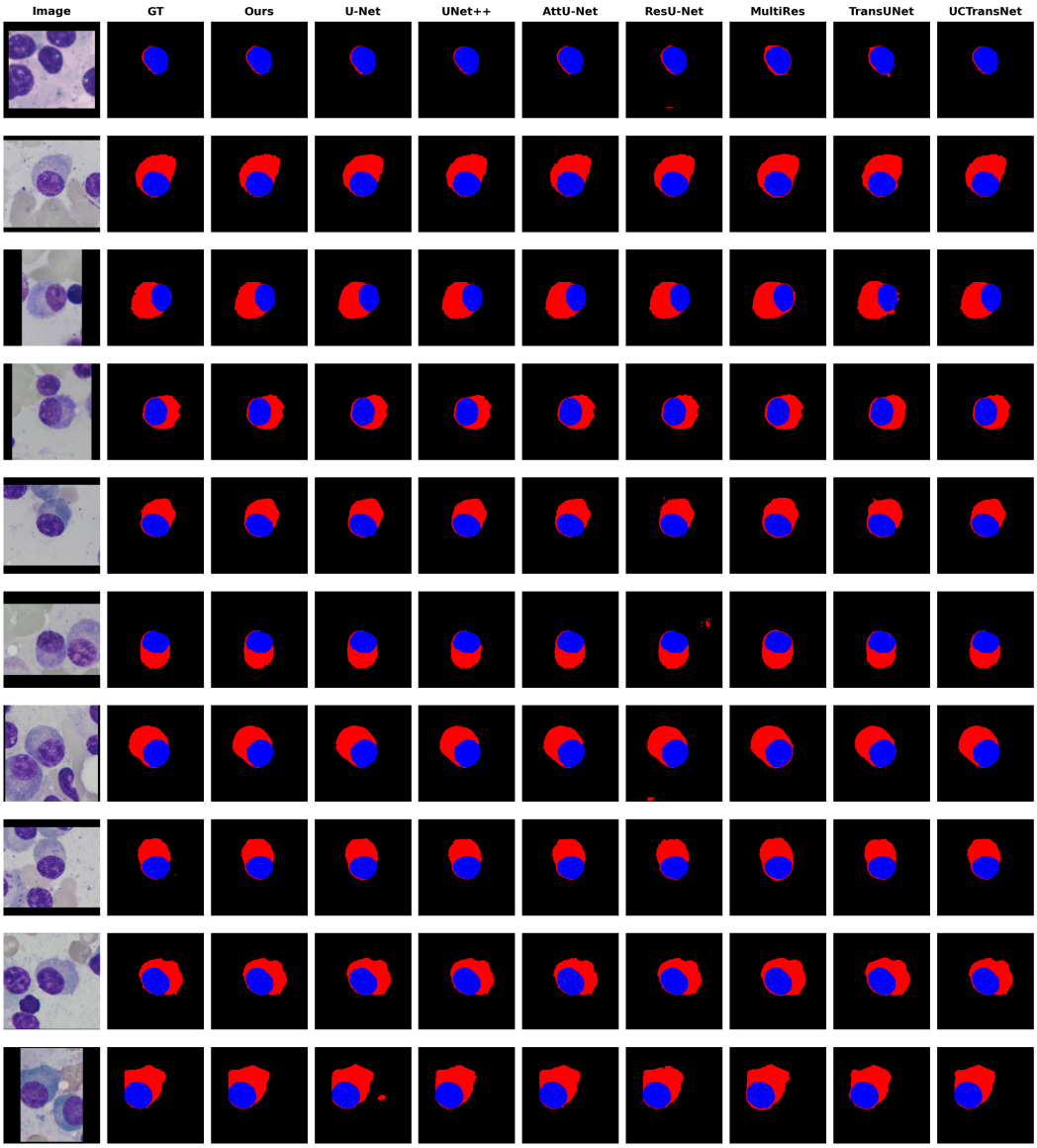

Figure 4. Qualitative comparisons on SegPC with representative U-Net-style baselines.

