RCCR-MVS: REVERSE CONTEXT CONDITIONED REGULARIZER FOR HIGH-RESOLUTION AERIAL MVS

Anonymous authors
Paper under double-blind review

ABSTRACT

While memory-efficient slice-wise recurrent regularization is essential for High-Resolution Aerial Multi-View Stereo (MVS), existing approaches either restrict state updates to unidirectional contexts or lack deep contextual integration across independent bidirectional passes. To use bidirectional evidence more effectively under controlled memory budgets, we propose a highly versatile Reverse-Context-Conditioned Regularizer (RCCR) that performs strongly across both single-stage and coarse-to-fine architectures. This asymmetric regularizer uses a lightweight reverse encoder to extract compressed context, which conditions a multi-scale forward recurrent encoder–decoder via a customized cell that uses an explicit evidence-state discrepancy mechanism for optimized sequential gating. We deploy this core mechanism using an efficient two-pass inference strategy that successfully avoids full volumetric storage. To formulate our complete **RCCR-MVS**, we augment its cascaded deployments with two complementary modules: Coverage-risk-guided Range Expansion (CRE) to prevent true depths from falling outside cascaded search intervals, and Spatial Evidence Support (SES) to mitigate noisy initial matching evidence. Extensive experiments demonstrate that RCCR-MVS achieves state-of-the-art performance among high-resolution aerial MVS methods on WHU-MVS and LuoJia-MVS while maintaining a favorable accuracy–resource trade-off. Furthermore, it demonstrates strong cross-domain and cross-resolution generalization on our newly constructed **MultiResXD-MVS** dataset and on the München and Vaihingen full-image scenarios.

1 INTRODUCTION

High-resolution aerial multiview stereo requires dense depth estimation over large image grids while preserving intricate scene details. While early learning-based approaches rely on dense cost volumes that scale restrictively with image dimensions and depth hypotheses (Yao et al., 2018), processing matching evidence recurrently along the depth axis provides a memory-efficient alternative (Yao et al., 2019). As compared in Figure 1(c), although expanding this slice-wise paradigm into a recurrent encoder–decoder architecture successfully incorporates multi-scale spatial features for high-resolution aerial imagery (Liu & Ji, 2020), its unidirectional recurrence inherently restricts sequential state updates to previously scanned depths. Because depth hypotheses represent competing geometric alternatives rather than strictly causal temporal events, integrating evidence from unscanned depths is crucial for resolving local ambiguities. Although existing bidirectional approaches access this context through symmetric independent passes that duplicate substantial computation (Fujitomi et al., 2021; Wei et al., 2024), they lack deep integration, necessitating a more efficient strategy that leverages a compact reverse representation to condition the primary multi-scale updates.

To use bidirectional evidence more effectively under controlled memory budgets, we propose a versatile **R**everse-**C**ontext-**C**onditioned **R**egularizer (RCCR) that delivers strong performance across both single-stage and coarse-to-fine architectures. This asymmetric regularizer employs a lightweight reverse encoder to scan the initial matching costs from the trailing hypotheses down to the current step and project its hidden state into a low-dimensional compressed reverse context. This context then conditions the gates and candidate states of a comprehensive multi-scale forward recurrent encoder–decoder. As further illustrated in Figure 1(a), we facilitate this deep contextual integration through a customized recurrent cell that utilizes an explicit evidence–state discrepancy mechanism, providing an optimized structural parameterization for sequential gating rules.

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.

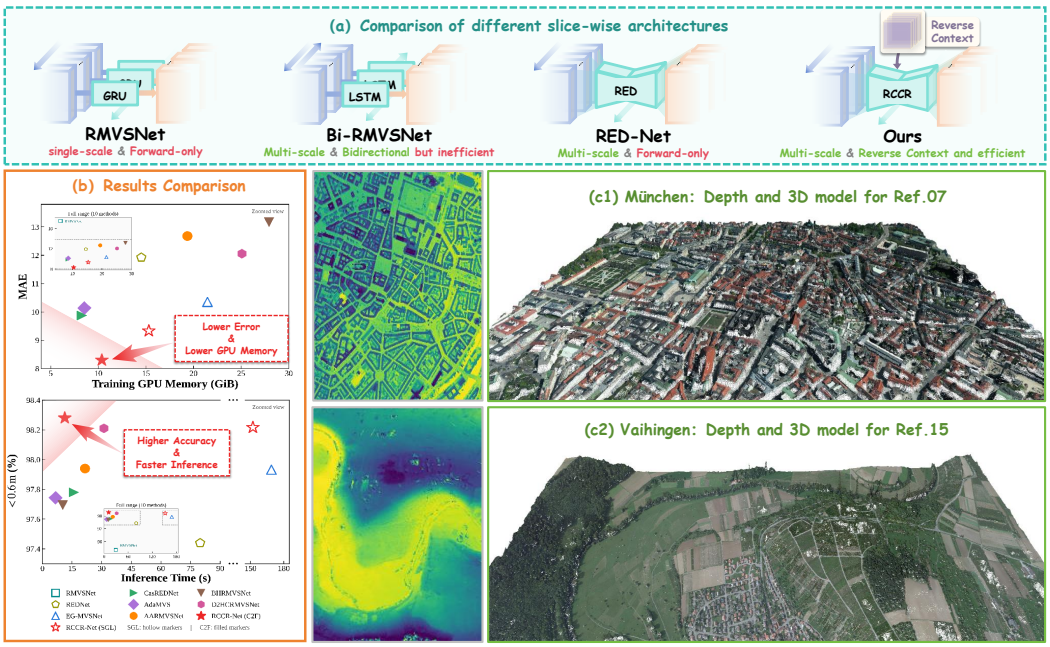

Figure 1: **Overview and performance of our approach.** (a) Architectural comparison of slice-wise recurrent cost regularizers. (b) Results comparison of RCCR-MVS against existing methods. (c) High-resolution point clouds and dense depth maps reconstructed by our pipeline.

To deploy this core regularizer effectively, we formulate a complete cascaded architecture named RCCR-MVS. As evidenced by the performance trends in Figure 1(b), this strategic formulation achieves a superior operational balance, yielding highly accurate predictions with significantly reduced memory allocation. Ultimately, as shown by the expansive point clouds and detailed depth maps in Figure 1(c), our entire pipeline delivers exceptionally robust and precise high-resolution scene reconstructions across complex large-scale aerial environments. Our primary contributions are summarized as follows:

- **Asymmetric bidirectional regularizer.** We propose a Reverse-Context-Conditioned Regularizer that employs a lightweight reverse encoder to extract compressed reverse context, explicitly conditioning the sequential updates of a multi-scale forward recurrent encoder–decoder through a customized cell utilizing an evidence and state discrepancy mechanism.
- **Enhanced cascaded framework.** We introduce Coverage-Risk-Guided Range Expansion to dynamically adjust narrowed search intervals and Spatial Evidence Support to inject structurally enhanced matching evidence prior to the core sequential regularization.
- **Efficient inference and broad generalization.** We implement a two-pass inference strategy that avoids full-volume storage. RCCR-MVS achieves state-of-the-art performance among High-Resolution Aerial MVS methods on WHU-MVS and LuoJia-MVS and demonstrates strong cross-domain and cross-resolution generalization on München, Vaihingen and our newly constructed MultiResXD-MVS.

2 Related Work

Slice-wise Recurrent Multi-View Stereo. R-MVSNet introduces sequential regularization of two-dimensional cost maps along the depth hypothesis axis using convolutional GRUs, which reduces the memory required for high-resolution depth inference (Yao et al., 2019). Subsequent methods enrich the spatial context within this sequential formulation, where D^{2}HC-RMVSNet and AA-RMVSNet combine recurrent regularization with multi-scale aggregation and adaptive feature extraction (Yan et al., 2020; Wei et al., 2021). Another line of work investigates information exchange across depth hypotheses by incorporating context from both scanning directions using bidirectional recurrent units (Fujitomi et al., 2021; Wei et al., 2024). Recently, RRT-MVS introduces recurrent

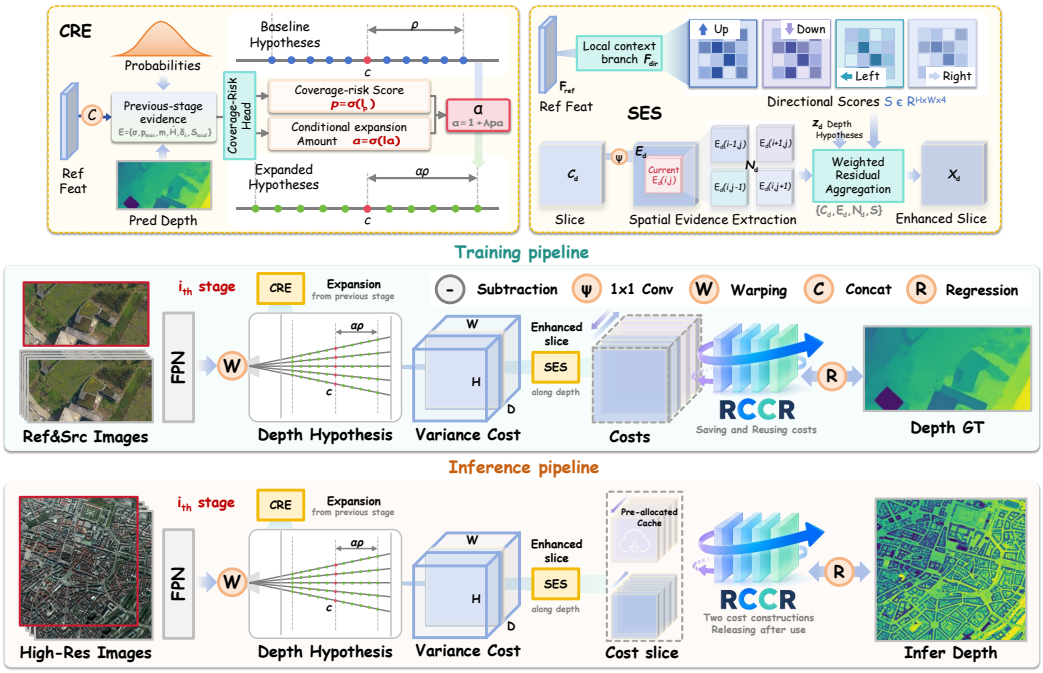

Figure 2: **RCCR-MVS architecture.** After feature extraction, each independent stage constructs a hypothesis grid and aggregates variance matching costs. Two complementary modules enhance this basic process. CRE dynamically adjusts the search intervals before feature warping, while SES structurally enhances the raw cost slices. The RCCR processes these enhanced slices to predict depth logits, yielding a soft-argmin depth map that centers the next cascade level.

self-attention over cost slices to combine spatial interactions with propagated information (Jiang et al., 2025c). Collectively, these developments motivate studying both the representation of cross-hypothesis context and the mechanisms through which it enters each recurrent update.

