RASE-Net: Residual Attention and Saliency Enhancement for Structure-Aware Monocular Depth Estimation

Bingxin Luo[1], Author Name[2] and Author Name[1,2]

[1] Department One, Institution One, City One, Country One
[2] Department Two, Institution Two, City Two, Country Two

E-mail: xxx@xxx.xx

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Monocular depth estimation remains fundamental to autonomous visual systems, yet structural fidelity under saliency-driven occlusions and boundary clutter poses persistent challenges. Traditional models often degrade near semantic transitions and under geometric variation, limiting their deployment in safety-critical applications. We introduce RASE-Net, a structure-aware refinement backbone that integrates residual saliency reweighting and foreground-guided modulation into a transformer decoder, enhancing edge continuity and obstacle awareness without modifying the base architecture or increasing computational demand. Extensive evaluations on NYU-v2, KITTI, and COCO2017 demonstrate consistent improvements across structure-sensitive metrics, maintaining depth coherence under cluttered and occlusion-rich conditions. These results position RASE-Net as a geometry-aware refinement module suitable for structure-critical depth estimation tasks.

Keywords: depth estimation, structure-aware refinement, residual attention, saliency enhancement

1. Introduction

Monocular depth estimation (MDE) is a core task in computer vision with wide-ranging applications in robotics, autonomous navigation, and scene understanding [1-3]. Although recent transformer-based architectures [4-6] have improved global scene representation, a key challenge remains: achieving consistent depth prediction while preserving geometric details at object boundaries and occlusions. Current approaches can be grouped into two categories. Transformer-based models [7] provide strong global context through long-range attention but often produce over-smoothed predictions in regions with complex structures, where fine-grained transitions are critical. Lightweight CNN models [8,9], while computationally efficient, often struggle to preserve spatial precision, especially near edges and high-gradient areas.

To examine these limitations, we adopt Depth Anything v2 (DAN-v2) [10] and FastDepth [11] as representative baselines for transformer-based and CNN-based architectures, respectively. DAN-v2, trained on diverse visual domains, delivers competitive zero-shot generalization but often lacks structural fidelity, particularly in saliency-dense or exposure-challenging scenarios. FastDepth, while achieving strong RMSE scores through lightweight convolutional design, struggles to maintain sharp geometric transitions and boundary integrity under complex spatial variations. Figure 1 summarizes these observations. In Figure 1(a), we compare DAN-v2, FastDepth, and our proposed method, RASE-Net, across five structure-aware metrics — Boundary Accuracy, Foreground δ_1, Edge-MAE, Grad-MAE, and RMSE — on NYU-v2, KITTI, and COCO2017. The persistent gaps, especially on boundary-sensitive metrics, indicate that structure-level errors persist across both architectural

paradigms. Figure 1(b) further illustrates qualitative comparisons under challenging lighting conditions: DAN-v2 predictions exhibit contour diffusion and foreground-background blending, FastDepth suffers from smoothing and

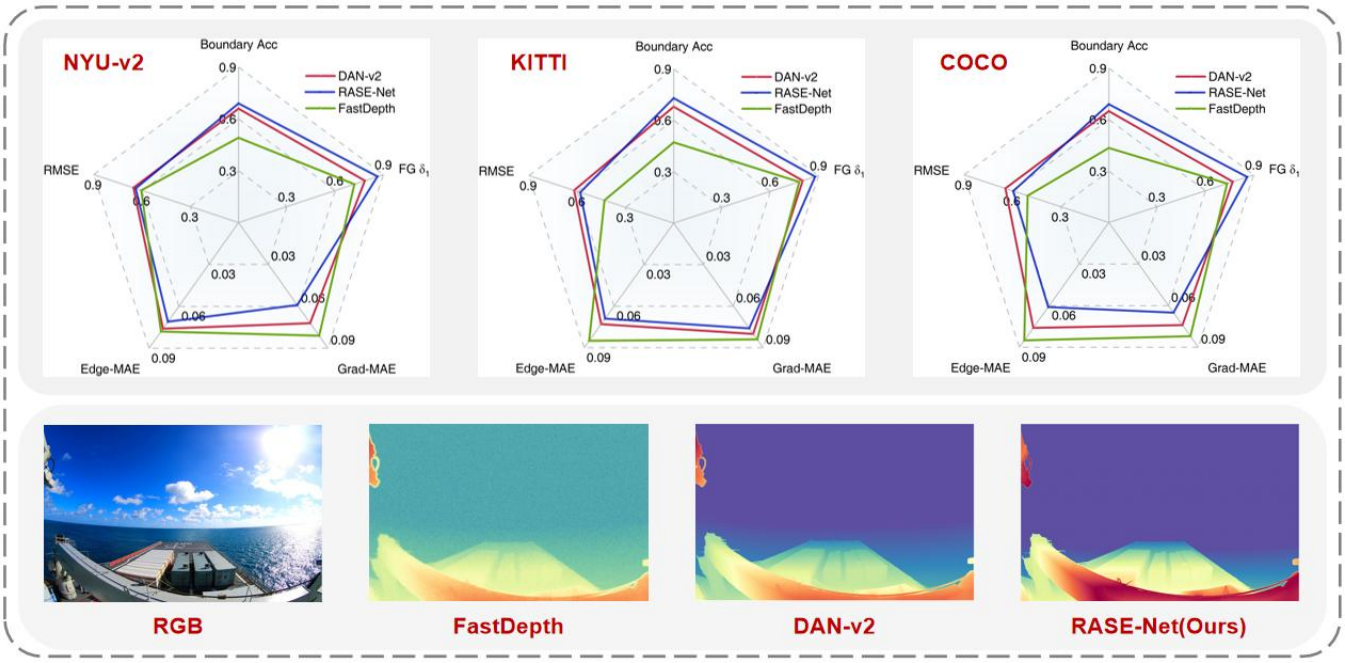

Figure 1. Structure-sensitive evaluation across domains and degradations. (a) RASE-Net surpasses the DAN-v2 baseline across five structure-aware indicators—Boundary Accuracy, Foreground δ_1, Edge-MAE, Grad-MAE, and RMSE—on NYU-v2, KITTI, and COCO2017, confirming its robustness under diverse geometric priors. (b) Under strong exposure and occlusion, DAN-v2 predictions exhibit boundary collapse and saliency smoothing, while RASE-Net restores gradient continuity and object-level contours. These visual and quantitative gaps motivate our refinement strategy (§3), which targets structural transitions and saliency-aware regions without altering the transformer backbone.

detail loss near object boundaries, while RASE-Net restores foreground layering, preserves contour integrity, and mitigates saliency diffusion under strong exposure and occlusion.

These results highlight the need for structure-aware refinement. Although recent methods [13,14] have proposed architecture-level modifications to address this, structural fidelity is still inconsistently handled. While hierarchical attention and decoder specialization [15-17] improve overall coherence, they often fail to restore fine-grained details near boundaries. Multi-scale fusion and contrastive strategies help mitigate context loss but cannot reliably recover spatial sharpness. This issue becomes particularly critical in safety-sensitive scenarios like UAV navigation, where small errors around structural edges can lead to planning failures [18]. Recent work has explored specialized techniques such as edge-aware losses [19], saliency-guided attention [20], and geometry-preserving constraints [21]. However, many of these require extra supervision or introduce complexity at inference, which limits deployment in real-time or resource-constrained systems [22]. Given these limitations, we build upon DAN-v2 to enhance structural fidelity, leveraging its scalable global reasoning while addressing localized degradation.

In this work, we propose Residual Attention Spatial Enhancement Network(RASE-Net), a lightweight refinement framework for enhancing structural fidelity in monocular depth estimation. Instead of replacing the backbone predictor, RASE-Net integrates two parameter-free modules atop transformer-based architectures to improve geometric consistency without incurring architectural modifications or inference overhead. The first is NoEffectSimAM, a residual-form attention regularizer that imposes structure-aware gradients during training while reverting to identity mapping at inference. The second is a Semantic Enhancement Unit (SEU), which refines predicted depth via post-hoc modulation guided by geometry-aligned saliency priors. These modules jointly recover occlusion transitions, foreground delineation, and high-frequency geometric cues, while preserving compatibility with the base model.

