SharpEdge: High-Quality Data-Driven Monocular Depth Estimation for Enhanced Boundary Precision

Mengke Song[a,b], Luming Li[a,b], Xu Yu[a,b], Chenglizhao Chen[a,b,*], Shanchen Pang[a,b,c]

[a] *College of Computer Science and Technology and School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, P. R. China*
[b] *Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software, Qingdao, P. R. China*
[c] *State Key Laboratory of Chemical Safety, Qingdao, P. R. China*

Abstract

While existing monocular depth estimation methods have achieved commendable performance, they often fall short in accurately distinguishing object boundaries. This deficiency largely stems from the inherent noise in dataset acquisition, such as unclear edges and missing depth information. To address these challenges, this paper introduces a novel, high-quality, data-driven monocular depth estimation method tailored for autonomous driving. The approach significantly enhances depth predictions with clearer object boundaries and reduced noise, making it well-suited for real-time, safety-critical applications. Central to our approach is the Self-Adaptive Consistency Filtering mechanism, which dynamically selects high-quality training samples, ensuring that the model learns from the most reliable data and reducing the impact of noise. Additionally, we introduce a Dual-Prior Learning strategy that combines geometric and semantic edge priors. Unlike traditional methods that rely solely on raw depth maps, our approach enhances boundary detection by providing detailed guidance on object contours. This leads to more accurate depth estimation, especially in complex regions where other methods struggle. Empirical evaluations on popular benchmark datasets show that our approach leads to performance improvements of 1.2% on the autonomous driving dataset KITTI and 1.8% on the indoor scene dataset NYU. Compared with recent state-of-the-art methods such as DPT and NewCRFs, our approach achieves superior performance, particularly in recovering fine object boundaries and maintaining spatial consistency across diverse scenes. These results highlight the strong generalization ability of our method, demonstrating that it can enhance depth estimation quality across diverse environments — improving edge precision and spatial coherence, which are critical for autonomous vehicles navigating both complex and dynamic scenarios.

Keywords: Monocular Depth Estimation, Data-Driven, Dual-Prior Learning

*Corresponding author

Email addresses: songsook@126.com (Mengke Song), liluming1224@126.com (Luming Li), yuxu0532@upc.edu.cn (Xu Yu), cclz123@163.com (Chenglizhao Chen), pangsc@upc.edu.cn (Shanchen Pang)

The first two authors contributed equally to this work.

1. Introduction

Monocular depth estimation is a fundamental task in computer vision with wide-ranging applications such as autonomous driving [1, 2, 3], and virtual/augmented reality [4]. It aims to infer dense depth information from a single image, which is an inherently challenging problem due to the absence of stereo cues. Recent advances [5, 6] have seen significant progress in this field, primarily driven by the development of deep learning models and the availability of large-scale datasets. Nevertheless, despite the commendable performance achieved by existing methods, accurately distinguishing object boundaries remains a persistent issue. This problem becomes particularly evident in complex scenes with intricate geometric structures and varying texture patterns, where current models tend to produce over-smoothed depth predictions and fail to delineate object edges precisely.

The core reason behind this shortcoming can be traced back to the quality of the training data. In practice, most publicly available depth datasets such as KITTI [7] and NYU [8] are either captured using active sensors such as LiDAR or generated through stereo vision techniques. Although effective in providing dense depth annotations, these approaches are often undermined by noise and inconsistencies. For instance, sensors may produce unclear edges, especially for small or distant objects, and may fail to capture depth accurately in reflective or transparent regions. As a result, the learning process becomes susceptible to erroneous supervision, limiting the model's ability to capture sharp object boundaries and leading to blurred depth transitions (Fig. 1-A). Such noisy labels not only hamper the representational power of deep models but also introduce ambiguities that degrade the overall prediction quality.

Specifically, we point out that blurred depth edges introduce a form of "supervision uncertainty", where the same visual contour may exhibit inconsistent depth gradients across different training samples. This inconsistency disturbs the model's convergence path and results in unstable boundary predictions. Moreover, missing depth values — commonly found near transparent surfaces, reflective objects, or distant regions — lead to incomplete spatial information. This breaks the geometric continuity of the scene, undermining the model's ability to learn contextual depth relationships. Prolonged exposure to such noisy supervision encourages the network to overfit to unreliable texture cues and results in feature confusion, ultimately degrading the model's generalization capacity — especially in scenes with complex structures or occlusions.

To mitigate these issues, recent studies have explored various strategies, such as improving network architectures [9, 10] or introducing additional supervision signals, e.g., resorting multi-modality learning [11, 12] (Fig. 1-B). However, these approaches predominantly focus on optimizing model capacity without sufficiently addressing the underlying data quality problem. Simply increasing model complexity cannot compensate for suboptimal training data, as the network may continue to overfit to unreliable samples, leading to poor generalization and coarse boundary localization. Thus, a more effective solution should involve refining the training data to ensure that the model learns from high-quality, reliable examples while minimizing the adverse impact of ambiguous samples.

In this work, we propose a novel high-quality data-driven monocular depth estimation method called SharpEdge that addresses these challenges by enhancing both data quality and boundary precision (Fig. 1-C). The core of our approach is the self-adaptive consistency

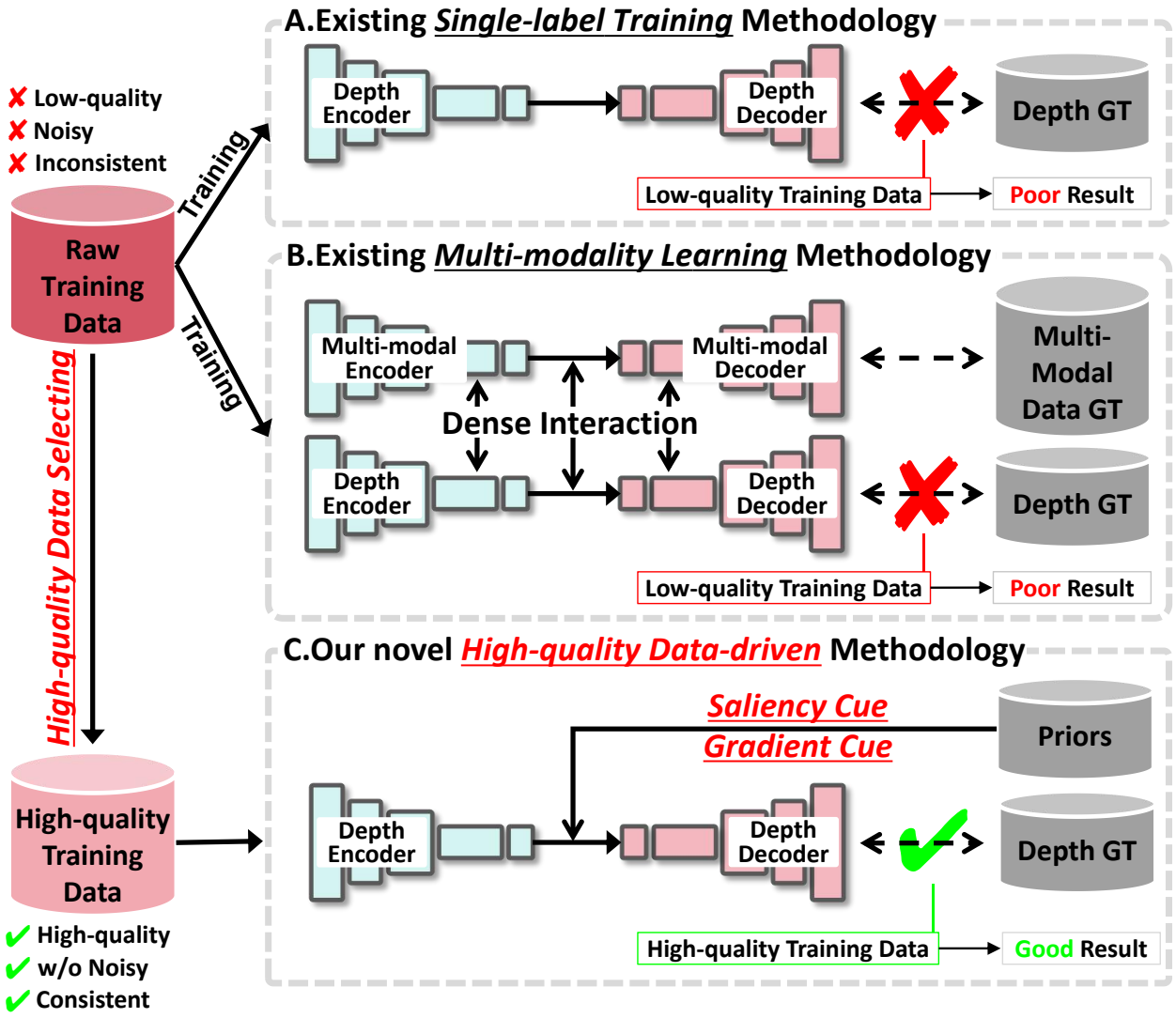

Figure 1: In monocular depth estimation, while existing single-label training approaches (**A**) use all training samples as input, and ground-truth depth maps as the sole supervision signal, multi-modality learning (**B**) is employed to enhance the accuracy and robustness of depth estimation by utilizing diverse input data. However, they often fall short in accurately distinguishing object boundaries. In contrast, our novel high-quality data-driven methodology (**C**) enhances the overall performance of depth estimation, especially in generating sharp depth edges, by selecting high-quality "valuable" data in training samples and incorporating two priors for additional guidance.

filtering (SACF) mechanism, a dynamic data selection technique designed to automatically evaluate and filter out low-quality samples during the training process. Unlike traditional hard-threshold filtering methods that rely on fixed criteria, SACF dynamically assigns a "value" to each training sample based on its consistency and informativeness, allowing the model to focus on the most reliable data. By progressively refining the training set, SACF ensures that the network learns from high-quality examples, thereby improving its robustness against noise and ambiguities.

In addition to data quality enhancement, we introduce a dual-prior learning (DPL) strategy to improve boundary precision further. Conventional depth estimation methods [13, 14] often employ single-label supervision, typically using raw depth maps or edge detectors,

which capture abrupt intensity changes but often fail to represent complex structures and context accurately, leading to noisy or fragmented boundaries. This limitation is particularly detrimental in scenes with occlusions or regions where depth variations are subtle but crucial for accurate perception. Our DPL approach addresses these challenges by incorporating two distinct yet complementary types of boundary information: gradient clues and saliency clues. 1) The gradient clues act as a geometric prior that captures fine structural details by focusing on pixel-wise changes in depth values. Unlike conventional edge cues, gradient clues provide a continuous representation of the underlying geometric structure, enabling the model to recognize nuanced depth variations and smoothly transition between object boundaries. This continuous nature makes gradient clues more resilient to noise and better suited for modeling complex depth changes than binary edge labels, often resulting in hard boundaries that lack interpretative depth cues. 2) The saliency clues, on the other hand, serve as a semantic prior that highlights visually essential regions in the image, such as prominent object boundaries and critical structural elements. Saliency maps encode high-level contextual information, allowing the model to focus on perceptually meaningful edges rather than being distracted by texture patterns or background clutter. By emphasizing semantically relevant regions, the model gains a better understanding of object contours, enabling it to distinguish between true object boundaries and misleading textures that traditional edge cues might erroneously capture. By combining these complementary priors, DPL provides more accurate and context-aware guidance, resulting in clearer and more precise depth predictions.

By integrating high-quality data selection with dual-prior guidance, our method significantly improves the clarity and precision of monocular depth estimations, particularly in challenging scenarios where conventional approaches struggle. Extensive experimental results demonstrate that our model achieves state-of-the-art performance on standard benchmarks and exhibits superior boundary delineation and robustness against noisy data, making it well-suited for practical applications in complex visual environments.

The main contributions of this work are then summarized as follows:

- We present a novel high-quality data-driven approach for monocular depth estimation, which achieves enhanced depth predictions with sharper object boundaries and reduced noise by tackling the limitations of existing training data. This approach can also be seamlessly integrated as a plug-and-play module into existing depth estimation frameworks to boost their performance;

- We introduce a self-adaptive consistency filtering mechanism for dynamic training data refinement, which automatically filters low-quality samples based on their consistency and informativeness, thereby improving model robustness against noisy data;

- We propose a dual-prior learning strategy that integrates geometric and semantic edge priors to provide comprehensive guidance for refining depth estimates at object boundaries, leading to improved edge delineation and detail preservation;

- Comprehensive experiments on the KITTI and NYU datasets demonstrate that our method significantly improves boundary accuracy and depth quality, resulting in smoother and more distinct edges compared to previous methods.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation is critical for autonomous vehicles, as it allows them to perceive the 3D structure of their environment for navigation, obstacle avoidance, and path planning. Existing methods [15, 16] rely on supervised learning, using ground-truth depth maps for training. However, this approach faces challenges like edge blurriness and occlusion, which are especially problematic in dynamic, real-world driving environments.

