iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83049

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

SkyRoad-CutDepth: A Framework for Sky-Aware and Road-Geometric Pretext Learning in Monocular Metric Depth Estimation

Amit Bhoir[1], Dr. Rekha Sharma[2]
[1]*Department of Computer Engineering, Thakur College of Engineering and Technology, Maharashtra, India*
[2]*Department of Computer Engineering, Thakur College of Engineering and Technology, Maharashtra, India*

Abstract: ***Monocular metric depth estimation is important for autonomous driving, road-scene understanding, navigation, and 3D perception from low-cost cameras. Recent foundation models such as Depth Anything V2, UniDepth, UniDepthV2, and Metric3D v2 have improved zero-shot generalization and metric depth prediction, while depth-specific augmentation methods such as CutDepth have shown that training strategies can improve depth learning without destroying object boundaries. However, dashcam imagery contains a difficult combination of road-plane geometry, horizon and vanishing-point structure, sky or infinity regions, and thin foreground objects such as poles, traffic signs, traffic lights, logos, wires, and distant roadside structures. This paper presents a theoretical and qualitative framework called SkyRoad-CutDepth for sky-aware and road-geometric pretext learning in metric monocular depth estimation. The work is motivated by qualitative outputs from a custom dashcam-dataset-trained Depth Anything V2 model with canonical-space transformations. The observations show that global road depth can become more consistent after custom training, but finite foreground objects located near the sky region or vanishing point can still be suppressed, partially missed, or assigned invalid depth. The proposed framework combines road-region priors, sky-region validity, sky-adjacent foreground preservation, boundary-aware losses, and region-guided CutDepth augmentation. By integrating these cues into a unified pretext-learning formulation, the framework aims to improve the geometric reliability of metric depth estimation in regions where current foundation models often produce over-smoothed, invalid, or semantically inconsistent depth predictions.***

Keywords: Monocular Depth Estimation, Metric Depth, Dashcam Imagery, Sky Segmentation, CutDepth, Road Geometry, Pretext Learning, Depth Anything V2.

I. INTRODUCTION

Depth estimation from a single RGB image is a highly ill-posed computer vision problem because the three-dimensional structure of a scene must be inferred from two-dimensional appearance cues. The problem becomes especially important in autonomous driving and road-scene perception, where camera-based systems must reason about the distance of vehicles, pedestrians, road signs, traffic lights, curbs, buildings, and distant road regions. Unlike stereo or LiDAR-based systems, monocular depth estimation can be deployed using low-cost dashcam sensors, but it must overcome scale ambiguity, sparse supervision, domain shift, and moving-object complexity [1], [2]. Recent research has significantly improved monocular depth estimation through transformer-based dense prediction [4], attention-based continuous pixel-wise prediction [5], metric depth transfer [6], camera-aware foundation models [7]-[9], large-scale pseudo-labeling [10], [11], and diffusion priors [23]-[25]. These models produce impressive depth maps in many real-world scenes. However, road-scene imagery presents a specific failure pattern, the lower part of the image often contains strong road-plane structure, while the upper part contains sky and infinity regions. Finite foreground objects, such as poles, signs, traffic lights, and wires, frequently appear against the sky. When models treat the upper image region as sky-like or geometrically invalid, these foreground objects may be weakened or omitted in the predicted depth map.

This issue is practically important. A road sign or traffic light may occupy only a small number of pixels, but it can still be safety-critical. Similarly, vanishing-point regions contain distant vehicles and infrastructure that are difficult to represent because of low resolution, sparse depth labels, and ambiguous sky-road transitions. A model may therefore produce visually smooth and globally consistent depth while still failing to preserve small finite structures near sky or horizon boundaries. Such local failures are not always captured by global metrics such as AbsRel, RMSE, or threshold accuracy, especially when the failed object occupies a small image area.

This paper proposes SkyRoad-CutDepth, a theoretical framework for sky-aware and road-geometric pretext learning in metric monocular depth estimation. The framework is based on a simple observation, dashcam depth learning should not treat all pixels equally. Road regions, pure sky, sky-adjacent foreground objects, and vanishing-point regions have different geometric meanings. Road regions should follow near-to-far structure, pure sky should be treated as undefined or down-weighted, foreground objects inside sky regions should remain valid finite-depth structures and horizon regions should be handled carefully because they combine sky, distant road, buildings, vegetation, signs and traffic structures.

The main contribution of this work is not a new benchmark result. Instead, the paper contributes a theoretically grounded design that connects five existing research directions. Metric depth foundation models, CutDepth augmentation, road-scene geometry, sky or infinity masking and semantic/boundary-aware depth learning. The framework is supported by qualitative evidence from dashcam RGB images and inference outputs obtained using a custom dashcam-dataset-trained Depth Anything V2 model with canonical-space transformations. These outputs show improved global depth consistency but still reveal missing or incomplete depth for poles and signs in sky-adjacent regions.

