Parameter-Efficient & Distortion-Aware Self-Supervised 360° Depth Estimation with VMamba, DCNv4, and BiTS

Ian, Hoi Yin Wong (MASc) | Supervisor: Dr. Xianke Lin | Department of Automotive and Mechatronics Engineering | Ontario Tech University

Scan for Interactive 3D

Motivation & Problem

Monocular 360° (omnidirectional) depth estimation holds strong potential for autonomous driving, robotics, and virtual reality, offering holistic spatial awareness without the cost, power, and calibration overhead of LiDAR or multi-camera rigs. Yet 360° depth remains far less explored than perspective estimation. While early methods relied on supervised training in simulation, the sim-to-real gap limits real-world generalization. Learning directly from real-world scenes captures authentic physical patterns, but dense ground-truth labeling is prohibitively expensive—making self-supervised learning (SSL) via view synthesis essential.

Despite its scalability, **self-supervised 360° depth estimation** is bottlenecked by **weak, ambiguous, and distorted supervisory signals.** Conventional methods compute photometric errors independently across adjacent frames and rely on hard minimum-reprojection heuristics. By rigidly selecting only one adjacent frame (I_{t-1} or I_{t+1}) with the lowest loss, this winner-take-all strategy discards complementary geometric constraints, causing mapping collisions, boundary tearing, and structural voids while ignoring symmetric motion dynamics.

Architecturally, equirectangular panoramas expose a trade-off between **global receptive fields** and **computational efficiency**. Standard CNNs rely on localized receptive fields and rigid grid sampling that fail under severe polar distortion, while Vision Transformers incur prohibitive $O(N^2)$ complexity.

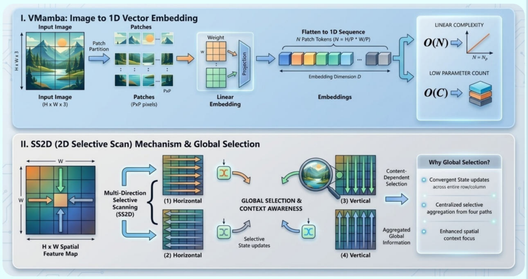
To overcome both bottlenecks with **zero runtime overhead**, we pair **Bilateral Temporal Softmasking (BiTS)**—a temperature-calibrated softmax weighting mechanism enforcing symmetric multi-frame supervision—with a **linear-complexity O(N) VMamba (SS2D)** encoder and **Deformable ConvNet v4 (DCNv4)** decoder for distortion-aware global geometry recovery. On our sequential 360° benchmark, our framework **closes 69.37% of the SL–SSL δ_1 accuracy gap** (lifting δ_1 from 90.51% to 94.79% toward 96.68%) while reconstructing sharp, orthogonal 3D room structures.

Architectural Design & Key Insights

VMamba Encoder (SS2D) | [Global Panoramic Context]

Mechanism: 2D Selective Scan (SS2D)

- **Patch Tokenization & 4-Way Serialization:** A 4×4 patch stem converts the 2D ERP image into N=H/4×W/4 tokens, unfolded along four directional paths (forward/backward, row/column) to preserve 2D panoramic spatial continuity.
- **Linear-Complexity Selective State Space:** Compresses global panoramic history into a compact 16-dimensional latent state at every step, achieving O(N) linear complexity.



360° Advantage & Impact: Global Distortion-Aware Encoding

- **Efficiency at High ERP Resolution:** At 512×1024 resolution (N=32,768 tokens), VMamba uses only 22.14M parameters—20.45% fewer than ConvNeXt (27.84M) and 19.64% fewer than Swin-T (27.55M).
- **Global Receptive Field for ERP Distortion:** VMamba links warped polar structures (±90°) directly to undistorted equatorial cues (0°) in a single pass, overcoming the spatial locality of ConvNeXt kernels and Swin-T windows.
- **Resolving SSL Reprojection Ambiguity:** Full-sphere context overcomes polar shape ambiguity and low-texture regions during self-supervised training, learning consistent 3D geometry across the entire panorama.

