ADAPTIVE DENSE RISK PREDICTION FOR CONSISTENT FEED-FORWARD 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent feed-forward 3D reconstruction models reconstruct dense 3D geometry from unconstrained multi-view images in a single forward pass. As a 3D foundation model, VGGT is increasingly used as a geometric prior in visual SLAM, dynamic scene reconstruction, and novel view synthesis, where VGGT confidence is crucial for point filtering and optimization weighting. However, VGGT confidence is learned jointly with geometric regression rather than directly supervised for depth accuracy, allowing high-confidence regions to contain severe depth errors. Such errors can survive filtering, receive large weights in optimization, and destabilize downstream tasks. We observe that the failures can be captured from VGGT outputs. Depth and point maps can reveal internal self-consistency of VGGT, and multi-view depth maps can capture geometric consistency across views. Motivated by this observation, we propose VERA, a **V**GGT **ER**ror-**A**ware dense risk prediction network that adaptively fuses two confidence maps and two consistency signals to predict a dense risk map of depth errors. Trained on nine datasets and evaluated on five sequence-disjoint test sets, VERA substantially improves depth error prediction metrics, reducing AUSE by 30.1% while improving AUROC by 28.6% and AUPR-Error by 81.3% over the best raw signal baseline. When used as an optimization weight, VERA reduces VGGT-Long odometry ATE RMSE by 13.7% and reduces Chamfer distance after confidence-weighted pose refinement by 16.6%, all without retraining or modifying the frozen foundation model.

1 INTRODUCTION

Feed-forward 3D reconstruction models directly predict dense 3D geometry from unconstrained RGB images in a single inference pass, enabled by the strong geometric priors of large-scale transformers (Wang et al., 2024). VGGT extends this paradigm as a 3D foundation model pretrained on diverse large-scale 3D datasets, predicting camera parameters, depth maps, point maps, and confidence maps from one to hundreds of views in a single forward pass (Wang et al., 2025). As VGGT outputs are increasingly used as geometric priors in visual SLAM, long-sequence reconstruction, streaming 4D scene reconstruction, and dense novel view synthesis (Jeon et al., 2026; Deng et al., 2026; Zhuo et al., 2026; Liu et al., 2025), its confidence maps have become an important signal for point filtering, geometric alignment, SLAM optimization, test-time adaptation, and learning control (Deng et al., 2026; Jeon et al., 2026; Tang et al., 2025; Lu et al., 2025; Chen et al., 2026).

However, a high VGGT confidence value does not necessarily imply a low depth or point-map error, as shown in Fig. 1. VGGT confidence is learned jointly with geometric regression heads, not directly from ground-truth prediction errors (Kendall & Gal, 2017). Such overconfident failures have also been observed in recent feed-forward reconstruction studies (Han et al., 2026; Lu et al., 2025). These failures can destabilize geometric refinement, as erroneous regions may survive confidence-based filtering and receive large optimization weights. We observe that such failures are detectable from VGGT outputs. Depth and point maps reveal internal self-consistency, while multi-view depth maps capture cross-view geometric consistency. These consistency signals complement raw confidence for reliable error prediction, as shown in Fig. 1.

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.

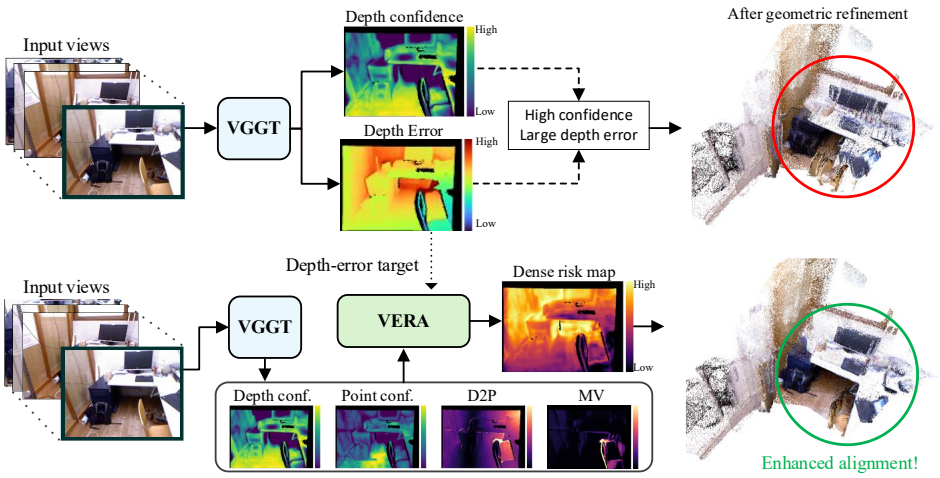


Figure 1: **Motivation and overview of our framework.** VGGT can assign high depth confidence to regions with large depth error, so overconfident failures persist during confidence-weighted geometric refinement. VERA predicts a dense risk map by fusing depth confidence, point confidence, Depth-to-Point (D2P) consistency, and Multi-View (MV) geometric consistency. Supervised by depth error, it identifies unreliable regions missed by raw confidence.

Motivated by this observation, we propose VERA, a **V**GGT **ER**ror-**A**ware dense risk prediction network that predicts a dense risk map by adaptively fusing two confidence maps and two consistency signals. VERA first encodes each signal with a convolutional pyramid encoder to capture local geometric structure at multiple scales. The four coarse features are fused via self-attention to incorporate complementary cues from the other signals. Given these enriched features, a spatial signal-wise gating adaptively determines the reliability of each signal at each location, allowing VERA to predict an accurate dense risk map of VGGT depth errors. VGGT itself remains frozen, so VERA acts as a lightweight reliability layer that adds 0.62M parameters to an approximately 1B-parameter foundation model. To the best of our knowledge, prior error-prediction methods have not jointly leveraged aleatoric uncertainty, consistency among different geometric predictions, and multi-view geometric consistency. VERA analyzes how these cues capture different failure modes and complement one another, and learns to adaptively weight their contributions.

To validate our approach, we train VERA on nine diverse 3D datasets and evaluate it on five sequence-disjoint held-out test sets. Relative to the best raw signal for each metric, VERA reduces AUSE by 30.1%, improves AUROC by 28.6%, and improves AUPR-Error by 81.3%. We further evaluate whether the risk map can replace VGGT confidence in downstream tasks. In VGGT-Long odometry, VERA reduces the average ATE RMSE by 13.7%, and in confidence-weighted pose refinement it reduces Chamfer distance by 16.6% while improving both accuracy and completeness.