To achieve robust structural modeling, we identify five desirable properties that a structure-aware depth model should exhibit: the ability to reliably estimate foreground regions across occlusions, maintain sharp transitions around object boundaries, reduce structural error sensitivity in

complex scenes, ensure consistency across diverse domains without fine-tuning, and retain structural fidelity under severe resolution degradation. These axes reflect the core challenges in deploying depth estimation under real-world visual degradation and geometric ambiguity.

Table 1 summarizes a comparative assessment across these properties. While FastDepth and DAN-v2 satisfy a subset of the criteria, RASE-Net fulfills all five, offering a more comprehensive solution for structure-sensitive depth perception.

Table 1. Comparison of representative methods across five structure-aware properties. A checkmark (✓) denotes satisfactory performance under the respective criterion, covering foreground modeling, boundary sharpness, error sensitivity, cross-set stability, and low-resolution robustness.

Preferable Properties	Foreground Modeling	Boundary Sharpness	Error Sensitivity	Cross-Set Stability	Low-Res Robustness
FastDepth	✗	✗	✗	✓	✓
DAN-v2	✓	✓	✗	✓	✗
RASE-Net(ours)	✓	✓	✓	✓	✓

2. Related work

2.1 Transformer-Based and Modular Approaches in Monocular Depth Estimation

Recent advances in monocular depth estimation (MDE) increasingly leverage transformer-based architectures to encode long-range dependencies and enforce global semantic consistency. Architectures such as DPT [23], AdaBins [24], and BinsFormer [25] employ structured tokenization and multi-scale attention to deliver strong scene-level predictions. Compared to earlier CNN-based designs (e.g., BTS [26], FastDepth), these models offer improved layout understanding but often blur object boundaries and thin structures due to spatial averaging.

To address these limitations, modular refinement strategies have emerged to decouple coarse prediction from structural correction. MiDaS [27] introduces task-specific heads for cross-domain adaptation, while GC-Depth [28] integrates geometric consistency constraints into the learning pipeline. Recent hybrid designs, including DepthFormer [29] and DFormer [30], fuse decoder-level semantics with geometry-preserving priors to recover localized details missed by global attention.

Collectively, these methods reflect a shift from monolithic prediction to compositional refinement, highlighting the growing need for structure-sensitive modules that can be flexibly attached without redesigning the full model pipeline.

2.2 Attention Mechanisms for Structural Reasoning

Attention mechanisms have become central to dense prediction tasks, enabling dynamic feature reallocation based on spatial or semantic cues. Classic channel-centric modules like SENet [31] and hybrid designs such as CBAM [32] adaptively modulate response strength across feature dimensions. More recent attention formulations, including DANet [33] and LiteHRNet [34], encode multi-scale structure via task-aligned attention fusion. SimAM [35] introduces a parameter-free mechanism based on neuron energy minimization, enabling training-time modulation without inference overhead. This class of residual-only attention modules offers strong architectural compatibility and deployment safety under tight efficiency constraints.

Beyond generic attention, a separate line of research targets structure-aware refinement. DepthPro [36] and SharpNet [37] introduce edge-guided losses and residual regularization to enhance contour fidelity, while GFNet [38] and structure-centric distillation strategies [39] improve spatial alignment across representations. While effective, many of these methods rely on segmentation maps, contrastive supervision, or task-specific branches — compromising scalability across domains.

These trends underscore a growing preference for lightweight, plug-in attention heads that restore boundary-level detail without architectural entanglement or inference penalties.

2.3 Structure-Aware Attention under Deployment Constraints

Deploying monocular depth estimation in aerial robotic systems introduces a distinct set of constraints—wide depth ranges, real-time latency requirements, frequent occlusions, and limited compute budgets. While LiDAR and stereo systems provide geometric accuracy, they remain impractical for lightweight UAVs due to power and payload constraints. Recent efforts mitigate these challenges through encoder streamlining (TinyDepthNet [40]), domain-adaptive learning for aerial visual priors [41], and motion-consistent attention tuned to UAV flight dynamics [42,43]. While effective, many such methods require staged supervision or synthetic fine-tuning, increasing system complexity and deployment cost.

A parallel direction explores plug-and-play refinement heads that operate after the main decoder stage. Methods

such as instance-level attention [44], saliency-guided sharpening [45], and residual enhancement [46] show promise in enhancing spatial transitions, particularly in occlusion-rich or saliency-dense scenes. However, these solutions often introduce inference-time branches or rely on external masks, limiting robustness across platforms.

These trends expose a persistent tension between architectural efficiency and structural fidelity in UAV-aligned perception. While prior work focuses on compression or modulation, few methods directly address the joint enhancement of foreground saliency and geometric discontinuity under real-time constraints. To this end, we pursue a decoupled refinement formulation that injects residual attention during training and applies geometry-aware saliency modulation post-prediction. The next section details the design of RASE-Net, which follows this dual-path strategy to selectively correct object-centric regions without modifying backbone inference.

3. Related work

3.1 Overall Framework

Figure 2 presents an overview of the proposed RASE-Net, which augments transformer-based monocular depth estimation with minimal architectural intrusion. The network integrates two lightweight plug-in modules: a geometry-preserving attention unit (NoEffectSimAM) and a saliency-gated enhancement unit (SEU), both appended post-DPTHead.

Unlike global attention layers that smooth over fine-grained transitions, our design explicitly targets structural artifacts introduced by attention diffusion—particularly around occlusion fronts and salient boundaries. RASE-Net adopts a residual-parallel enhancement strategy, where each module independently refines depth predictions along distinct structural axes. NoEffectSimAM applies parameter-free weighting based on energy-derived neuron activation, enabling structure-aware training gradients without inference overhead. It operates in residual form, maintaining identity mappings during testing to avoid disrupting backbone pathways. In parallel, SEU selectively modulates predicted depth maps by isolating salient foregrounds using geometry-informed priors, followed by subtraction-based edge sharpening.