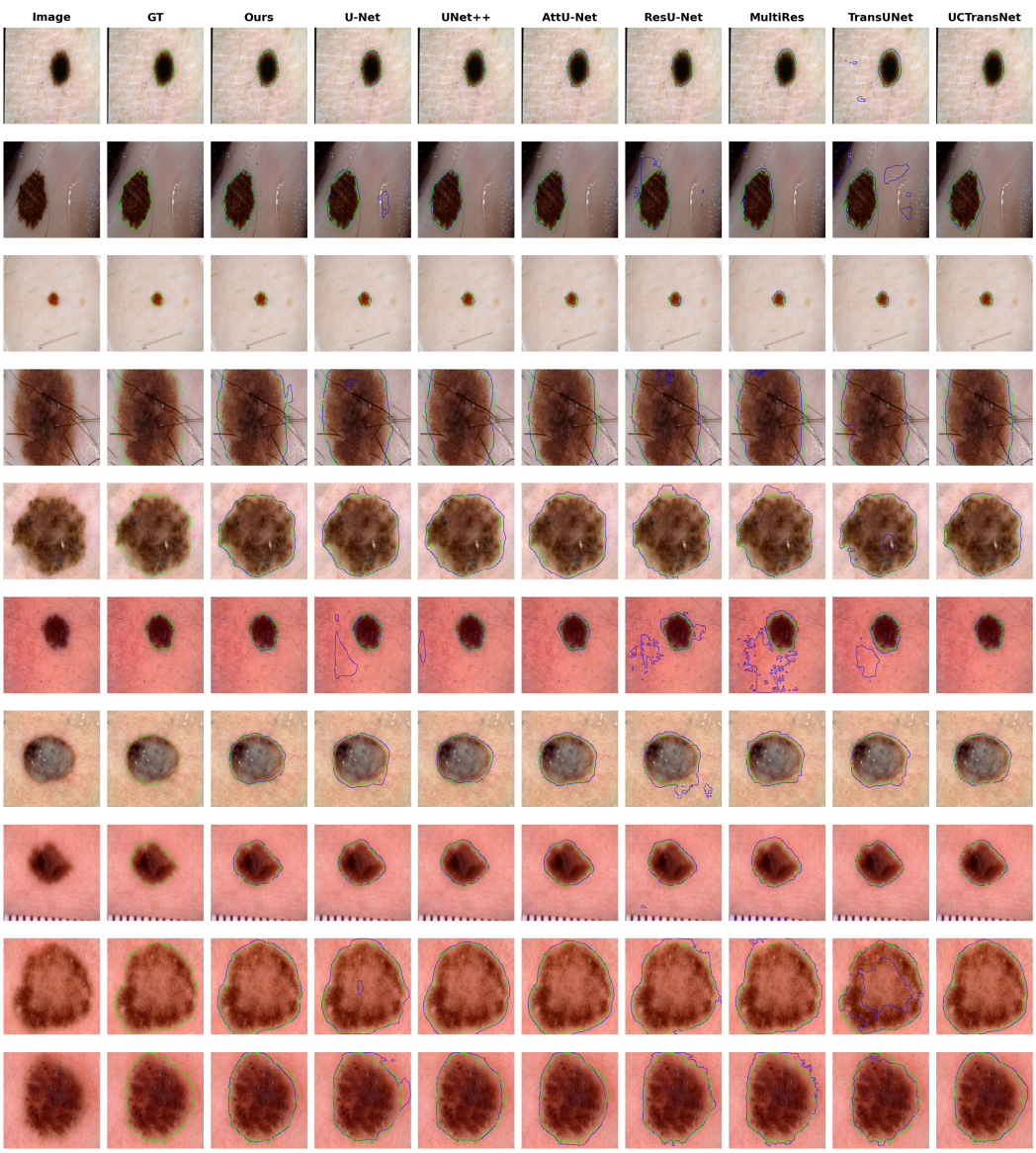

Figure 5. Qualitative comparisons on ISIC with representative U-Net-style baselines.



Figure 6. Qualitative comparisons on GLaS with representative U-Net-style baselines.