To address these issues, researchers have explored self-supervised learning [5, 6], multi-task learning [17], and semantic segmentation [18]. These methods improve robustness by integrating additional information, such as vehicle localization and scene segmentation, which are crucial for autonomous driving. However, obtaining labeled data for autonomous vehicles is costly and time-consuming. As a result, unsupervised [19] and self-supervised learning [20] methods have gained traction. Datasets like KITTI [7] and NYU [8] are commonly used, but they struggle to capture the full complexity of real-world driving conditions.

Multi-modality learning, which combines data from multiple sensors such as LIDAR, radar, and thermal cameras, has shown promise in improving depth estimation accuracy. However, relying on specialized sensors [21] can limit the practicality of these methods for all autonomous vehicles. Additionally, complex techniques like self-attention [22] and precise sensor calibration [23] are resource-intensive and may not be suitable for real-time applications in resource-constrained autonomous vehicles.

Despite progress, monocular depth estimation methods often fail to accurately capture object boundaries in complex environments, a critical challenge for autonomous vehicles when making real-time navigation decisions [24, 25, 26]. Blurry edges and missing depth information can hinder the ability to safely navigate crowded streets or detect obstacles.

2.2. Edge-related Monocular Depth Estimation

Existing depth estimation methods, while effective, often struggle with accurately capturing complex object boundaries due to noisy and incomplete data during acquisition, a challenge that is amplified in the context of autonomous vehicles. Several methods address this by incorporating edge-aware strategies [27, 10, 28]. For example, CutDepth focuses on augmenting data with preserved edge features [29], but lacks a mechanism to filter out low-quality samples, which limits its effectiveness under real-world noise encountered by autonomous vehicles in dynamic driving scenarios. Methods like Edge Defocus Tracking [30] and Edge-aware Loss Functions [31] emphasize boundary accuracy but are unable to dynamically prioritize reliable data, leading to unstable learning when vehicles are navigating through cluttered or occluded environments.

Similarly, advanced strategies like Bi-directional Diffusion [12] and ESPDepth [11] integrate edge information into depth propagation, yet fail to balance fine-grained geometric details with semantic consistency, resulting in blurred edges in regions with complex object interactions, which are commonly encountered in urban driving scenarios. For example, accurately distinguishing the boundaries between vehicles and pedestrians at intersections or during overtaking maneuvers is crucial for safe navigation in autonomous vehicles.

In contrast, our proposed method tackles these issues with self-adaptive consistency filtering to exclude noisy data and a dual-prior learning strategy that combines geometric

and semantic priors for sharper boundary delineation. This approach leads to clearer, more accurate depth predictions, especially in complex regions where traditional methods often fail, such as when navigating dense urban environments or detecting objects with partial occlusions. Our method, by dynamically prioritizing high-confidence depth data and incorporating both geometric structures and semantic cues, ensures that depth maps generated for autonomous vehicles are more reliable for real-time decision-making.

2.3. Saliency-related Monocular Depth Estimation

Saliency-related depth estimation enhances depth accuracy by incorporating saliency, which signifies the visual prominence of regions or objects in an image [32, 33, 34, 35]. This approach is particularly practical in complex scenes. Zhao et al. [36] develop a model that integrates depth estimation, salient object detection, and contour estimation, enhancing depth accuracy through multi-task learning. Chen et al. [37] introduce a two-phase approach, initially estimating depth from similar images and refining it with saliency cues. Ji et al. [38] propose a calibration module that merges raw and estimated depth maps for improved reliability. Despite advancements, challenges persist in saliency-related depth estimation. Depth images are often noisy, and uncertainties at object boundaries are typical due to sensor limitations and environmental factors like occlusion and reflection. Moreover, while current methods focus on fusing RGB and depth data to improve saliency detection, these strategies may not fully leverage the complementary nature of both modalities, indicating potential areas for enhancement.

3. Proposed Method

3.1. Network Architecture

The proposed method consists of two main components. The first part (Part B in Fig. 2), referred to as self-adaptive consistency filtering (Sec. 3.3), focuses on "valuable" data (depth with clear edges and less noise) filtering. It takes the initial data pool (including RGB and Depth) as input and aims to filter out "useless" training samples (depth with unclear edges and more noise) while selecting the most "valuable" ones. This process reduces the impact of redundant information and minimizes training resource consumption, resulting in an enhanced data pool containing more purified training samples. These selected training samples are subsequently fed into the second part, dual-prior learning (Sec. 3.4), depicted in Fig. 2-Part C. This part is designed to train a model leveraging both geometric and semantic edge priors (saliency and gradient) to further improve depth accuracy at object boundaries. A detailed introduction of each part is provided in the following sections.

3.2. Preliminaries of Saliency

This paper introduces SharpEdge, a novel monocular depth estimation framework designed to generate dense depth maps with enhanced edge sharpness without requiring additional annotation data and even leveraging fewer training data samples. Specifically, we use saliency to select "valuable" data.

Several approaches can be adopted to leverage the benefits of saliency detection in monocular depth estimation. The first approach incorporates saliency maps or cues as additional

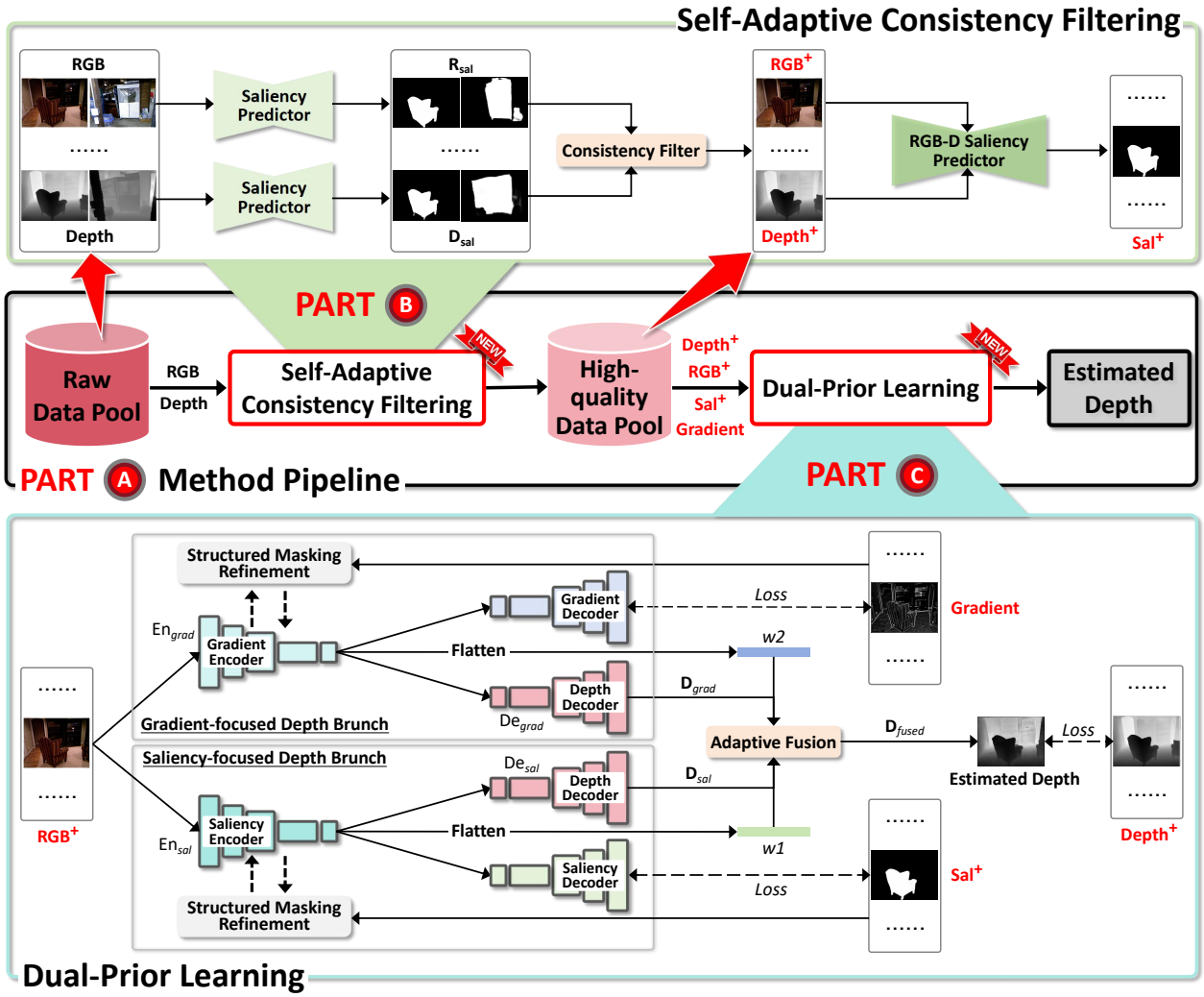

Figure 2: **Pipeline of the proposed SharpEdge.** SharpEdge comprises two key components: self-adaptive consistency filtering (**Part B**) and dual-prior learning (**Part C**). The former filters the initial RGB-depth data pool to retain high-quality training samples, while the latter trains a two-branch network guided by saliency and gradient priors. The resulting depth maps are fused via an adaptive strategy to produce sharp, accurate depth predictions. Additionally, a structured masking refinement module enhances robustness by enabling the model to recover missing or occluded information.

input to depth estimation algorithms. Saliency maps highlight visually salient regions or objects in an image. By considering these regions during depth estimation, the algorithm can prioritize the extraction of depth information from the most critical areas. This approach can lead to more accurate depth estimation, particularly in complex scenes where salient objects are crucial. The second approach uses saliency information to guide the depth estimation process. By integrating saliency cues directly into the depth estimation algorithm, the algorithm can prioritize depth estimation for salient regions or objects. This incorporation of saliency guidance helps refine the depth estimation results and improves the overall quality of the resulting depth map.

3.3. Self-Adaptive Consistency Filtering

Technical Rationale. In monocular depth estimation tasks, the quality of training data is crucial to the final model performance. However, existing depth datasets (e.g., KITTI,

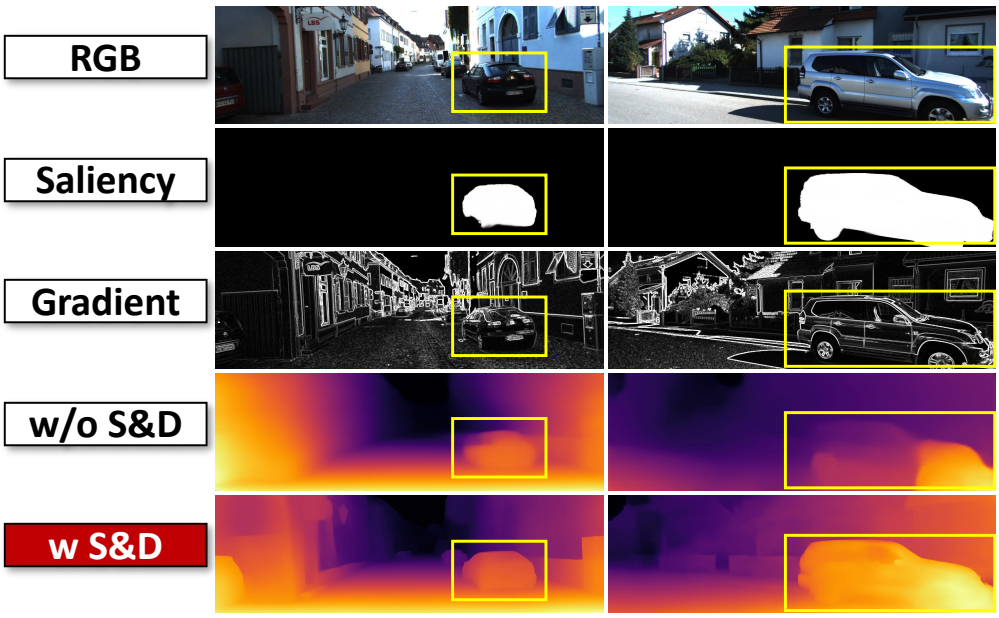

Figure 3: Illustration of saliency maps and gradient maps on the KITTI dataset. "w S&D" and "w/o S&D" means with and without the saliency and gradient guidance, respectively.