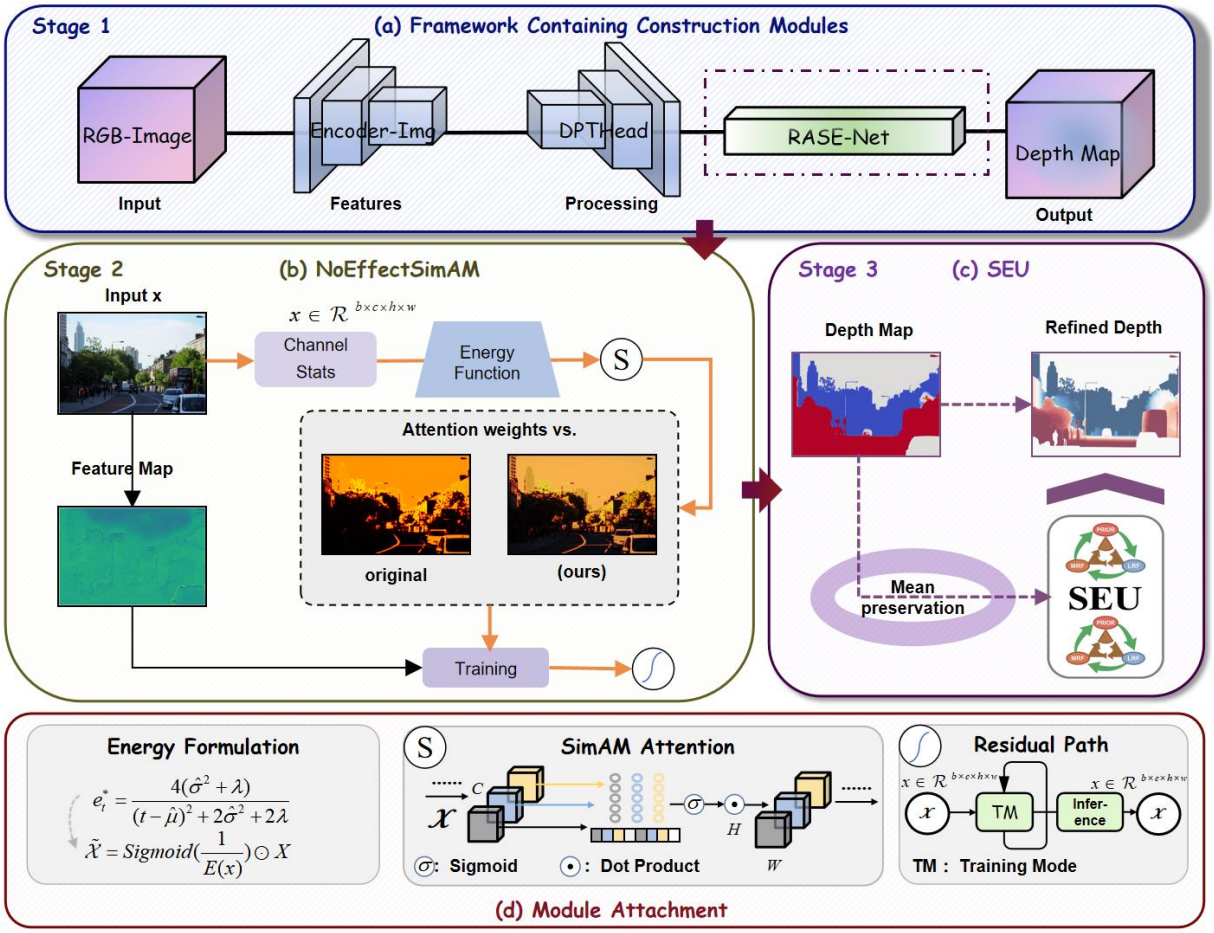

Figure 2. Overview of the proposed RASE-Net architecture. The figure illustrates our residual-parallel refinement design, composed of two lightweight modules appended post-DPTHead. (a) Presents the full monocular depth estimation

pipeline with transformer backbone and RASE-Net enhancement head. (b) Visualizes the internal structure of NoEffectSimAM, which computes attention weights based on neuron energy to guide training-time regularization without inference overhead. (c) Depicts the SEU module, which enhances the predicted depth map by selectively refining salient object regions based on geometry-informed priors. (d) Details the residual attachment and energy-based weighting formulation used to inject structure-aware gradients during training while preserving backbone integrity.

Both modules are fused with the decoder output in a residual fashion, enhancing contour delineation and instance transitions. The design maintains full architectural compatibility and inference stability, enabling RASE-Net to serve as a transferable refinement strategy for UAV-centric or structure-critical perception systems.

3.2 NoEffectSimAM

We aim to introduce attention-driven regularization into the training process without altering the model's inference-time behavior. To this end, we design a structure-aware attention simulation module, termed NoEffectSimAM. As depicted in Figure 2(b), the module is positioned after the final decoder fusion block and before the depth prediction head. During training, NoEffectSimAM injects structured gradients that encourage spatial discrimination and boundary awareness. At inference, it reverts to an identity function, thereby ensuring computational consistency and deployment efficiency.

Let $x \in \mathcal{R}^{N\times C\times H\times W}$ denote an intermediate feature tensor, we compute the per-channel spatial mean μc and squared deviation $d_{i,j}^{(c)}$ as:

$$\mu_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_{i,j}^{(c)} \tag{1}$$

$$d_{i,j}^{(c)} = (x_{i,j}^{(c)} - \mu_c)^2 \tag{2}$$

Motivated by the closed-form derivation of SimAM, we adopt an energy-based normalization strategy to construct our spatial attention weights. Unlike the original SimAM, which targets neuron-level saliency using channel-level activation energy, our formulation emphasizes feature-level spatial regularization. Specifically, we construct the attention map $\mathcal{A}tt_{i,j}^{(c)}$ for each spatial location (i,j) in channel c as:

$$\mathcal{A}tt_{i,j}^{(c)} = \frac{4\cdot(\frac{1}{HW}\sum_{i,j} d_{i,j}^{(c)})^2 + \varepsilon}{(d_{i,j}^{(c)})^2 + \varepsilon} \tag{3}$$

Here, ε is a small constant for numerical stability. This formulation emphasizes high-energy spatial positions — typically corresponding to semantically and geometrically significant regions — as soft spatial priors. To avoid interference at inference, we adopt a residual-modulation scheme controlled by a learnable scalar $\alpha_c \in [0,1]$, yielding the output:

$$\hat{x}_{i,j}^{(c)} = (1-\alpha_c)\cdot x_{i,j}^{(c)} + \alpha_c \cdot (x_{i,j}^{(c)} \cdot \mathcal{A}tt_{i,j}^{(c)}) \tag{4}$$

We initialize $\alpha \ll 1$ (e.g.,10-6)such that the network gradually learns whether and how strongly to activate attention. The partial derivative with respect to α_c is:

$$\frac{\partial \hat{x}_{i,j}^{(c)}}{\partial \alpha_c} = -x_{i,j}^{(c)} + x_{i,j}^{(c)} \cdot \mathcal{A}tt_{i,j}^{(c)} \tag{5}$$

This ensures meaningful gradients even when the forward path approximates identity. During inference, we explicitly set α_c=0, eliminating attention modulation entirely, and yielding $\hat{x}_{i,j}^{(c)} = x_{i,j}^{(c)}$. This guarantees zero runtime overhead and full compatibility with the original forward path. This formulation preserves training-time attention for spatial regularization while ensuring inference-time transparency. It supports latency-sensitive deployment scenarios, particularly where architectural stability is critical. Figure 2(d) illustrates the complete structure of NoEffectSimAM, including the energy-driven attention mechanism and the modulation pathway. Combined with Figure 2(b), these subfigures convey the architectural and computational role of NoEffectSimAM as a lightweight, training-only structural prior injector.

```
# NoEffectSimAM - training-time residual attention
# X: feature map [N, C, H, W]; alpha: modulation scalar

def residual_attention(X, alpha)
       mu = X.mean(dim=[2, 3], keepdim = True)
       # channel-wise mean
       delta = (X - mu) . pow(2)
       # spatial deviation
       energy = delta . sum(dim = [2, 3], keepdim = True)
       # variance term
       att = 4 * (delta + ε )
       # Eq (3) attention
       out = (1 - alpha) * X + alpha * X * att
       # Eq (4) output
       return out
```

Figure 3. A PyTorch-style implementation of NoEffectSimAM. The routine computes spatial attention via energy normalization and applies residual blending. Inference-time output is identity, ensuring zero-overhead deployment.

To complete the formulation, we include the forward logic of NoEffectSimAM. The code instantiates Equations (3)–(5), applying spatial energy normalization and residual modulation during training, while reverting to identity at inference.

3.3 Semantic Enhancement Unit

While NoEffectSimAM provides structure-aware training guidance at the feature level, it does not directly optimize the final depth prediction. To address residual geometric ambiguities — particularly around foreground boundaries — we propose the SEU: a lightweight, fully differentiable refinement module that selectively modulates the depth map using geometry-informed saliency priors. SEU operates as a post-decoder plug-in and is compatible with standard encoder – decoder pipelines.

As shown in Figure 4, SEU consists of three sequential stages: (1) saliency prior estimation, (2) confidence-based mask generation, and (3) residual contrast modulation. These stages are designed to sharpen spatial transitions near occlusions and recover local detail commonly smoothed by dense regression.