Our contributions are threefold. First, we identify overconfident failures in VGGT, where severe depth errors can still receive high confidence and affect downstream filtering or optimization. Second, we show that two additional consistency signals can reveal failures missed by confidence alone, and propose VERA, a VGGT error-aware dense risk prediction network that combines two confidence maps and two consistency signals. Third, we show that VERA improves depth-error prediction on held-out sequences while providing more reliable weights for long-sequence odometry and confidence-weighted pose refinement.

2 Related Work

2.1 Feed-forward 3D reconstruction and applications

Classical 3D reconstruction pipelines process images through sequential stages of feature matching, camera pose estimation, and bundle adjustment (Triggs et al., 2000; Schönberger & Frahm, 2016). Although deep learning has improved task-specific modules such as keypoint detection and image

matching (DeTone et al., 2018; Sarlin et al., 2020), these multi-stage pipelines remain prone to error accumulation and propagation across stages. Wang et al. (2024) introduced a feed-forward model that directly regresses dense point maps from uncalibrated image pairs in a single forward pass, without requiring known camera parameters. Subsequent works scaled feed-forward reconstruction to arbitrary numbers of views (Yang et al., 2025; Wang et al., 2026), and recent 3D foundation models further extend this paradigm by jointly predicting camera parameters, depth maps, point maps, and confidence maps after pretraining on large-scale diverse 3D datasets (Wang et al., 2026; 2025).

We focus on VGGT as a representative 3D foundation model whose outputs are increasingly adopted as geometric priors in applications, including visual SLAM (Jeon et al., 2026), long-sequence reconstruction (Deng et al., 2026), streaming 4D scene reconstruction (Zhuo et al., 2026), and dense novel view synthesis (Liu et al., 2025). In these applications, confidence is often used directly as a reliability score, for example in confidence-weighted point alignment (Deng et al., 2026), keyframe selection and optimization weighting (Jeon et al., 2026), and online memory updates during test-time reconstruction (Chen et al., 2026). Confidence from feed-forward reconstruction models is also used for point map regression weighting (Tang et al., 2025) and test-time self calibration through multi view geometric consistency (Lu et al., 2025). VGGT confidence was not originally designed as a calibrated estimator of realized depth error, and is not directly supervised to match the geometric error of each prediction. The mismatch matters because these follow-up works use raw confidence to decide which predictions are retained, how strongly they influence optimization, and how the system state evolves over time, making it a pipeline-level rather than a model-level limitation. We address this gap by adopting two additional consistency signals to complement VGGT confidence to predict dense risk maps for more reliable geometric refinement.

2.2 Confidence and uncertainty estimation

Confidence and uncertainty have long been used in depth and geometric vision to identify reliable predictions. In stereo matching and multi-view stereo (MVS), confidence is estimated from matching ambiguity, photometric consistency, or geometric agreement across views, and learning-based methods train confidence networks that combine spatial coherence with multi-view consistency, underscoring the role of geometric agreement in depth reliability (Li et al., 2020; 2021).

Feed-forward reconstruction models learn confidence differently through aleatoric uncertainty-based objectives (Kendall & Gal, 2017). In this formulation, the predicted uncertainty modulates the regression loss, allowing the model to downweight inherently ambiguous regions such as textureless surfaces, occlusions, or specular reflections (Kendall & Gal, 2017; Klodt & Vedaldi, 2018). Following this work, VGGT learns confidence as part of their depth and point map regression losses (Wang et al., 2025). Such confidence is useful for training, but it is not directly optimized to rank the errors of depth or 3D point predictions. This gap becomes important when VGGT confidence is used for filtering, weighting, or alignment. We demonstrate that the predicted risk maps improve performance in downstream tasks compared to using confidence directly.

3 Method

Fig. 2 illustrates VERA, a VGGT error-aware dense risk prediction network that estimates where VGGT depth predictions are likely to fail. VERA is designed to combine learned confidence with geometric consistency cues, capture both local structure and cross-signal context, and adaptively determine which cues are reliable at each location. Through this error-aware fusion, VERA predicts a dense risk map of VGGT depth errors from depth confidence, point confidence, depth-to-point (D2P) residual, and multi-view (MV) geometric residual.

3.1 Input signals

Given N images $\{I_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$, VGGT outputs a depth map $D_i \in \mathbb{R}^{H\times W}$, a point map $P_i \in \mathbb{R}^{H\times W\times 3}$, camera intrinsics $K_i \in \mathbb{R}^{3\times 3}$, camera extrinsics $T_i \in \mathbb{R}^{3\times 4}$, and confidence maps $C_i^{\text{depth}}, C_i^{\text{point}} \in \mathbb{R}^{H\times W}$ for each view i. Following DUSt3R (Wang et al., 2024), VGGT learns confidence through an aleatoric uncertainty loss (Kendall & Gal, 2017).

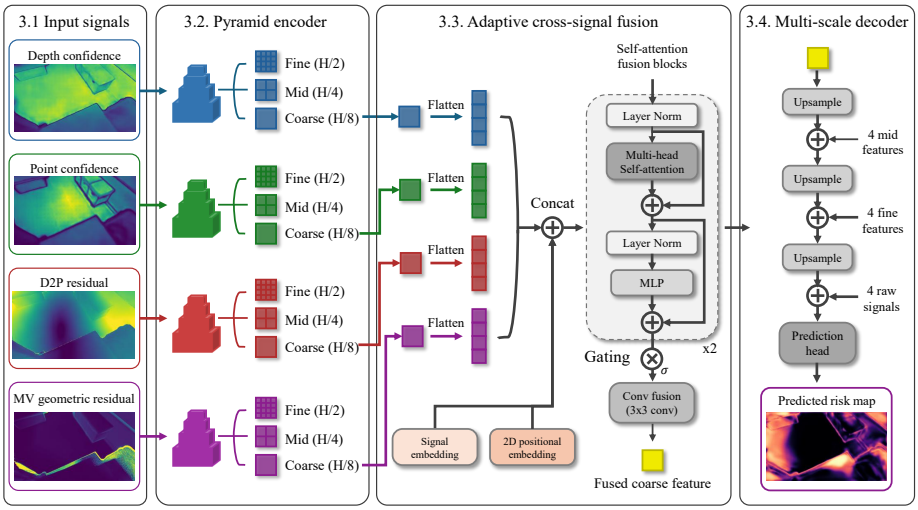


Figure 2: **Overview of VERA**, a VGGT error-aware dense risk prediction network. VERA fuses two confidence maps and two consistency signals with adaptive cross-signal fusion and a multi-scale decoder to predict a dense risk map of VGGT depth errors.