Multi-View Stereo for Aerial Reconstruction. Aerial MVS must accommodate extensive scene coverage alongside sharp depth transitions. Following the foundational recurrent encoder–decoder architecture and WHU benchmark introduced by RED-Net (Liu & Ji, 2020), subsequent research has focused on improving visual consistency and structural awareness. These advancements include adaptive cross-view aggregation (Liu et al., 2023; Zhang et al., 2024b), the integration of explicit geometric priors such as building edges and monocular depth cues (Zhang et al., 2024a; Mao et al., 2024; Liu et al., 2025), and the deployment of hierarchical attention and dynamic propagation mechanisms (Wang et al., 2026a; Zhang et al., 2025; 2026). Together, these developments highlight the critical importance of structural detail preservation and robust scene-level consistency.

Cost Regularization for High-Resolution Reconstruction. Cost regularization aggregates spatial and depth evidence to resolve correspondence ambiguities (Yao et al., 2018). To capture broader context, recent architectures expand beyond standard 3D convolutions by employing geometric alignment (Wu et al., 2024), attention mechanisms (Cao et al., 2024; Jiang et al., 2025c), and state-space modeling (Li et al., 2026). For high-resolution aerial imagery, balancing this expansive context against computational limits is crucial. Consequently, efficient approaches use multi-scale recurrent encoder–decoders (Liu & Ji, 2020), lightweight multi-stage architectures (Liu et al., 2023), and explicit structural guidance through edge or cross-view feature fusion (Zhang et al., 2024a;b). These studies collectively establish that effective large-scale regularization requires a strict balance between contextual modeling and computational efficiency. Unifying these research trajectories, RCCR-MVS builds upon the slice-wise framework to enable robust high-resolution aerial depth estimation. Specifically, our core innovation directly targets the regularization stage by introducing an asymmetric bidirectional architecture that efficiently integrates compressed reverse context and structurally enhanced geometric evidence.

3 METHOD

3.1 REVERSE-CONTEXT-CONDITIONED REGULARIZATION

Since depth hypotheses represent competing geometric alternatives rather than strictly causal events, information from subsequent hypotheses provides essential context for resolving local ambiguities. To exploit this efficiently, RCCR employs an asymmetric architecture where a lightweight reverse encoder extracts a compressed reverse context from trailing hypotheses. As illustrated in Figure 3, this compact representation explicitly conditions the gating mechanisms and candidate states of the primary multi-scale forward regularizer.

Compressed Reverse Context Encoding. Starting with an initial zero state $b_D = 0$, the reverse encoder traverses the enhanced matching costs from hypotheses $D-1$ down to 0:

$$b_d = \mathrm{ConvGRU}_b\big(\psi(X_d), b_{d+1}\big), \qquad s_d = \phi(b_d). \tag{1}$$

Here, the input projection ψ reduces the high-dimensional cost slice to a compact representation, which a reverse ConvGRU then processes to update the hidden state b_d. Subsequently, the output projection ϕ generates the low-dimensional compressed reverse context $s_d \in \mathbb{R}^{H\times W\times K}$. Because the context is extracted immediately after updating b_d with the current slice X_d, its receptive field is inherently current-inclusive, effectively summarizing the geometric evidence from X_d through X_{D-1}. This alignment ensures that the forward regularizer receives geometrically relevant compressed context from the current and subsequent hypotheses without duplicating the computationally expensive multi-scale feature extraction.

To efficiently deploy this reverse context within a slice-wise inference paradigm, we adopt a two-pass execution strategy. During the initial reverse pass, the network computes the slice costs on the fly, updates the reverse state, and caches only the lightweight K-channel context s_d. In our experiments, we set this channel dimension K to four to guarantee a minimal memory footprint. During the subsequent forward pass, the same costs are reconstructed and the aligned reverse context is consumed alongside them. While caching this context requires storing $KDHW$ elements per sample and the hypothesis grids demand depth-dependent memory, this controlled deployment successfully injects the reverse context without the prohibitive memory overhead of instantiating and retaining a full-resolution 4D cost volume.

RCCR Cell. The RCCR cell drives the forward state propagation by seamlessly integrating the multi-scale spatial features $x_d^{(\ell)}$, the previous hidden state $h_{d-1}^{(\ell)}$, and the bilinearly resized reverse context $s_d^{(\ell)}$. Omitting the spatial scale superscript ℓ for notational clarity, the cell first computes an explicit evidence and state discrepancy e_d which acts strictly as an internal structural parameterization rather than a supervised prediction error. This discrepancy, alongside the primary spatial inputs and the reverse context, feeds into a unified projection to derive the update gate u_d and reset gate r_d. These gates subsequently regulate the formation of the candidate state $\widetilde{h}_d$ and the final updated memory h_d:

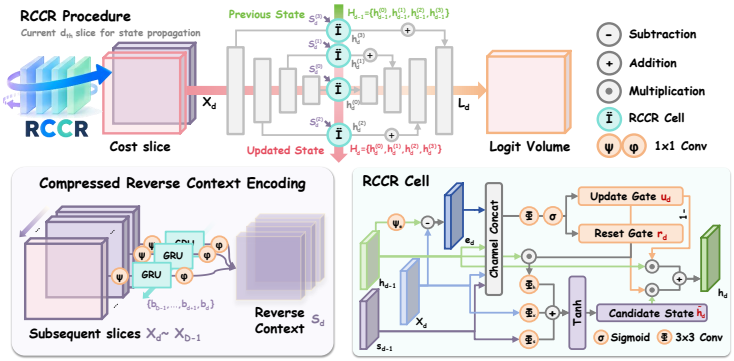

Figure 3: **RCCR Procedure.** The reverse encoder extracts the compressed context s_d. The RCCR cell updates its hidden state by integrating this context with spatial features and the previous state.

$$e_d = x_d - \psi_e(h_{d-1}) \quad [u_d, r_d] = \sigma(\Phi([x_d, h_{d-1}, e_d, s_d]))\,, \tag{2}$$

$$\widetilde{h}_d = \tanh[\Phi_x(x_d) + \Phi_h(r_d \odot h_{d-1}) + \Phi_s(s_d)] \tag{3}$$

$$h_d = (1-u_d)\odot h_{d-1} + u_d \odot \widetilde{h}_d. \tag{4}$$

Unlike standard recurrent architectures, our customized cell actively utilizes the compressed reverse context s_d to explicitly condition both the sequential gating rules and the candidate state generation. Following this context-conditioned update at each independent scale, the decoder sequentially fuses these multi-scale states to predict the final scalar logit o_d. Crucially, these intermediate decoder-fused features are never written back into the recurrent memory, thereby ensuring that the state transitions operate strictly along the depth axis to preserve a clean separation between multi-scale spatial fusion and depth-wise contextual propagation.

3.2 Two Extensions Supporting RCCR

Coverage-Risk-Guided Range Expansion. Because depth estimation relies on a convex combination of sampled hypotheses a cascaded stage cannot recover true depths outside its predefined candidate interval. CRE addresses this limitation by predicting where the baseline interval needs to widen before cost construction. For an enabled destination stage $t > 1$, let $\bar{z}_d$ represent the baseline hypotheses. We define the baseline center c, which naturally inherits the preceding estimate, and the half-range ρ, which specifies the standard cascaded span as:

$$c = \frac{\bar{z}_0 + \bar{z}_{D-1}}{2}, \qquad \rho = \frac{\bar{z}_{D-1} - \bar{z}_0}{2}. \tag{5}$$

As illustrated in Figure 2, we introduce a lightweight convolutional coverage-risk head $\mathcal{F}_{\text{risk}}$ to dynamically expand this range. By processing concatenated previous-stage uncertainty statistics E and projected reference features $\psi(F_{\text{ref}})$, this module directly predicts a risk score p and an expansion amount a. Architectural details for extracting these statistics are deferred to Appendix A.2. These outputs determine a pixel-wise expansion factor α bounded by a hyperparameter λ set to 0.5, which proportionally stretches the hypotheses around the center:

$$p, a = \mathcal{F}_{\text{risk}}([E, \psi(F_{\text{ref}})]), \qquad \alpha = 1 + \lambda p a, \qquad z_d = c + \alpha(\bar{z}_d - c). \tag{6}$$

Because the total hypothesis count remains constant, this wider coverage naturally increases sample spacing without adding memory overhead. To isolate this module from the main depth gradients, the applied expansion factor is detached from the computational graph and supervised exclusively by an auxiliary objective. Finally, this expansion mechanism is inherently bypassed at the initial stage where no preceding estimate exists.