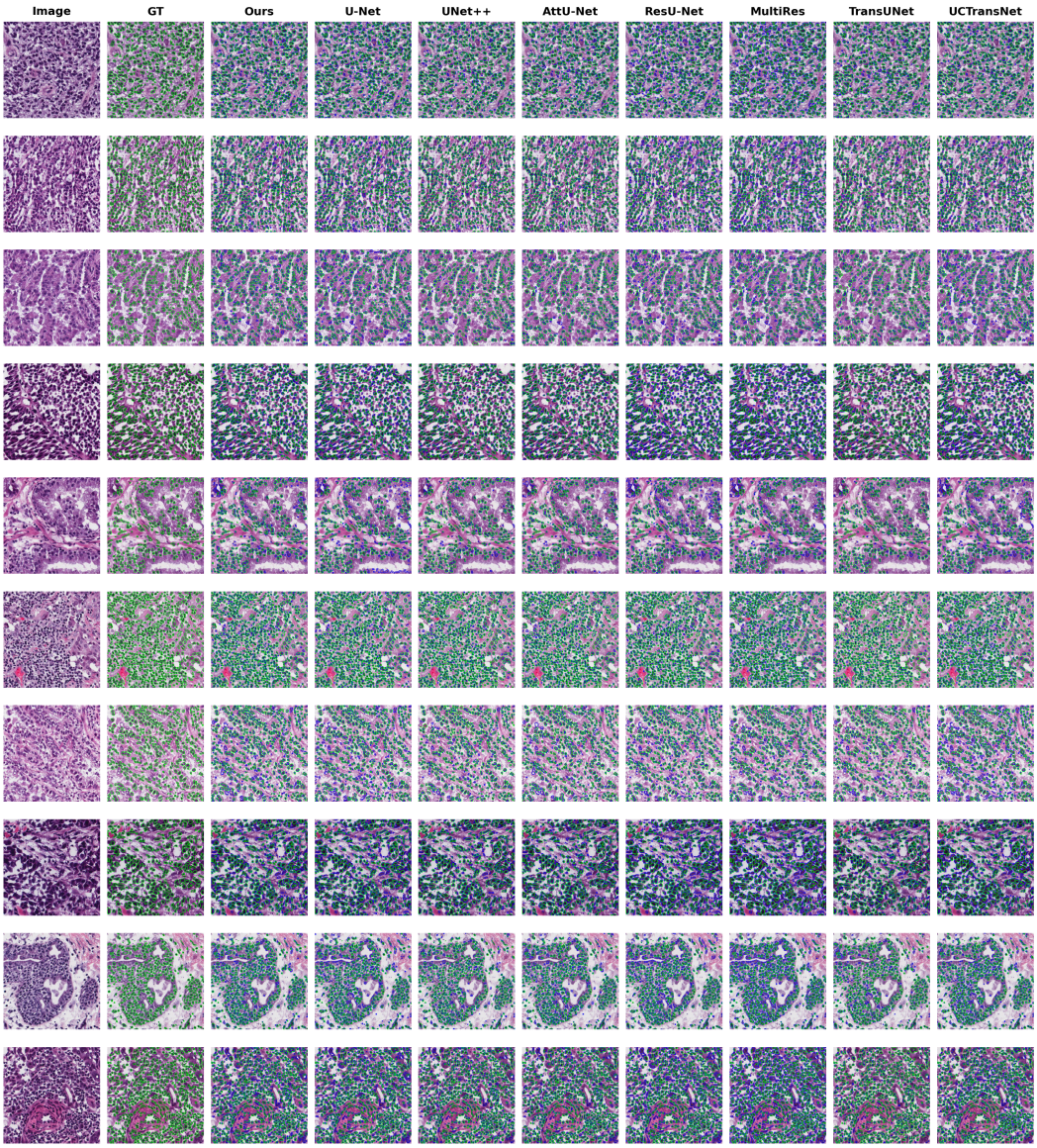

Figure 7. Qualitative comparisons on MoNuSeg with representative U-Net-style baselines.