II. RELATED WORK

A. *Monocular Metric Depth Estimation*

Prior surveys have shown that monocular depth estimation has progressed from early learning-based methods to modern dense-prediction models [1], [3]. However, metric depth prediction from a single RGB image remains ill-posed because absolute scale cannot be uniquely recovered without additional cues such as camera intrinsics, geometric priors, object-scale assumptions, or suitable training distributions. DPT introduced transformer-based dense prediction and demonstrated the value of global context for depth and other pixel-level tasks [4]. TransDepth also used transformer attention to improve continuous pixel-wise prediction [5]. ZoeDepth combined relative and metric depth to achieve zero-shot transfer across depth domains [6]. More recent methods such as Metric3D v2 directly address metric ambiguity by using canonical camera-space transformation and large-scale multi-camera training [7]. UniDepth and UniDepthV2 further emphasize camera-conditioned depth features and metric 3D scene reconstruction from a single image [8], [9].

B. *Foundation Models and Large-Scale Pseudo-Labeling*

Depth Anything and Depth Anything V2 demonstrate the value of large-scale unlabeled data, synthetic data, teacher-student training, and pseudo-labeling for robust monocular depth estimation [10], [11]. Depth Anything V2 is particularly relevant because it produces strong depth predictions and can be fine-tuned using metric depth labels. However, large-scale learning does not automatically solve every local geometric failure. In dashcam scenes, thin poles, small signs, and logos against the sky can still be underrepresented because they occupy limited pixel area and may conflict with the learned sky-background prior.

The qualitative samples used in this paper come from a custom dashcam-dataset-trained Depth Anything V2 model with canonical-space transformations. This means that the observed errors are not simply due to the absence of camera normalization. Instead, they suggest that camera normalization and global fine-tuning may still require additional sky-object and road-geometric priors.

C. *CutDepth and Depth-Specific Augmentation*

CutDepth was proposed as an edge-aware data augmentation method for depth estimation [12]. Instead of using only ordinary image augmentation, CutDepth pastes part of the depth map onto the RGB input during training. The method is useful because monocular depth estimation is pixel-wise and ordinary image transformations may damage geometric consistency. Later variants such as Vertical CutDepth in global-local path networks introduced direction-aware depth augmentation [13].

Although CutDepth is promising, the original formulation is not explicitly aware of road-scene structure. A random depth patch may be reasonable in generic indoor or outdoor data but may be inappropriate in dashcam imagery if it mixes finite road depth into pure sky or destroys horizon geometry. Therefore, this paper argues that CutDepth can be made more suitable for road scenes by using road masks, sky masks, object masks, boundary masks, and vanishing-point regions as augmentation constraints.

D. *Self-Supervised and Semantic-Guided Depth*

Self-supervised monocular depth methods use image reconstruction, pose estimation, and temporal consistency. Monodepth2 introduced minimum reprojection, auto-masking, and full-resolution multi-scale sampling to reduce artifacts and handle violations of camera-motion assumptions [14]. SQLDepth further learns fine-grained scene structure priors from ego-motion and reports strong performance on KITTI and Cityscapes [15]. These methods show that geometric pretext tasks can improve depth learning without requiring dense labels.

Semantic guidance has also been used to handle dynamic objects and improve object-aware depth. SGDepth uses semantic segmentation to guide self-supervised depth estimation in scenes containing moving cars and pedestrians [17]. Semantic pre-training for monocular depth estimation, shows that semantic auxiliary tasks can support depth learning during training [18]. Fine-grained semantics-aware representation learning also shows that local geometry and semantic boundaries can improve depth detail [19]. These works support the use of sky segmentation and foreground-object preservation as pretext tasks.

E. Sky and Infinity Regions

Sky regions are special because they do not correspond to ordinary finite 3D scene surfaces in the same way as road, vehicles, buildings, or poles. MoGe explicitly handles undefined geometry by predicting a valid region mask for infinity regions such as sky [16]. This is directly relevant to the proposed framework. However, pure sky masking alone is insufficient for dashcam depth because finite objects may be surrounded by sky. If the model simply removes or down-weights the whole upper image region, it can suppress road signs, traffic-light poles, lamp posts, and wires. Therefore, this paper distinguishes pure sky from sky-adjacent finite foreground objects.