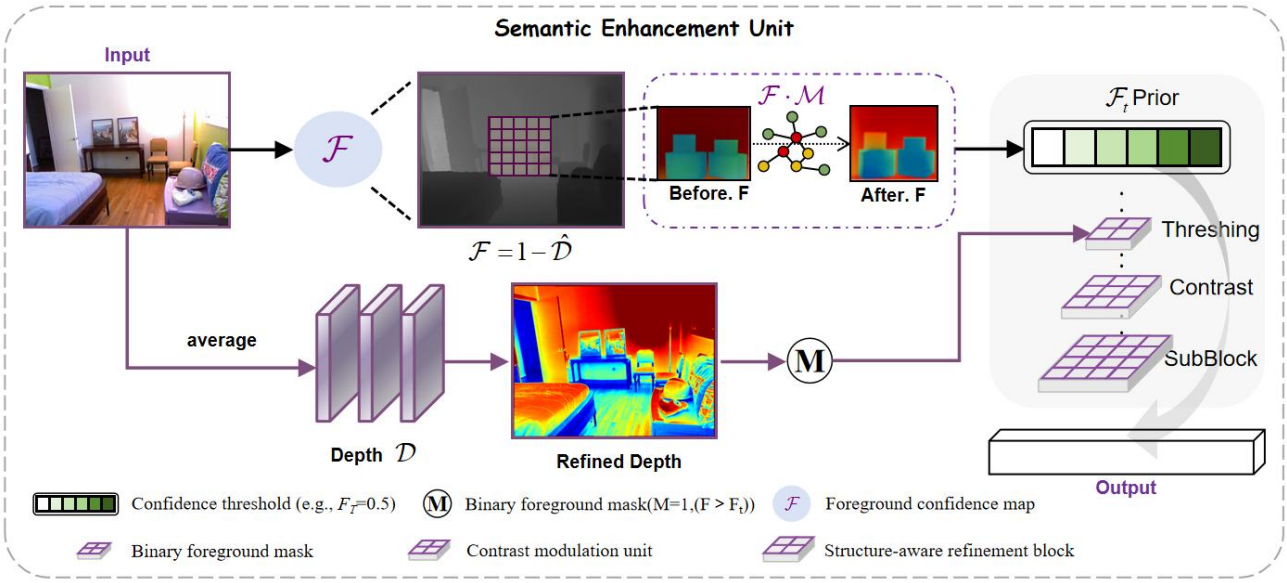

Figure 4. Overview of the Semantic Enhancement Unit (SEU). Given an inverted depth map, a soft confidence prior is computed and adaptively modulated via a three-stage refinement: thresholding, contrast modulation unit, and structure-aware refinement block. The entire module operates post-hoc and is fully differentiable.

Given a predicted depth map $\mathcal{D} \in \mathcal{R}^{H \times W}$, we compute a soft foreground confidence prior $\mathcal{F}_{i,j} \in [0,1]^{H \times W}$ via inverted min – max normalization:

$$\mathcal{F} = 1 - \mathcal{N}(\mathcal{D}), \mathcal{N}(\mathcal{D}) = \frac{\mathcal{D}_{i,j} - \min(\mathcal{D})}{\max(\mathcal{D}) - \min(\mathcal{D})}, \tag{6}$$

Here, $\mathcal{F}_{i,j} \in [0,1]^{H \times W}$ encodes pixel-wise spatial proximity to the camera, with higher values representing closer objects. This formulation serves as a structure-sensitive proxy for foregroundness without requiring any semantic annotations. Compared to parametric attention maps, this normalization-based prior is interpretable, efficient, and robust to scene variation. To localize refinement, we derive a binary foreground mask $\mathcal{M} \in \{0,1\}^{H \times W}$ using a fixed threshold $\tau \in (0,1)$:

$$\mathcal{M}_{i,j} = \Pi(\mathcal{F} > \tau) \tag{7}$$

The mask acts as a discrete selector, focusing enhancement on salient regions where structural boundaries are more likely to be ambiguous or smoothed. This separation of prior and mask enables SEU to jointly encode both gradient magnitude and spatial relevance. The final depth refinement is formulated as contrast-aware residual suppression:

$$\hat{\mathcal{D}} = \mathcal{D} - \beta \cdot \mathcal{M} \odot (1 - \mathcal{F}) \tag{8}$$

Where $\beta \in \mathcal{R}^{+}$ modulates the enhancement strength, and $\odot$ denotes element-wise multiplication. The term $(1 - \mathcal{F})$ ensures that refinement is stronger in spatially uncertain (low-confidence) foreground regions, while M confines the adjustment to salient zones. This formulation enhances object contours without disrupting global layout or stable background predictions. Unlike decoder-based correction schemes that require auxiliary branches or supervision, SEU performs structure-aware enhancement in a fully unsupervised, geometry-driven manner. It introduces no additional parameters unless explicitly enabled, remains fully differentiable, and imposes negligible runtime overhead.

To complement the above formulation, we include the forward routine of SEU in Figure 5. The snippet outlines the core operations — depth normalization, foreground masking, and residual contrast modulation — as defined in Equations (6)–(8). By decoupling saliency-guided contrast modulation

from the main prediction path, SEU offers a plug-and-play refinement strategy applicable to both training and inference. Compatible with CNN and transformer backbones, it provides a lightweight yet effective mechanism for restoring structural fidelity in dense depth estimation.

```
# SEU - post-processing depth refinement
# D: predicted depth [H, W]
# beta: contrast strength
# tau: binarization threshold

def seu_forward(D, beta, tau):
        # normalize and invert depth Eq (6)
        D_norm = (D - D.min()) / (D.max() - D.min())
        F = 1 - D_norm
        # confidence prior
        M = (F > tau).float()
        # thresholding
        # generate binary foreground mask Eq(7)
        D_refined = D - beta * M * (1 - F)
        # apply residual contrast-based correction Eq (8)
        return D_refined
```

Figure 5. A PyTorch-style implementation of the Semantic Enhancement Unit. The routine normalizes and inverts depth to compute a saliency prior, constructs a binary foreground mask, and applies residual contrast modulation. All operations are differentiable and inference-efficient.

4. Experiments

4.1 Experiments Setup

Dataset and Protocol. Our evaluations primarily focus on the NYU-v2 benchmark, which provides dense indoor layouts with diverse spatial configurations. This dataset contains 50K RGB-D pairs collected in cluttered scenes with occlusions, object boundaries, and lighting variations — characteristics that pose significant challenges for structure-aware depth estimation. To assess cross-domain generalization under heterogeneous spatial priors, we further evaluate on KITTI and COCO2017. The KITTI dataset offers large-scale street-view depth maps with sparse LiDAR supervision and strong perspective distortion. In contrast, COCO2017 introduces heavy occlusions, complex object layouts, and scene clutter, enabling robustness analysis under dense appearance variation.

Evaluation Metrics. We evaluate depth estimation performance using four global error metrics—RMSE, MAE, Log RMSE, and AbsRel—to assess the pixel-level consistency between predictions and ground truth. To capture structural fidelity beyond global alignment, we introduce a structure-aware evaluation protocol that focuses on geometry-sensitive regions and local discontinuities. Specifically, we report: (1) δ_1, δ_2, and δ_3, which measure the ratio of pixels within progressively relaxed relative error thresholds $\{1.25, 1.25^2, 1.25^3\}$; (2) δ_1(Fg), a foreground-restricted variant of δ_1 that emphasizes accuracy in object-centric regions; and (3) EdgeAcc and Grad-MAE, which quantify contour sharpness and gradient-level consistency, respectively. All evaluations are conducted under zero-shot settings without fine-tuning.

Implementation Details. All experiments are conducted using the DAN-v2 ViT-S backbone pretrained with DINOv2. This variant offers a favorable balance between training efficiency and representational power, and is consistently used across all comparisons and ablations. Multi-scale features from four transformer stages are aggregated via a DPT-style decoder, with output dimensions of {256, 512, 1024, 1024}. The proposed RASE-Net refinement head—composed of NoEffectSimAM and SEU—is appended after the decoder output. During training, NoEffectSimAM applies residual-form spatial weighting as structure-aware regularization, without altering inference behavior. At inference, SEU operates in a post-hoc fashion using a fixed saliency threshold ($\tau = 0.5$) and contrast scaling factor ($\beta = 0.3$). All experiments are conducted on a single NVIDIA RTX 4070 TiS GPU without any post-processing or test-time augmentation.

4.2 Comparison with Baseline

We benchmark RASE-Net under zero-shot settings against DAN-v2, using identical encoder-decoder configurations to isolate the impact of our proposed structure-aware refinement modules. This setup ensures that observed differences stem solely from architectural augmentations, without the confounding effect of pretraining or fine-tuning.

Table 2. Quantitative comparison on the NYUv2 benchmark. All methods are evaluated under consistent training protocols. Error metrics (↓) include RMSE, MAE, Log RMSE, and AbsRel; accuracy metrics (↑) include δ_1, δ_2, and δ_3. These metrics quantify global alignment between predicted and ground-truth depth across the full image domain.