The two confidence maps provide the learned reliability estimates of the depth and point-map heads. Neither is directly supervised to rank actual depth errors, so confidence can remain high even in erroneous regions. To capture such failures, we complement them with two geometric consistency signals derived from VGGT outputs.

The depth-to-point map residual tests the internal self-consistency among VGGT's independently predicted depth, camera, and point map. From the predicted depth and camera parameters, we unproject each pixel into 3D and compare it with the point map directly predicted by VGGT. Let $T_i^{wc} = (R_i^{wc}, t_i^{wc})$ denote the camera-to-world transformation for view i. For a pixel (u, v), we compute $P_i^{\text{unproj}}(u,v) = R_i^{wc}(D(u,v)K^{-1}[u,v,1]^\top) + t_i^{wc}$ and $R^{\text{D2P}}(u,v) = \|P_i^{\text{VGGT}}(u,v) - P_i^{\text{unproj}}(u,v)\|_2$. A large depth-to-point residual indicates that the predicted depth, camera, and point map are mutually inconsistent, which can reveal errors missed by raw confidence.

The multi-view residual measures the consistency of VGGT depth predictions across neighboring views. For each target view i, we reproject the source depths from a local temporal window $\mathcal{S}_i = \{i-2, i-1, i+1, i+2\}$ and compute $R_i^{\text{MV}} = \text{mean}_{j\in\mathcal{S}_i} \left|D_{j\to i}^{\text{reproj}} - D_i\right|$. over valid reprojections. We evaluate multiple configurations and select the window of radius 2, which preserves predictive accuracy while reducing computational cost. Detailed results are provided in Appendix A. A large residual indicates cross-view geometric inconsistency, which can reveal errors that are difficult to detect from a single-view confidence map alone. However, the informativeness of the MV residual varies across regions with view overlap and scene geometry. Therefore, R^{MV} is useful but context-dependent, and should be interpreted jointly with the other signals.

Overall, the confidence maps encode VGGT's learned reliability, while the residuals encode internal and cross-view geometric consistency, providing complementary evidence for depth-error risk. Since these signals have different scales, distributions, and failure modes, we normalize each signal with signal-wise log normalization and spatial z-score normalization, producing two normalized channels per signal that encode both absolute magnitude and relative spatial structure.

3.2 Pyramid Encoder

Raw input signals provide per-pixel evidence of depth-error risk, but reliable risk prediction requires local context. VERA therefore encodes each signal with a signal-wise convolutional pyramid encoder (Lin et al., 2017), producing feature maps $F_\ell \in \mathbb{R}^{C\times H/2^\ell \times W/2^\ell}$ for $\ell \in \{1, 2, 3\}$ at fine, mid, and coarse resolutions. The coarse feature provides spatially correlated information for efficient cross-signal fusion, since depth errors often form region-level patterns rather than isolated pixel noise,

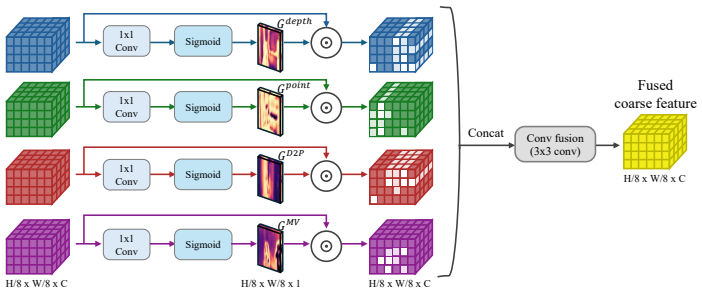


Figure 3: **Signal-wise spatial gating module.** The output features from the self-attention blocks are separated into four signal branches, where each branch predicts a spatial gate to modulate its feature map. The gated feature maps are then concatenated and fused to form the fused coarse feature F_{fused}.

while the fine and mid-level features are preserved as skip features so that the decoder can recover object boundaries, thin structures, and small erroneous regions. The four encoders share the same architecture but maintain separate weights to learn signal-specific reliability features.

3.3 Adaptive cross-signal fusion

Each input signal captures a different aspect of depth-error risk, but no single signal is reliable everywhere. To let each signal be interpreted in the context of the others, VERA performs cross-signal fusion at the coarse feature level, where local error evidence has already been spatially summarized. The coarse feature maps from the four signals are flattened and concatenated, each augmented by a learnable signal embedding and a 2D positional embedding (Dosovitskiy et al., 2021), then processed by $L{=}2$ self-attention blocks (Xiong et al., 2020) that exchange information both within and across signals.

VERA then estimates which signal should be trusted at each spatial location. The features are split back into four signal-specific feature maps and passed to a signal-wise spatial gating module (Fig. 3). For each signal $k \in \{\text{depth}, \text{point}, \text{D2P}, \text{MV}\}$, an independent $1{\times}1$ convolution followed by a sigmoid produces a spatial gate $G^k \in [0,1]^{H/8 \times W/8}$. Each gate modulates the corresponding feature map by element-wise multiplication. The gated features are then concatenated and fused by a $3{\times}3$ convolution to produce the fused coarse feature F_{fused}.

Because the gates are predicted after self-attention, they determine the reliability of each signal from contextual evidence at both the same signal and the others, rather than from its raw magnitude alone. VERA can therefore suppress a signal when it produces a misleading response and emphasize it when it provides useful evidence. Fig. 4 visualizes the learned gates on scenes from Replica and PointOdyssey (Straub et al., 2019; Zheng et al., 2023). In both examples the gates suppress residual signals that are high in low-error regions. For example, the D2P residual is suppressed in the left example, while the MV residual is suppressed in the right example. This shows that the gating module learns per-signal and per-location selectivity from the risk prediction objective, rather than relying on a hand-crafted threshold or a fixed signal preference.