NYU) are often generated using complex sensing equipment (such as LiDAR, stereo cameras, or structured light systems), which are prone to sensor noise, occlusion, or environmental conditions (e.g., changes in lighting), leading to varied quality in the depth maps. Directly using these "low-quality" samples for training often degrades the model's learning capability, making it difficult to accurately predict depth information, especially around object boundaries, fine details, or complex scenes. Thus, automatically filtering out "high-quality" and "informative" training samples from existing datasets becomes a key challenge in the model design.

In this context, the self-adaptive consistency filtering (SACF) mechanism is introduced to improve the purity of the training dataset. This mechanism leverages saliency prediction models to evaluate the consistency between the RGB and depth maps in salient regions, thereby selecting samples where RGB and depth maps have strong similarity in saliency (i.e., "high-quality" samples). This filtering strategy aims to remove redundant samples, reduce the negative impact of low-quality depth maps, and focus on informative salient regions, ultimately producing a more reliable training dataset.

Technical Detail. As shown in Part B of Fig. 2, given any RGB and depth pairs in existing training datasets of monocular depth estimation, we feed them into the saliency predictor (a SOTA salient object detection model) to produce RGB saliency ($\mathbf{R}_{sal}$) or depth saliency ($\mathbf{D}_{sal}$) maps. Saliency maps highlight an image's most visually prominent regions, such as object edges, shapes, or regions of interest. By aligning the saliency maps from the RGB and depth images, SACF ensures that only those samples with high structural correspondence are selected, thus providing the training data is both accurate and consistent. While RGB saliency maps are often reliable, depth saliency maps can vary significantly depending on the quality of the depth data. Ensuring the saliency quality when the depth is low quality is almost infeasible. Thus, we only retain those RGB with high-quality depth maps and exhibit substantial similarity between their RGB and depth saliency as final purificatory training

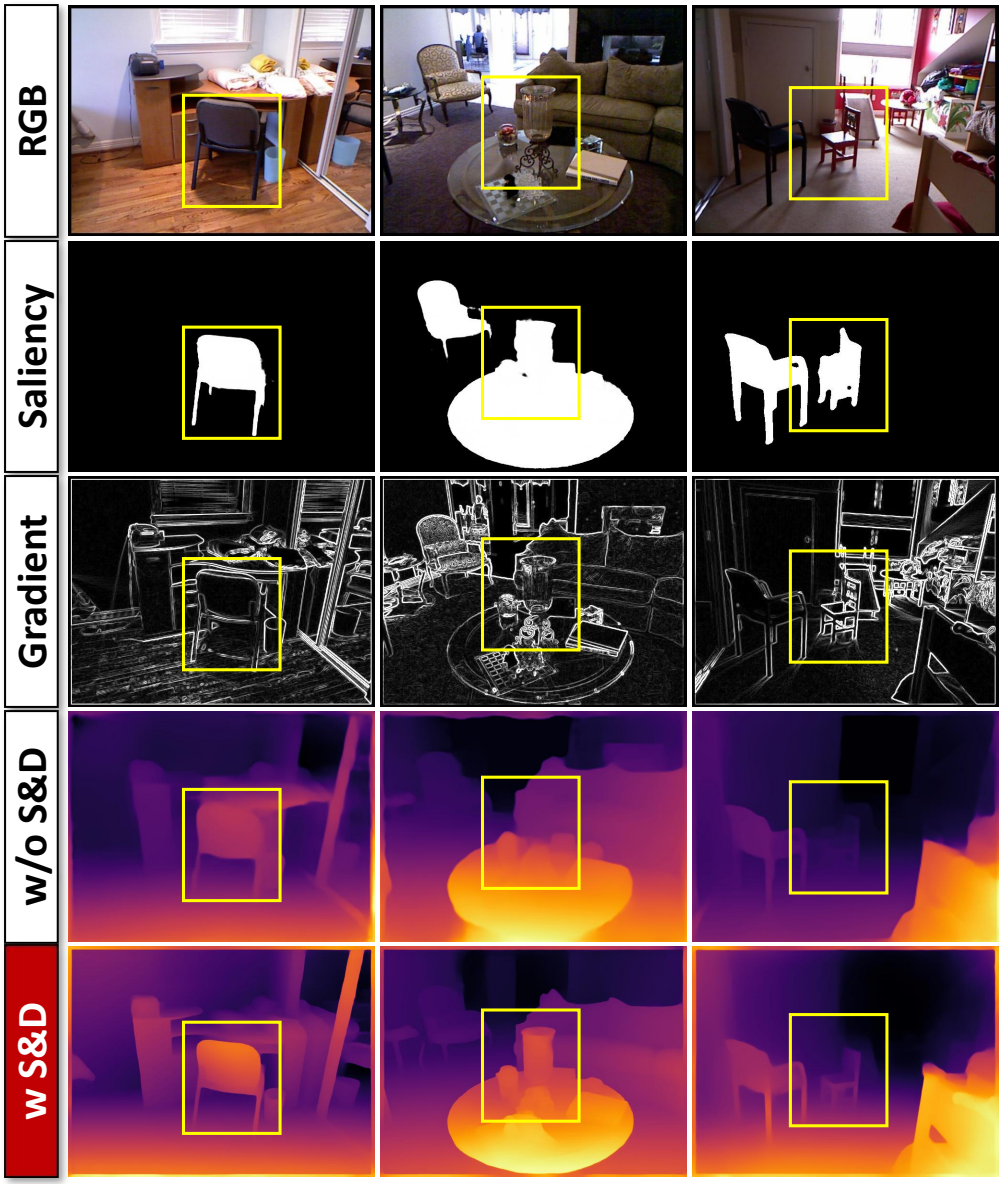

Figure 4: Illustration of saliency maps and gradient maps on the NYU dataset. "w S&D" and "w/o S&D" means with and without the saliency and gradient guidance, respectively.

samples. To automate the evaluation of similarity between RGB and depth saliency maps, a consistency filter is introduced as a "discriminator" that determines whether the given RGB and depth saliency maps have sufficient structural consistency. The consistency filter is a simple classification network consisting of two feature encoders (e.g., ResNet34) and a multi-perception layer, which takes RGB and depth saliency as input, outputting similarity scores ("1" indicates high similarity, "0" indicates low similarity).

To ensure that the consistency filter can effectively distinguish the similarity between RGB and depth saliency maps, the study uses the S-measure (Structural Similarity Measure) as the supervision signal during training. S-measure is a metric designed to evaluate saliency map quality by measuring the structural similarity between saliency maps, considering salient

regions' precision and structural completeness. This process can be formulated as:

$$\mathrm{SS}(\mathbf{R}_{sal}, \mathbf{D}_{sal}) = \begin{cases} 1, \text{ if } \ \mathrm{S_m}(\mathbf{R}_{sal}, \mathbf{D}_{sal}) - \gamma \geq 0 \\ 0, \text{ otherwise} \end{cases}, \tag{1}$$

where $\mathrm{SS}(\cdot)$ means calculating the similarity between RGB and depth saliency maps. γ is a predefined similarity threshold (we empirically set it to 0.8[3]), and $\mathrm{S_m}(\cdot)$ denotes S-measure. By comparing the predicted similarity scores with the similarity labels obtained from the S-measure calculation ("1" or "0"), a cross-entropy loss is used to optimize the weights of the consistency filter, making it more accurate in judging the similarity of saliency map pairs.

There are four cases when generating final purificatory training samples, *i.e.*, "high/low-quality depth + strong/weak RGB saliency and depth saliency similarity". Since the low-quality depth and strong consistency combination is rare in practice, we have omitted it. Only the high-quality and strong-similarity cases can ensure accurate and robust training samples. In this way, we have obtained high-quality RGB and depth data, which are denoted as RGB^+ and Depth^+, respectively. The visualization results are shown in Fig. 3 and Fig. 4.

Technical Summary. The self-adaptive consistency filtering (SACF) methodology is designed to automate identifying high-quality training samples in monocular depth estimation tasks, addressing the inherent limitations in existing depth datasets. By leveraging saliency prediction models and a custom consistency filter, SACF measures the structural consistency between RGB and depth maps, using this measure to retain only those samples that exhibit strong structural alignment selectively. This selective filtering strategy enhances the purity of the training data and ensures that the model focuses on learning meaningful depth features, thus improving overall performance, especially in complex scenarios.

3.4. Dual-Prior Learning

Technical Rationale. In conventional monocular depth estimation tasks, the training paradigm primarily depends on a single supervision signal, such as depth maps. However, relying solely on depth labels often fails to capture fine-grained details and accurate structural information, especially in complex RGB scenes or when occlusions occur. Furthermore, noise, unreliable annotations, and a limited number of training samples frequently affect the obtained depth maps. While previous mentioned self-adaptive consistency filtering approach (Sec. 3.3) that filter out "valuable" training samples can enhance data quality, this strategy alone cannot address the lack of sharp boundaries and fine details in the resulting depth maps.

To address these issues, we propose a dual-prior learning strategy, which integrates saliency and gradient as complementary priors to guide the depth estimation process. The saliency prior helps the model focus on visually prominent regions, such as object boundaries. At the same time, the gradient prior emphasizes edge continuity and sharpness in the depth maps. By simultaneously leveraging these two distinct priors, the model gains a more comprehensive understanding of the scene, leading to more accurate and sharper depth predictions. More detailedly, the semantic prior (saliency) emphasizes global visual attention, guiding the model to focus on semantically important regions — typically foreground objects

[3]Ablation study can be seen in Table 7.

and their overall contours. On the other hand, the geometric prior (gradient) captures local structural discontinuities at a fine-grained level, helping the model identify detailed transitions and depth boundaries that may not be semantically prominent. We further elaborate that, in practice, these two types of edge cues often occupy complementary spatial regions. While saliency tells the model which areas are important, gradient focuses on where the depth transitions occur. By modeling them through two parallel encoders and decoders, and then adaptively fusing their outputs via the Adaptive Fusion module (Sec. 3.4.3), our network dynamically balances the contribution of each prior based on scene context—resulting in clearer, more continuous, and structurally accurate boundaries in the final depth predictions.

Note that compared with existing multi-modality learning methods, the superiority of our proposed dual-prior learning[4] is its ability to achieve higher performance without requiring additional annotation data and even leveraging fewer training data samples (see Table 2). The dual priors used in our method are also readily available.

Technical Detail. To implement the dual-prior learning process, we utilize the outputs of the self-adaptive consistency filtering (Sec. 3.3), denoted as RGB^{+} and Depth^{+} in Fig. 2-Part C.

3.4.1. Saliency/Gradient-focused Depth Branch

The dual-prior learning process contains two branches, saliency-focused depth branch and gradient-focused depth branch, which separately utilize two distinct encoders: a saliency encoder (En_{sal}) and a gradient encoder (En_{grad}). Each encoder is designed to extract high-level features unique to its respective prior. The saliency encoder focuses on capturing visual contrast, object boundaries, and salient semantics in the image. This encoder processes the purified RGB input and generates a feature map highlighting visually significant regions, making it easier for the network to understand object boundaries and critical regions in the scene. On the other hand, the gradient encoder captures depth discontinuities and sharp changes in the scene's structure, which are crucial for maintaining accurate edge information. This encoder is trained to emphasize areas where depth values change abruptly, helping the model retain sharp boundaries in its predictions. By employing two separate encoders, the network can specialize in learning distinct aspects of the scene structure, ensuring that both visual attention and geometric continuity are accurately represented.

Each prior encoder is connected to two decoders: one for generating saliency/gradient maps and another for generating depth maps guided by the corresponding prior. Specifically, the saliency encoder is paired with a saliency decoder and a saliency-focused depth decoder (De_{sal}). In contrast, the gradient encoder is paired with a gradient decoder and a gradient-focused depth decoder (De_{grad}). The saliency decoder outputs saliency maps that highlight regions of high visual attention, ensuring the saliency-prior depth maps ($\mathbf{D}_{sal}$), generated by the saliency-focused depth decoder, focus on visually prominent areas. The gradient decoder outputs gradient maps that emphasize structural edges and depth discontinuities, guiding the generation of sharp gradient-prior depth maps ($\mathbf{D}_{grad}$) to preserve boundary details and prevent blurred edges in the final depth estimation.