Methods	*Lower is better* ↓				*Higher is better* ↑		
	RMSE	MAE	Log RMSE	AbsRel	δ_1	δ_2	δ_3
DAN-v2(Paper)	0.592	0.150	0.205	0.129	0.830	0.980	0.992
DAN-v2	0.653	0.473	0.118	0.168	0.746	0.940	0.977

Ours	<u>0.639</u>	<u>0.462</u>	0.175	0.193	<u>0.758</u>	0.917	0.955

Compared to the reproduced DAN-v2 baseline, RASE-Net achieves a relative +1.6% improvement in δ_1 and a −2.1% reduction in RMSE, suggesting enhanced depth fidelity with lower large-error dispersion. These improvements are primarily driven by the two proposed modules: the SEU refines saliency-dense foreground regions through residual contrast correction, while NoEffectSimAM introduces structure-aware regularization at the feature level without affecting inference. Slight degradations in AbsRel, log RMSE, and δ_2/δ_3 can be observed, particularly in low-texture background regions. We attribute these changes to the model's capacity reallocation, favoring near-field detail preservation and edge continuity over global smoothing—a trade-off aligned with our design intent to improve geometric expressiveness near occlusion boundaries.

Notably, our reproduced DAN-v2 results slightly diverge from the original paper, likely due to preprocessing inconsistencies or implementation gaps. However, all comparisons are conducted under strictly consistent settings, and our reported gains are measured relative to the re-evaluated baseline. To better quantify structure-aware performance, we introduce additional metrics in Section 4.4 targeting boundary precision and foreground reliability. Individual module effects and synergistic gains are further analyzed in Section 4.3.

4.3 Ablation Study

To assess the individual contributions of each proposed module, we conduct controlled ablation experiments under the zero-shot setup on NYU-v2, using DAN-v2 (ViT-S) as the backbone. Beyond global error metrics (e.g., RMSE), we emphasize structure-aware indicators—Foreground Accuracy (δ_1(Fg)), Edge Accuracy (EA),—to quantify improvements in semantically critical and geometrically complex regions.

Table 3. Ablation analysis of progressive module insertion. We evaluate the individual and joint contributions of NoEffectSimAM and SEU through a stepwise insertion strategy. Both modules contribute independently, with NoEffectSimAM improving Grad-MAE and SEU refining saliency boundaries.

Methods				RMSE ↓	*Higher is better* ↑		
DAN-v2	+NoEffectSimAM	+SEU	RASE-Net		δ_1	δ_1（Fg）	EdgeAcc
✓	-	-	-	0.653	0.746	0.940	0.512
✓	✓	-	-	0.653	0.746	0.946	0.546
✓	-	✓	-	0.639	0.753	0.957	0.537
✓	-	-	✓	0.639	<u>0.758</u>	0.957	<u>0.562</u>

Table 3 presents the results of progressive module insertion. Compared with the DAN-v2 baseline, both NoEffectSimAM and SEU individually improve structure-aware metrics without incurring additional model parameters. Specifically, NoEffectSimAM improves Grad-MAE and δ_1, validating its training-time regularization efficacy. SEU contributes larger gains in δ_1(Fg) and EA, reflecting stronger boundary refinement and geometric modulation. When combined into RASE-Net, the two modules collectively yield superior performance across all axes, revealing a complementary dynamic between training-guided structure injection and test-time saliency refinement.

To visualize structure-aware behavior under challenging conditions, Figure 6 presents predictions from progressive module integration. Inputs are deliberately selected from low-resolution scenes, where spatial aliasing often amplifies depth ambiguity and occlusion-induced collapse—common failure cases in monocular estimation. From left to right, we show depth maps from DAN-v2 , +NoEffectSimAM, +SEU (i.e., full RASE-Net), and structural edge maps from DAN-v2 and RASE-Net.

The visual trajectory aligns with our ablation trends: NoEffectSimAM enhances spatial regularization and edge continuity through energy-driven priors, while SEU sharpens foreground layering and suppresses saliency diffusion. Notably, RASE-Net consistently recovers sharper transitions and semantic separation—especially along occlusion fronts and fine contours—even under strong compression. This suggests stronger geometric fidelity in resolution-constrained settings.

To evaluate cross-domain robustness, Table 4 reports ablation results on KITTI and COCO2017 without fine-tuning. Despite domain and style shifts, trends persist: NoEffectSimAM improves δ_1 and Grad-MAE, while SEU refines δ_1(Fg) and edge accuracy. RASE-Net achieves the best trade-off between global and structure-aware fidelity under domain shifts.

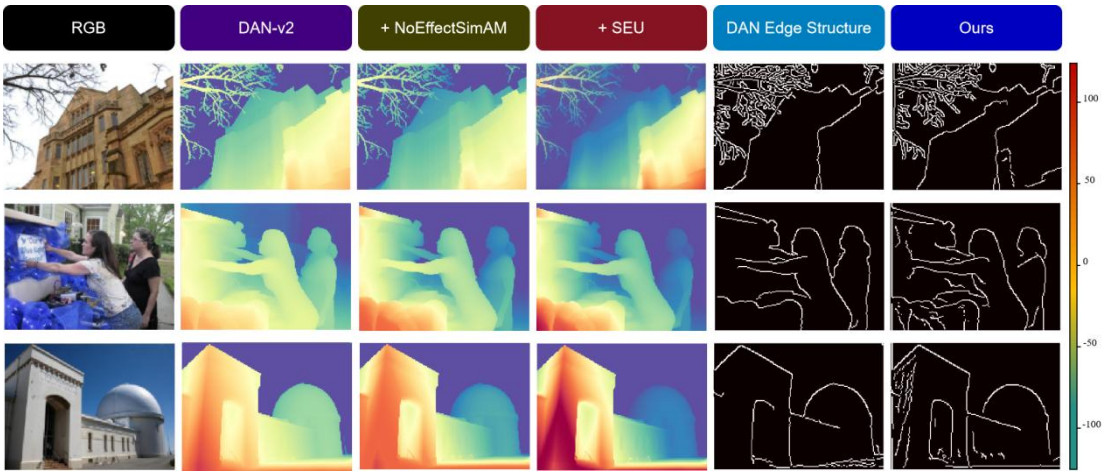

Figure 6. Qualitative visualization of progressive module integration. From left to right: input RGB, depth predictions from DAN-v2, +NoEffectSimAM, +SEU (i.e., full RASE-Net), and structural edge maps from DAN-v2 and RASE-Net. All samples are low-resolution to intensify structural degradation.

Table 4. Cross-domain ablation on KITTI and COCO2017. NoEffectSimAM and SEU consistently improve boundary accuracy and foreground fidelity. The full model maintains structure-aware advantages under diverse visual domains.

Dataset	Method	RMSE	δ_1	δ_1（Fg）	EdgeAcc
KITTI	DAN-v2	0.412	0.791	0.821	0.486
	+ NoEffectSimAM	0.410	0.795	0.829	0.519
	+ SEU	0.408	0.802	0.845	0.505
	RASE-Net	<u>0.403</u>	<u>0.810</u>	<u>0.853</u>	<u>0.537</u>
COCO2017	DAN-v2	0.496	0.743	0.794	0.421
	+ NoEffectSimAM	0.493	0.748	0.806	0.456
	+ SEU	0.490	0.755	0.821	0.442
	RASE-Net	<u>0.487</u>	<u>0.763</u>	<u>0.834</u>	<u>0.468</u>

To further quantify our claim of efficiency, Table 5 reports model size, runtime, and FPS to assess computational efficiency. While introducing structural refinement, RASE-Net preserves real-time throughput (FPS > 30) and maintains identical parameter count. The observed latency increase is marginal (<0.3 ms), confirming that the proposed design yields structure-aware gains without incurring inference overhead.

Table 5. Efficiency analysis of all variants. RASE-Net maintains identical parameter count and real-time performance, introducing negligible overhead during inference.

Method	Params(M)	FPS	Runtime(ms)
DAN-v2	98.3	31.2	32.1
+ NoEffectSimAM	98.3	31.2	32.1
+ SEU	98.3	31.1	32.3
RASE-Net	<u>98.3</u>	<u>31.0</u>	<u>32.3</u>

4.4 Structure-Aware Evaluation and Visualization

Standard depth estimation metrics often fail to reveal prediction deficiencies in spatially critical regions, where object boundaries, saliency discontinuities, and fine-grained geometric transitions intersect. To address this gap, we introduce a structure-aware evaluation protocol that decomposes model behavior across interpretable spatial

axes. Our analysis jointly examines foreground saliency, signed depth errors, and contour fidelity across diverse visual conditions.