3.4 Multi-scale decoder

The fused coarse feature F_{fused} is upsampled by a U-Net-style decoder (Ronneberger et al., 2015) to the input resolution. The decoder uses skip connections from the mid- and fine-scale encoder features of all four signals, allowing the coarse fused representation to recover local spatial details. A lightweight prediction head outputs a dense risk map $r \in [0,1]^{H \times W}$, where larger values indicate higher predicted risk.

3.5 Training loss

We train VERA with binary depth-error targets instead of directly regressing continuous depth errors. Empirically, regression or ranking objectives were unstable due to heavy-tailed raw signals and ambiguous low-to-medium error pixels. We therefore convert the per-pixel absolute depth error into

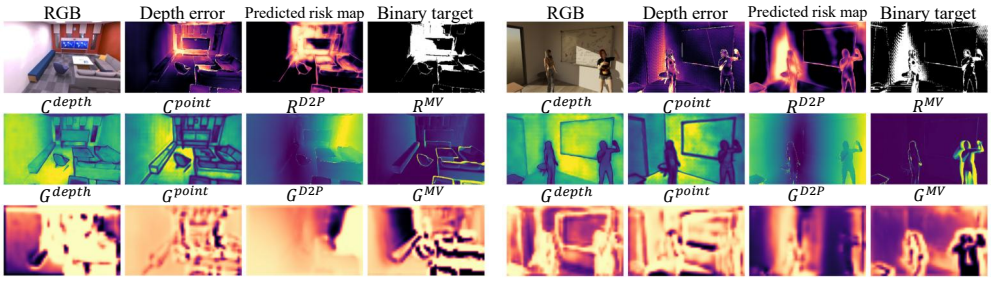


Figure 4: Gate visualization on sample scenes from Replica (left) and PointOdyssey (right). Top row: RGB image, depth error, predicted risk map, and binary target, where the top 20% of pixels with the largest depth error are white. Middle row: the four input signals. Bottom row: the learned gates, where brighter values correspond to larger gate values.

a binary target $y \in \{0,1\}^{H\times W}$ as $y(u,v) = \mathbf{1}\{e(u,v) \geq \tau\}$, where τ is set so that the top 20% highest-error pixels in each image are treated as risk positives.

VERA predicts a soft risk map $r \in [0,1]^{H\times W}$, where larger values indicate higher depth-error risk. We optimize it with a combination of focal loss (Lin et al., 2020) and soft Dice loss (Milletarì et al., 2016), $\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}(r,y) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(r,y)$. Focal loss downweights easy pixels and focuses learning on hard examples, which matters under the strong imbalance between risky and non-risky pixels, while soft Dice loss encourages overlap between the predicted risk region and the target high-error region, promoting spatially coherent risk maps. Detailed loss definitions and hyperparameters are given in Appendix B.

4 Experiments

4.1 Experimental settings

Datasets. Following VGGT's multi-dataset training setup, we train VERA on nine datasets covering object- and scene-level, indoor and outdoor, and synthetic and real-world settings: CO3Dv2 (Reizenstein et al., 2021), MVS-Synth (Huang et al., 2018), PointOdyssey (Zheng et al., 2023), ScanNet (Dai et al., 2017), Replica (Straub et al., 2019), Hypersim (Roberts et al., 2021), Virtual KITTI (Gaidon et al., 2016), KITTI (Geiger et al., 2013), and TUM RGB-D (Sturm et al., 2012). We evaluate on CO3Dv2, ScanNet, MVS-Synth, Hypersim, and PointOdyssey, which contain sufficient independent sequences for validation and test sets. All splits are sequence-disjoint, and each sample contains 5–30 randomly selected views. Results are averaged over five random repetitions per sequence.

Implementation Details. VERA is trained with the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 1×10^{-4} with cosine decay. The encoder channel width is set to C=32, and the number of self-attention fusion blocks is set to L=2. Our experiments are conducted on a single NVIDIA RTX 4090 GPU. Details are provided in the Appendix C.

Evaluation Metrics. We evaluate error ranking using the Area Under the Sparsification Error curve (AUSE) (Ilg et al., 2018; Poggi et al., 2020), which compares the sparsification curve obtained by removing pixels according to predicted risk with the oracle curve obtained using ground-truth errors. We evaluate high-error detection using AUROC and AUPR-Error (Hanley & McNeil, 1982; Zhu et al., 2022; Davis & Goadrich, 2006; Boyd et al., 2013), where the top 20% of pixels ranked by ground-truth depth error in each image are treated as positives. Lower AUSE and higher AUROC and AUPR-Error indicate better performance.

4.2 Depth error prediction

We evaluate VERA against four raw signal baselines, C^{depth}, C^{point}, R^{D2P}, and R^{MV}, and two simple learned baselines. The Simple MLP and Simple CNN receive the same four input signals, preprocessing, error supervision, loss, and risk target as VERA, isolating the effect of the proposed

Table 1: Depth error prediction on the five sequence-disjoint test sets. (a) Comparison with raw signals and simple learned baselines. Results are reported as mean ± standard deviation across the test sets. (b) Signal ablation using different input combinations for VERA. Confidence denotes both depth and point confidence maps. Consistency denotes both D2P and MV residuals.

(a) Comparison with baselines.

Method	AUSE ↓	AUROC ↑	AUPR-Error ↑
C^{depth}	0.272 ± 0.111	0.686 ± 0.089	0.362 ± 0.098
C^{point}	0.292 ± 0.142	0.685 ± 0.081	0.372 ± 0.090
R^{D2P}	0.367 ± 0.123	0.677 ± 0.067	0.386 ± 0.098
R^{MV}	0.603 ± 0.133	0.594 ± 0.060	0.263 ± 0.051
Simple MLP	0.243 ± 0.096	0.800 ± 0.060	0.544 ± 0.106
Simple CNN	0.226 ± 0.102	0.810 ± 0.057	0.555 ± 0.103
VERA	**0.190** ± 0.085	**0.882** ± 0.052	**0.700** ± 0.094

(b) Signal ablation.

Signals	AUSE ↓	AUROC ↑	AUPR-Error ↑
Confidence only	0.209	0.847	0.613
Consistency only	0.222	0.855	0.647
Confidence + D2P	0.207	0.865	0.665
VERA	**0.190**	**0.882**	**0.700**

Table 2: VGGT-Long odometry results on KITTI odometry sequences measured by ATE RMSE in meters. Bold and underline indicate the best and second best method for each column. Avg.* averages all sequences except sequence 01.