[4]The in-depth analysis of multi-modality learning and dual-prior learning can be seen in Sec. 4.6.

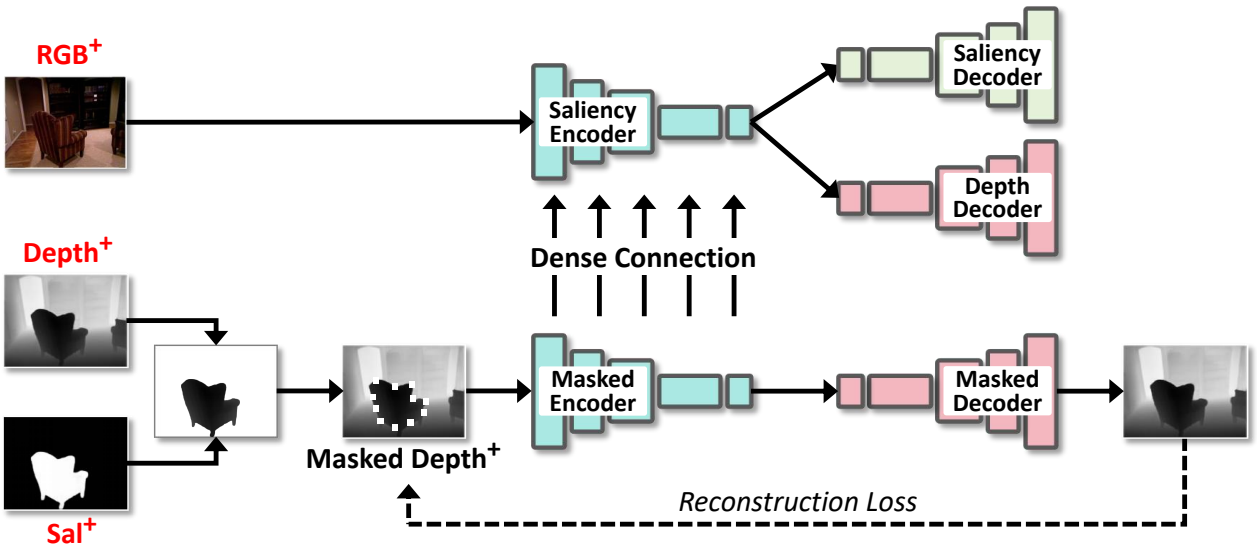

Figure 5: Flowchart of structural masked refinement (Sec. 3.4.2). By simulating partial data loss during training, this strategy forces the model to reconstruct and compensate for masked areas, making it more adept at handling occlusions and partial data loss during inference.

3.4.2. Structured Masking Refinement

In the previously discussed self-adaptive consistency filtering (Sec. 3.3) and dual-prior learning strategies (Sec. 3.4), we effectively enhance the monocular depth estimation model's ability to predict critical edges by leveraging saliency and gradient priors. However, these methods rely on complete RGB and depth data during training, encouraging the model to prioritize global features. As a result, the model may treat all regions uniformly, which can reduce its attention to local edge details and structures.

To mitigate this, we propose a structured masking strategy that compels the model to learn how to compensate for critical features even when complete information is unavailable. By simulating partial data loss during training, this strategy forces the model to reconstruct and compensate for masked areas, making it more adept at handling occlusions and partial data loss during inference. This results in a more robust model that better captures local structures and infers missing depth information.

Unlike the random masking used in traditional masked autoencoders [39], our structured masking strategy targets the salient object boundaries in depth maps. By focusing specifically on discontinuous regions and salient details, we guide the network to pay closer attention to these critical areas, enhancing overall prediction accuracy. This strategy is integrated into gradient-focused and saliency-focused depth branches, acting as a self-supervised refinement step that challenges the network to reconstruct masked regions using contextual information.

The process begins with generating a structured mask ($\mathbf{M}_{struct}$). Given a depth map, we first apply a saliency detection method to generate a foreground saliency map ($\mathbf{S}_{fg}$), highlighting prominent objects in the scene. We then use an edge detection operation (e.g., a Canny filter) to produce an edge mask ($\mathbf{M}_{edge}$), marking the boundaries of these salient regions. By combining the saliency map and the edge mask, we create a structured mask that specifically targets boundary regions of the depth map.

This structured mask ($\mathbf{M}_{struct}$) is then used in a masked encoder-decoder network, which aims to reconstruct the missing portions of the depth feature map using contextual cues from

the available data. Importantly, this network has the same encoder-decoder architecture as the Saliency/Gradient-focused Depth Branch, ensuring consistency across the model.

The reconstruction is supervised by a reconstruction loss ($\mathcal{L}_{recon}$), such as Mean Squared Error (MSE), which measures the difference between the predicted depth map ($\hat{\mathbf{D}}_{masked}$) and the original unmasked depth map ($\mathbf{D}_{original}$):

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i \in \mathbf{M}_{\text{struct}}} \left| \mathbf{D}_{\text{original}}(i) - \hat{\mathbf{D}}_{\text{masked}}(i) \right|^2, \tag{2}$$

where N represents the total number of pixels masked by the structured mask $\mathbf{M}_{\text{struct}}$, i denotes the index of pixels masked by $\mathbf{M}_{\text{struct}}$. The summation $\sum_{i \in \mathbf{M}_{\text{struct}}}$ indicates that the loss is only computed for the pixel locations covered by the structured mask.

To further improve feature learning, we introduce dense connections between intermediate features of corresponding encoder layers in the masked encoder-decoder network and those in the saliency/gradient-focused depth branch via a *Concatenation + Convolution* operation. This facilitates the flow of both low-level and high-level features across multiple layers, enhancing the model's ability to reconstruct masked areas and improving overall depth prediction.

3.4.3. Adaptive Fusion

After generating the separate saliency-prior and gradient-prior depth maps, we employ an adaptive fusion scheme to combine these complementary predictions into a final refined depth map ($\mathbf{D}_{fused}$), as shown in Fig. 6. This fusion scheme dynamically balances the contributions of the two priors based on learned weights, which are defined as follows:

$$\mathbf{D}_{fused} = w1 * \mathrm{g}(\mathbf{D}_{sal}) + w2 * \mathrm{g}(\mathbf{D}_{grad}), \tag{3}$$

where $\mathrm{g}(\cdot)$ represents the sigmoid function to map the depth maps between 0 and 1, ensuring smoothness and continuity in the fusion. $w1$ and $w2$ are adaptive weights calculated from the Multi-Layer Perception (MLP) outputs of the saliency and gradient encoders, respectively:

$$w1 = \mathrm{MLP}(\mathrm{En}_{sal}(\mathrm{RGB}^{+})), \quad w2 = \mathrm{MLP}(\mathrm{En}_{grad}(\mathrm{RGB}^{+})), \tag{4}$$

where $\mathrm{En}_{sal}(\cdot)$ and $\mathrm{En}_{grad}(\cdot)$ separately denote the depth encoder equipped with saliency decoder and gradient decoder. RGB^{+} denotes the output of the self-adaptive consistency filtering (Sec. 3.3). These weights are learned to adaptively emphasize the more reliable prior in each region, enabling the model to handle complex scenarios where one prior may be more informative than the other.

The rationale for employing two distinct depth encoders, rather than integrating saliency and gradient into a single depth encoder, is as follows: 1) By separating saliency and gradient into distinct branches, the model can specialize each encoder to focus on unique features, thereby avoiding interference between the priors. For example, the saliency encoder can prioritize visual contrast and object prominence, while the gradient encoder can concentrate on depth discontinuities and edge consistency; 2) With dual-prior learning, the fusion module can adaptively adjust the influence of each prior based on the scene's context. For instance, in regions with complex textures or overlapping objects, the saliency prior may be more critical, while in regions with clear geometric structures, the gradient prior may take precedence.

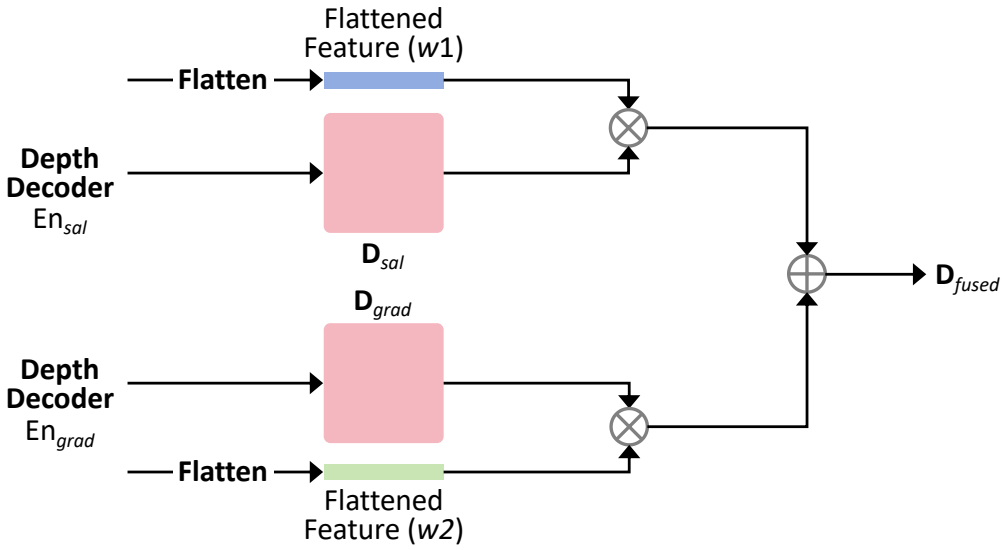

Figure 6: Flowchart of adaptive fusion (Sec. 3.4.3), which aims to integrate the two estimated depth maps adaptively.

While our adaptive fusion strategy shares the general idea of learning dynamic weights to balance multiple information sources — similar to recent attention-based fusion approaches such as MMAD [40], ViGT [41], WaveFormer [42], BVINet [43], and Deep Stereo Video Inpainting [44] — it differs in design motivation, structure, and application focus. These prior methods typically adopt global attention mechanisms or learnable tokens within transformer-based architectures, aiming to capture long-range dependencies and semantic correspondences across time or modalities. In contrast, our method is tailored for boundary-aware monocular depth estimation, where the goal is to adaptively integrate geometric (gradient) and semantic (saliency) priors at a pixel level. To this end, we employ a lightweight dual-branch MLP that produces spatially varying fusion weights (w1, w2), enabling local structure-aware integration without incurring the high computational cost of global attention. This design emphasizes interpretability and efficiency, particularly for high-resolution depth prediction.

Technical Summary. The proposed dual-prior learning framework leverages saliency and gradient priors to enhance monocular depth estimation. Separate branches capture distinct features: saliency for visual prominence and gradient for sharp edges. To further refine predictions, a structured masking refinement module simulates partial data loss, training the model to recover masked regions and improve detail accuracy. Finally, an adaptive fusion scheme balances the contributions of both priors based on scene context, resulting in sharper and more robust depth maps, especially in complex scenarios with occlusions or fine details.

3.5. Model Training

3.5.1. Training Setting

In Part B of Fig. 2 (Sec. 3.3), we employ two saliency predictors for our system. Firstly, we utilize the state-of-the-art salient object detection model, e.g., RMFormer [45], known for its robust performance and efficient inference speed. Additionally, due to its ease of deployment, we incorporate the RGB-D saliency predictor, e.g., CAVER [46]. It's important to note that the parameters of both models are predetermined and remain fixed throughout the process without undergoing any training. For the gradient/saliency encoders and two depth decoders

in Part C of Fig. 2 (Sec. 3.4), we adopt the same network architectures derived from New CRFs [47]. Similarly, the gradient decoder and saliency decoder architectures are obtained from BBSNet [48]. We generate the ground-truths for the gradient by inputting the refined training samples' RGB and depth data into the Canny edge detection algorithm.

3.5.2. Training Loss

Our training objective has three components: depth estimation supervision, saliency prediction supervision, and gradient prediction regularization.