Figures 7-9 present qualitative comparisons across NYU-v2, KITTI, and COCO2017. Each visualization follows a consistent layout: input RGB, estimated foreground masks, signed depth difference maps, and predictions from DAN-v2 and our RASE-Net. These datasets span a spectrum of occlusion complexity and depth topology, from indoor layouts to urban streetscapes and crowded public scenes. On NYU-v2, RASE-Net produces cleaner depth transitions and sharper foreground delineations in complex urban scenes. On KITTI, it preserves long-range consistency across vehicle contours and building facades under real-world motion. On COCO2017, it maintains semantic alignment despite umbrella occlusions, dense human overlap, and low-light conditions—scenarios where baseline methods exhibit substantial structural degradation.

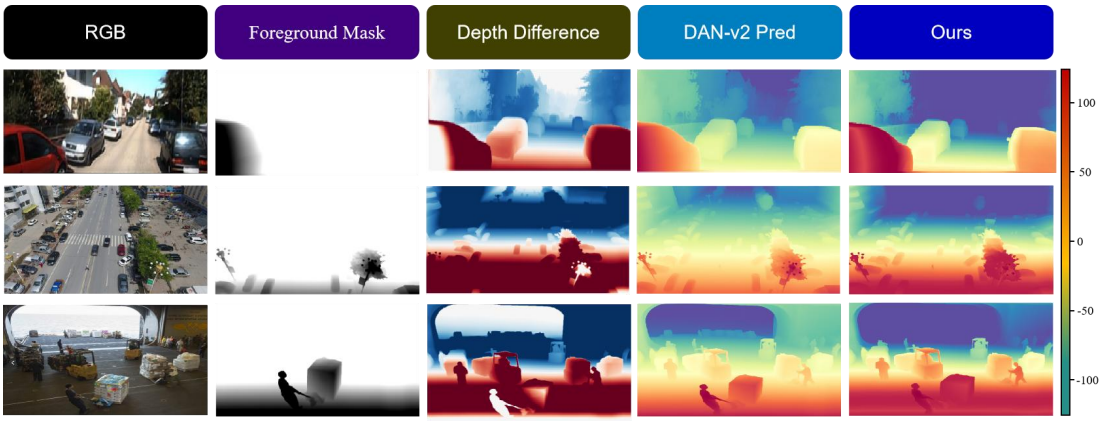

Figure 7. Structure-aware visualization on NYU-v2. RASE-Net produces cleaner depth transitions and more compact saliency masks in cluttered street-level scenes. Differences localize around structural boundaries, indicating refined geometry alignment.

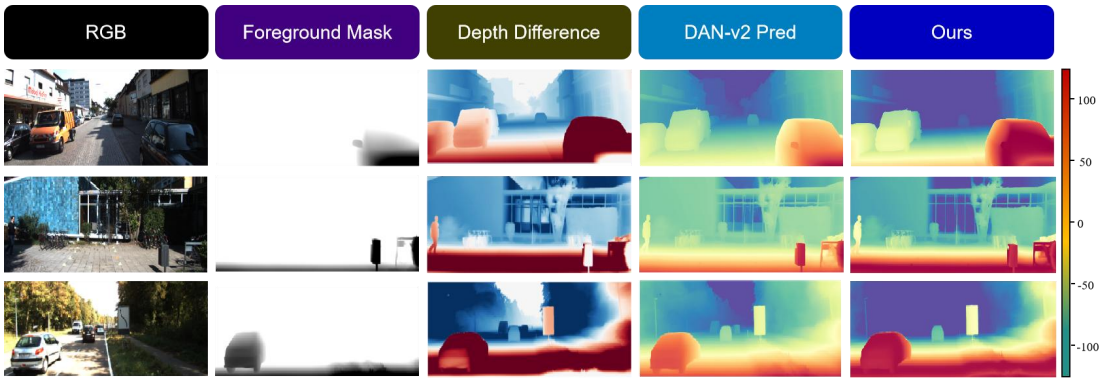

Figure 8. Structure-aware visualization on KITTI. Our method preserves long-range consistency and reduces structural flattening across vehicle and facade contours under real-world depth variation.

Specifically, RASE-Net consistently yields sharper and more localized saliency transitions, forming instance-aware masks along silhouettes and occlusion fronts. Compared to DAN-v2, whose activations often diffuse into background clutter under low lighting, RASE-Net preserves structure-aligned priors without compromising contextual integrity. These improvements manifest especially under high-curvature regions, facade edges, and occlusion-dense environments. Difference maps highlight directional depth errors. Positive deviations—where RASE-Net predicts shallower depths—frequently emerge near object contours, suggesting improved spatial anchoring. Conversely, negative bands indicate baseline overestimation in under-segmented regions. These asymmetries confirm RASE-Net's bias toward geometry-conforming corrections, particularly under occlusion and coarse backgrounds.

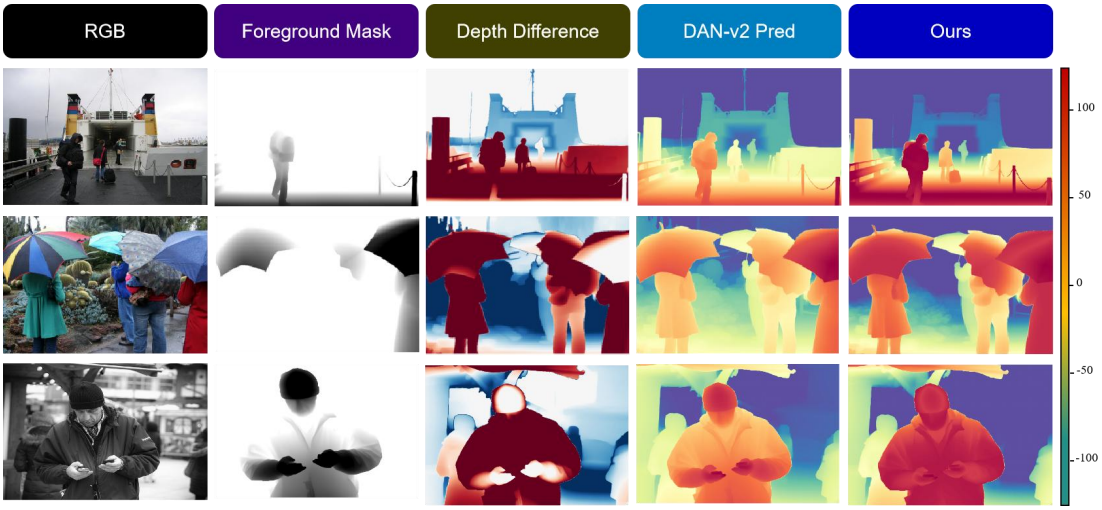

Figure 9. Structure-aware visualization on COCO2017. Under heavy occlusion and dense object overlap, RASE-Net produces sharper saliency masks and more localized depth corrections. Discrepancies in difference maps highlight improved alignment along instance contours, with RASE-Net preserving structure under umbrella occlusion, body coalescence, and low-light clutter.

To complement qualitative analysis, we assess four structure-aware metrics—Foreground Accuracy (δ_1), Edge Alignment (EA), Gradient MAE, and High-Error Rate (HER)—over NYU-v2, KITTI, and COCO2017 (Fig. 10). RASE-Net achieves consistent gains across all metrics, improving boundary localization and reducing structural noise. Lower HER scores further indicate reduced failure rates under geometrically ambiguous regions. These results validate the robustness of our design in preserving fine-grained structure across heterogeneous domains.