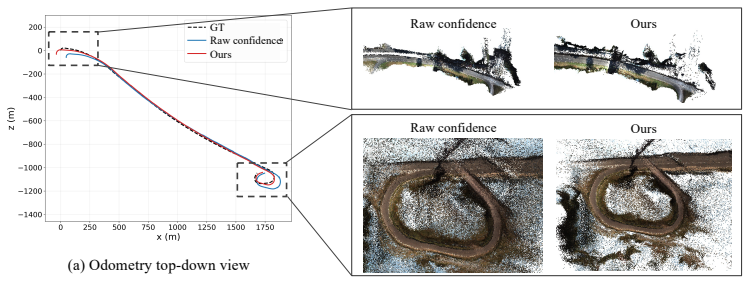
Method	00	01	02	03	04	05	06	07	08	09	10	Avg.	Avg.*
Uniform	11.44	69.66	<u>54.86</u>	9.13	**3.24**	10.87	7.21	2.93	**55.90**	33.92	23.58	25.70	21.31
C^{point}	<u>10.83</u>	<u>66.61</u>	55.80	**5.58**	<u>3.92</u>	<u>10.13</u>	<u>6.93</u>	<u>2.85</u>	57.52	<u>31.15</u>	<u>22.06</u>	<u>24.85</u>	<u>20.68</u>
Ours	**9.73**	**37.81**	**52.90**	<u>7.04</u>	4.58	**9.73**	**5.35**	**2.47**	<u>56.55</u>	**29.77**	**20.13**	**21.46**	**19.83**

architecture. The Simple MLP applies a stack of 1×1 convolutions to each pixel independently, whereas the Simple CNN uses four 3×3 convolutional layers to capture local spatial context. Table 1a shows that VERA achieves the best average performance on all three metrics. Relative to the best raw signal for each metric, VERA reduces AUSE from 0.272 to 0.190, improves AUROC from 0.686 to 0.882, and improves AUPR-Error from 0.386 to 0.700. Both simple learned baselines improve over raw depth confidence on all three metrics, but remain below VERA. Compared with the Simple CNN, VERA reduces AUSE from 0.226 to 0.190 and improves AUROC and AUPR-Error by 0.072 and 0.145, respectively. These results show that VERA outperforms simple predictors trained with the same input signals and error supervision. Per-dataset results are reported in Appendix D.

We further analyze the contribution of the input signals by comparing VERA variants trained with different signal combinations in Table 1b. The confidence-only and consistency-only variants outperform the raw-signal baselines, showing that the proposed architecture effectively extracts predictive information from each signal group. Combining the confidence maps with D2P signal further improves risk prediction, while VERA achieves the best overall performance by jointly using all four signals. Together with the MLP and CNN comparisons, these results demonstrate the effectiveness of both the proposed architecture and the use of complementary confidence and geometric consistency signals.

4.3 Odometry and geometric refinement

VGGT-Long Odometry. We evaluate whether VERA's risk map improves long-sequence reconstruction on KITTI when VERA has been trained wihtout KITTI dataset. We convert the predicted risk map $\hat{r}$ into a confidence map through temperature-scaled inversion, $\tilde{C} = \sigma(-\hat{r}/T)$ such that pixels with higher predicted risk receive lower weights. This confidence replaces raw VGGT point confidence in the confidence-weighted point filtering and chunk alignment of VGGT-Long (Deng et al., 2026). We compare uniform weighting, raw VGGT point confidence, and confidence derived from VERA using ATE RMSE on KITTI odometry sequences 00 to 10. As shown in Table 2, VERA reduces mean ATE RMSE from 24.85m with raw point confidence to 21.46m, a 13.7% reduction.



(a) Odometry top-down view

Figure 5: VGGT-Long odometry result on KITTI sequence 01. Left: top down trajectory comparison between ground truth, raw confidence weighting, and VERA refined confidence weighting. Right: point cloud visualizations at two trajectory regions, comparing the baseline with raw confidence and the result with refined confidence.

Table 3: Reconstruction quality after confidence-weighted pose refinement, averaged over eight Replica scenes. All metrics are measured in meters.

Weighting	Chamfer ↓	Accuracy ↓	Completeness ↓
Uniform	0.1825	0.1794	0.1856
C^{point}	0.2023	0.2033	0.2013
VERA	**0.1686**	**0.1663**	**0.1710**

VERA also reduces the mean ATE RMSE from 20.68m to 19.83m, a 4.2% reduction, when excluding sequence 01. These results show that VERA improves average odometry accuracy over both uniform weighting and raw point confidence, despite being trained without KITTI data. Fig. 5 shows a qualitative result on sequence 01, where VERA produces a trajectory closer to the ground truth and cleaner local geometry. This supports the quantitative results, showing that better error-aware weighting reduces drift and improves geometric consistency over long sequences.

Confidence-Weighted Pose Refinement. We evaluate whether the confidence derived from VERA improves geometric consistency when used to weight pose refinement. Camera poses are optimized using weighted geometric residuals while the predicted depth maps remain fixed. We compare uniform weighting, raw VGGT point confidence, and VERA-derived confidence on eight Replica scenes using Chamfer distance, accuracy, and completeness. Table 3 shows that VERA achieves the best average performance across all three metrics. Compared with raw point confidence, VERA reduces the average Chamfer, accuracy, and completeness distances by 16.6%, 18.2%, and 15.0%, respectively. VERA also achieves the lowest Chamfer and completeness distances in every scene, demonstrating consistent improvements in the resulting geometry. Fig. 6 presents qualitative comparisons on room0 and room2 scenes. Raw point confidence can assign large weights to inaccurate regions, resulting in noisy or inconsistent geometry, whereas VERA suppresses these regions and produces cleaner, more coherent reconstructions. These results demonstrate that VERA improves not only depth-error prediction but also the geometry obtained after pose refinement. Per-scene quantitative results and additional qualitative pose-refinement results on other datasets are provided in Appendix F.

4.4 Signal contribution analysis

Table 1b shows that the four signals are jointly useful on average. To analyze how each input signal contributes to the final prediction, we remove one input signal at a time and visualize the change in the predicted risk map. Fig. 7 presents two examples. In the first example, C^{depth} mainly supports high-risk detection, while C^{point} helps preserve low-risk regions and suppress false positives. In the second example, the residuals R^{D2P} and R^{MV} play a more prominent role, indicating that internal and cross-view geometric inconsistencies are important cues for predicting depth-error risk. The contribution of each signal is thus spatially localized, and VERA adaptively combines confidence and residuals according to the local failure mode.