F. Road-Scene Datasets

KITTI is one of the most widely used autonomous-driving datasets and includes camera, LiDAR, and road-scene data for several perception tasks [21]. Cityscapes provides high-quality pixel-level semantic labels for urban street scenes and is useful for road, sky, vehicle, and traffic-object segmentation [20]. Virtual KITTI 2 provides synthetic driving data with RGB, depth, class segmentation, instance segmentation, optical flow, scene flow, and variations in weather and camera configuration [22]. These datasets demonstrate that depth, semantics, camera geometry, and weather/domain variation are naturally connected in road-scene research.

III. QUALITATIVE FAILURE ANALYSIS

This study examines qualitative outputs from a custom dashcam-dataset-trained Depth Anything V2 model with canonical-space transformations to analyze depth prediction behavior in road-scene imagery. The analysis shows that although global road depth remains visually consistent, thin foreground structures near sky and vanishing-point regions are often weakly represented or partially suppressed. These observations highlight a recurring failure mode in current metric depth prediction and provide the basis for the proposed sky-road guided framework. Two representative examples are shown in Fig. 1 and Fig. 2.

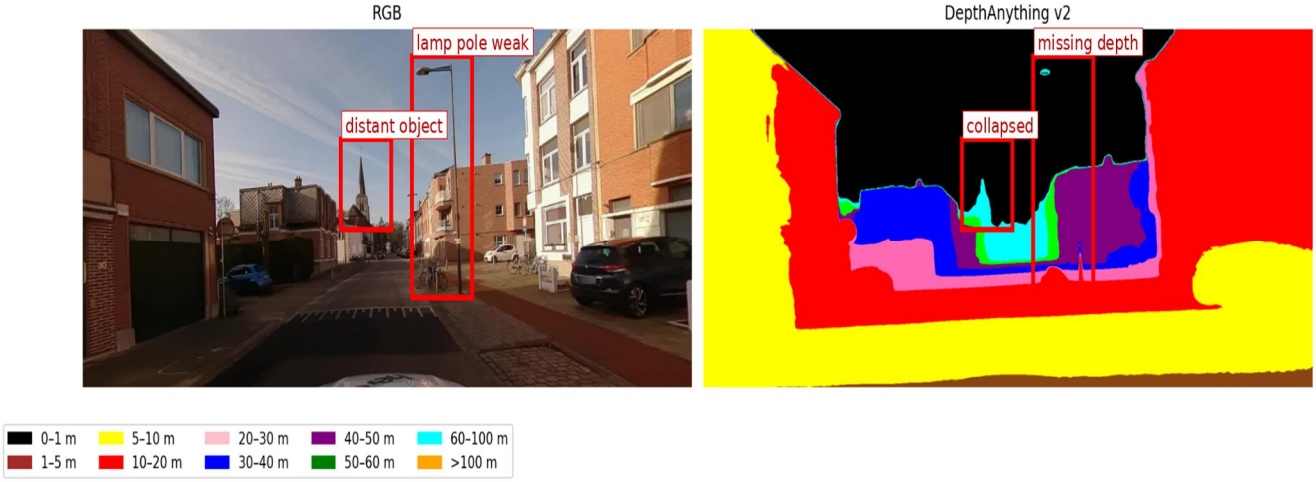


Fig. 1 Qualitative failure case showing weak preservation

In Fig. 1, the RGB image contains a road scene with buildings, a lamp pole, and distant structures near the vanishing point. The predicted depth map represents large regions such as road, nearby buildings, and sidewalks using coherent depth bands. However, the lamp pole in the sky region is not preserved as a clear finite-depth structure in the corresponding output. The region near the horizon also becomes compressed into broad depth categories. This shows that a model may learn useful global depth consistency while still losing thin foreground structures against the sky.

Fig. 2 Qualitative failure case showing partial suppression of thin objects

In Fig. 2, the scene contains traffic congestion, traffic-light poles, a logo/sign, and roadside objects under a cloudy sky. The depth output captures nearby vehicles and broad road-side regions, but several thin structures are weakly represented. The sign/logo appears only as a partial depth blob, and the traffic-light line and poles are not clearly separated from the surrounding sky-background region. This example motivates an explicit sky-object preservation mechanism rather than a simple sky mask.