Spatial Evidence Support. SES augments each variance slice before reverse encoding and forward regularization by integrating reference-guided directional scores with local spatial evidence. As depicted in Figure 2, the module first processes the reference features F_{ref} through a local context branch $\mathcal{F}_{\text{dir}}$ to generate a depth-independent directional score tensor $S \in \mathbb{R}^{H \times W \times 4}$. This tensor supplies one independent structural prior for each of the four axis-aligned neighbors and is evaluated only once per cascaded stage. Concurrently, at each depth step d, a bias-free projection maps the high-dimensional matching cost slice C_d into a compact low-dimensional embedding $E_d \in \mathbb{R}^{H \times W \times J}$. For any central pixel $\mathbf{p}$, the embeddings of its four spatial neighbors are gathered to form a local evidence set $\mathcal{N}_d(\mathbf{p})$ that inherently captures the surrounding matching cost distribution:

$$S = \mathcal{F}_{\text{dir}}(F_{\text{ref}}), \tag{7}$$

$$\mathcal{N}_d(\mathbf{p}) = \{E_d(\mathbf{q}) \mid \mathbf{q} \in \text{neighbors of } \mathbf{p}\}. \tag{8}$$

Finally, a Weighted Residual Aggregation mechanism $\mathcal{F}_{\text{agg}}$ seamlessly integrates these extracted components. This operation dynamically weighs local evidence differences using the directional scores S and an explicit discrepancy penalty between the central embedding E_d and its spatial neighborhood $\mathcal{N}_d$. Crucially, it incorporates the depth hypotheses z_d to enforce geometric compatibility and prevent feature blurring across depth boundaries. These weighted residuals are then projected and added back to the original slice, yielding the enhanced output X_d:

$$X_d = \mathcal{F}_{\text{agg}}(C_d, E_d, \mathcal{N}_d, S, z_d). \tag{9}$$

To ensure the initial preservation of the raw matching costs, the internal output projection within $\mathcal{F}_{\text{agg}}$ is strictly zero-initialized, while detailed mathematical formulations of this aggregation are deferred to the Appendix A.3. The resulting enhanced slice X_d is subsequently supplied to both the reverse and forward passes of the regularization module.



Figure 4: Spatial distribution of MultiResXD-MVS.

3.3 Training Objective

The network is supervised by a stagewise smooth L1 depth loss $\mathcal{L}_{\text{depth}}$ (Gu et al., 2020). For stages $t \in \mathcal{T}_{\text{CRE}}$ enabling range expansion, we derive a binary overflow label and a continuous expansion target from ground-truth depth deviations. These supervise three auxiliary components: a binary cross-entropy loss $L_p^{(t)}$ for the risk logit, a smooth L1 loss $L_a^{(t)}$ for the positive-pixel expansion amount, and a smooth L1 loss $L_s^{(t)}$ for their product. Class-balanced averaging prevents background gradient dominance. The total objective combines the primary depth loss and these auxiliary terms using uniform weights $\beta_p = \beta_a = \beta_s = 0.05$:

$$\mathcal{L} = \mathcal{L}_{\text{depth}} + \frac{1}{|\mathcal{T}_{\text{CRE}}|} \sum_{t \in \mathcal{T}_{\text{CRE}}} \left(\beta_p L_p^{(t)} + \beta_a L_a^{(t)} + \beta_s L_s^{(t)} \right). \tag{10}$$

4 Experiments

4.1 Experimental Setup

Datasets. We benchmark on WHU (Liu & Ji, 2020) and Luojia-MVS (Li et al., 2023) for standard training and evaluation. To assess cross-domain generalization, we utilize München and Vaihingen (Haala, 2013), evaluating both original-scale tiles and quarter-resolution full images. Furthermore, as shown in Figure 4, we construct **MultiResXD-MVS** from real UAV imagery of campus and village, comprising 343 unique sample groups across 49 geographic locations. Featuring image sizes from 768×384 to 3072×1536, this dataset evaluates generalization under three complementary protocols comprising expanding native coverage, evaluating identical central pixels, and resizing fixed spatial extents. More dataset details are provided in Appendix A.4 and A.5.

Evaluation Protocol and Implementation Details. Following Liu & Ji (2020), we report MAE for errors below $10\,\text{m}$ alongside threshold accuracies of $< 0.6\,\text{m}$ and < 3 intervals, utilizing a unified depth spacing for fair cross-architecture evaluation. In all tables, light orange highlights our RCCR-MVS and bold red denotes the best results. We conduct experiments using dual 32GB NVIDIA RTX 5090 GPUs. Coarse-to-fine methods consistently employ 48, 32, and 8 hypotheses with interval ratios of 4, 2, and 1. Most single-stage baselines operate on a 0.5 spatial scale with a 2.0 interval factor, whereas R-MVSNet uses 0.25 and 1.0.

4.2 Experimental Results

Evaluation on the WHU and Luojia-MVS Datasets. As detailed in Table 1 and Appendix A.6 Table 6, our approach consistently achieves superior performance in all view configurations on the WHU and Luojia-MVS datasets, while maintaining a highly compact memory footprint and competitive inference speed. For single-stage models, conditioning updates on compressed reverse context resolves local ambiguities, yielding substantial accuracy improvements over the unidirectional baseline. Furthermore, our cascaded pipeline outperforms all competitors, which confirms the synergy of our extensions: CRE reduces the risk of true depths falling outside the search intervals, while SES mitigates noisy initial matching evidence. Ultimately, integrating reverse context with structurally enhanced evidence establishes a robust solution for high-resolution aerial reconstruction.

Table 1: **Comparison on WHU and Luojia with 5 views and generalization on München and Vaihingen tiles.** Inference efficiency is measured at 768 × 384 resolution. Methods marked with * and gray-shaded results are quoted from original papers; "–" denotes unreported results.

Method	WHU 5 views			Luojia 5 views			München tiles		Vaihingen tiles		Efficiency	
	MAE	< 0.6m	< 3intv.	MAE	< 0.6m	< 3intv.	MAE	< 0.6m	MAE	< 0.6m	GPU/GB	Time/s
Full-volume methods												
RingMo-Aerial* Diao et al. (2025)	0.0970	98.10	98.70	0.1220	98.10	96.50	–	–	–	–	–	–
HDC-MVSNet* Li et al. (2023)	0.0860	98.00	98.10	0.1210	98.30	96.60	–	–	–	–	–	–
SG-MVS Wang et al. (2026b)	0.0891	98.15	96.90	0.1815	95.31	93.32	0.7393	74.53	0.5520	80.21	1.007	0.09
MonoMVSNet Jiang et al. (2025b)	0.1270	96.91	96.84	0.1675	96.34	93.47	0.9557	64.20	1.1725	56.00	0.889	0.08
MVSMamba Jiang et al. (2025a)	0.1268	96.86	97.47	0.1531	96.82	96.82	1.2780	49.94	1.2108	56.10	0.712	**0.05**
RRT-MVS Jiang et al. (2025c)	0.1087	97.45	96.02	0.1287	97.91	94.25	0.7285	75.02	0.5904	78.59	1.759	0.35
Single-stage slice-wise methods												
R-MVSNet Yao et al. (2019)	0.1751	95.35	96.55	0.1933	95.38	95.04	1.8739	38.54	2.7270	18.64	**0.113**	0.88
RED-Net Liu & Ji (2020)	0.1192	97.44	98.90	0.1313	98.00	98.46	0.7888	71.50	0.7823	66.68	0.115	0.78
EG-MVSNet Zhang et al. (2024a)	0.1035	97.93	98.85	0.1524	97.39	98.16	0.7230	75.01	0.9656	63.16	0.163	2.21
RCCR-MVS (Single-stage)	0.0933	98.22	**98.97**	0.1426	97.85	98.45	0.8831	70.87	0.8020	66.53	0.243	1.0517
Coarse-to-fine slice-wise methods												
CasREDNet Gu et al. (2020)	0.0987	97.78	98.31	0.1297	97.93	97.89	0.8377	72.84	1.1603	62.03	0.238	0.2145
AdaMVS Liu et al. (2023)	0.1014	97.74	98.01	0.1670	96.72	96.77	0.7269	75.40	0.6216	75.55	0.341	0.2015
AARMVSNet Wei et al. (2021)	0.1268	97.94	98.32	0.1422	97.70	97.75	0.8955	66.82	1.2059	53.29	0.718	0.4203
BH-RMVSNet Wei et al. (2024)	0.1317	97.69	98.12	0.1567	97.35	97.44	0.9008	64.86	0.9709	59.99	0.550	0.2985
D^2HCRMVSNet Yan et al. (2020)	0.1206	98.21	98.52	0.1033	98.79	98.77	1.0061	61.39	1.1794	52.47	0.708	0.3626
RCCR-MVS (Coarse-to-fine)	**0.0830**	**98.28**	98.70	**0.0838**	**99.09**	**99.02**	**0.6820**	**76.64**	**0.5144**	**81.14**	0.284	0.3267

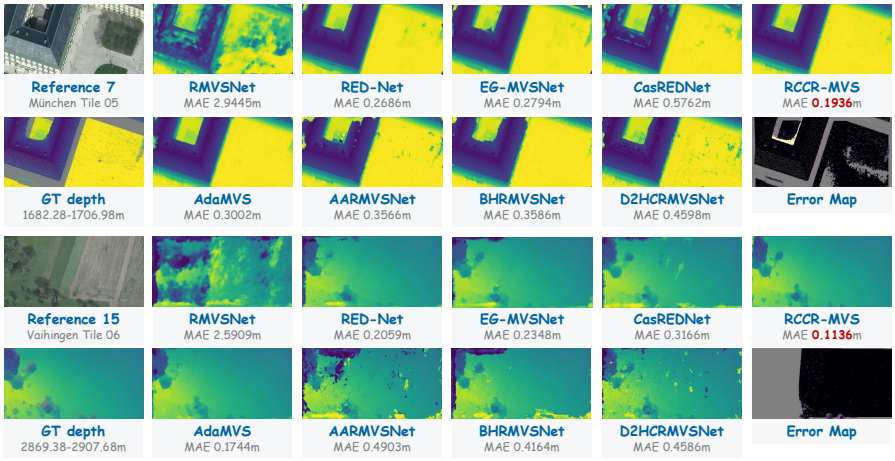

Figure 5: Qualitative comparison on München and Vaihingen tiles.