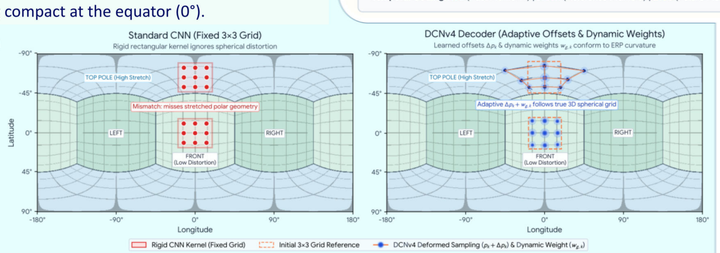
Deformable ConvNet (DCN) Decoder | [Spherical Distortion Adaptation]

Mechanism: Learned Sampling Offsets (Δp_k)

- **Spatial Offset Adaptation (Δp_k):** DCNv4 replaces rigid 3×3 CNN grids with continuous 2D offsets (Δp_k) that stretch laterally near poles (±90°) to match spherical curvature while remaining compact at the equator (0°).

360° Advantage & Impact: Lightweight Distortion-Aware Reconstruction

- **85% Decoder Parameter Reduction:** Decoder parameters drop by 85.0% (17.52M → 2.62M), cutting total parameters by 37.7% (39.77M → 24.77M), and compute by 38.5% (82.10 → 50.53 GFLOPs).
- **Lossless Fidelity (δ_1) & Clean Boundaries:** Adaptive sampling preserves full depth accuracy against the heavy CNN baseline (94.11% vs. 94.17% δ_1) while eliminating polar stretching artifacts.



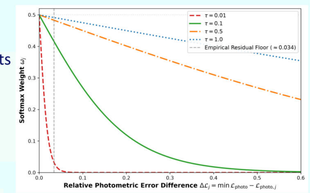
Bilateral Temporal Softmasking (BiTS) | [Training Stability & Guidance]

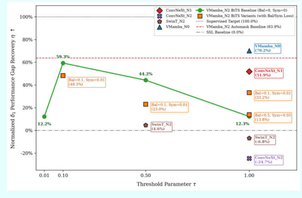
Mechanism: Temperature-Scaled Temporal Softmax Weighting (w_j)

- **Soft-Weighted Temporal Aggregation:** Replaces hard minimum selection (Automasking) with temperature-scaled softmax weights (w_j) computed from relative photometric differences ($\Delta L_j = \min L_{photo} - L_{photo,j}$) across adjacent frames ($j \in \{r-1, r+1\}$).
- **Balanced 50–50 Statistical Mean Signal:** Forms a balanced 50–50 split across similar adjacent frames—stabilizing gradients on textureless indoor surfaces (walls, floors, ceilings) without an auxiliary smoothness loss while suppressing occlusions.

360° Advantage & Impact: Tunable Stability & SL–SSL Gap Closure

- **Differentiable Occlusion & Edge Sensitivity (τ):** Lower τ sharpens weight transitions along depth edges to match Automasking selectivity while preserving smooth, differentiable gradients for stable training.
- **59% SL–SSL Performance Gap Closure:** BiTS boosts CNN self-supervised δ_1 accuracy from 90.51% to 94.17% (vs. 96.88% SL upper bound), closing up to 59.32% of the performance gap with zero inference overhead.
- **Cross-Backbone Stability:** Unlike hard Automasking which degrades on ConvNeXt and Swin Transformer despite reaching 94.45% δ_1 on VMamba. BiTS generalizes reliably across all three backbones with predictable τ-tuning behavior.