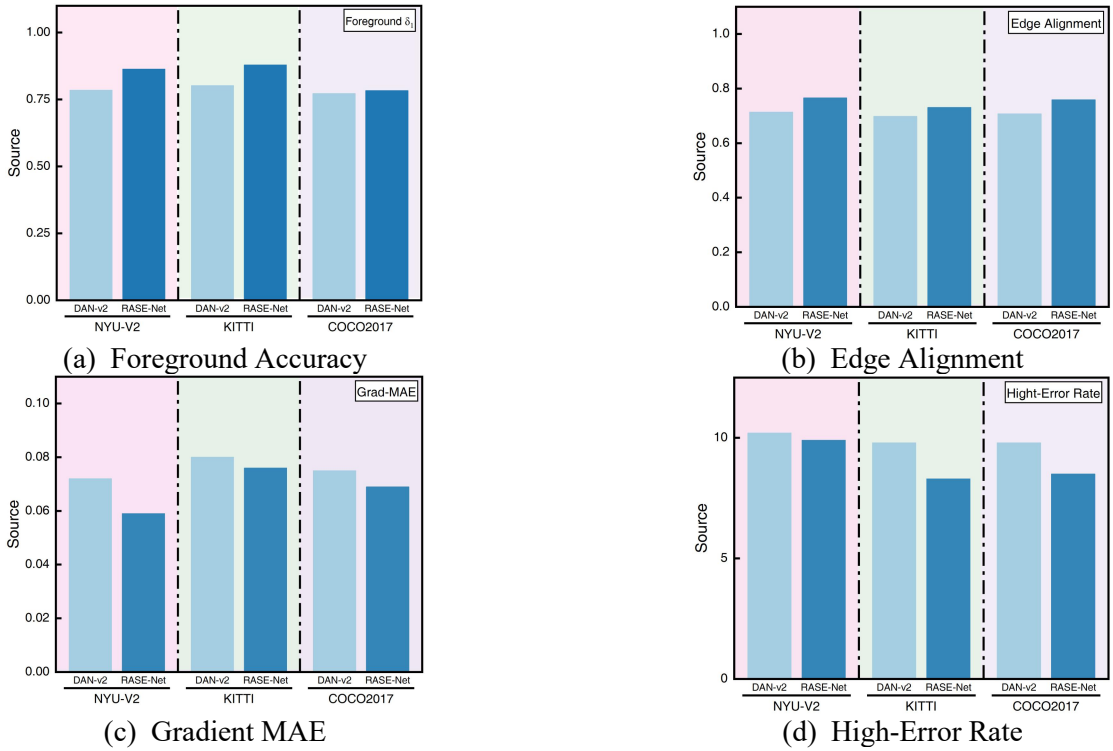

Figure 10. Cross-dataset comparison of structure-aware metrics. δ_1, EA, Grad-MAE, and HER are reported on NYU-v2, KITTI, and COCO2017. RASE-Net exhibits stable trends across metrics and datasets, indicating improved structural regularity under diverse conditions.

Table 6. Cross-dataset comparison of HER, GME, and EA. We report structure-aware metrics on NYU-v2, KITTI, and COCO2017 to assess generalization across domains. Compared with prior models, RASE-Net yields stable gains in gradient smoothness and edge alignment while maintaining a competitive HER, supporting its structural consistency under diverse input conditions.

Method	NYU-v2			KITTI			COCO2017		
	HER	GME	EA	HER	GME	EA	HER	GME	EA
MiDas[47]	10.1	0.074	0.710	9.9	0.082	0.715	9.8	0.080	0.705
DPT	10.0	0.072	0.730	9.7	0.080	0.721	9.7	0.078	0.720
DAN-v1[48]	10.5	0.076	0.722	9.6	0.080	0.721	9.6	0.079	0.712
DAN-v2	10.2	0.072	0.746	9.4	0.078	0.740	9.6	0.076	0.732
RASE(Ours)	9.6	0.060	0.775	8.8	0.076	0.742	9.0	0.071	0.760

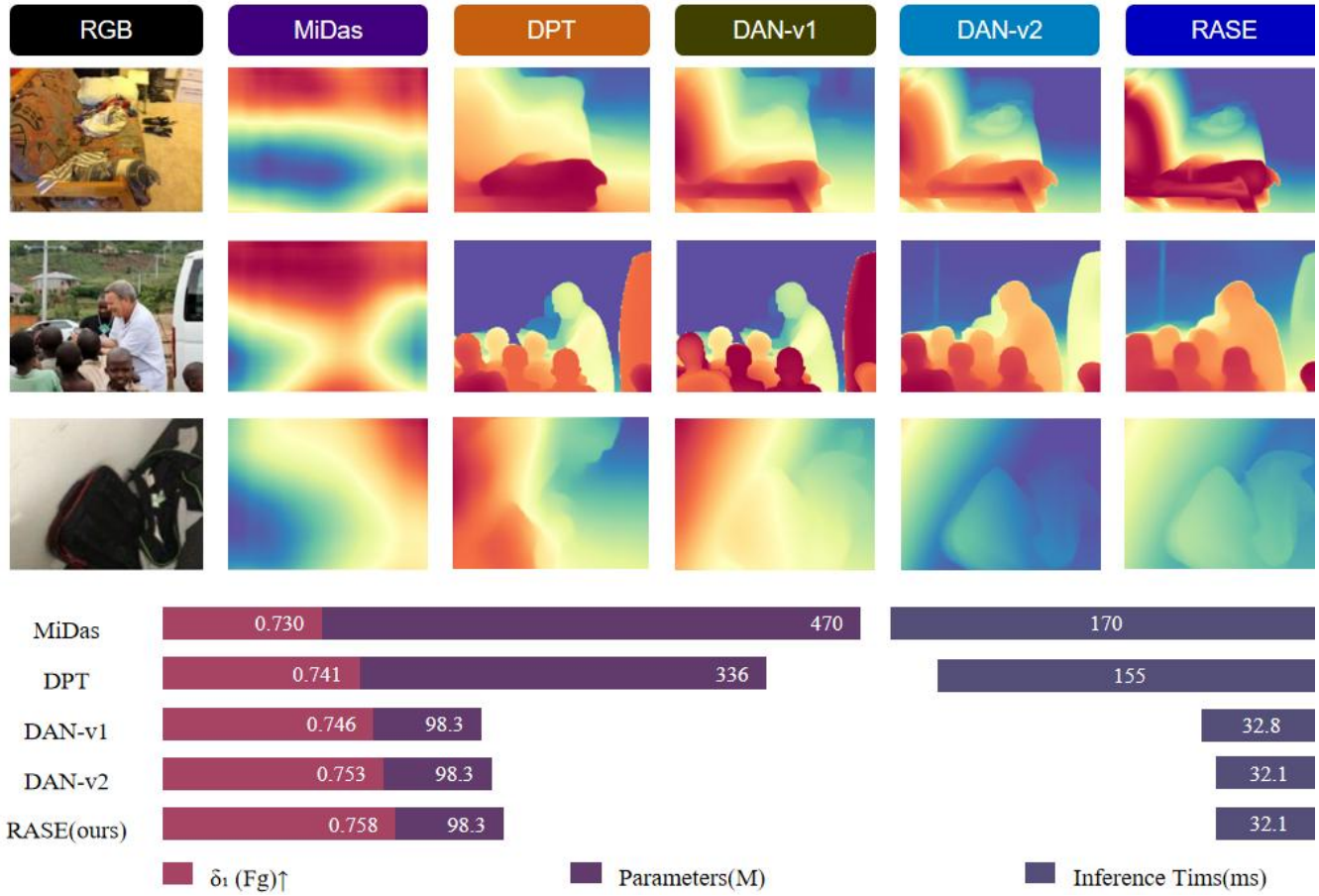

Figure 11. Structure-aware comparison under resolution constraints. Under degraded input quality, RASE-Net preserves sharper boundaries and finer geometry compared to prior methods, achieving the highest δ_1(Fg) without incurring additional latency or parameters.

To further assess model generality and structural robustness, we conduct a cross-dataset comparison of three structure-aware metrics—HER, GME, and EA—on NYU-v2, KITTI, and COCO2017 (Table 6). Across all datasets, RASE-Net exhibits the lowest GME, indicating improved geometric smoothness along gradients. On NYU-v2 and COCO2017, RASE-Net also achieves the highest EA (0.775, 0.760), suggesting more accurate alignment near object boundaries. While HER remains comparable to DAN-series models, the combination of lower GME and higher EA supports more consistent contour modeling across heterogeneous domains. Notably, DAN-v2 demonstrates competitive HER, but underperforms in EA on KITTI and GME on COCO2017, revealing tradeoffs in boundary fidelity.

Figure 11 further examines model performance under resolution degradation. Three representative low-resolution images are used to evaluate MiDas, DPT, DAN-v1, DAN-v2, and our RASE-Net. Despite the reduced visual quality, RASE-Net retains compact structural delineation, especially

at object silhouettes and occlusion fronts. Quantitatively, it yields the highest δ_1(Fg) (0.758) while maintaining the same parameter size (98.3M) and latency (32.1ms) as DAN-v2. In contrast, MiDas and DPT, although faster, exhibit blurred or over-smoothed predictions, leading to lower δ_1(Fg) (0.730 and 0.741, respectively). Together, these results validate the structural advantages of RASE-Net, particularly in scenarios with limited resolution or complex object layouts, without incurring computational overhead.

5. Conclusion

We propose RASE-Net, a residual attention refinement framework for monocular depth estimation, designed to reconcile the saliency – geometry conflict while enhancing structure-aware prediction. The network integrates two lightweight components: a residual attention unit for feature-level regularization and a saliency-guided enhancement block for output adaptation. Both modules operate in a plug-and-play manner without modifying backbone inference or introducing additional complexity. Comprehensive evaluations on NYU-v2, KITTI, and COCO2017 demonstrate that RASE-Net consistently enhances foreground delineation, contour sharpness, and geometric fidelity, while preserving real-time throughput. These results establish RASE-Net as an effective structure-aware augmentation for monocular depth models across diverse visual conditions.

Future work will explore scale-aware saliency modulation and task-adaptive structural priors to further enhance generalization and robustness. Given the improved spatial consistency and foreground fidelity of RASE-Net, we also see potential for applying the framework to downstream 3D reconstruction tasks, such as supporting more reliable depth-based scene modeling for autonomous perception systems.

Acknowledgements

xxx

References

[1] Valenti F, Giaquinto D, Musto L, et al. Enabling computer vision-based autonomous navigation for unmanned aerial vehicles in cluttered gps-denied environments[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 3886-3891.

[2] Saxena A, Chung S, Ng A. Learning depth from single monocular images[J]. Advances in neural information processing systems, 2005, 18.

[3] Chen H C. Monocular vision-based obstacle detection and avoidance for a multicopter[J]. IEEE Access, 2019, 7: 167869-167883.

[4] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12179-12188.

[5] Bhat S F, Alhashim I, Wonka P. Adabins: Depth estimation using adaptive bins[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4009-4018.

[6] Li Z, Wang X, Liu X, et al. Binsformer: Revisiting adaptive bins for monocular depth estimation[J]. IEEE Transactions on Image Processing, 2024.

[7] Yao J, Wu T, Zhang X. Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn[J]. arxiv preprint arxiv:2308.08333, 2023.

[8] Wofk D, Ma F, Yang T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 6101-6108.

[9] Lee J H, Han M K, Ko D W, et al. From big to small: Multi-scale local planar guidance for monocular depth estimation[J]. arxiv preprint arxiv:1907.10326, 2019.

[10] Yang L, Kang B, Huang Z, et al. Depth anything v2[J]. Advances in Neural Information Processing Systems, 2024, 37: 21875-21911.

[11] Wofk D, Ma F, Yang T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 6101-6108.

[12] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]//Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer Berlin Heidelberg, 2012: 746-760.

[13] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer vision – ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. Springer International Publishing, 2014: 740-755.

[15] Huang X, Fan L, Zhang J, et al. Real time complete dense depth reconstruction for a monocular camera[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 32-37.

[16] Ruhkamp P, Gao D, Chen H, et al. Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation[C]//2021 International Conference on 3D Vision (3DV). IEEE, 2021: 837-847.

[17] Ranftl R, Lasinger K, Hafner D, et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 44(3): 1623-1637.

[18] Butt M Z, Nasir N, Rashid R B A. A review of perception sensors, techniques, and hardware architectures for autonomous low-altitude UAVs in non-cooperative local obstacle avoidance[J]. Robotics and Autonomous Systems, 2024, 173: 104629.

[19] Zuo S, Xiao Y, Wang X, et al. Structure perception and edge refinement network for monocular depth estimation[J]. Computer Vision and Image Understanding, 2025: 104348.

[20] He Z, Zhang Y, Mu J, et al. LiteGfm: A Lightweight Self-supervised Monocular Depth Estimation Framework for Artifacts Reduction via Guided Image Filtering[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 8903-8912.
[21] Li P, Wu P, Yan X, et al. GeoDC: Geometry-Constrained Depth Completion With Depth Distribution Modeling[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.
[22] Zhang J. Survey on Monocular Metric Depth Estimation[J]. arxiv preprint arxiv:2501.11841, 2025.
[23] Zhang L, Lu J, Zheng S, et al. Vision transformers: From semantic segmentation to dense prediction[J]. International Journal of Computer Vision, 2024, 132(12): 6142-6162.
[24] Miclea V C, Nedevschi S. SemanticAdaBins-Using Semantics to Improve Depth Estimation based on Adaptive Bins in Aerial Scenarios[C]//2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). IEEE, 2024: 01-06.
[25] Li Z, Wang X, Liu X, et al. Binsformer: Revisiting adaptive bins for monocular depth estimation[J]. IEEE Transactions on Image Processing, 2024.
[26] Lee J H, Han M K, Ko D W, et al. From big to small: Multi-scale local planar guidance for monocular depth estimation[J]. arxiv preprint arxiv:1907.10326, 2019.
[27] Birkl R, Wofk D, Müller M. Midas v3. 1--a model zoo for robust monocular relative depth estimation[J]. arxiv preprint arxiv:2307.14460, 2023.
[28] Li S, Luo Y, Zhu Y, et al. Enforcing temporal consistency in video depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1145-1154.
[29] Li Z, Chen Z, Liu X, et al. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation[J]. Machine Intelligence Research, 2023, 20(6): 837-854.
[30] Yin B, Zhang X, Li Z, et al. Dformer: Rethinking rgbd representation learning for semantic segmentation[J]. arxiv preprint arxiv:2309.09668, 2023.
[31] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[32] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[33] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.
[34] Yu C, Xiao B, Gao C, et al. Lite-hrnet: A lightweight high-resolution network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10440-10450.
[35] Yang L, Zhang R Y, Li L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]//International conference on machine learning. PMLR, 2021: 11863-11874.
[36] Bochkovskii A, Delaunoy A Ã Ḡ, Germain H, et al. Depth pro: Sharp monocular metric depth in less than a second[J]. arxiv preprint arxiv:2410.02073, 2024.
[37] Ramamonjisoa M, Lepetit V. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.
[38] Dong X, Garratt M A, Anavatti S G, et al. Lightweight monocular depth estimation with an edge guided network[C]//2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV). IEEE, 2022: 204-210.
[39] Chen R, Luo H, Zhao F, et al. Structure-Centric Robust Monocular Depth Estimation via Knowledge Distillation[C]//Proceedings of the Asian Conference on Computer Vision. 2024: 2970-2987.
[40] Cheng Z, Zhang Y, Yu Y, et al. TinyDepth: Lightweight self-supervised monocular depth estimation based on transformer[J]. Engineering Applications of Artificial Intelligence, 2024, 138: 109313.
[41] Sikdar A, Gurunath P, Udupa S, et al. SAGA: Semantic-Aware Gray color Augmentation for Visible-to-Thermal Domain Adaptation across Multi-View Drone and Ground-Based Vision Systems[J]. arxiv preprint arxiv:2504.15728, 2025.
[42] Gaigalas J, Perkauskas L, Gricius H, et al. A Framework for Autonomous UAV Navigation Based on Monocular Depth Estimation[J]. Drones, 2025, 9(4): 236.
[43] Pirvu M, Robu V, Licaret V, et al. Depth distillation: unsupervised metric depth estimation for UAVs by finding consensus between kinematics, optical flow and deep learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3215-3223.
[44] Jiang W, Li Y, Yi Z, et al. Multi-instance imbalance semantic segmentation by instance-dependent attention and adaptive hard instance mining[J]. Knowledge-Based Systems, 2024, 304: 112554.
[45] Li Y, Zhang H, Jia W, et al. Saliency guided naturalness enhancement in color images[J]. Optik, 2016, 127(3): 1326-1334.
[46] Kim J H, Lee J S. Deep residual network with enhanced upscaling module for super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 800-808.
[47] Ghysels E, Sinko A, Valkanov R. MIDAS regressions: Further results and new directions[J]. Econometric reviews, 2007, 26(1): 53-90.
[48] Yang L, Kang B, Huang Z, et al. Depth anything: Unleashing the power of large-scale unlabeled data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 10371-10381.