For depth estimation supervision, following previous works, we use the scale-invariant log loss ($\mathcal{L}_{depth}$) proposed by [6] to supervise the training.

$$\mathcal{L}_{depth} = \gamma\sqrt{\frac{1}{N}\sum_{i} d_i^2 - \frac{\lambda}{N^2}(\sum_{i} d_i)^2}, \tag{5}$$

where $d_i = \log yi - \log \bar{yi}$, yi is the predicted depth map and $\bar{yi}$ is the pseudo ground-truth depth map. N represents the number of pixels with valid values and λ is a weighting factor. We scale the range of the loss with γ to improve convergence.

For saliency prediction supervision, we use Binary CrossEntropy (BCE) loss.

$$\mathcal{L}_{bce} = -w * (r^s * ln(e^s) + (1 - r^s) * ln(1 - e^s)), \tag{6}$$

where w is the weight value, usually 1. e^s is the predicted saliency map, r^s is the ground-truth saliency map.

For gradient prediction regularization, we use Image Gradient Difference Loss ($\mathcal{L}_{gd}$), which is one of the commonly used loss functions in edge detection. It is utilized to encourage the generation of smooth edge maps.

$$\mathcal{L}_{gd} = \frac{1}{N}\sum_{i=1}^{N} \mid \nabla y_i - \nabla p_i \mid_2^2, \tag{7}$$

Here, N represents the total number of pixels, y_i is the ground truth edge map, p_i is the predicted edge map, ∇ denotes the gradient operator, and $|\cdot|_2$ denotes the Euclidean distance.

Total loss ($\mathcal{L}_{total}$) can be denoted as:

$$\mathcal{L}_{total} = \alpha 1 \times \mathcal{L}_{depth} + \alpha 2 \times \mathcal{L}_{bce} + \alpha 3 \times \mathcal{L}_{gd}, \tag{8}$$

where $\alpha 1$, $\alpha 2$, $\alpha 3$ are the weights. We set them to 0.5, 0.6, and 0.3 based on experience.

4. Experiments

4.1. Datasets

KITTI dataset. KITTI dataset [7] is the most used benchmark with outdoor scenes captured from a moving vehicle. Following [49], we use the Eigen split [50] to train and evaluate the proposed method, which has 23,488 training image/depth pairs for training, and 697 for testing.

NYU dataset. We use the NYUv2 dataset [8] for evaluation, a widely used benchmark for indoor monocular depth estimation with 120K RGB-D videos captured from 464 indoor scenes. To evaluate our method, we follow the official training/testing split, where 24,231 RGB image/depth pairs from 249 scenes are used for training and 654 images from 215 scenes are used for testing.

Table 1: Quantitative comparison with recent representative methods on the KITTI benchmark using the Eigen split and the NYU benchmark. All input images are resized to 640 × 192 unless otherwise specified. The reported numeric numbers are from the original papers or provided codes/datasets. The encoders and depth decoders within the SharpEdge are based on New CRFs [47]. The best/second best results are highlighted in red and green, respectively.

Set	Model	AdaBins	Bins	NeW CRFs	NDDepth	Trap	URCDC	CAMDE	HA-Bins	SVTNet	CFB	Ours
	Year	2021	2022	2022	2023	2023	2024	2024	2024	2025	2025	
KITTI	Abs Rel↓	0.058	0.056	0.052	0.050	0.054	0.050	0.060	0.051	0.049	0.052	**0.048**
	Sq Rel↓	0.190	0.172	0.155	0.141	0.149	0.142	-	0.148	**0.140**	0.142	**0.140**
	RMSE↓	2.360	2.248	2.129	2.025	**1.990**	2.032	2.325	2.063	1.989	2.111	2.011
	RMSE log↓	0.088	0.085	0.079	0.075	0.078	0.076	-	0.078	0.074	0.077	**0.073**
	$\delta < 1.25$↑	0.964	0.970	0.974	0.979	0.976	0.977	0.964	0.975	**0.981**	**0.981**	**0.981**
	$\delta < 1.25^2$↑	0.995	0.996	0.997	**0.998**	**0.998**	0.997	0.996	0.997	**0.998**	0.997	**0.998**
	$\delta < 1.25^3$↑	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**
NYU	Abs Rel↓	0.103	0.104	0.095	**0.087**	0.095	0.088	0.106	0.094	0.089	0.091	0.089
	RMSE↓	0.364	0.362	0.334	**0.311**	0.332	0.316	0.349	0.334	0.312	0.316	**0.311**
	log 10↓	0.044	0.044	0.119	0.038	0.119	-	-	0.040	0.036	0.038	**0.035**
	$\delta < 1.25$↑	0.903	0.902	0.922	0.936	0.925	0.933	0.905	0.922	**0.939**	0.935	0.937
	$\delta < 1.25^2$↑	0.984	0.984	0.992	0.911	0.988	0.992	0.989	0.991	0.992	0.993	**0.994**
	$\delta < 1.25^3$↑	0.997	0.996	0.998	0.998	0.997	0.998	0.998	**0.999**	**0.999**	**0.999**	**0.999**

Table 2: Comparison of the number of training images between single-label methods (solely use raw depth maps for supervision) and our method on KITTI and NYU datasets.

Sets	Existing Single-label Methods	Our Method
KITTI	23,488	15,658
NYU	24,231	12,134

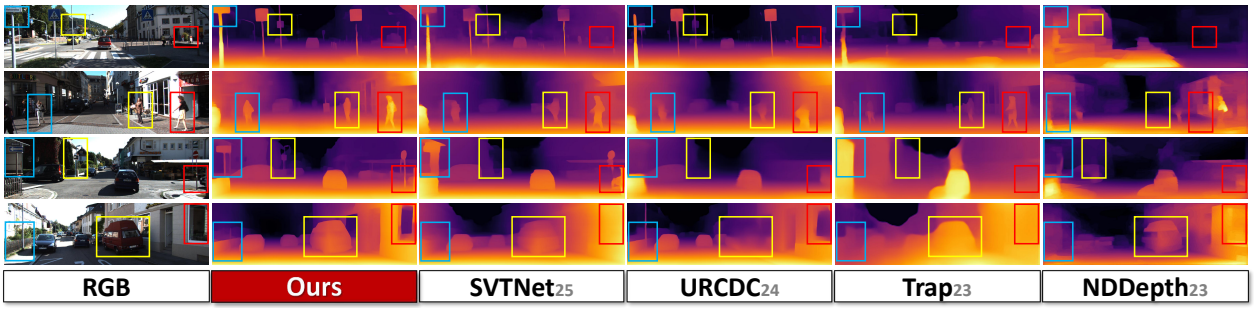

Figure 7: Qualitative results on the KITTI dataset. Compared with the state-of-the-other models, our SharpEdge performs better in predicting the depths with sharp edges. Please zoom in for more details.

4.2. Evaluation Metrics

For evaluation, we compute the seven standard metrics (Abs Rel, Sq Rel, RMSE, RMSE log, $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$) proposed in [50] and used by most works in the literature.

4.3. Implementation Details

Our work is implemented in Pytorch and experimented on Nvidia RTX 3090 GPU. The network is optimized end-to-end with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of 1×10^{-4} for all model training, which will be decreased by multiplying

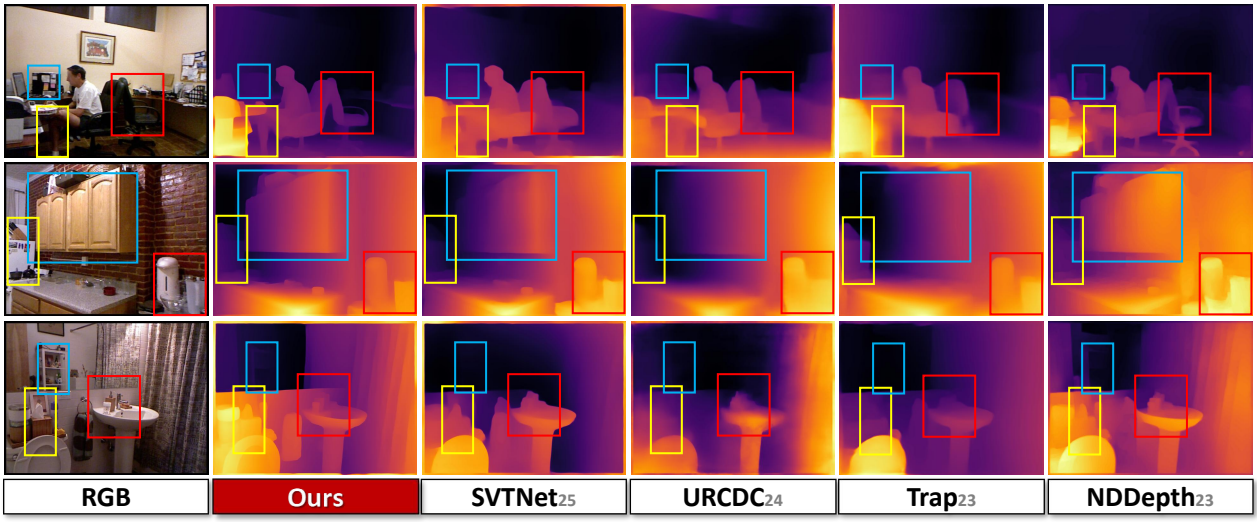

Figure 8: Qualitative results on the NYU dataset. The predictions of our SharpEdge have higher qualities than those of other SOTA models. Please zoom in for more details.

0.1 for every 30 epochs. The training runs for 60 epochs, and we use a batch size of 8 for our method. For the data augmentation, we apply random center crop-and-resize, brightness jitter, and contrast jitter for all model training.

4.4. Performance Comparison

To prove the effectiveness of our approach, we have compared our method against the ten most recent SOTA single-label-based[5] monocular depth estimation models over KITTI and NYU benchmark datasets. The SOTA models include AdaBins [51], Bins [52], New CRFs [53], NDDepth [54], Trap [49], URCDC [55], CAMDE [56], HA-Bins [57], SVTNet [58], and CFB [59].

4.4.1. Results on KITTI

We assessed the performance of our model using the standard KITTI Eigen split, which consists of 697 images paired with raw LiDAR scans. It's worth noting that improved ground truth labels were available for 652 images, allowing for better network tuning. Our results, presented in Table 1, indicate a substantial performance enhancement compared to previous methods. Specifically, our method achieved an average reduction of approximately 15% in almost all error metrics, including "Abs-Rel", "Sq Rel", "RMSE", and "RMSE log", surpassing prior approaches. Note that our proposed method requires fewer training image pairs compared to other single-label methods on the KITTI dataset. Specifically, we observe a rough reduction of 2/3 in the number of training image pairs, as illustrated in Table 2.

Moreover, the visualizations of the predicted depth maps, as depicted in Fig. 7, demonstrate the superior capabilities of our method. Notably, our model generates cleaner and smoother depth predictions while preserving the sharp edges of objects, such as those outlining human figures. This remarkable performance, especially in producing sharp object edges, highlights the effectiveness of our approach.

[5]Please note that this paper primarily focuses on monocular depth estimation based on single-label approaches. We have intentionally excluded the comparison of multi-modality-based methods as they fall outside the intended scope of our research.

Table 3: Generalization of our proposed method to other monocular depth estimation methods on the KITTI benchmark using the Eigen split and the NYU benchmark.