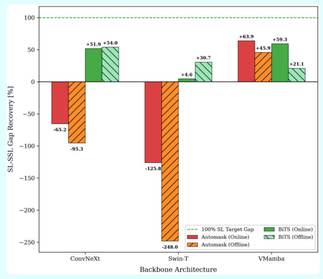


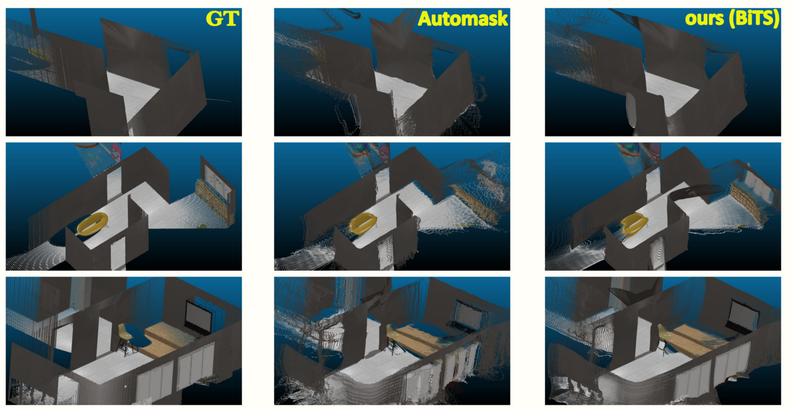


Experimental Validation & Quantitative Results

BiTS Cross-Backbone Validation | [Quantitative Benchmarks]

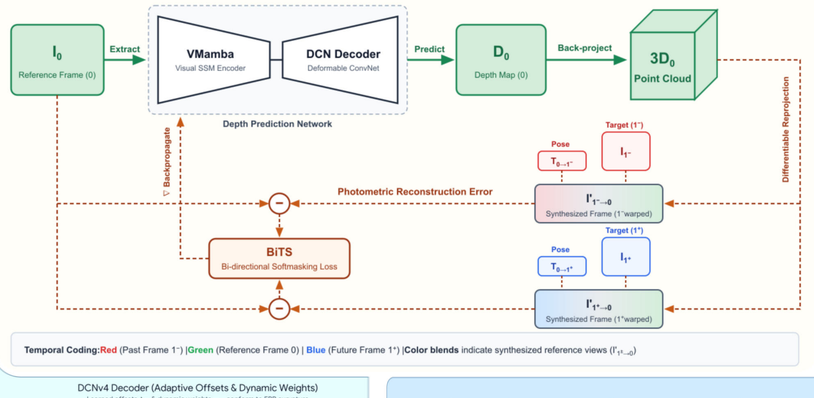
- **Universal Positive Gap Recovery:** BiTS recovers the SL–SSL δ_1 gap across every backbone—achieving +51.9% / +54.0% (online/offline) on ConvNeXt (88.93%→92.27% / 93.67%→95.76%), +4.6% / +30.7% on Swin-T (90.44%→90.71% / 94.72%→95.82%), and +59.3% / +21.1% on VMamba (90.51%→94.17% / 96.07%→96.53%).
- **Implicit Global Loss-Distribution Awareness:** Hard binary Automasking collapses locally restricted backbones (−65.2% / −95.3% on ConvNeXt; −125.8% / −248.0% on Swin-T), whereas BiTS leverages image-wide error statistics to implicitly encode the global loss distribution and preserve smooth gradients across all pixels.
- **Global State-Space Momentum:** VMamba's global receptive field propagates panoramic context across the entire 360° frame, providing sufficient optimization momentum to escape hard-masking local minima while BiTS further lowers offline AbsRel from 4.93% to 3.85%.