TABLE I OBSERVED FAILURE CATEGORIES IN QUALITATIVE OUTPUTS

Failure Category	Observed Depth-Map Behaviour	Underlying Estimation Challenge	Proposed Framework
Sky-adjacent foreground suppression	Poles, signs, or traffic lights are missed or partially represented.	Small object width, weak texture, sky-dominated local context, and insufficient semantic separation between non-finite sky pixels and finite foreground pixels.	Introduce sky-object validity masks and boundary-preserving supervision to separate pure sky from finite foreground structures.
Vanishing-point depth collapse	Distant road objects merge into broad depth bands near the horizon.	Low pixel resolution, sparse supervision, and perspective compression.	Apply road-geometric ordering and horizon-aware consistency to preserve near-to-far structure in distant road regions.
Thin-structure boundary degradation	Poles, wires, traffic-light arms, and sign edges lose contour sharpness in the predicted depth map.	Region-dominated depth losses favor large surfaces such as road, buildings, and vehicles, reducing sensitivity to thin structures.	Use thin-object emphasis, edge-aware depth consistency, and region-guided CutDepth sampling around high-risk foreground boundaries.
Pure-sky depth invalidity	Sky regions receive arbitrary, saturated, or semantically inconsistent depth values.	Sky does not correspond to a finite object surface in metric depth estimation and therefore violates standard dense finite-depth assumptions.	Down-weight or mask pure-sky pixels while preserving valid finite-depth supervision for foreground objects projected against sky.

The failure categories in Table I are important because they affect road-scene objects and may be hidden by average depth metrics. A small traffic sign can be crucial even if it occupies less than one percent of the image. Therefore, the proposed framework is designed around region-specific validity rather than uniform pixel-level treatment.

IV. PROPOSED SKYROAD-CUTDEPTH FRAMEWORK

SkyRoad-CutDepth is formulated as a backbone-agnostic training framework for improving metric monocular depth estimation in dashcam imagery. The framework can be integrated with depth backbones such as Depth Anything V2, UniDepth-style metric models, or other transformer-based depth networks. Its central idea is to combine sky-road semantic priors, foreground-object validity, boundary-aware supervision, and region-guided CutDepth augmentation within a unified pretext-learning strategy.

A. Region Definitions

Let I be an RGB dashcam image and D be the predicted metric depth map. The framework assumes four conceptual masks: M_{road} for road pixels, M_{sky} for pure sky or infinity pixels, M_{obj} for finite foreground objects such as poles, signs, vehicles, traffic lights, and boards, and M_{bnd} for object boundaries, sky-object boundaries, and road-object boundaries.

These masks may be manually annotated, generated using a semantic segmentation model, or obtained from datasets such as Cityscapes or Virtual KITTI 2 during future experiments.

The important distinction is that M_{sky} should not remove foreground objects that happen to be located inside the sky background. Therefore, the valid-depth mask can be conceptualized as:

$$M_{valid} = (1 - M_{sky}) \text{ OR } M_{obj},$$

where pure sky is down-weighted or excluded from metric depth supervision, but finite foreground objects remain valid. This simple rule addresses the observed failure in which road signs and poles are treated as sky-like regions.

B. Road-Geometric Pretext Tasks

Road scenes provide strong geometric priors. For most forward-facing dashcam images, road pixels near the bottom of the image are closer, while road pixels near the horizon are farther. Although this is not universally true for steep roads or unusual camera poses, it is a useful weak prior. The road-geometric pretext component therefore encourages near-to-far ordering, smooth road-plane depth, and horizon-aware consistency.

For road pixels divided into near, middle, and far bands, the expected ordering can be written as:

$$\text{mean } (D_{near}) < \text{mean } (D_{middle}) < \text{mean } (D_{far}).$$

This ordering should not be applied to all objects because vehicles, poles, and road signs violate smooth road-plane geometry. It is applied only to road-region pixels. In addition, road-plane smoothness is encouraged using a local gradient penalty on road pixels while preserving discontinuities at object boundaries.