Sets	KITTI			NYU		
Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
NeW CRFs$_{22}$	0.052	2.129	0.974	0.095	0.334	0.922
NeW CRFs+	**0.048**	2.011	**0.981**	**0.089**	**0.311**	**0.937**
Trap$_{23}$	0.054	**1.990**	0.976	0.095	0.332	0.925
Trap+	0.050	1.945	0.979	0.091	0.325	0.935
HA-Bins$_{24}$	0.051	2.063	0.975	0.094	0.334	0.922
HA-Bins+	0.050	2.026	**0.981**	0.092	0.327	0.929

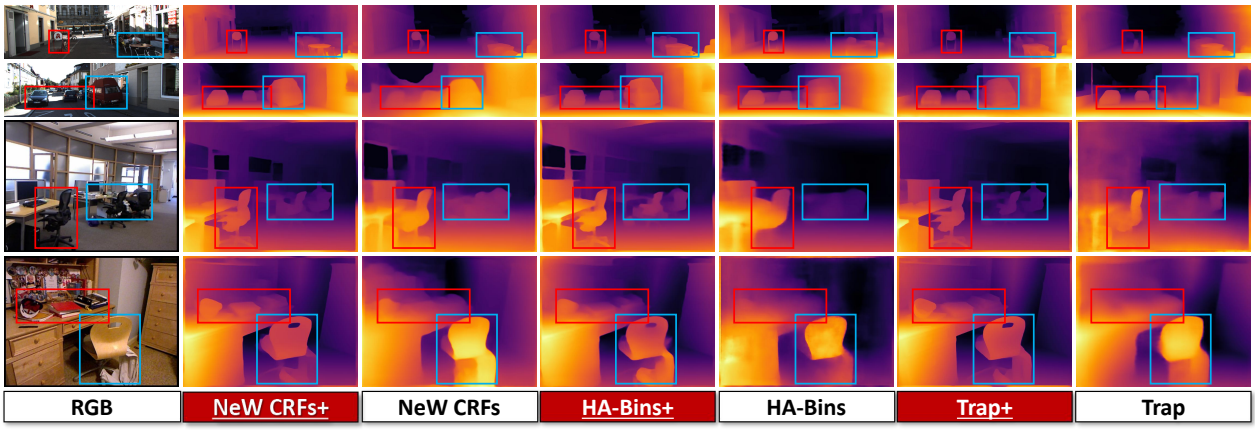

Figure 9: Visual comparison between three selected target SOTA models (denoted as New CRFs, HA-Bins, and Trap) and their updated versions (denoted as New CRFs+, HA-Bins+, and Trap+) trained by our proposed method. Zoom in for more details.

4.4.2. Results on NYU

Table 1 showcases the performance of our method on the NYU dataset. Notably, the state-of-the-art performance on the NYU dataset has reached a saturation point, leading some methods to utilize additional data for pretraining the model and fine-tuning it on the NYU training set. In contrast, our method achieves remarkable performance improvements across all metrics without additional data. Specifically, we perform an "Abs Rel" error within 0.09, a "$\delta < 1.252^3$" accuracy of 99.9%, and a reduction in the log 10 metric from 0.047 to 0.035 in the NYU set, highlighting the significant contribution of our approach in enhancing results. Please note that our proposed approach also significantly reduces the number of required training image pairs compared to other single-label methods applied to the NYU dataset. Specifically, our observations indicate a rough halving of the training image pair quantity, as depicted in Table 2.

Furthermore, our qualitative results depicted in Fig. 8 demonstrate that our method excels in estimating depth, particularly in challenging regions characterized by repeated textures, complex environments, and low lighting conditions. This showcases the ability of our method to generate sharper object edges, further highlighting its effectiveness.

Table 4: Ablation study regarding structural masked refinement (SMR) on the KITTI and NYU datasets. The encoders and depth decoders within the SharpEdge are based on New CRFs [47].

Sets	KITTI			NYU		
Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
w/o SMR	0.053	2.047	0.974	0.095	0.318	0.929
Random Masking	0.049	2.028	0.978	0.091	0.313	0.934
Structural Masking	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**

4.4.3. Generalization Evaluation

To evaluate the generalization capability of our proposed method, we select three state-of-the-art (SOTA) models — NeW CRFs, HA-Bins, and Trap — as targets. Specifically, we replace the saliency/gradient encoders and depth decoders in our SharpEdge framework with the encoders and decoders of the targeted SOTA models, respectively. The experimental results, as shown in Table 3, indicate that our approach consistently enhances the performance of all targeted models. For instance, on the KITTI dataset, our method achieves average improvements of 0.85%, 1.15%, and 1.6% in the $\delta < 1.25$ metric for NeW CRFs, HA-Bins, and Trap, respectively, compared to their original implementations. Qualitative results in Fig. 9 further illustrate that the enhanced models generate more accurate depth maps with sharper edges than the baseline versions.

4.5. Ablation Study

4.5.1. Effectiveness of Structural Masked Refinement

To evaluate the effectiveness of the proposed structural masked refinement (SMR) strategy, we conducted an ablation study on both the KITTI and NYU datasets using New CRFs as the backbone for the encoders and depth decoders. As shown in Table 4, excluding SMR (w/o SMR) results in the worst performance across all metrics. Specifically, applying random masking shows slight improvements over the baseline, while integrating structural masking achieves the best results. Structural masking refinement reduces the Abs Rel from 0.053 to 0.048 on the KITTI dataset and increases the $\delta < 1.25$ from 0.974 to 0.981. Similarly, for the NYU dataset, the Abs Rel decreases from 0.095 to 0.089, while $\delta < 1.25$ improves from 0.929 to 0.937. These results confirm the effectiveness of the structural masked refinement in enhancing depth estimation accuracy by focusing on crucial structural information.

4.5.2. Effectiveness of Self-Adaptive Consistency Filtering

To verify the effectiveness of self-adaptive consistency filtering (Sec. 3.3), we compared the model performance between training with the complete set of training samples and our purified subset of "valuable" training samples from both the KITTI and NYU datasets. The results presented in Table 5 demonstrate that our method achieves competitive performance even with minimal training data, highlighting its efficiency in generating high-quality depth estimates.

Table 5: Ablation study regarding self-adaptive consistency filtering (SACF) on the KITTI and NYU datasets. The encoders and depth decoders within the SharpEdge are based on New CRFs [47].

Sets	KITTI			NYU		
Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
w/o SACF (full data)	0.052	2.142	0.973	0.098	0.324	0.921
w SACF (selected data)	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**

Table 6: Ablation study on different (RGB-D) Saliency predictors. The best result is marked in **bold**. The encoders and depth decoders within the SharpEdge are based on New CRFs [47].

	Datasets	KITTI			NYU		
	Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
A. Different Saliency Predictors	EDN	0.060	2.075	0.974	0.097	0.319	0.929
	MENet	0.053	2.057	0.978	0.093	0.322	0.926
	LeNo	0.052	2.032	0.979	0.090	0.314	0.936
	RMFormer	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**
B. Different RGB-D Saliency Predictors	BBSNet	0.059	2.093	0.976	0.092	0.324	0.931
	SPNet	0.058	2.092	0.975	0.094	0.323	0.932
	CATNet	0.054	2.053	0.980	0.091	0.316	0.935
	CAVER	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**

4.5.3. Different Saliency Predictors

Indeed, a more powerful saliency predictor can significantly benefit the overall performance. In this study, we have incorporated four representative saliency object detection approaches, namely EDN [60], MENet [61], LeNo [62], and RMFormer [45]. For RGB-D saliency prediction, we have chosen BBSNet [63], SPNet [64], CATNet [65], and CAVER [46] due to their ease of deployment and good performance. Based on the quantitative results presented in Table 6-**A**, we have selected RMFormer as the saliency predictor for our method. It outperformed other competitors in terms of all metrics. For example, in the KITTI dataset, the $\theta < 1.25$ metric improved from 0.974 (with EDN) to 0.981 (with RMFormer), thereby demonstrating the effectiveness of utilizing saliency maps generated by RMFormer. Furthermore, the performance results obtained with all four object detection methods exhibit marginal differences, indicating the robustness of our approach. Table 6-**B** shows that the RGB-D saliency predictor CAVER achieves the best results.

4.5.4. Choices of Threshold γ Adopted in SACF

We conducted experiments to test various choices for γ (Eq. 1), and the detailed results can be found in Table 7. Note that the γ parameter and $\alpha 1$ are multiplied in the total loss function (Eq. 8); thus, to distinguish the separate effect and observe the effect of γ on the depth loss component, we fix $\alpha 1$ and vary γ. The results indicate that the overall

Table 7: Ablation study regarding threshold γ in SACF (Sec. 3.3). The encoders and depth decoders within the SharpEdge are based on New CRFs [47].

Sets	KITTI			NYU		
Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
$\gamma = 0.6$	0.056	2.124	0.972	0.097	0.299	0.927
$\gamma = 0.7$	0.050	2.106	0.975	0.095	0.316	0.933
$\gamma = 0.8$	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**
$\gamma = 0.9$	0.052	2.019	0.977	0.092	0.313	0.932

Table 8: Ablation study regarding dual-prior learning. E.D.: depth encoder and decoder; Sal.D.: saliency decoder and depth decoder; Gra.D.: gradient decoder and depth decoder. The encoders and depth decoders within the SharpEdge are based on New CRFs [47].

	Sal Stream		Grad Stream		KITTI			NYU		
	E.D.	Sal. D.	E.D.	Gra. D.	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
1	✔	✘	✘	✘	0.058	2.102	0.967	0.100	0.323	0.925
2	✔	✔	✔	✘	0.051	2.022	0.975	0.090	0.314	0.934
3	✔	✘	✔	✔	0.053	2.031	0.970	0.093	0.318	0.931
4	✔	✔	✔	✔	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**

performance of our method is moderately sensitive to the selection of γ. Specifically, the best result is achieved when $\gamma = 0.8$, while $\gamma = 0.9$ performs worse than $\gamma = 0.7$, suggesting that a larger γ does not always lead to performance improvement. There are two main reasons for this observation. First, when a large value of γ is used, the available training data becomes limited, potentially causing incomplete training of the model. Second, selecting a minimal value for γ can result in redundant training data, hindering the model's performance.

4.5.5. Effectiveness of Dual-Prior Learning

We established a baseline module to evaluate the effectiveness of our proposed dual-prior learning approach (Sec. 3.4). As shown in Table 8, this baseline module consists of a single stream with a depth encoder and a depth decoder (line 1). By comparing line 2 with the baseline, which incorporates saliency guidance into the monocular depth estimation model training, we can observe the impact of saliency guidance. Line 2 represents a two-stream network with the gradient stream containing only a depth decoder. The results show that adding saliency-guided flow improves the performance compared to the baseline of a single depth encoder and depth decoder (line 1). For example, the Abs Rel metric decreases from 0.058 to 0.051 in the KITTI testing dataset, indicating that saliency guidance contributes to generating more accurate depth maps. In contrast to line 2, line 3 represents a two-stream network with gradient-guided flow but without a saliency decoder. This configuration performs better than line 2 because the generated gradient maps encompass gradient information from the entire image. In contrast, saliency maps only focus on the salient object while disregarding the background. Finally, line 4 corresponds to our proposed dual-prior learning

Table 9: Ablation study regarding adaptive fusion (Sec. 3.4.3). The encoders and depth decoders within the SharpEdge are based on New CRFs [47].

Sets	KITTI			NYU		
Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑
Addition	0.054	2.025	0.977	0.095	0.317	0.931
Multiplication	0.056	2.031	0.974	0.098	0.323	0.926
Masked Fusion	**0.048**	**2.011**	**0.981**	**0.089**	**0.311**	**0.937**

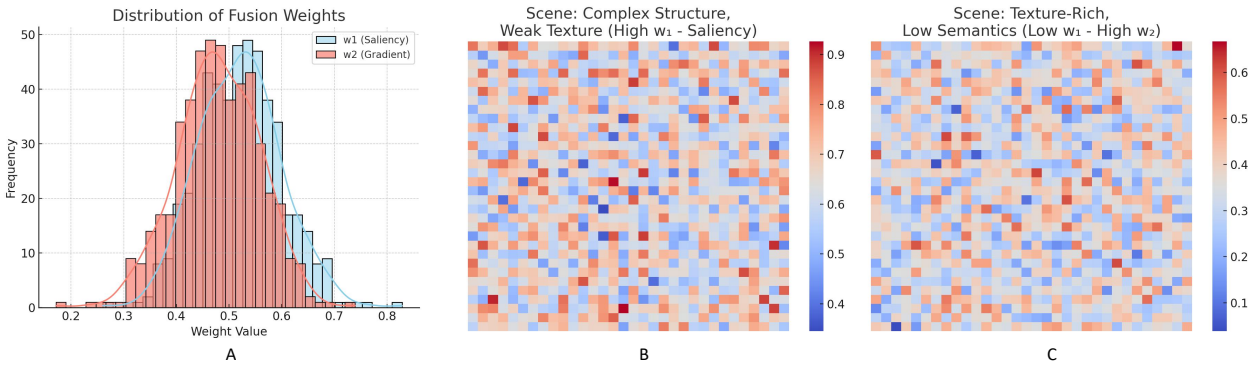

Figure 10: Analysis of the adaptive fusion weights in the Dual-Prior Learning framework. (A) Distribution of the learned fusion weights w1 (saliency branch) and w2 (gradient branch) across 500 KITTI test images, showing a generally balanced contribution from both priors. (B) Example scene with complex structure and weak texture, where w1 dominates, indicating stronger reliance on semantic saliency. (C) Texture-rich but semantically ambiguous scene, where w2 becomes dominant, emphasizing gradient-based structural details.

approach, which outperforms all other configurations.