Generalization on the München and Vaihingen Datasets. As detailed in Table 1 and Appendix A.6 Table 6, our RCCR-MVS secures the best performance in seven of eight metrics across both evaluation protocols. In tile settings, our method outperforms all competitors, recovering sharper object boundaries and more consistent planar regions, as shown in Figure 5. This robust cross-domain generalization extends to the Vaihingen full-image setting. While the München full-image setting is more challenging, our method still retains the highest threshold accuracy. Ultimately, although unbounded full-image reconstruction remains difficult for all evaluated methods, RCCR-MVS provides consistently strong large-scale depth estimates, with further qualitative analysis provided in the Appendix A.7, Figures 11 and 12.

Generalization across Input Resolutions on MultiResXD-MVS Dataset. Evaluated across three MultiResXD-MVS protocols in Table 2, our RCCR-MVS minimizes error among coarse-to-fine methods across all input sizes, proving robust against severely constrained spatial context. The single-stage variant also performs strongly across resolutions, achieving the lowest MAE under the fixed-center and resized protocols, while RED-Net obtains slightly lower errors at 3× and 4× under expanding native coverage. Overall, these results indicate that RCCR remains effective across diverse spatial extents and resolutions, with detailed trend lines provided in Appendix A.8.

4.3 Analysis for RCCR-MVS

Table 2: MAE (m) on MultiResXD-MVS under three evaluation protocols and four input scales.

Method	Native: expanding coverage				Native: same center pixels				Resized: same coverage			
	1×	2×	3×	4×	1×	2×	3×	4×	1×	2×	3×	4×
R-MVSNet	2.2697	1.8470	1.7933	1.7715	2.2692	1.6379	1.7337	1.7678	1.2407	1.3540	1.5153	1.7715
RED-Net	1.3567	0.8727	**0.7362**	**0.7219**	1.3664	0.6317	0.6072	0.5780	0.4925	0.5244	0.5945	0.7619
EG-MVSNet	1.5228	0.9289	0.7631	0.7596	1.5323	0.5952	0.5806	0.5675	0.5070	0.5465	0.6075	0.7796
RCCR-MVS (SGL)	**1.3337**	**0.8530**	0.7464	0.7431	**1.3336**	**0.5770**	**0.5690**	**0.5442**	**0.4779**	**0.5039**	**0.5785**	**0.7431**
CasREDNet	1.4300	1.1938	1.1937	1.2166	1.4295	1.0686	1.1075	1.1242	0.6973	0.7839	0.9509	1.2166
AdaMVS	1.3750	1.0803	0.8678	0.8641	1.3748	0.8499	0.8548	0.8404	0.5060	0.5829	0.7098	0.9041
AARMVSNet	1.4495	1.0645	0.9567	0.9342	1.4490	0.8340	0.8386	0.8248	0.6820	0.7267	0.7631	0.9342
BH-RMVSNet	2.0818	1.3102	1.0616	0.9898	2.0813	0.8405	0.8719	0.8839	0.6201	0.6679	0.7729	0.9898
D^2HCRMVSNet	1.6172	1.1473	1.0330	1.0137	1.6169	0.9817	0.9700	0.9407	0.6533	0.6930	0.7868	1.0137
RCCR-MVS (C2F)	**1.2659**	**0.9741**	**0.8525**	**0.8577**	**1.2656**	**0.7553**	**0.7668**	**0.7573**	**0.4361**	**0.5189**	**0.6342**	**0.8577**

Table 3: Ablation of proposed components across different datasets.

Components			WHU 5 views		Luojia 5 views		München tiles		MultiResXD-1x	
RCCR	CRE	SES	MAE	< 0.6 m	MAE	< 0.6 m	MAE	< 0.6 m	MAE	< 0.6 m
×	×	×	0.0987	97.78	0.1297	97.93	0.8377	72.84	0.6973	79.56
✓	×	×	0.0913	98.01	0.1079	98.76	0.7064	75.61	0.4559	86.36
✓	✓	×	0.0856	98.20	0.0998	98.63	0.6932	76.12	0.4323	86.94
✓	×	✓	0.0891	98.04	0.1006	98.79	0.6878	75.83	**0.4252**	87.05
✓	✓	✓	**0.0830**	**98.28**	**0.0838**	**99.09**	**0.6820**	**76.64**	0.4361	**87.31**

Ablation and Analysis. As detailed in Table 3, we evaluate the proposed components on trained datasets WHU and Luojia-MVS alongside unseen generalization targets München and MultiResXD-MVS. The RCCR alone consistently improves all metrics, confirming its fundamental benefit to baseline accuracy and robust generalization. Adding CRE or SES independently improves performance across complex urban layouts. The complete pipeline achieves the best performance on seven of the eight metrics, demonstrating strong synergy between CRE and SES.

Effectiveness of RCCR. Fig. 6 compares the baseline with RCCR-only, without CRE or SES, on challenging WHU-MVS and MultiResXD-MVS scenes. RCCR-only suppresses spurious probability peaks, better aligns posterior means with ground truth, and reduces high-error regions in the final depth maps. Influence matrices and slice-26 perturbation curves reveal responses in both earlier and later output logits, whereas the baseline responds only at and after the perturbed slice. Aggregate results in Appendix A.10, Table 8, further show MAE reductions of 7.57% and 34.63%, with improvements on 97.94% and 87.76% of the evaluated images, respectively.

Effectiveness of CRE and SES. As validated in Appendix A.11, Tables 9 and 10, our architectural extensions provide targeted benefits. Adaptive CRE minimizes MAE by expanding search intervals strictly based on predictive uncertainty, optimally balancing out-of-bounds recovery with dense sampling. Complementarily, SES preserves intricate structures near discontinuities by explicitly restricting cross-boundary evidence mixing prior to slice-wise regularization.

4.4 Visualization of 3D point cloud reconstructions

To validate generalization, we deploy the trained RCCR-MVS to reconstruct point clouds across various unseen domains. As showcased by the scenes of München and Vaihingen in Figure 1(c) alongside the MultiResXD-MVS layouts in Figure 7, the generated reconstructions exhibit structural completeness. RCCR-MVS captures the village details in Figure 7(a1–a2) while preserving sharp building footprints and vegetation across the campus regions in Figure 7(b1–b3). These results illustrate the ability of RCCR-MVS to scale to high-resolution aerial environments without domain-specific fine-tuning.

5 Conclusion

We presented RCCR-MVS, a memory-efficient slice-wise recurrent framework for High-Resolution Aerial MVS. At its core, an asymmetric Reverse-Context-Conditioned Regularizer avoids full volumetric storage by utilizing compressed reverse context to explicitly condition a multi-scale forward pass. Complemented by CRE and SES to prevent interval overflows and mitigate initial matching noise, this paradigm achieves a superior accuracy-efficiency balance. Extensive evaluations show strong performance on standard benchmarks, along with exceptional generalization

on our newly constructed MultiResXD-MVS dataset and large-scale full-image scenarios. Moving forward, extending this framework to unbounded city-scale environments and incorporating visual foundation models represent promising future directions.

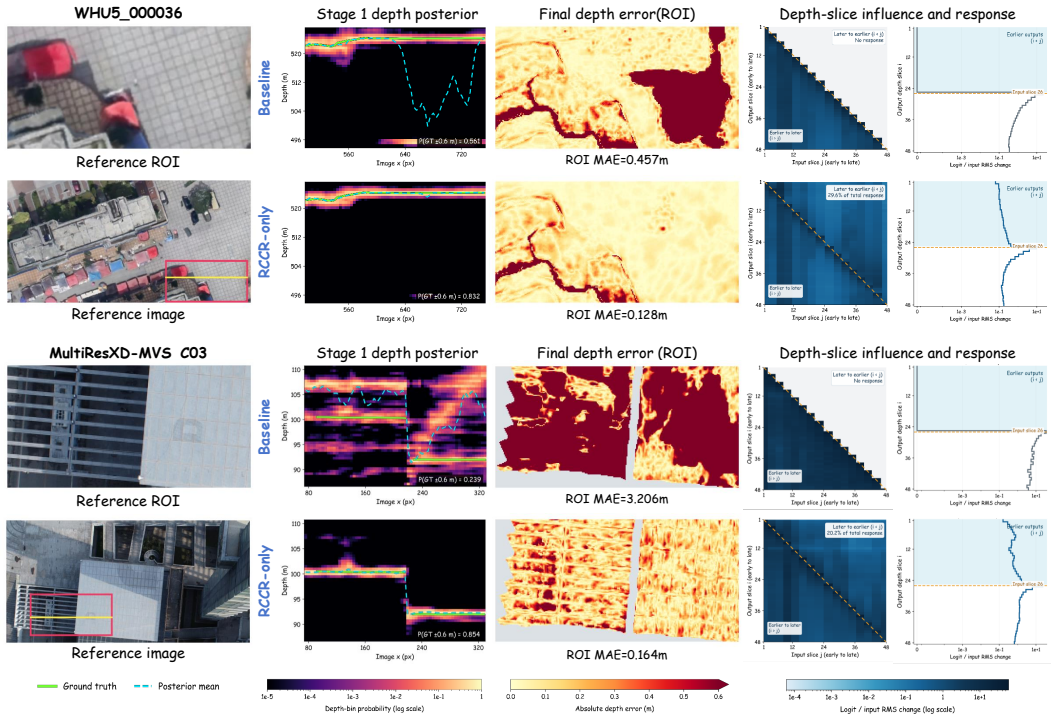

Figure 6: Qualitative comparison and depthwise influence analysis between the baseline and our RCCR-only variant on two challenging WHU-MVS and MultiResXD-MVS scenes.