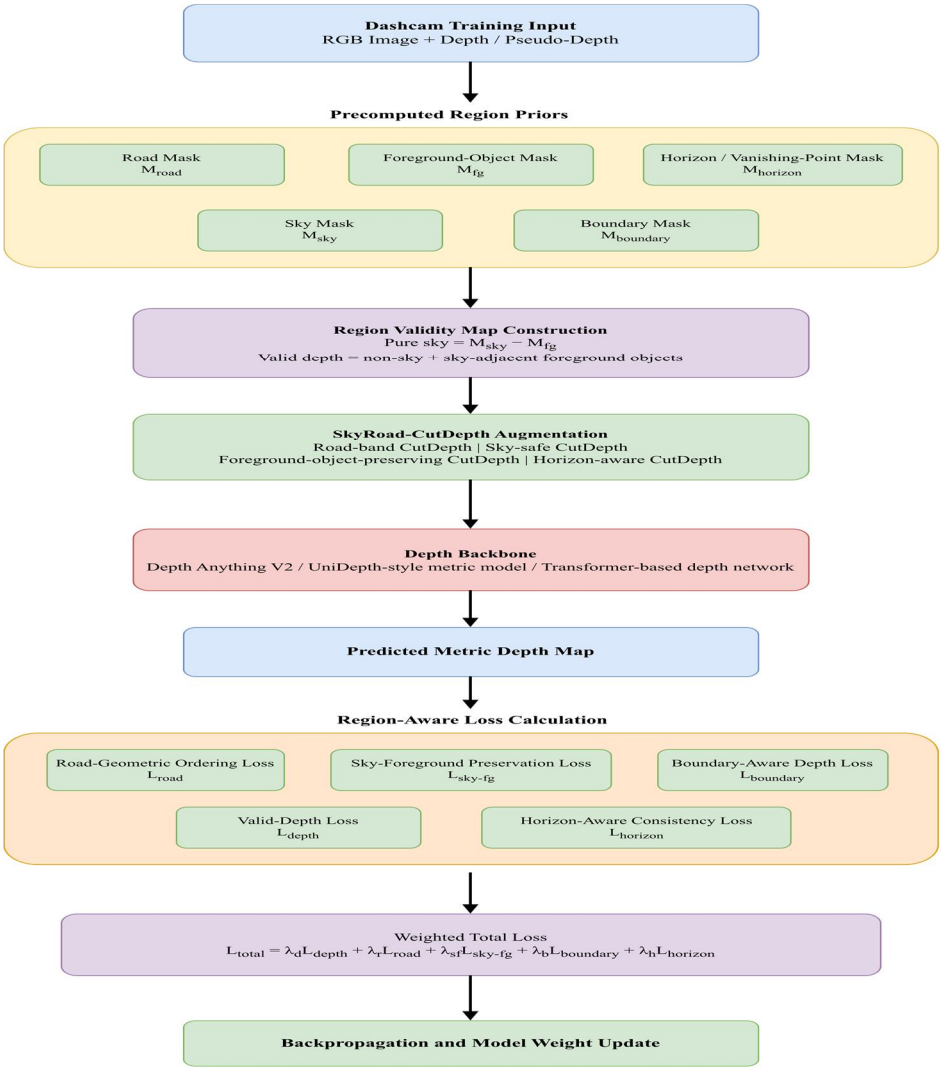


Fig. 3 Overview of the proposed SkyRoad-CutDepth framework

C. Sky-Aware Foreground Preservation

The sky-aware component has two responsibilities. First, it prevents the model from learning arbitrary finite depths for pure sky. Second, it protects foreground objects that appear against the sky. Therefore, sky segmentation is used not only for masking but also for object preservation. A naive approach would set all sky-region pixels as invalid. SkyRoad-CutDepth instead introduces a sky-object region where object pixels remain valid even when surrounded by sky.

The sky-object preservation loss can be expressed conceptually as:

$$L_{skyobj} = L_{valid\text{-}object} + L_{thin\text{-}object} + L_{sky\text{-}boundary},$$

where $L_{valid\text{-}object}$ penalizes invalid depth on finite foreground objects, $L_{thin\text{-}object}$ gives extra importance to poles, signs, and traffic lights, and $L_{sky\text{-}boundary}$ preserves sharp depth discontinuities between sky and foreground structures.

D. Region-Guided CutDepth

The proposed augmentation modifies standard CutDepth by making it region-aware. Instead of selecting arbitrary patches, the framework uses road, sky, object, and boundary priors. Road-band CutDepth applies augmentation within near, middle, or far road bands without destroying the monotonic road-depth structure. Sky-safe CutDepth avoids pasting finite road depth into pure sky. Object-preserving CutDepth oversamples or protects patches around signs, poles, traffic lights, and other thin objects. Horizon-aware CutDepth treats the transition between road, buildings, vegetation, and sky carefully because this is where vanishing-point depth collapse often occurs.

Table II Region-Guided Cutdepth Rules

Region	Region-Specific Training Constraint	Intended Effect on Depth Learning
Road surface	Apply band-wise CutDepth within geometry-compatible near, middle, and far road zones.	Encourages stable near-to-far road-depth consistency.
Pure sky	Avoid finite-depth patch insertion into sky-only regions and down-weight metric supervision.	Prevents the model from learning arbitrary or unstable metric-depth responses for non-finite sky regions.
Sky-adjacent foreground boundary	Protect or oversample patches containing signs, poles, traffic lights, and wires.	Improves the representation of small and thin foreground objects.
Horizon and vanishing-point region	Use controlled patch size and preserve object boundaries.	Reduces distant object collapse.
Large dynamic objects	Exclude vehicles, pedestrians, and cyclists from road-plane smoothness constraints.	Avoids incorrect planar assumptions.