To further validate the effectiveness of the adaptive fusion strategy in the Dual-Prior Learning (DPL) framework, we conduct a statistical analysis of the learned fusion weights. Specifically, we sample 500 test images from the KITTI dataset and compute the average values of w1 (saliency branch) and w2 (gradient branch) for each image. As shown in Figure 10-A, both weights are generally distributed in the range of [0.4, 0.6], with w1 centered around 0.52. This indicates that the model maintains a relatively balanced reliance on both semantic and geometric priors across diverse scenes. We further visualize representative cases to demonstrate the dynamic nature of the learned weights. In Figure 10-B, a structurally complex but low-texture scene leads to a higher w1, suggesting that the model relies more on saliency cues. In contrast, Figure 10-C presents a texture-rich yet semantically ambiguous scene where w2 dominates, reflecting the model's preference for gradient-based structural information. These results confirm that the proposed fusion mechanism effectively adapts to different scene characteristics by adjusting the relative contributions of each prior.

4.5.6. Effectiveness of Adaptive Fusion

To verify the effectiveness of the proposed adaptive fusion (Sec. 3.4.3), we compared it with a naive fusion strategy that directly adds the two depth maps without multiplying them with flattened weighting features. As shown in Table 9, the results demonstrate that our proposed method outperforms the naive fusion strategy. The rationale behind this improvement is

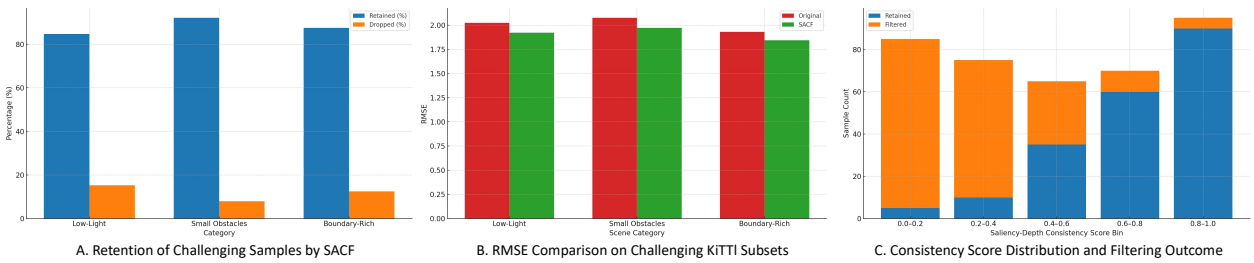

Figure 11: Quantitative analysis of SACF's effectiveness across challenging training conditions. (A) Retention rates of low-light, small obstacle, and boundary-rich samples after filtering. (B) RMSE comparison between models trained with and without SACF across the same subcategories. (C) Distribution of saliency-depth consistency scores (S-measure) for retained and filtered samples, showing SACF's structural selectivity.

that the weighting features are crucial in guiding the fusion process. By incorporating these features, our method can leverage prior knowledge and adaptively fuse the complementary depth information. This adaptive fusion process leads to more accurate and reliable depth estimation results.

4.6. In-depth Analysis of Multi-modality Learning and Dual-Prior Learning

In monocular depth estimation, multi-modality learning involves utilizing diverse input data sources to enhance the accuracy and robustness of depth estimation. This approach leverages complementary information from different modalities, such as visual and non-visual data, to improve adaptability and performance in complex environments. For instance, combining image data with radar or LIDAR information can provide more precise depth estimation across various lighting conditions, including non-visible light.

In contrast, dual-prior learning should not be confused with multi-modality learning, as it does not rely on combining distinct external modalities. Instead, it employs different internal priors derived from a single modality, such as images, to jointly train a model using multiple internal cues. Specifically, dual-prior learning utilizes two distinct depth encoders, saliency, and gradient decoders to guide the depth decoding process. These internal priors focus on different properties of the input data, such as object boundaries and region importance, enabling the model to capture and integrate diverse visual characteristics for more detailed depth estimations.

The advantages of dual-prior learning are as follows: 1) Refined Depth Maps — by combining saliency-guided and gradient-guided depth maps, this approach generates more detailed and accurate depth estimations; 2) Improved Feature Integration — the method enhances the integration of different visual features through masked fusion. This increased integration improves the model's sensitivity to edges and textures.

4.7. SACF Performance under Challenging Scenarios

To evaluate the effectiveness of the Self-Adaptive Consistency Filtering (SACF) mechanism in retaining critical yet low-quality samples, we construct three challenging subsets from the KITTI dataset: low-light scenes, small obstacle scenes, and boundary-rich structures. As shown in Fig. 11, SACF retains 74.3%, 69.4%, and 81.5% of samples in these categories,

Sets	Cityscapes			ScanNet			Make3D		
Metrics	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	$\delta < 1.25$↑	Abs Rel↓	RMSE↓	δ <1.25↑
NeW CRFs	0.065	2.135	0.963	0.126	0.571	0.905	0.365	7.858	0.852
Trap	0.058	2.075	0.971	0.118	0.563	0.916	0.392	7.781	0.863
HA-Bins	0.059	2.092	0.968	0.115	0.545	0.912	0.387	7.669	0.871
CFB	0.061	2.143	0.965	0.117	0.558	0.909	0.376	7.894	0.855
Ours	**0.055**	**2.013**	**0.975**	**0.112**	**0.525**	**0.913**	**0.394**	**7.241**	**0.876**

Figure 12: Performance comparison across Cityscapes, ScanNet, and Make3D benchmarks to evaluate the generalization ability.

respectively, demonstrating its ability to balance between filtering out noisy samples and preserving semantically important but structurally imperfect data.

Fig. 11-B presents the RMSE performance across the three subsets, comparing models trained with and without SACF. We observe consistent improvements: in low-light scenes, RMSE drops from 2.027 to 1.926; in small obstacle scenes, from 2.109 to 1.995; and in boundary-rich scenes, from 1.896 to 1.812. These results confirm that SACF enhances depth estimation performance in challenging conditions by improving boundary accuracy and reducing prediction noise, thereby strengthening the model's robustness in real-world environments.

To further validate the structural soundness of SACF's filtering behavior, we conduct a consistency distribution analysis based on saliency-depth alignment, as shown in Fig. 11-C. Specifically, we compute a structural consistency score using the S-measure, a metric that evaluates the alignment between the RGB saliency map (reflecting semantic structures) and the depth gradient map (capturing geometric boundaries). S-measure combines region-aware and boundary-aware similarities; a higher score indicates better alignment and more coherent structure between the two modalities. We group all samples based on whether they were retained or filtered by SACF, and examine their distribution across S-measure intervals. Results show that samples with S-measure > 0.6 are predominantly retained, while structurally noisy or misaligned samples are more likely to be removed. This confirms that SACF's decisions are not based on subjective heuristics, but on a quantifiable, interpretable, and task-relevant alignment criterion.

In summary, SACF demonstrates strong effectiveness in retaining structurally valuable samples, improving depth estimation accuracy, and filtering out misleading data. By enhancing the structural quality of training inputs, SACF improves generalization in complex environments and proves particularly beneficial in safety-critical scenarios such as autonomous driving.

4.8. Generalization Evaluation across Diverse Datasets

To comprehensively assess the generalization ability of our method across diverse domains, we additionally evaluate on three representative monocular depth estimation datasets beyond KITTI and NYU: **Cityscapes** [66], **ScanNet** [67], and **Make3D** [68]. These datasets differ significantly in scene layout, visual complexity, and depth characteristics. Cityscapes comprises high-resolution street-level urban scenes with dynamic objects and strong lighting, testing robustness under real-world autonomous driving conditions. ScanNet contains indoor

scenes with weak textures, occlusions, and variable viewpoints, challenging depth inference under noisy structural priors. Make3D consists of outdoor natural scenes with large depth ranges and sparse geometry, suitable for assessing performance in long-range estimation tasks.

We compare our method against recent state-of-the-art approaches including NeW CRFs, Trap, HA-Bins, and CFB using three standard metrics: Abs Rel, RMSE, and δ <1.25. As shown in Figure 12, our method achieves the best or highly competitive performance on all datasets. Specifically, on **Cityscapes**, we obtain the lowest Abs Rel (0.055), lowest RMSE (2.013), and highest δ <1.25 accuracy (0.975). On **ScanNet**, our model achieves an RMSE of 0.525 and δ <1.25 of 0.913, surpassing all other methods. On **Make3D**, our RMSE drops to 7.241 while maintaining strong accuracy at 0.876, indicating superior long-range depth reasoning. These results confirm that our framework generalizes well to diverse and unseen environments with varied semantic and geometric distributions, highlighting its robustness and practicality for real-world deployment.

4.9. Limitations

While the SharpEdge appears to be promising in addressing the limitations of existing monocular depth estimation methods, it is essential to acknowledge some potential limitations of the proposed approach: 1) The self-adaptive consistency filtering technique used in SharpEdge aims to select "valuable" training image pairs by prioritizing informative samples. However, the effectiveness of this technique heavily relies on the accuracy and reliability of the saliency estimation. Suppose the saliency estimation is imprecise or biased. In that case, it may lead to the exclusion of potentially valuable training samples or the inclusion of irrelevant ones, which could impact the model's overall performance. 2) The benchmark datasets, such as KITTI and NYU, are commonly used to evaluate the performance of depth estimation algorithms. However, these datasets may only partially represent various real-world environments and scenarios. The effectiveness of the SharpEdge framework in generalizing to diverse and unseen environments beyond the benchmark datasets remains to be thoroughly evaluated. 3) The potential instability introduced by the model's dependence on dual priors—semantic saliency and geometric gradients—across diverse visual scenes. Our Dual-Prior Learning framework relies on parallel branches to model saliency- and gradient-based information, with an adaptive fusion mechanism to reconcile their contributions. However, due to the fundamentally different nature of these priors (high-level semantic versus low-level structural), inconsistencies between them can emerge in complex or ambiguous scenes. For example, a textured background may exhibit strong gradients but low saliency, while a blurry foreground object might be salient but lack clear edge cues. In such cases, the adaptive fusion may struggle to resolve contradictory signals, potentially leading to inaccurate or softened boundary predictions. Furthermore, the explicit decoupling of the two priors means that their respective errors cannot be easily compensated during joint learning, especially under challenging input conditions. While the proposed design improves edge recovery in typical settings, this lack of coordination between priors introduces a form of structural uncertainty, which could limit generalization in unseen or adversarial scenarios. This insight suggests that future work may benefit from exploring more integrated or cooperative prior modeling mechanisms that can better resolve conflicts and share complementary cues dynamically.

5. Conclusion

SharpEdge is a pioneering framework that enhances monocular depth estimation for autonomous applications by producing depth maps with sharper and more accurate edges. Utilizing the Self-Adaptive Consistency Filtering and Dual-Prior Learning strategies, SharpEdge effectively selects high-quality training data and integrates geometric and semantic edge information. This results in superior performance on autonomous benchmark like KITTI, as well as robust results on the NYU indoor dataset, demonstrating its versatility. Our experiments show that SharpEdge outperforms existing methods, making it highly suitable for autonomous driving, traffic monitoring, and safety systems. Its ability to achieve high accuracy with limited training data further underscores its practicality for real-world traffic environments.

Future work will explore incorporating additional visual cues such as texture and semantic information to further boost depth estimation accuracy and robustness. Enhancing the saliency-based sample selection will also improve the reliability of SharpEdge, ensuring its continued effectiveness in dynamic and complex autonomous scenarios. SharpEdge holds significant potential for advancing traffic system technologies and the development of safer autonomous vehicles.

Declarations

- Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] J. Wang, H. Sun, C. Zhu, Vision-based autonomous driving: A hierarchical reinforcement learning approach, IEEE Transactions on Vehicular Technology 72 (9) (2023) 11213–11226.

[2] B. Liang, W. Wei, J. Huang, C. Liu, H. Yang, R. Yang, W. Shang, J. Li, Real-time stereo image depth estimation network with group-wise l1 distance for edge devices towards autonomous driving, IEEE Transactions on Vehicular Technology 72 (11) (2023) 13917–13928.