Framework Pipeline: Self-Supervised 360° Depth Estimation

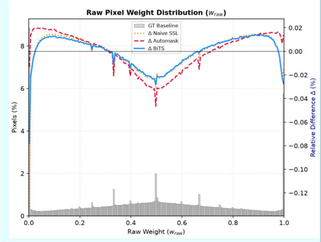
VMamba Encoder with Deformable Convolution Decoder & BiTS Loss — Inference Flow (Feedforward) — Training Flow (Loss & Gradient)

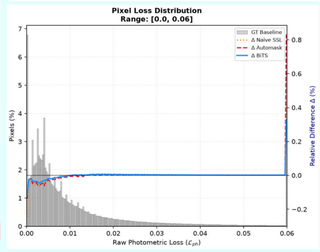


Temporal Coding: **Red** (Past Frame t⁻) | **Green** (Reference Frame 0) | **Blue** (Future Frame t⁺) | **Color blends** indicate synthesized reference views ($I'_{t^+\to 0}$)

BiTS Statistical Weighting Dynamics | [Loss Distribution Analysis]

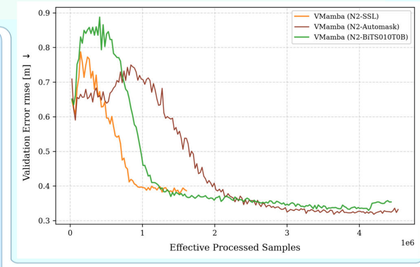
- **Natural Mid-Range Weight Alignment (w_{raw}):** In the Raw Pixel Weight Distribution, ground-truth weights naturally center around $w_{raw} \approx 0.50$, where BiTS achieves lower relative-difference deviation from the GT baseline than Automasking (σ=0.0076 vs. 0.0084)
- **Low-Loss Pixel Utilization ($L_{ph} \leq 0.01$):** In the Pixel Loss Distribution, roughly 64% of ground-truth pixels exhibit low photometric loss (≤0.01). BiTS closely tracks this GT profile via smooth weight calibration, whereas rigid 0/1 Automasking induces training fluctuations.
- **High-Loss Outlier Suppression ($L_{ph} \geq 0.06$):** At the upper tail of the Pixel Loss Distribution (≥0.06), Automasking produces >0.8% more high-error outlier pixels than the GT baseline, while BiTS cuts this excess to roughly 0.3% by smoothly down-weighting corrupted regions.





BiTS Training Dynamics & Convergence Stability | [Optimization Trajectories]

- **Calibrated Domain-Adaptation Phase:** During early training (<0.5M samples), BiTS moderates initial convergence as its image-wise statistics adapt to the panoramic loss landscape, avoiding the rigid binary-masking shortcut impact (0.65 - 0.75m RMSE plateau) that trap locally restricted backbones.
- **Sustained Late-Stage Convergence Momentum:** Once optimization enters the flat late-stage loss plane (L_{val} ≈0.034), raw photometric gradients weaken and cause baseline SSL to oscillate and stall early (1.37M samples), whereas BiTS's relative error weighting synthesizes steady directional gradients guiding the model smoothly down to 0.33m RMSE (94.17% δ_1 at 4.13M samples).
- **Lowest Training Loss Variance (σ_{train}):** BiTS achieves the lowest training loss variance across all methods ($\sigma_{train} = 0.49\times10^{-4}$), reducing optimization noise by 56.3% relative to baseline SSL (1.12×10^{-4}) and 14.0% relative to Automasking (0.57×10^{-4}).



Qualitative 360° Comparisons & Key Takeaways

- **Textureless Wall Stability (Row 1):** On low-texture walls, Automasking suffers from noisy signal fluctuations and overestimates wall-to-ceiling corner depth, whereas BiTS's tempered weighting suppresses outlier noise to produce smooth, accurate planar surfaces.
- **Edge Sensitivity & Temperature (τ) Trade-Off (Rows 2–3):** High edge sensitivity allows Automasking to capture intricate structures like the curved sofa (Row 2) and thin chair legs (Row 3), while tuning BiTS's temperature parameter (τ) progressively sharpens structural edges—such as doorframes—toward Automasking's boundary fidelity without losing planar smoothness.

[RGB | Ground Truth | VMamba + CNN | VMamba + DCNv4 (Automask) | Ours: VMamba + DCNv4 (BiTS)]