E. Total Conceptual Objective

The overall training objective for future implementation can be written as:

$$L_{total} = L_{depth} + \lambda_1 L_{road} + \lambda_2 L_{sky} + \lambda_3 L_{skyobj} + \lambda_4 L_{boundary} + \lambda_5 L_{cutdepth}.$$

Here, L_{depth} is the main metric depth loss, L_{road} enforces road-plane consistency and near-to-far ordering, L_{sky} handles pure sky validity, L_{skyobj} preserves finite foreground structures in sky-background regions, $L_{boundary}$ sharpens semantic and depth discontinuities, and $L_{cutdepth}$ ensures consistency under region-guided CutDepth augmentation. The lambda terms are weighting factors that should be selected empirically in future work.

V. METHODOLOGY

The proposed methodology can be implemented in four stages. The first stage collects dashcam RGB images and either obtains or predicts semantic masks for road, sky, foreground objects, and boundaries. The second stage generates qualitative failure maps by comparing RGB-visible objects with their corresponding predicted depth representation. The third stage applies SkyRoad-CutDepth rules during training. The fourth stage trains or fine-tunes a depth backbone using the total loss formulation described above.

A. Step-Wise Training Strategy

1) Step 1: Estimate semantic and geometric masks from dashcam images. Road, sky, and object masks can come from a segmentation model or from datasets with semantic labels. Horizon or vanishing-point regions can be estimated geometrically or approximated using road-sky transition cues.
2) Step 2: Generate region-specific augmentation candidates. Patches are categorized as road patches, sky patches, object patches, boundary patches, or horizon patches. Each patch type has a different CutDepth rule.
3) Step 3: Train auxiliary pretext heads. The model predicts road mask, sky mask, boundary mask, and optionally horizon or road-band labels. These heads can be used during training and removed during deployment, similar to semantic pre-training strategies.
4) Step 4: Fine-tune the metric depth model using region-guided losses. Pure sky receives down-weighted depth supervision, while finite foreground objects near sky receive additional boundary and validity penalties.

B. Why the Framework Is Needed

The qualitative examples show that a dashcam dataset based custom-trained model can learn coherent road and building depth but still miss thin structures near the sky. This suggests that global metric consistency and local foreground preservation are not equivalent. A model can score well in broad scene regions while still failing on small but important road objects. Therefore, a region-aware learning strategy is justified.

The framework also avoids a limitation of simple sky masking. MoGe and related geometry estimation works show that infinity regions should be handled carefully [16]. However, a road-scene model must go one step further by distinguishing pure sky from finite objects projected against the sky. SkyRoad-CutDepth therefore uses sky segmentation as a validity prior, not as a hard removal rule.

C. Difference from Standard CutDepth

Standard CutDepth is depth-aware but not scene-aware. It improves data variation without destroying edge features [12], but it does not explicitly know whether a patch lies on road, sky, object, or horizon. SkyRoad-CutDepth changes the role of augmentation from random regularization to structured geometric regularization. The patch selection process becomes guided by semantic and geometric constraints.

D. Difference from General Semantic Guidance

Semantic-guided depth methods usually use semantic information to handle dynamic objects, improve segmentation-depth consistency, or sharpen object boundaries [17]-[19]. The proposed framework is narrower and more failure-specific. It focuses on sky-adjacent foreground objects and vanishing-point regions in dashcam imagery. This makes the contribution targeted rather than a general multi-task segmentation-depth architecture.

E. Expected Evaluation Plan

Future experimental validation should use standard depth metrics such as AbsRel, SqRel, RMSE, RMSE log, δ_1, δ_2, and δ_3. However, the central failure mode requires additional region-specific metrics. These may include road-region AbsRel, sky-object boundary F1, thin-object depth error, horizon-region depth error, and object-preservation recall for poles, signs, and traffic lights. Without such local metrics, improvements in safety-relevant regions may be hidden by global averages.

VI. FUTURE WORK

The present study focuses on the conceptual formulation and qualitative analysis of the proposed SkyRoad-CutDepth framework. The qualitative examples are used to identify recurring failure patterns in dashcam metric depth prediction, particularly around sky-adjacent foreground objects and vanishing-point regions. Future work will extend this formulation through quantitative training, ablation studies, and benchmark evaluation to measure its effect on global depth accuracy, foreground-object preservation, and horizon-region consistency. The framework relies on semantic and geometric priors such as road, sky, foreground-object, boundary, and horizon masks. Therefore, the quality of these priors can influence the effectiveness of the proposed training strategy. Future implementations should compare different sources of region priors, including manually annotated masks, off-the-shelf segmentation models, and dataset-provided semantic labels. Additional evaluation under night scenes, heavy rain, fog, overexposure, steep road slopes, and non-forward-facing camera views would further clarify the generality of the framework.