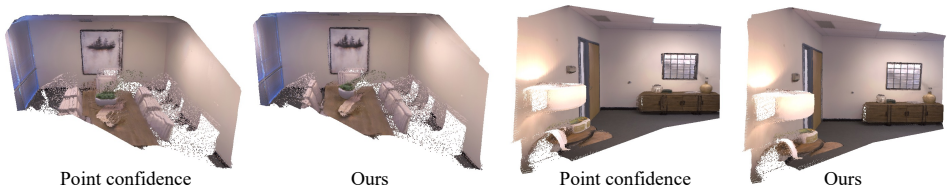


Figure 6: Confidence-weighted pose refinement on Replica. Reconstruction results refined using raw VGGT point confidence and VERA-derived confidence.

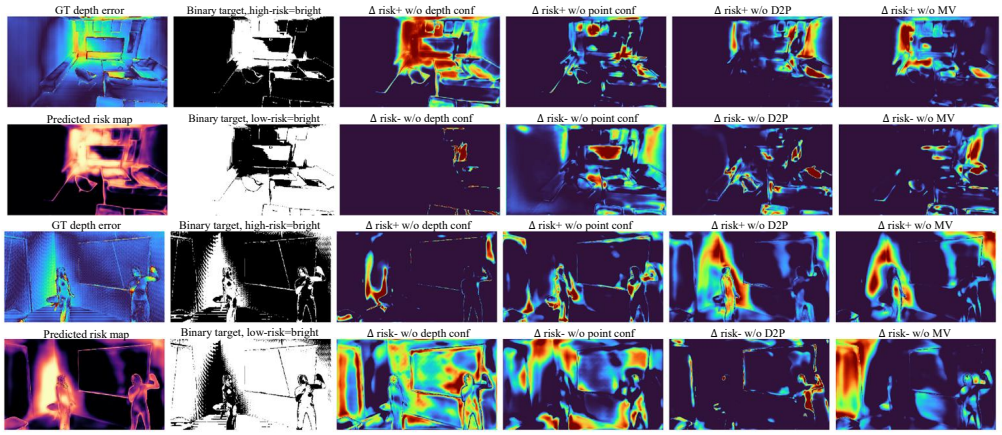


Figure 7: Signal contribution analysis on two datasets, each occupying two rows. The first two columns show the depth-error map, the predicted risk map, and the binary targets, where bright pixels indicate high-risk regions in the first row and low-risk regions in the second. The remaining columns show leave-one-signal-out changes in the predicted risk map. The Δrisk^{+} and Δrisk^{-} maps highlight where the removed signal contributes to high-risk and low-risk prediction, respectively.

5 Limitations and Discussion

VERA currently focuses on VGGT, so its input design is tied to VGGT outputs. Extending it to other feed-forward 3D foundation models and training it in an unsupervised manner are important future directions. Error detection is also limited by the information contained in the four input signals. Errors that remain confident and geometrically consistent may not be detected. Learning-based adaptation settings such as TTT3R (Chen et al., 2026) remain to be explored. More broadly, enhanced VGGT confidence can improve the reliability of downstream 3D reconstruction tasks, but it should be carefully validated before deployment in safety-critical applications.

6 Conclusion

We introduced VERA, a VGGT error-aware dense risk prediction network that addresses overconfident failures in raw VGGT confidence. VERA predicts dense depth-error risk by combining depth confidence, point confidence, depth-to-point consistency, and multi-view geometric consistency. The resulting risk map provides a more reliable estimate for point filtering and optimization weighting. Trained on nine datasets and evaluated on five sequence-disjoint test sets, VERA substantially improves depth-error prediction over individual raw signals and simple neural network baselines. We further demonstrate that the predicted risk improves VGGT-Long odometry and confidence-weighted pose refinement. These results highlight the importance of reassessing confidence from feed-forward 3D foundation models before using it as a reliability signal in downstream geometric pipelines.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with code implementation and improving the readability of the manuscript. We have not used generative AI tools for developing the research ideas, the methodology, and the experiments, and data generation are not applicable to this work. Additionally, we used generative AI tools for grammar checking. We have reviewed all AI-assisted work. We reviewed all AI-assisted text to ensure that it preserved the intended meaning, and verified all AI-generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training objectives, input-signal construction, evaluation metrics, and implementation details in the main paper and appendices. Additional per-dataset and per-scene results, hyperparameter settings, and downstream evaluation protocols are also included in the appendices. Codes will be released upon publication.

REFERENCES

Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III*, pp. 451—-466, 2013.

Xingyu Chen, YUE CHEN, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, pp. 50694–50718, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233—-240, 2006.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences. In *ICRA*, 2026.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 337–33712, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4340–4349, 2016.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, pp. 1231–1237, 2013.

Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, pp. 29–36, 1982.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830, 2018.

Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, pp. 677—-693, 2018.

Jinwoo Jeon, Dong-Uk Seo, Eungchang Mason Lee, and Hyun Myung. Aim-slam: Dense monocular slam via adaptive and informative multi-view keyframe prioritization with foundation model. In *ICRA*, 2026.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5580–5590, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Maria Klodt and Andrea Vedaldi. Supervising the new with the old: Learning sfm from sfm. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, pp. 713—-728, 2018.

Zhaoxin Li, Wangmeng Zuo, Zhaoqi Wang, and Lei Zhang. Confidence-based large-scale dense multi-view stereo. *IEEE Transactions on Image Processing*, pp. 7176–7191, 2020.

Zhaoxin Li, Xiaoge Zhang, Kangkan Wang, Hao Jiang, and Zhaoqi Wang. High accuracy and geometry-consistent confidence prediction network for multi-view stereo. *Computers & Graphics*, pp. 148–159, 2021.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 318–327, 2020.

Yang Liu, Chuanchen Luo, Zimo Tang, Junran Peng, and Zhaoxiang Zhang. Vggt-x: When vggt meets dense novel view synthesis. *arXiv preprint arXiv:2509.25191*, 2025.

Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. Lora3d: Low-rank self-calibration of 3d geometric foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3224–3234, 2020.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902, 2021.