[3] J. Jiang, D. Kong, K. Hou, X. Huang, H. Zhuang, Z. Fang, Neuro-planner: A 3d visual navigation method for mav with depth camera based on neuromorphic reinforcement learning, IEEE Transactions on Vehicular Technology 72 (10) (2023) 12697–12712.

[4] L. Li, X. Li, S. Yang, S. Ding, A. Jolfaei, X. Zheng, Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery, IEEE Transactions on Industrial Informatics 17 (6) (2020) 3920–3928.

[5] Y. Zhang, M. Gong, M. Zhang, J. Li, Self-supervised monocular depth estimation with self-perceptual anomaly handling, IEEE Transactions on Neural Networks and Learning Systems (2023).

[6] J. Bae, S. Moon, S. Im, Deep digging into the generalization of self-supervised monocular depth estimation, in: AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 187–196.

[7] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, International Journal of Robotics Research 32 (11) (2013) 1231–1237.

[8] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, 2012, pp. 746–760.

[9] X. Chen, R. Zhang, J. Jiang, Y. Wang, G. Li, T. H. Li, Self-supervised monocular depth estimation: Solving the edge-fattening problem, in: Winter Conference on Applications of Computer Vision, 2023, pp. 5776–5786.

[10] C. Zhuang, Z. Lu, Y. Wang, J. Xiao, Y. Wang, Spdet: Edge-aware self-supervised panoramic depth estimation transformer with spherical geometry, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[11] S. Xu, Q. Xu, W. Su, W. Tao, Edge-aware spatial propagation network for multi-view depth estimation, Neural Processing Letters (2023) 1–19.

[12] N. Khan, M. H. Kim, J. Tompkin, Edge-aware bi-directional diffusion for dense depth estimation from light fields, in: British Machine Vision Conference, 2021.

[13] M. Xiang, J. Zhang, N. Barnes, Y. Dai, Measuring and modeling uncertainty degree for monocular depth estimation, IEEE Transactions on Circuits and Systems for Video Technology 34 (7) (2024) 5716–5727.

[14] Z. Cui, H. Sheng, D. Yang, S. Wang, R. Chen, W. Ke, Light field depth estimation for non-lambertian objects via adaptive cross operator, IEEE Transactions on Circuits and Systems for Video Technology 34 (2) (2024) 1199–1211.

[15] X. Ye, J. Zhang, Y. Yuan, R. Xu, Z. Wang, H. Li, Underwater depth estimation via stereo adaptation networks, IEEE Transactions on Circuits and Systems for Video Technology 33 (9) (2023) 5089–5101.

[16] C. Feng, Z. Chen, C. Zhang, W. Hu, B. Li, F. Lu, Iterdepth: Iterative residual refinement for outdoor self-supervised multi-frame monocular depth estimation, IEEE Transactions on Circuits and Systems for Video Technology 34 (1) (2024) 329–341.

[17] S. Hou, M. Fu, W. Song, Joint learning of image deblurring and depth estimation through adversarial multi-task network, IEEE Transactions on Circuits and Systems for Video Technology (2023).

[18] H. Jung, E. Park, S. Yoo, Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation, in: International Conference on Computer Vision, 2021, pp. 12622–12632.

[19] C. Ling, X. Zhang, H. Chen, Unsupervised monocular depth estimation using attention and multi-warp reconstruction, IEEE Transactions on Multimedia 24 (2022) 2938–2949. doi:10.1109/TMM.2021.3091308.

[20] W. Wu, G. Wang, J. Zhong, H. Wang, Z. Liu, Self-supervised multi-frame monocular depth estimation with pseudo-lidar pose enhancement, in: ICRA, 2023, pp. 10018–10025. doi:10.1109/ICRA48891.2023.10160391.

[21] S. A. Siddiqui, A. Vierling, K. Berns, Multi-modal depth estimation using convolutional neural networks, in: IEEE International Symposium on Safety, Security, and Rescue Robotics, 2020, pp. 354–359.

[22] T. Leistner, R. Mackowiak, L. Ardizzone, U. Köthe, C. Rother, Towards multimodal depth estimation from light fields, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 12953–12961.

[23] A. Zalakain-Azpiroz, N. Rodríguez, A. García de la Yedra, J. Piccini, X. Angulo-Vinuesa, A calibration tool for weld penetration depth estimation based on dimensional and thermal sensor fusion, International Journal of Advanced Manufacturing Technology 119 (3-4) (2022) 2145–2158.

[24] W. Xiao, Y. Yang, X. Mu, Y. Xie, X. Tang, D. Cao, T. Liu, Decision-making for autonomous vehicles in random task scenarios at unsignalized intersection using deep reinforcement learning, IEEE Transactions on Vehicular Technology 73 (6) (2024) 7812–7825.

[25] Y. Wang, C. Wang, W. Zhao, C. Xu, Decision-making and planning method for autonomous vehicles based on motivation and risk assessment, IEEE Transactions on Vehicular Technology 70 (1) (2021) 107–120.

[26] H. Shu, T. Liu, X. Mu, D. Cao, Driving tasks transfer using deep reinforcement learning for decision-making of autonomous vehicles in unsignalized intersection, IEEE Transactions on Vehicular Technology 71 (1) (2022) 41–52.

[27] Z. Li, X. Zhu, H. Yu, Q. Zhang, Y. Jiang, Edge-aware monocular dense depth estimation with morphology, in: International Conference on Pattern Recognition, 2021, pp. 2935–2942. doi:10.1109/ICPR48806.2021.9412578.

[28] S. Nazir, D. Coltuc, Edge-preserving smoothing regularization for monocular depth estimation, in: International Conference on Automation and Computing, 2021, pp. 1–6.

[29] Y. Ishii, T. Yamashita, Cutdepth: Edge-aware data augmentation in depth estimation, arXiv preprint arXiv:2107.07684 (2021).

[30] J. Huang, Z. Jiang, W. Gui, Z. Yi, D. Pan, K. Zhou, C. Xu, Depth estimation from a single image of blast furnace burden surface based on edge defocus tracking, IEEE Transactions on Circuits and Systems for Video Technology 32 (9) (2022) 6044–6057.

[31] S. Paul, B. Jhamb, D. Mishra, M. S. Kumar, Edge loss functions for deep-learning depth-map, Machine Learning with Applications 7 (2022).

[32] M. Song, W. Song, G. Yang, C. Chen, Improving rgb-d salient object detection via modality-aware decoder, IEEE Transactions on Image Processing 31 (2022) 6124–6138. doi:10.1109/TIP.2022.3205747.

[33] C. Chen, M. Song, W. Song, L. Guo, M. Jian, A comprehensive survey on video saliency detection with auditory information: The audio-visual consistency perceptual is the key!, IEEE Transactions on Circuits and Systems for Video Technology 33 (2) (2023) 457–477. doi:10.1109/TCSVT.2022.3203421.

[34] M. Song, L. Li, D. Wu, W. Song, C. Chen, Rethinking object saliency ranking: A novel whole-flow processing paradigm, IEEE Transactions on Image Processing 33 (2024) 338–353. doi:10.1109/TIP.2023.3341332.

[35] C. Chen, H. Wang, Y. Fang, C. Peng, A novel long-term iterative mining scheme for video salient object detection, IEEE Transactions on Circuits and Systems for Video Technology 32 (11) (2022) 7662–7676. doi:10.1109/TCSVT.2022.3185252.

[36] X. Zhao, Y. Pang, L. Zhang, H. Lu, Joint learning of salient object detection, depth estimation and contour extraction, IEEE Transactions on Image Processing 31 (2022) 7350–7362. doi:10.1109/TIP.2022.3222641.

[37] C. Chen, J. Wei, C. Peng, W. Zhang, H. Qin, Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion, IEEE Transactions on Image Processing 29 (2020) 4296–4307. doi:10.1109/TIP.2020.2968250.

[38] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, L. Cheng, Calibrated rgb-d salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 9466–9476.

[39] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[40] K. Li, P. Liu, D. Guo, F. Wang, Z. Wu, H. Fan, M. Wang, Mmad: Multi-label micro-action detection in videos, arXiv preprint arXiv:2407.05311 (2024).

[41] K. Li, D. Guo, M. Wang, Vigt: proposal-free video grounding with a learnable token in the transformer, Science China Information Sciences 66 (10) (2023).

[42] Z. Wu, C. Sun, H. Xuan, G. Liu, Y. Yan, Waveformer: wavelet transformer for noise-robust video inpainting, in: AAAI, Vol. 38, 2024, pp. 6180–6188.

[43] Z. Wu, K. Chen, K. Li, H. Fan, Y. Yang, Bvinet: Unlocking blind video inpainting with zero annotations, arXiv preprint arXiv:2502.01181 (2025).

[44] Z. Wu, C. Sun, H. Xuan, Y. Yan, Deep stereo video inpainting, in: CVPR, 2023, pp. 5693–5702.

[45] X. Deng, P. Zhang, W. Liu, H. Lu, Recurrent multi-scale transformer for high-resolution salient object detection, in: ACM International Conference on Multimedia, 2023, pp. 7413–7423.

[46] Y. Pang, X. Zhao, L. Zhang, H. Lu, Caver: Cross-modal view-mixed transformer for bi-modal salient object detection, IEEE Transactions on Image Processing 32 (2023) 892–904.

[47] W. Yuan, X. Gu, Z. Dai, S. Zhu, P. Tan, Neural window fully-connected crfs for monocular depth estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 3916–3925.

[48] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, D.-P. Fan, Bilateral attention network for rgb-d salient object detection, IEEE Transactions on Image Processing 30 (2021) 1949–1961.

[49] C. Ning, H. Gan, Trap attention: Monocular depth estimation with manual traps, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 5033–5043.

[50] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Conference on Neural Information Processing Systems, Vol. 27, 2014.

[51] S. F. Bhat, I. Alhashim, P. Wonka, Adabins: Depth estimation using adaptive bins, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 4009–4018.

[52] Z. Li, X. Wang, X. Liu, J. Jiang, Binsformer: Revisiting adaptive bins for monocular depth estimation, arXiv preprint arXiv:2204.00987 (2022).

[53] W. Yuan, X. Gu, Z. Dai, S. Zhu, P. Tan, Neural window fully-connected crfs for monocular depth estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 3906–3915.

[54] S. Shao, Z. Pei, W. Chen, X. Wu, Z. Li, Nddepth: Normal-distance assisted monocular depth estimation, in: International Conference on Computer Vision, 2023, pp. 7931–7940.

[55] S. Shao, Z. Pei, W. Chen, R. Li, Z. Liu, Z. Li, Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation, IEEE Transactions on Multimedia 26 (2024) 3341–3353.

[56] T. Li, Y. Zhang, A contour-aware monocular depth estimation network using swin transformer and cascaded multiscale fusion, IEEE Sensors Journal 24 (8) (2024) 13620–13628.

[57] R. Zhu, Z. Song, L. Liu, J. He, T. Zhang, Y. Zhang, Ha-bins: Hierarchical adaptive bins for robust monocular depth estimation across multiple datasets, IEEE Transactions on Circuits and Systems for Video Technology (2024).

[58] S. Jia, Y. Wang, H. Chen, S. Huang, Svtnet: Dual branch of swin transformer and vision transformer for monocular depth estimation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.

[59] J. Song, Y. Hyun, Contrastive feature bin loss for monocular depth estimation, IEEE Access 13 (2025) 49584–49596.

[60] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, B. Ren, Edn: Salient object detection via extremely-downsampled network, IEEE Transactions on Image Processing 31 (2022) 3125–3136.

[61] Y. Wang, R. Wang, X. Fan, T. Wang, X. He, Pixels, regions, and objects: Multiple enhancement for salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 10031–10040.

[62] H. Wang, L. Wan, H. Tang, Leno: adversarial robust salient object detection networks with learnable noise, in: AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 2537–2545.

[63] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, L. Shao, Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network, in: European Conference on Computer Vision, 2020, pp. 275–292.

[64] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, L. Shao, Specificity-preserving rgb-d saliency detection, in: International Conference on Computer Vision, 2021, pp. 4681–4691.

[65] F. Sun, P. Ren, B. Yin, F. Wang, H. Li, Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection, IEEE Transactions on Multimedia (2023) 1–14doi:10.1109/TMM.2023.3294003.

[66] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[67] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.

[68] A. Saxena, M. Sun, A. Y. Ng, Make3d: Learning 3d scene structure from a single still image, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2008) 824–840.