Overall, SkyRoad-CutDepth is designed as a modular training strategy that can be integrated with existing monocular depth backbones. This makes the framework practical for future implementation because it modifies the supervision and augmentation process rather than requiring a completely new depth foundation model.

VII. CONCLUSION

This paper presented SkyRoad-CutDepth, a theoretical framework for sky-aware and road-geometric pretext learning in metric monocular depth estimation for dashcam imagery. The work was motivated by qualitative outputs from a custom dashcam-dataset-trained Depth Anything V2 model with canonical-space transformations. The outputs show that global road and building depth can be consistent while thin foreground objects near sky and vanishing-point regions remain partially suppressed or invalid.

The proposed framework combines road-plane priors, sky validity, foreground-object preservation, boundary-aware losses, and region-guided CutDepth augmentation. The central idea is that pure sky, road, horizon, and sky-adjacent foreground objects should not be treated uniformly during training. Pure sky can be down-weighted, road can be constrained by near-to-far geometry, and finite objects such as poles, signs, traffic lights, and wires should be protected through sky-object validity and boundary preservation. Although full quantitative validation is future work, the proposed framework provides a focused and literature-supported direction for robust metric depth estimation in road scenes.

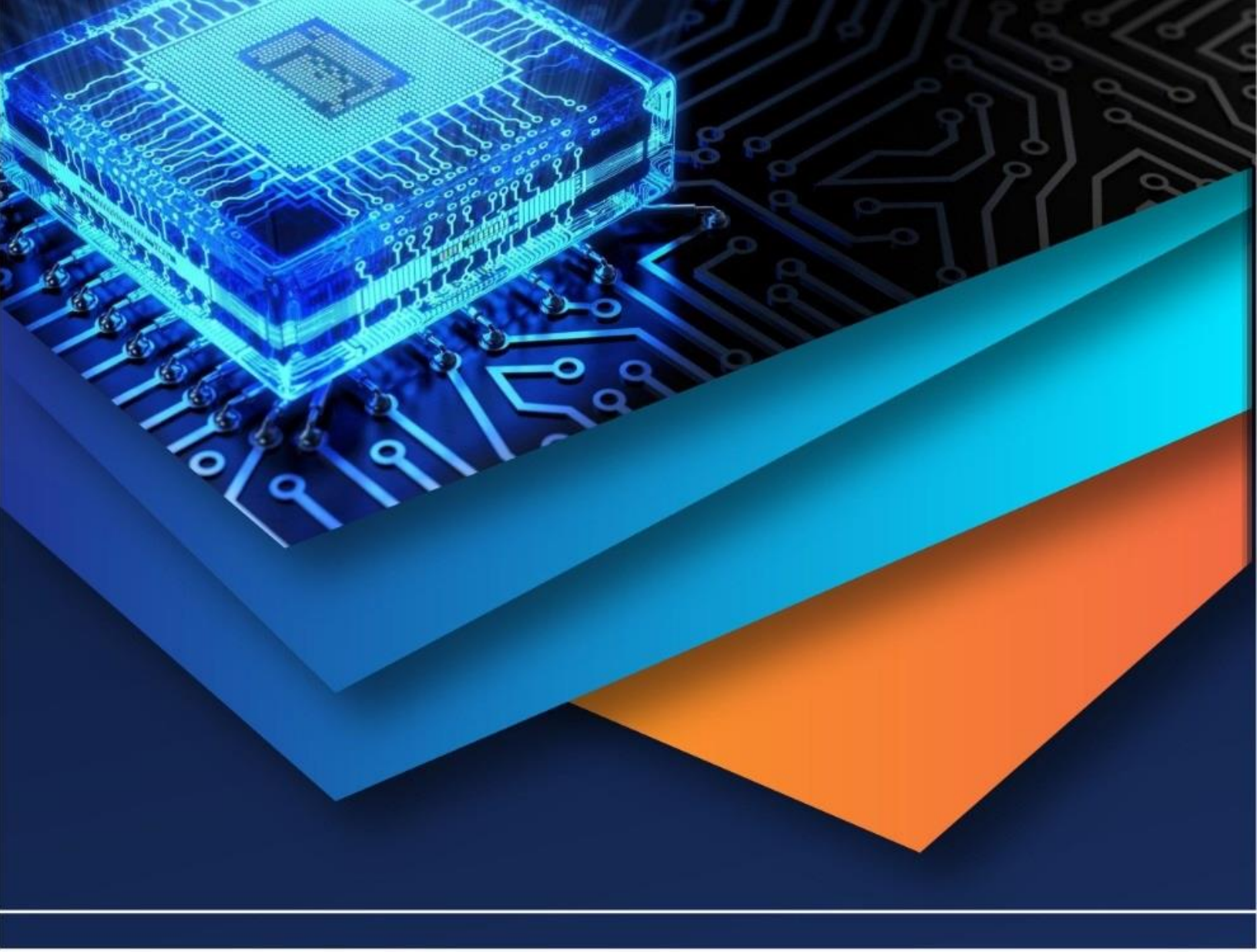
VIII. ACKNOWLEDGMENT

The author thanks the open-source computer vision community for providing public depth-estimation models and research resources. The qualitative analysis in this paper uses dashcam RGB images and depth outputs generated for research discussion. Author names, affiliation details, and acknowledgments may be updated before final submission.