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

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng Ren, Shobhit Verma, Anton Clarkson, Ming Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesús Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke Strasdat, Renzo De Nardi, Michael Goesele, S. Lovegrove, and Richard A. Newcombe. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580, 2012.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5283–5293, 2025.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. *Vision Algorithms: Theory and Practice*, pp. 298–372, 2000.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pp. 518—-536, 2022.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *ICLR*, 2026.

A COMPUTATIONAL COST

We evaluate the computational overhead of VERA on a CO3D scene with 30 input views. Runtime is averaged over 10 runs after excluding warm-up iterations. Both the depth-to-point (D2P) and multi-view (MV) residuals are computed using batched GPU implementations. Table 4 compares different configurations for computing the MV residual. Because the performance of depth error prediction is comparable, we select the radius-2 window which reduces the MV runtime from $0.0867\,\mathrm{s}$ to $0.0424\,\mathrm{s}$ compared with the full-resolution all-pairs configuration. For the whole pipeline, VGGT inference requires $0.8277\,\mathrm{s}$, while the D2P and MV residual computations require $0.0144\,\mathrm{s}$ and $0.0424\,\mathrm{s}$, respectively. The resulting end-to-end runtime is $0.8845\,\mathrm{s}$, corresponding to $1.07\times$ the runtime of VGGT inference alone. Thus, VERA introduces only 6.9% runtime overhead over the VGGT backbone.

Table 4: Comparison of multi-view residual configurations. MV runtime measures the standalone multi-view consistency residual computation. Depth error prediction metrics are averaged over the five test sets.

MV configuration	MV runtime (s)	AUSE / AUROC / AUPR-Error
All pairs, full resolution	0.0867	0.189 / 0.880 / 0.696
All pairs, 1/2 resolution	0.0403	0.192 / 0.881 / 0.698
All pairs, 1/4 resolution	0.0431	0.188 / 0.879 / 0.694
Radius 2, full resolution (VERA)	**0.0424**	**0.190 / 0.882 / 0.700**
Radius 3, full resolution	0.0441	0.190 / 0.882 / 0.697

B TRAINING DETAILS

Table 5 summarizes the training hyperparameters used for VERA. We train VERA as a dense binary risk prediction model, where high-risk pixels are defined by the per-image ranking of absolute depth errors. For each pixel (u, v), we compute the absolute depth error as

$$e(u, v) = |D(u, v) - D^{\mathrm{gt}}(u, v)|. \tag{1}$$

The binary risk target is defined as

$$y(u, v) = \mathbf{1}\{e(u, v) \geq \tau\}, \tag{2}$$

where τ is chosen such that the top $\rho = 20\%$ valid pixels with the largest absolute depth errors are treated as positives. The network predicts a soft risk map $r \in [0, 1]^{H \times W}$, where larger values indicate higher predicted depth error risk.

Table 5: Training hyperparameters used for the signal-token fusion model.

Category	Hyperparameter	Value
Optimization	Optimizer	Adam
	Learning rate	1×10^{-4}
	Epochs	300
	LR schedule	Cosine decay
Risk target	Error type	Absolute depth error
	Positive ratio ρ	0.2
Loss	Main loss	Focal loss
	Auxiliary loss	Soft Dice loss
	Focal α	0.75
	Focal γ	2.0
	Dice weight λ_{dice}	0.05
	Dice smooth s	1.0

C MODEL STRUCTURE AND PARAMETERS

As shown in Table 6, the proposed model receives two confidence maps and two consistency residuals as input. Signal-specific encoders first extract features from each signal independently, preserving the individual signal statistics before fusion. Cross-signal interaction is then performed at the coarse token level using lightweight self-attention, and the fused representation is decoded with multi-scale features to produce full-resolution risk logits. Table 7 summarizes the trainable parameters of the proposed model. The model contains only 0.62M trainable parameters, with most parameters allocated to the signal-specific encoders, transformer fusion blocks, and multi-scale decoder. The spatial signal gates add only negligible overhead.

Let $r(u, v) \in [0, 1]$ denote the predicted probability that pixel (u, v) is risky, and let $y(u, v) \in \{0, 1\}$ denote the corresponding binary high-error target. We optimize the predicted risk map using a combination of focal and soft Dice losses. To address the imbalance between risky and non-risky pixels, we use a masked focal loss (Lin et al., 2020):

$$\begin{aligned}\mathcal{L}_{\text{focal}} = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} &\Big[\alpha y(u,v) + (1-\alpha)(1-y(u,v))\Big] \\ &\cdot \Big[1 - \big\{y(u,v)r(u,v) + (1-y(u,v))(1-r(u,v))\big\}\Big]^{\gamma} \\ &\cdot \text{BCE}(r(u,v), y(u,v))\,,\end{aligned} \tag{3}$$

where Ω is the set of valid pixels, α controls class balancing, and γ is the focusing parameter.

We additionally use a soft Dice loss (Milletarì et al., 2016) to encourage spatial overlap between the predicted risk map and the binary high-error target:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{(u,v)\in\Omega} r(u,v)y(u,v) + s}{\sum_{(u,v)\in\Omega} r(u,v) + \sum_{(u,v)\in\Omega} y(u,v) + s}, \tag{4}$$

where s is a smoothing constant that ensures numerical stability.

Table 6: Model architecture.

Stage	Feature size
Input signals	$4 \times (B, 1, H, W)$
Preprocessing	$4 \times (B, 2, H, W)$
Pyramid encoder	$4 \times \{(B, C, H/2^{\ell}, W/2^{\ell})\}_{\ell=1}^{3}$
Coarse tokenization	$(B, 4H_cW_c, D)$
Token fusion	$(B, 4H_cW_c, D)$
Spatial gating	$4 \times (B, C, H_c, W_c)$
Decoder	$(B, 1, H, W)$

Table 7: Trainable parameter.

Stage	Component	Params
Pyramid encoder	Depth-confidence encoder	74.6K
	Point-confidence encoder	74.6K
	Depth-to-point residual encoder	74.6K
	Multi-view residual encoder	74.6K
Adaptive cross-signal fusion	Signal embedding	0.4K
	Input token projection	3.2K
	Transformer fusion blocks	149.6K
	Output token projection	3.1K
	Spatial signal gates	0.1K
	Coarse convolutional fusion	46.1K
Multi-scale decoder	Multi-scale decoder	122.4K
Total	**All trainable components**	**0.62M**

D DETAILED EXPERIMENTAL RESULTS

Table 8 presents the per-dataset results. VERA substantially improves AUROC and AUPR-Error across all five test datasets while achieving the best AUSE on three of them. On CO3Dv2 and MVS-Synth, point confidence and depth confidence, respectively, yield slightly lower AUSE, whereas VERA maintains comparable AUSE while substantially improving both AUROC and AUPR-Error. This indicates that VERA largely preserves the overall error-ranking ability while significantly improving the identification of high-error regions. These results demonstrate that combining confidence and residual signals enables VERA to compensate for the limitations of individual signals, providing more robust and discriminative depth-error risk estimates across diverse datasets.

Table 8: Per-dataset depth error prediction results on the five test sets, reported as AUSE / AUROC / AUPR-Error. Lower is better for AUSE, higher is better for AUROC and AUPR-Error. Best per dataset and metric in bold.

Dataset	VERA	C^{depth}	C^{point}	R^{D2P}	R^{MV}
CO3Dv2	**0.159 / 0.890 / 0.734**	0.254 / 0.689 / 0.348	0.149 / 0.757 / 0.483	0.296 / 0.650 / 0.358	0.407 / 0.620 / 0.286
ScanNet	**0.335 / 0.784 / 0.519**	0.389 / 0.561 / 0.226	0.341 / 0.617 / 0.270	0.474 / 0.609 / 0.297	0.627 / 0.546 / 0.222
MVS-Synth	0.072 / **0.932 / 0.781**	**0.069** / 0.834 / 0.527	0.107 / 0.807 / 0.478	0.158 / 0.804 / 0.574	0.819 / 0.695 / 0.344
Hypersim	**0.185 / 0.921 / 0.762**	0.344 / 0.643 / 0.327	0.384 / 0.622 / 0.319	0.459 / 0.672 / 0.324	0.613 / 0.524 / 0.198
PointOdyssey	**0.200 / 0.880 / 0.702**	0.304 / 0.700 / 0.384	0.480 / 0.623 / 0.311	0.448 / 0.652 / 0.376	0.549 / 0.588 / 0.266
Average	**0.190 / 0.882 / 0.700**	0.272 / 0.686 / 0.362	0.292 / 0.685 / 0.372	0.367 / 0.677 / 0.386	0.603 / 0.594 / 0.263

E ADDITIONAL EXPERIMENTS

Generalization to an unseen dataset. To evaluate cross-dataset generalization, we exclude KITTI entirely from training, train on the remaining eight datasets, and evaluate on all 155 KITTI windows. We select KITTI because the remaining training datasets contain no real-world outdoor driving data. As shown in Table 9, VERA outperforms every individual raw signal even in an unseen outdoor domain. This provides quantitative evidence of cross-dataset generalization to an unseen real-world driving domain.

Table 9: Depth error prediction on KITTI when KITTI is excluded from training.

Signal	AUSE ↓	AUROC ↑	AUPR-Error ↑
C^{depth}	0.234	0.762	0.411
C^{point}	0.217	0.793	0.519
R^{D2P}	0.295	0.764	0.495
R^{MV}	0.433	0.702	0.337
VERA, KITTI unseen	**0.154**	**0.873**	**0.645**

F ADDITIONAL DOWNSTREAM RESULTS

Table 10 reports the per-scene results summarized in Table 3. VERA maintains stable reconstruction quality in cases where raw point confidence assigns unreliable weights and substantially degrades the refined geometry. Figure 8 further presents qualitative pose-refinement results on additional datasets, where VERA similarly produces cleaner and more coherent geometry than raw VGGT confidence.

Table 10: Per-scene reconstruction quality after confidence-weighted pose refinement on Replica, reported as Chamfer / Accuracy / Completeness. All metrics are measured in meters and lower is better.

Scene	Uniform	C^{point}	VERA
office0	0.2197 / 0.2208 / 0.2185	0.2195 / 0.2193 / 0.2197	**0.1954 / 0.1927 / 0.1980**
office1	0.2011 / 0.1898 / 0.2124	0.2035 / 0.1925 / 0.2146	**0.1872 / 0.1738 / 0.2006**
office2	0.1988 / 0.2021 / 0.1954	0.2008 / 0.2046 / 0.1970	**0.1866 / 0.1907 / 0.1826**
office3	0.2268 / 0.2327 / 0.2210	0.2339 / 0.2421 / 0.2256	**0.2076 / 0.2026 / 0.2126**
office4	0.1708 / 0.1804 / 0.1611	0.1669 / 0.1701 / 0.1637	**0.1551 / 0.1537 / 0.1565**
room0	0.1612 / 0.1440 / 0.1784	0.2214 / 0.2204 / 0.2225	**0.1518 / 0.1412 / 0.1624**
room1	0.1484 / **0.1398** / 0.1569	0.1471 / 0.1415 / 0.1526	**0.1390** / 0.1494 / **0.1286**
room2	0.1334 / **0.1256** / 0.1412	0.2250 / 0.2356 / 0.2144	**0.1264** / 0.1260 / **0.1269**
Average	0.1825 / 0.1794 / 0.1856	0.2023 / 0.2033 / 0.2013	**0.1686 / 0.1663 / 0.1710**

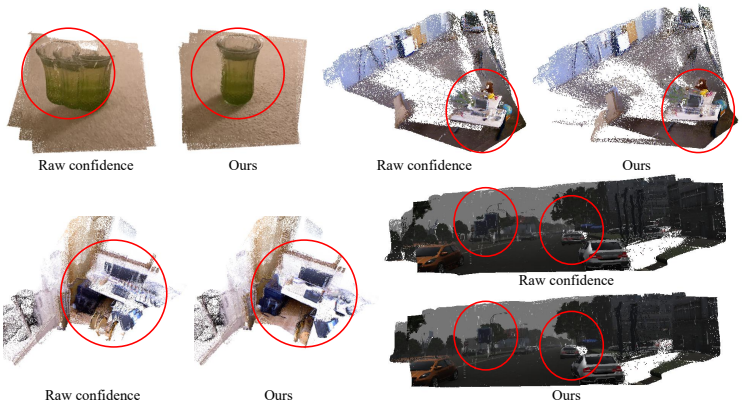


Figure 8: Additional qualitative comparisons of confidence-weighted pose refinement across datasets. Each example compares reconstructions refined using raw VGGT confidence and the VERA risk map. Red circles highlight regions where raw confidence leads to noisy or inconsistent geometry, while the risk map produces cleaner and more coherent reconstructions.