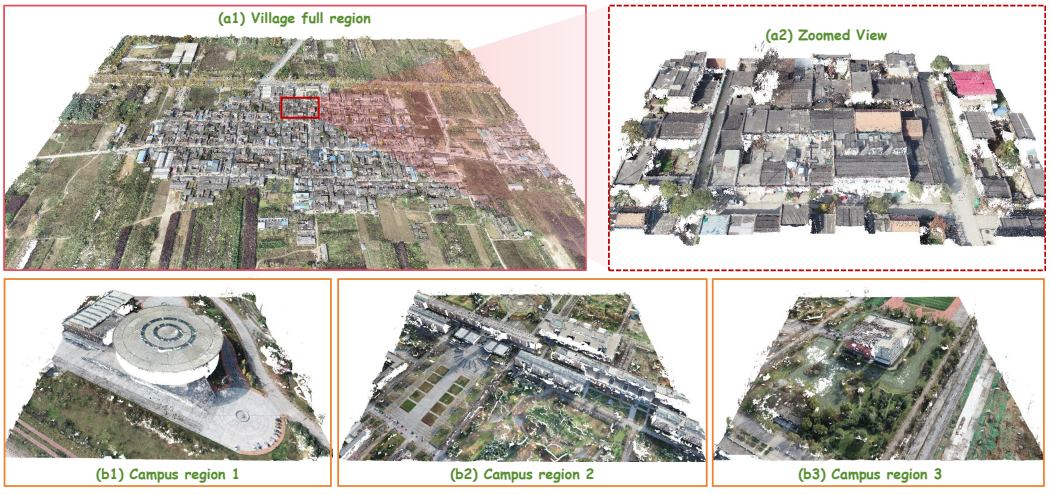

Figure 7: Point cloud reconstructions on the MultiResXD-MVS using direct inference with RCCR-MVS, showcasing (a1–a2) a village scene with local details and (b1–b3) three campus regions.

AI Assistance Disclosure

Generative AI tools, including large language models, were used to assist with English-language editing and manuscript revision, including improving clarity and conciseness, checking grammar and terminology consistency, and identifying potential inconsistencies in cross-references, tables,

and experimental descriptions. All AI-assisted content was manually reviewed and verified by the authors against the corresponding implementations, experimental records, and cited sources. The authors take full responsibility for the final content of this work.

REFERENCES

Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wXWfvSpYHh.

Wenhui Diao, Haichen Yu, Kaiyue Kang, Tong Ling, Di Liu, Yingchao Feng, Hanbo Bi, Libo Ren, Xuexue Li, Yongqiang Mao, and Xian Sun. RingMo-Aerial: An aerial remote sensing foundation model with affine transformation contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):10900–10913, 2025. doi: 10.1109/TPAMI.2025.3602237.

Taku Fujitomi, Seiya Ito, Naoshi Kaneko, and Kazuhiko Sumi. Bi-directional recurrent mvsnet for high-resolution multi-view stereo. In *International Conference on Machine Vision and Applications*, 2021. URL https://www.mva-org.jp/Proceedings/2021/papers/P3-9.pdf.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2495–2504, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Gu_Cascade_Cost_Volume_for_High-Resolution_Multi-View_Stereo_and_Stereo_Matching_CVPR_2020_paper.html.

Norbert Haala. The landscape of dense image matching algorithms. In Dieter Fritsch (ed.), *Photogrammetric Week '13*, pp. 271–284, Berlin/Offenbach, Germany, 2013. Wichmann.

Jianfei Jiang, Qiankun Liu, Hongyuan Liu, Haochen Yu, Liyong Wang, Jiansheng Chen, and Huimin Ma. MVSMamba: Multi-view stereo with state space model. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-4846.

Jianfei Jiang, Qiankun Liu, Haochen Yu, Hongyuan Liu, Liyong Wang, Jiansheng Chen, and Huimin Ma. MonoMVSNet: Monocular priors guided multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27806–27816, 2025b.

Jianfei Jiang, Liyong Wang, Haochen Yu, Tianyu Hu, Jiansheng Chen, and Huimin Ma. Rrt-mvs: Recurrent regularization transformer for multi-view stereo. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3994–4002, 2025c. doi: 10.1609/aaai.v39i4.32418. URL https://ojs.aaai.org/index.php/AAAI/article/view/32418.

Jiayi Li, Xin Huang, Yujin Feng, Zhen Ji, Shulei Zhang, and Dawei Wen. A hierarchical deformable deep neural network and an aerial image benchmark dataset for surface multiview stereo reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. doi: 10.1109/TGRS.2023.3234694.

Liangliang Li, Guihua Liu, Feng Xu, Wenjin Liao, and Wenjie Zhao. Mamba-enhanced multi-view stereo: Geometry-aware feature fusion and 3d cost volume regularization. *Applied Soft Computing*, 195:115074, 2026. doi: 10.1016/j.asoc.2026.115074. URL https://www.sciencedirect.com/science/article/pii/S1568494626005223.

Jin Liu and Shunping Ji. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6050–6059, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Liu_A_Novel_Recurrent_Encoder-Decoder_Structure_for_Large-Scale_Multi-View_Stereo_Reconstruction_CVPR_2020_paper.html.

Jin Liu, Jian Gao, Shunping Ji, Chang Zeng, Shaoyi Zhang, and JianYa Gong. Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:42–60, 2023. doi: 10.1016/j.isprsjprs.2023.08.015. URL https://www.sciencedirect.com/science/article/pii/S0924271623002289.

Yimei Liu, Yakun Ju, Yuan Rao, Hao Fan, Junyu Dong, Feng Gao, and Qian Du. Aerial multiview stereo via adaptive depth range inference and normal cues. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 2025. doi: 10.1109/TGRS.2025.3578255. URL https://arxiv.org/abs/2506.05655.

Yong-Qiang Mao, Hanbo Bi, Liangyu Xu, Kaiqiang Chen, Zhirui Wang, Xian Sun, and Kun Fu. SDL-MVS: View space and depth deformable learning paradigm for multiview stereo reconstruction in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024. doi: 10.1109/TGRS.2024.3464574. URL https://arxiv.org/abs/2405.17140.

Lina Wang, Jiayu Zhang, Ziqing Wang, Shuheng Liu, and Jiangfeng She. Hierarchical attention and dynamic consistency filtering for robust aerial multi-view stereo. *Knowledge-Based Systems*, 333:114973, 2026a. doi: 10.1016/j.knosys.2025.114973. URL https://www.sciencedirect.com/science/article/pii/S0950705125020118.

Shichao Wang, Mingxing Jia, Dapeng Niu, and Jiaxu Zhao. 3D reconstruction of aerial images with symmetrical gradient integral regression multi-view stereo network. *Advanced Engineering Informatics*, 69:104049, 2026b. doi: 10.1016/j.aei.2025.104049.

Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *IEEE/CVF International Conference on Computer Vision*, pp. 6187–6196, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Wei_AA-RMVSNet_Adaptive_Aggregation_Recurrent_Multi-View_Stereo_Network_ICCV_2021_paper.pdf.

Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Bidirectional hybrid lstm based recurrent neural network for multi-view stereo. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3062–3073, 2024. doi: 10.1109/TVCG.2022.3165860. URL https://ieeexplore.ieee.org/document/9754239/.

Jiang Wu, Rui Li, Haofei Xu, Wenxun Zhao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. GoMVS: Geometrically consistent cost aggregation for multi-view stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20207–20216, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_GoMVS_Geometrically_Consistent_Cost_Aggregation_for_Multi-View_Stereo_CVPR_2024_paper.html.

Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123490647.pdf.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, pp. 767–783, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yao_Yao_MVSNet_Depth_Inference_ECCV_2018_paper.html.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5525–5534, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Yao_Recurrent_MVSNet_for_High-Resolution_Multi-View_Stereo_Depth_Inference_CVPR_2019_paper.pdf.

Song Zhang, ZhiWei Wei, WenJia Xu, LiLi Zhang, Yang Wang, JinMing Zhang, and JunYi Liu. Edge aware depth inference for large-scale aerial building multi-view stereo. *ISPRS Journal*

of Photogrammetry and Remote Sensing, 207:27–42, 2024a. doi: 10.1016/j.isprsjprs.2023.11.020. URL https://www.sciencedirect.com/science/article/abs/pii/S0924271623003258.

Wei Zhang, Qiang Li, Yuan Yuan, and Qi Wang. Visual consistency enhancement for multiview stereo reconstruction in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024b. doi: 10.1109/TGRS.2024.3482697. URL https://doi.org/10.1109/TGRS.2024.3482697.

Wei Zhang, Zhigang Yang, Qiang Li, and Qi Wang. Semantic-guided multiview stereo reconstruction for aerial image. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. doi: 10.1109/TGRS.2025.3585623. URL https://doi.org/10.1109/TGRS.2025.3585623.

Wei Zhang, Yihang Wu, Shengkai Yu, Songhua Li, Qiang Li, and Qi Wang. GPR-MVS: Global propagation regularization for large-scale multiview stereo. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026. doi: 10.1109/TGRS.2026.3710063. URL https://doi.org/10.1109/TGRS.2026.3710063.

A Appendix

A.1 Preliminaries

Cascaded multi-view stereo. Given a reference image and multiple source views cascaded MVS progressively estimates depth across multiple spatial resolutions (Gu et al., 2020). At each stage source features are warped into the reference camera frustum at sampled depth hypotheses to construct a raw matching cost C_d using channel-wise variance (Yao et al., 2018). After a regularization module processes these sequential costs into scalar logits o_d the final continuous depth map is computed through a probability-weighted expectation.

Slice-wise recurrent regularization. To avoid memory-intensive volumetric convolutions for high-resolution images recurrent architectures process matching costs sequentially along the depth dimension to enable a memory-efficient slice-wise inference strategy (Yao et al., 2019). Rather than instantiating a full cost volume the matching cost C_d and the soft-argmin expectation are computed on the fly for each hypothesis d. Building upon this paradigm the recurrent encoder–decoder integrates a spatial multi-scale structure where a unidirectional cell at each scale ℓ updates its recurrent state via $h_d^{(\ell)} = \mathrm{ConvGRU}_\ell(x_d^{(\ell)}, h_{d-1}^{(\ell)})$ (Liu & Ji, 2020). A decoder subsequently fuses these updated multi-scale states to predict the current logit o_d which effectively propagates geometric context from previously scanned hypotheses while keeping overall GPU memory consumption strictly bounded.

A.2 Detailed Architecture of the Coverage-Risk Head

This appendix provides the full mathematical and structural details of the risk evaluation function $\mathcal{F}_{\text{risk}}$ introduced in Section 3.2. The coverage-risk head is designed to combine spatial context from the current reference features with explicit uncertainty statistics derived from the preceding depth posterior. From the previous posterior distribution π_j^- and its corresponding depth hypotheses z_j^-, we extract four pixel-wise statistics to quantify the estimation uncertainty. These metrics comprise the depth standard deviation σ_z, the maximum probability $p_{\max}$, the top-two probability margin m, and the normalized entropy $\bar{\mathcal{H}}$:

$$\begin{aligned} \sigma_z^2 &= \sum_j \pi_j^- (z_j^- - \widehat{z}^-)^2, \\ p_{\max} &= \pi_{(1)}^-, \qquad m = \pi_{(1)}^- - \pi_{(2)}^-, \\ \bar{\mathcal{H}} &= -\frac{\sum_j \pi_j^- \log \pi_j^-}{\log D^-}, \end{aligned} \tag{11}$$

where the numerical subscripts denote the indices of the highest and second-highest probabilities. These statistical maps are resized first to the original image resolution and then bilinearly adapted to the destination feature resolution to perfectly align with the target sampling grid. To capture local geometric fluctuations, we compute the local span v_{span} and deviation v_{dev} of the baseline center within a 3×3 spatial neighborhood $\mathcal{N}_3(\mathbf{p})$, utilizing replicate padding at the image boundaries:

$$\begin{aligned} v_{\text{span}}(\mathbf{p}) &= \max_{\mathbf{q}\in\mathcal{N}_3(\mathbf{p})} c(\mathbf{q}) - \min_{\mathbf{q}\in\mathcal{N}_3(\mathbf{p})} c(\mathbf{q}), \\ v_{\text{dev}}(\mathbf{p}) &= \left| c(\mathbf{p}) - \mathrm{mean}_{\mathbf{q}\in\mathcal{N}_3(\mathbf{p})}\, c(\mathbf{q}) \right|. \end{aligned} \tag{12}$$

To ensure numerical stability across varying scene scales, we normalize the standard deviation and the local variation metrics by the baseline half-range bounded by a minimal epsilon of 10^{-6}. The concatenated six-channel statistical input q is strictly bounded between zero and ten to prevent anomalous gradients:

$$q = \mathrm{clip}\big([\sigma_z/\rho_\epsilon,\ p_{\max},\ m,\ \bar{\mathcal{H}},\ v_{\text{span}}/\rho_\epsilon,\ v_{\text{dev}}/\rho_\epsilon],\ 0,\ 10\big). \tag{13}$$

A lightweight convolutional network subsequently processes these statistics alongside the detached image features. Specifically, an input projection reduces the detached reference features $F_0^{(t)}$ to four channels. This representation is concatenated with the detached statistical vector q to form a

ten-channel input. The network sequentially applies a standard three-by-three convolution mapping ten channels to sixteen, a SiLU activation, a sixteen-channel depthwise convolution, another SiLU activation, and a final projection to yield the two output logits:

$$\begin{aligned} \mathbf{v}_0 &= [\mathcal{P}_{\text{ref}}(\text{sg}(F_0^{(t)})), \text{sg}(q)], \\ \mathbf{v}_1 &= \text{SiLU}(\mathcal{C}_{3\times3}(\mathbf{v}_0)), \\ [\ell_p, \ell_a] &= \mathcal{P}_{\text{out}}(\text{SiLU}(\mathcal{D}_{3\times3}(\mathbf{v}_1))) \,. \end{aligned} \tag{14}$$

Each enabled interface maintains its own independent set of weights for this head. To ensure stable early training, the risk-output bias ℓ_p is initialized to negative four and the amount-output bias ℓ_a to zero, combined with exceptionally small output weights. This configuration guarantees that the predicted expansion amount starts near zero and the risk probability remains very low, keeping the initial candidate range safely close to the unexpanded baseline.

A.3 Detailed Formulation of the Weighted Residual Aggregation

This appendix provides the complete mathematical formulation of the *Weighted Residual Aggregation* mechanism $\mathcal{F}_{\text{agg}}$ introduced in Section 3.2. The core objective of this mechanism is to dynamically weigh and aggregate local evidence differences based on both image-guided directional priors and geometric compatibility.

Because the identical cost slice index may represent vastly different metric depths at adjacent pixels near object boundaries, we first define a local depth gap g_d derived from the depth hypotheses z_d to prevent feature blurring across depth discontinuities. This gap is computed based on the absolute metric depth difference between the central pixel $\mathbf{p}$ and its neighbor $\mathbf{q}$ normalized by the local depth spacing Δ_d at the central pixel:

$$\begin{aligned} \Delta_d(\mathbf{p}) &= \frac{|z_{d_+}(\mathbf{p}) - z_{d_-}(\mathbf{p})|}{d_+ - d_-}, \\ g_d(\mathbf{p}, \mathbf{q}) &= \frac{|z_d(\mathbf{p}) - z_d(\mathbf{q})|}{\max(\Delta_d(\mathbf{p}), 10^{-6})}, \end{aligned} \tag{15}$$

where d_- and d_+ represent the valid adjacent slice indices constrained by the hypothesis boundaries. A neighbor $\mathbf{q}$ is considered valid only if it falls within the spatial image bounds and strictly satisfies a geometric compatibility threshold $g_d \leq \gamma$. We denote the subset of valid neighbors for a given pixel by $\mathcal{V}_d(\mathbf{p})$.

To further quantify the intrinsic feature discrepancy between adjacent pixels, we calculate a relative evidence difference δ_d. This term is defined by the absolute distance between the low-dimensional embeddings E_d normalized by their respective channel-wise magnitudes combined with a small constant to ensure numerical stability:

$$\delta_d(\mathbf{p}, \mathbf{q}) = \frac{\frac{1}{J}\|E_d(\mathbf{q}) - E_d(\mathbf{p})\|_1}{\frac{1}{J}\|E_d(\mathbf{p})\|_1 + \frac{1}{J}\|E_d(\mathbf{q})\|_1 + 10^{-4}}. \tag{16}$$

For the joint weighting strategy, we formulate an unnormalized neighbor score v_d by integrating three distinct components. It positively incorporates the reference-guided directional score S extracted by the context branch while penalizing both the capped geometric gap and the relative evidence difference:

$$v_d(\mathbf{p}, \mathbf{q}) = S_{\text{dir}(\mathbf{q}-\mathbf{p})}(\mathbf{p}) - \min(g_d(\mathbf{p}, \mathbf{q}), 10) - \delta_d(\mathbf{p}, \mathbf{q}). \tag{17}$$

The central pixel is inherently assigned an unnormalized score of zero to anchor the normalization process. For each valid neighbor, the final aggregation weight w_d is obtained via a softmax-like function applied over the valid neighborhood set:

$$w_d(\mathbf{p}, \mathbf{q}) = \frac{\exp(v_d(\mathbf{p}, \mathbf{q}))}{1 + \sum_{\mathbf{r}\in\mathcal{V}_d(\mathbf{p})} \exp(v_d(\mathbf{p}, \mathbf{r}))}. \tag{18}$$

Invalid neighbors strictly receive zero weight and the constant one in the denominator explicitly accounts for the self-contribution of the central pixel. Finally, the aggregation block computes the

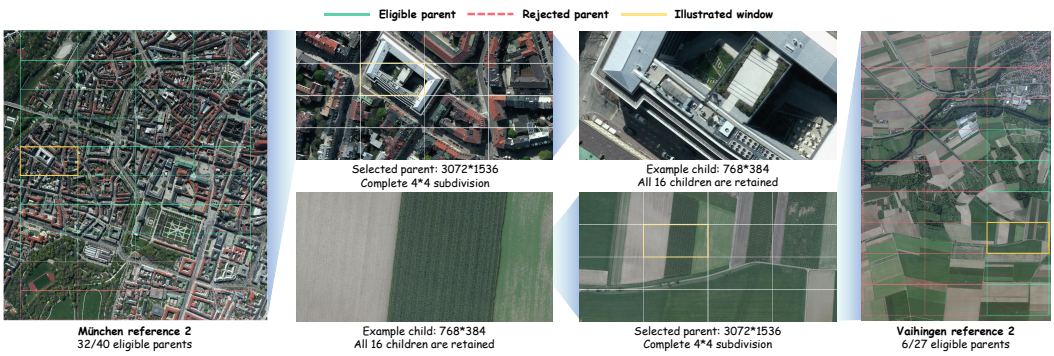

Figure 8: München and Vaihingen tile preparation process. Green windows indicate eligible parents containing sufficient valid depth pixels while dashed red windows denote rejected regions. Every eligible 3072×1536 parent window uniformly contributes sixteen 768×384 child tiles for native-resolution evaluation.

Table 4: Frozen München and Vaihingen evaluation sets detailing the number of extracted references, parent windows, and final native tiles evaluated.

Dataset	References	Candidate parents	Eligible parents	Native tiles
München	9	360	284	4,544
Vaihingen	30	810	571	9,136

spatially weighted residual sum R_d and injects it back into the original matching cost slice C_d via a learned output projection W_{out} scaled by a fixed hyperparameter η:

$$R_d(\mathbf{p}) = \sum_{\mathbf{q} \in \mathcal{V}_d(\mathbf{p})} w_d(\mathbf{p}, \mathbf{q})[E_d(\mathbf{q}) - E_d(\mathbf{p})],$$
$$X_d = C_d + \eta W_{\text{out}} R_d. \tag{19}$$

As discussed in the main text, initializing the weights of W_{out} to zero guarantees that the residual term initially vanishes allowing the module to act as a stable identity mapping at the beginning of the training phase.

A.4 Selection of München and Vaihingen Evaluation Samples

The München and Vaihingen datasets provide high-resolution aerial imagery capturing complex urban and rural landscapes from which we select 9 and 30 reference images, respectively. We evaluate model performance on these datasets through two distinct protocols comprising full-image and tile-level evaluations. For the full-image protocol, we downsample the original large-scale rasters by a scale factor of 0.25 yielding evaluable inputs of 3536×3888 pixels for München and 2352×3600 pixels for Vaihingen. Conversely, the tile-level protocol evaluates models at their native resolution by placing a non-overlapping grid of 3072×1536 parent windows over the reference images. As illustrated in Figure 8 and detailed in Table 4, parent windows containing at least fifty percent valid depth pixels are retained and subsequently subdivided into complete sets of sixteen 768×384 children tiles. Across both protocols, the geometric search ranges are dynamically bounded using the first and ninety-ninth percentiles of the valid reference depths to ensure consistent metric evaluation without strictly testing the recovery of unknown global depth boundaries.

A.5 Construction of MultiResXD-MVS

MultiResXD-MVS is constructed from 2020 nadir UAV imagery capturing diverse campus and village environments to evaluate high-resolution depth estimation and spatial generalization. The initial source data comprises 5472×3648 pixel sensor images accompanied by photogrammetric digital surface models with a grid spacing of 0.05 meters, which provide the fundamental surface-depth references. Candidate locations are systematically sampled across visually delineated building clusters using a dense uniform grid to ensure comprehensive coverage of complex roof structures. As



Figure 9: Complementary MultiResXD-MVS input families comparing native crops that enlarge spatial coverage at a fixed pixel sampling rate against resized inputs that vary pixel sampling over identical spatial extents.

Table 5: MultiResXD-MVS input pairings across the native and resized evaluation families.

Input size	Native crops	Fixed-coverage resizing
768×384	$1\times$ crop	Scale factor 1/4
1536×768	$2\times$ crop	Scale factor 1/2
2304×1152	$3\times$ crop	Scale factor 3/4
3072×1536	$4\times$ crop	Identical to native $4\times$

illustrated in Figure 4, strict geographic distance constraints and geometric validity filtering yield forty-nine distinct evaluation locations comprising 18 campus and 31 village scenes. At each retained location, we configure one reference and four source images fixed across all resolution variants, yielding 343 unique five-view input instances. We establish depth hypothesis bounds using the minimum and maximum camera depths from the underlying sparse reconstruction structure, then expand outward to safely encompass the full geometric search space.

To thoroughly evaluate model generalization across varying resolutions and spatial extents, we generate two complementary families of input variants detailed in Table 5 and illustrated in Figure 9. The native family evaluates expanding spatial coverage by progressively nesting image crops around a common center on the undistorted full-resolution pixel grid, thereby preserving the inherent camera pixel density. Conversely, the resized family evaluates the impact of interpolation and changing spatial sampling densities by progressively downsampling the largest available five-view parent crop using area interpolation. Notably, both families share the largest resolution setting of 3072×1536, bridging the two evaluation paradigms. Across all variants, evaluation validity masks strictly require positive depth projections into at least two source images coupled with strong agreement against the underlying photogrammetric surface reference to ensure high-fidelity quantitative scoring, as depicted in Figure 10.

A.6 Supplementary Quantitative Results

Table 6 reports additional quantitative results on WHU and Luojia-MVS using three input views, as well as full-image evaluation results on unseen München and Vaihingen datasets. MAE is reported

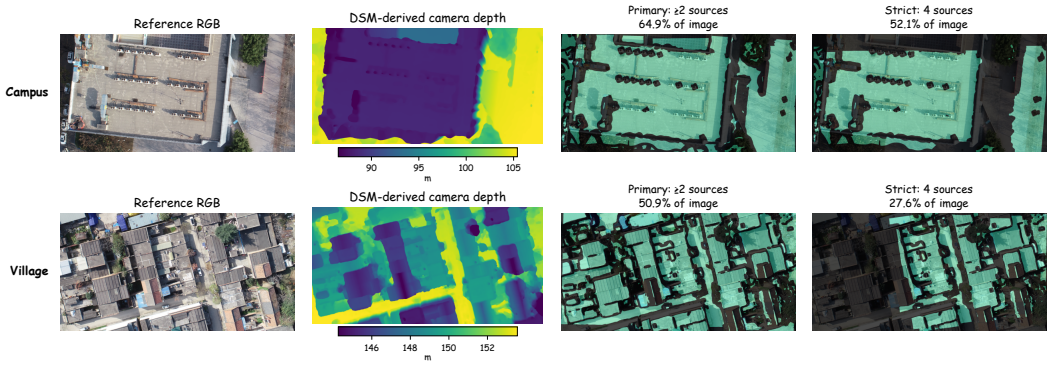

Figure 10: Digital surface model-derived reference depths and binary evaluation support masks for the largest native input setting, with the evaluation masks enforcing geometric visibility.

Table 6: **Additional comparison on WHU and Luojia-MVS with 3 views and generalization on full-image München and Vaihingen evaluation.** Methods marked with * and gray-shaded results are quoted from original papers. VFM denotes foundation models; "–" denotes unreported results.

Method	Mode	WHU 3 views			Luojia 3 views			München full		Vaihingen full	
		MAE	< 0.6 m	< 3 intv.	MAE	< 0.6 m	< 3 intv.	MAE	< 0.6 m	MAE	< 0.6 m
Full-volume methods											
RingMo-Aerial* Diao et al. (2025)	VFM	0.1540	97.40	98.30	0.0950	98.70	97.80	–	–	–	–
HDC-MVSNet* Li et al. (2023)	C2F	0.1210	98.10	97.70	0.0890	98.70	97.80	–	–	–	–
SG-MVS Wang et al. (2026b)	C2F	0.1232	97.59	98.23	0.1650	96.96	95.71	–	–	–	–
MonoMVSNet Jiang et al. (2025b)	C2F	0.1635	96.09	95.87	0.1461	96.83	94.65	–	–	–	–
MVSMamba Jiang et al. (2025a)	C2F	0.1537	96.32	96.91	0.1377	97.16	97.15	–	–	–	–
RRT-MVS Jiang et al. (2025c)	C2F	0.1331	97.06	95.15	0.1187	98.20	94.85	–	–	–	–
Single-stage slice-wise methods											
R-MVSNet Yao et al. (2019)	SGL	0.1931	94.82	95.96	0.1704	96.50	96.19	2.8257	13.95	4.0030	5.14
RED-Net Liu & Ji (2020)	SGL	0.1306	97.15	98.84	0.1414	97.83	98.34	**1.5566**	32.80	2.9246	11.86
EG-MVSNet Zhang et al. (2024a)	SGL	0.1154	97.59	98.69	0.1141	98.53	98.89	1.6773	30.82	2.5336	15.09
RCCR-MVS	SGL	0.1034	98.08	**98.86**	0.0905	99.07	**99.29**	1.7310	25.81	3.1297	8.80
Coarse-to-fine slice-wise methods											
CasREDNet Gu et al. (2020)	C2F	0.1047	97.87	98.33	0.0950	98.82	98.79	1.7685	30.63	2.4762	16.95
AdaMVS Liu et al. (2023)	C2F	0.1192	97.58	98.03	0.1071	98.01	98.39	1.9004	32.08	2.8037	14.36
AARMVSNet Wei et al. (2021)	C2F	0.1454	97.47	97.87	0.1174	98.45	98.46	1.8904	27.34	2.6440	14.34
BH-RMVSNet Wei et al. (2024)	C2F	0.1630	96.87	97.23	0.1572	97.52	97.42	2.5766	15.10	2.8686	12.05
D^2HCRMVSNet Yan et al. (2020)	C2F	0.1481	97.36	97.78	0.1062	98.72	98.71	2.0229	24.44	2.8171	12.50
RCCR-MVS	C2F	**0.0983**	**98.11**	98.54	**0.0826**	**99.08**	99.04	1.6311	**32.81**	**2.2728**	**18.95**

in meters, and threshold accuracies are expressed as percentages. These results complement the five-view and tile-based evaluations presented in Table 1.

A.7 Additional Full-Image Qualitative Results

Figures 11 and 12 supplement the main text with full image predictions for the München and Vaihingen datasets, allowing for the detailed inspection of both local structural fidelity and global spatial consistency across expansive scenes. Specifically, the München visualizations reveal distinct performance variations in recovering densely arranged building boundaries, while the Vaihingen examples highlight differences across broad agricultural fields and complex vegetation. Although our method captures the large-scale geometric layout, localized depth artifacts around extremely fine structures occasionally remain visible, illustrating the inherent challenges of unbounded full-image aerial reconstruction and providing valuable qualitative context for the dataset-level quantitative results.

A.8 Additional Visualizations for MultiResXD-MVS Generalization

Figure 13 supplements the quantitative results from the main text by visualizing the error trends across the three complementary MultiResXD-MVS evaluation protocols. These curves clearly illustrate the performance trajectories as input sizes increase. Notably, the stable, flat trajectories of

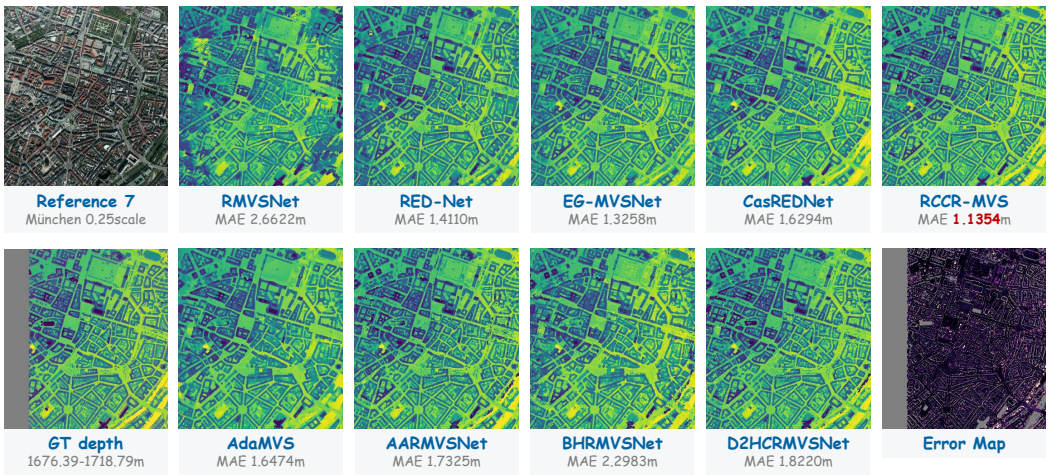

Figure 11: Full-image qualitative comparison on München reference 7 at input scale 0.25. The panels show the reference image, ground-truth depth, method predictions, and an absolute-error map. MAE annotations are specific to this reference image.

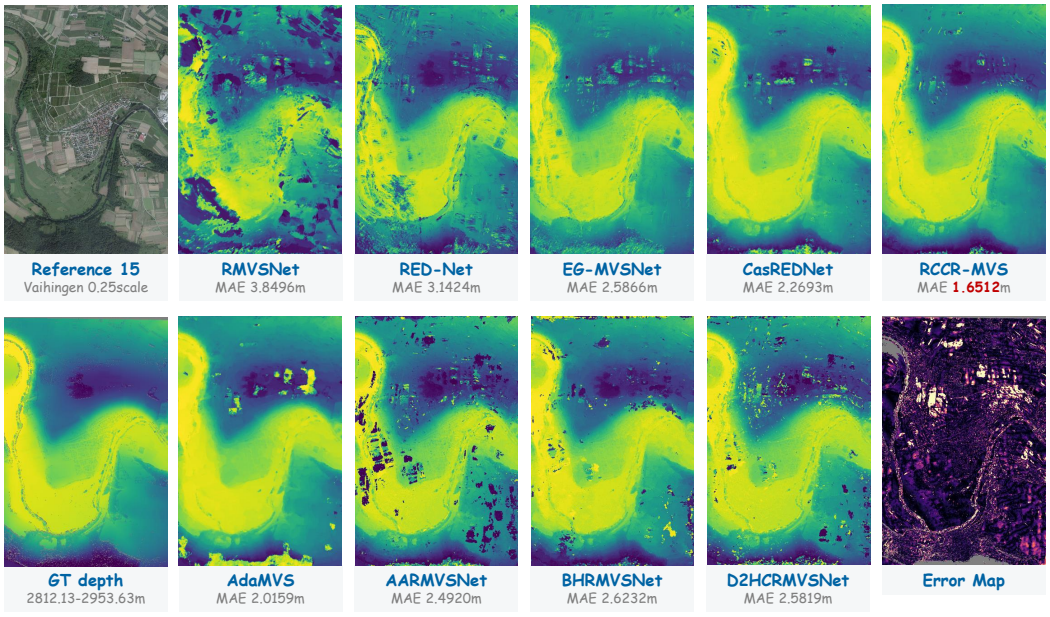

Figure 12: Full-image qualitative comparison on Vaihingen reference 15 at input scale 0.25. The panels show the reference image, ground-truth depth, method predictions, and an absolute-error map. MAE annotations are specific to this reference image.

RCCR-MVS under the fixed-center protocol visually confirm its robust ability to leverage expanding spatial context without degrading central-region accuracy.

A.9 Efficiency Analysis

As detailed in Table 7, our RCCR-MVS achieves the lowest MAE while maintaining a competitive computational footprint. Although extracting bidirectional context introduces additional operations compared to strict baselines, our resource consumption remains substantially lower than heavily parameterized volumetric architectures, especially on extremely high-resolution inputs. This practical efficiency stems directly from our sequential two-pass inference strategy. By computing matching costs on the fly to extract compressed reverse context and recomputing them during the forward pass, we completely avoid memory-prohibitive full-volume storage. While caching these lightweight

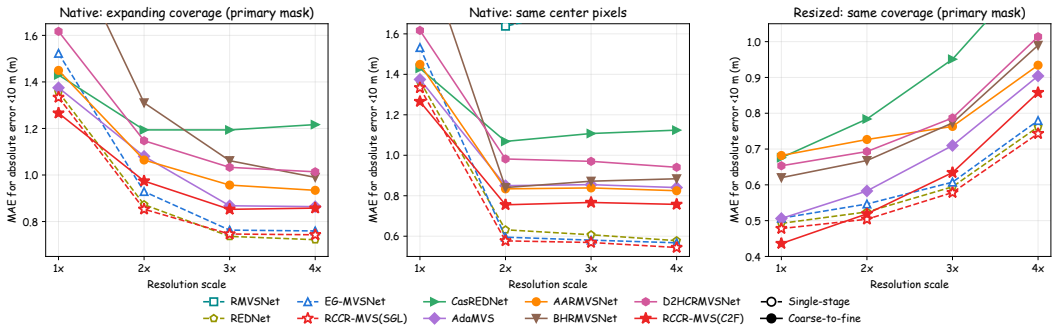

Figure 13: Generalization error trends on MultiResXD-MVS under varying input sizes: (left) expanding coverage on the native grid; (middle) evaluating identical central pixels with increasing context; (right) varying resolution at fixed coverage. Dashed and solid lines indicate single-stage and coarse-to-fine methods, respectively.

Table 7: Accuracy and computational efficiency comparison.

Method	WHU MAE	GPU GB	GFLOPs	Infer GPU 768 × 384	Infer time 768 × 384	Infer GPU 3536 × 3888	Infer time 3536 × 3888
R-MVSNet	0.1751	**5.93**	163.91	**0.113**	0.8834	**4.75**	28.9050
RED-Net	0.1192	14.51	341.04	0.115	0.7836	**4.75**	79.8093
EG-MVSNet	0.1035	21.46	713.63	0.163	2.2095	7.14	168.8595
CasREDNet	0.0987	8.27	220.70	0.238	0.2145	10.86	15.8728
AdaMVS	0.1014	8.53	**147.40**	0.341	**0.2015**	16.10	**6.6398**
AARMVSNet	0.1268	19.35	766.38	0.718	0.4203	28.88	21.5607
BH-RMVSNet	0.1317	27.92	1225.06	0.550	0.2985	25.04	10.2599
D^{2}HCRMVSNet	0.1206	25.07	1351.98	0.708	0.3626	25.80	31.1064
RCCR-MVS	**0.0830**	10.36	349.01	0.284	0.3267	13.37	11.2907

summaries requires linear memory scaling, this controlled recomputation successfully unlocks deep bidirectional integration for large-scale aerial imagery within standard hardware constraints.

A.10 Aggregate Evaluation of RCCR

Table 8: Aggregate comparison of Baseline and RCCR-only without CRE or SES on the WHU-MVS 5-view test set and MultiResXD-MVS resized_1x dataset. P_{GT} denotes the mean first-stage probability mass within 0.6 m of ground truth.

Dataset	Method	MAE (m) ↓	< 0.6 m (%) ↑	P_{GT} (%) ↑	Improved (%) ↑
WHU	Baseline	0.0987	97.78	79.91	–
WHU	RCCR-only	**0.0913**	**98.01**	**82.39**	97.94
MultiResXD-1x	CasREDNet	0.6973	79.56	57.03	–
MultiResXD-1x	RCCR-only	**0.4559**	**86.36**	**65.72**	87.76

Table 8 complements the qualitative analysis with aggregate results on WHU-MVS and MultiResXD-MVS. Compared with Baseline, RCCR-only reduces MAE by 7.57% and 34.63%, respectively, while improving first-stage probability mass near the ground-truth depth and final threshold accuracy. Moreover, MAE improves on 97.94% and 87.76% of the evaluated images, indicating that the gains extend beyond the selected visualization examples.

A.11 Detailed Analysis of CRE and SES

This section provides detailed quantitative evaluations of our auxiliary modules on the complete WHU-MVS five-view test set. Table 9 compares our adaptive CRE strategy against fixed-scaling baselines, demonstrating that dynamically allocating expansion strictly based on local risk mini-

Table 9: Ablation of the CRE module on the WHU dataset. The variables denote the mean expansion factor $\bar{s}$, ground truth coverage C, and the recovery rate R of true depths originally falling outside the unexpanded range.

Method	Stage 2			Stage 3			MAE (m) ↓
	$\bar{s}$	C (%) ↑	R (%) ↑	$\bar{s}$	C (%) ↑	R (%) ↑	
Fixed matched	1.00565	99.697	1.704	1.01816	98.299	2.248	0.09297
Fixed 1.5×	1.50000	**99.817**	**60.414**	1.50000	**98.850**	**36.153**	0.08884
Adaptive CRE	1.00646	99.812	32.094	1.01651	98.768	25.409	**0.08563**

Table 10: Ablation of the SES module on the WHU dataset. All variants retain bidirectional regularization while disabling range expansion. Region masks are derived strictly from ground truth depths and reference images. Texture-based regions may naturally overlap depth-based regions.

Variant	Overall MAE	Edge MAE	Flat MAE	Weak MAE	Rich MAE
Uniform	0.08797	0.56961	0.03505	0.05659	0.16463
Context	0.08688	0.56226	0.03465	0.05609	**0.16204**
Evidence	0.08735	0.56226	0.03526	0.05642	0.16309
w/o hard mask	0.08698	0.56569	**0.03421**	0.05621	0.16235
w/o geometry	0.08755	0.57013	0.03430	0.05629	0.16342
Full SES	**0.08681**	**0.55749**	0.03561	**0.05599**	0.16224

mizes the overall mean absolute error. Furthermore, Table 10 breaks down the performance of SES across distinct geometric regions. The complete geometry-aware aggregation yields particularly pronounced improvements near severe depth discontinuities, confirming its robust capability to prevent feature blurring across structural boundaries.