REFERENCES

[1] A. Bhoi, “Monocular depth estimation: A survey,” arXiv preprint arXiv:1901.09402, Jan. 2019, doi: 10.48550/arXiv.1901.09402.
[2] Y. Ming, X. Meng, C. Fan, and H. Yu, “Deep learning for monocular depth estimation: A review,” Neurocomputing, vol. 438, pp. 14-33, May 2021, doi: 10.1016/j.neucom.2020.12.089.
[3] T. Ehret, “Monocular depth estimation: A review of the 2022 state of the art,” Image Processing On Line, vol. 13, pp. 38-56, 2023, doi: 10.5201/ipol.2023.459.
[4] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 12179-12188, doi: 10.1109/ICCV48922.2021.01196.
[5] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, “Transformer-based attention networks for continuous pixel-wise prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 16269-16279, doi: 10.1109/ICCV48922.2021.01596.
[6] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, “ZoeDepth: Zero-shot transfer by combining relative and metric depth,” arXiv preprint arXiv:2302.12288, Feb. 2023, doi: 10.48550/arXiv.2302.12288.
[7] M. Hu et al., “Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation,” arXiv preprint arXiv:2404.15506, Apr. 2024, doi: 10.48550/arXiv.2404.15506.
[8] L. Piccinelli et al., “UniDepth: Universal monocular metric depth estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 10106-10116, doi: 10.1109/CVPR52733.2024.00963.
[9] L. Piccinelli, C. Sakaridis, Y.-H. Yang, M. Segu, S. Li, W. Abbeloos, and L. Van Gool, “UniDepthV2: Universal monocular metric depth estimation made simpler,” arXiv preprint arXiv:2502.20110, Feb. 2025, doi: 10.48550/arXiv.2502.20110.
[10] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth Anything: Unleashing the power of large-scale unlabeled data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 10371-10381, doi: 10.1109/CVPR52733.2024.00987.
[11] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, “Depth Anything V2,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 37, pp. 21875-21911, 2024, doi: 10.52202/079017-0688.
[12] Y. Ishii and T. Yamashita, “CutDepth: Edge-aware data augmentation in depth estimation,” arXiv preprint arXiv:2107.07684, Jul. 2021, doi: 10.48550/arXiv.2107.07684.
[13] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, “Global-local path networks for monocular depth estimation with vertical CutDepth,” arXiv preprint arXiv:2201.07436, Jan. 2022, doi: 10.48550/arXiv.2201.07436.
[14] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 3827-3837, doi: 10.1109/ICCV.2019.00393.
[15] Y. Wang, Y. Liang, H. Xu, S. Jiao, and H. Yu, “SQLDepth: Generalizable self-supervised fine-structured monocular depth estimation,” in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 6, pp. 5713-5721, 2024, doi: 10.1609/aaai.v38i6.28383.
[16] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang, “MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 5261-5271, doi: 10.1109/CVPR52734.2025.00496.
[17] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, “Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 582-600, doi: 10.1007/978-3-030-58565-5_35.
[18] P. Rottmann et al., “Improving monocular depth estimation by semantic pre-training,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2021, doi: 10.1109/IROS51168.2021.9636546.

[19] H. Jung, E. Park, and S. Yoo, "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 12642-12652, doi: 10.1109/ICCV48922.2021.01241.
[20] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 3213-3223, doi: 10.1109/CVPR.2016.350.
[21] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp. 3354-3361, doi: 10.1109/CVPR.2012.6248074.
[22] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," arXiv preprint arXiv:2001.10773, Jan. 2020, doi: 10.48550/arXiv.2001.10773.
[23] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. Caye Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 9492-9502, doi: 10.1109/CVPR52733.2024.00907.
[24] S. Patni, A. Agarwal, and C. Arora, "ECoDepth: Effective conditioning of diffusion models for monocular depth estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 28285-28295, doi: 10.1109/CVPR52733.2024.02672.
[25] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 5729-5739, doi: 10.1109/ICCV51070.2023.00527.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY