EG UK Computer Graphics & Visual Computing (2024)
A. Slingsby and D. Hunter (Editors)

A Stereo-Integrated Novel View Synthesis Pipeline for the Enhancement of Road Surface Reconstruction Dataset

M. Zhan[1], T. Morley[1] and M. Turner[1]

[1]The University of Manchester, England, UK

Abstract

This proposal outlines a novel view synthesis pipeline designed for road reconstruction in autonomous driving scenarios that leverages virtual camera technology to synthesise images from unvisited camera poses, thereby enhancing and expanding current datasets. It consists of three main steps: data acquisition, data preprocessing and fusion, and then importantly interacting with new 3D view synthesis with geometric priors. The modular design allows each component to be independently optimised and upgraded, ensuring flexibility and adaptability to various datasets and task requirements. The proposed approach aims to improve the robustness, realism, and photometric consistency of novel view synthesis, effectively handling dynamic scenes and varying lighting conditions. Additionally, this research plans to open-source a low-cost stereo camera hardware solution with the included software implementation.

CCS Concepts
• ***Computing methodologies*** → ***Reconstruction; 3D imaging;*** *Computational photography;*

1. Introduction

The rapid development of autonomous vehicles has imposed higher demands on safety and ride comfort. Road feedback is a critical factor in the interaction between vehicles and the physical environment, directly affecting vehicle performance and comfort. Traditional vehicle navigation systems primarily rely on static maps and road construction records for road surface information [BHLL23]. However, these data sources often lack precision, failing to capture critical details such as minor road damage, complex gradient changes, or hidden obstacles [GUGK17]. These deficiencies limit the precision of autonomous vehicles' decision-making and dynamic planning capabilities, potentially hindering their ability to effectively respond to sudden situations and complex road conditions, thereby impacting overall driving safety and ride comfort.

Significant progress has been made in research utilising stereo cameras, LiDAR, and radar sensors to collect road surface data [GLSU13, CBL*20, COR*16]. These sensors can provide high-precision road data, enabling accurate reconstruction of road geometry. However, several key issues persist with these methods. Firstly, existing data primarily relies on fixed-position onboard cameras, making it difficult to capture road details from different perspectives, resulting in inconsistent performance of the same perception algorithm across different vehicles, thereby affecting the overall performance of autonomous systems [DTP21]. Secondly, repeated data collection with various sensor configurations is costly, making it impractical for real-world applications [PNC*22]. Additionally, ensuring data consistency under varying temporal and spatial conditions poses a significant challenge.

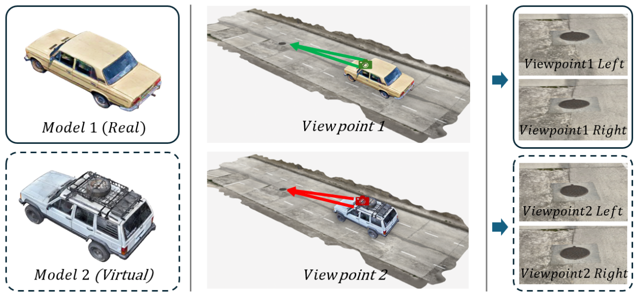

Figure 1: *This diagram illustrates the process of generating image data using virtual cameras combined with novel view synthesis. The diagram is divided into two sections: real images (yellow car) and virtual synthesised images (blue car). Real images are obtained from real-world data collection, while the synthesised images are collected by positioning virtual cameras at different poses within a 3D scene. Source of 3D models: [pol]*

To overcome these issues, several studies propose using virtual cameras combined with multi-view stereo (MVS) technology to capture new viewpoint image data (see Figure 1) [YLL*20,

DOI: 10.2312/cgvc.20241225

LHH16]. Although these methods have made significant progress in geometric reconstruction, they still exhibit noticeable deficiencies in texture realism and lighting consistency. With the advent of Novel View Synthesis (NVS) methods such as Neural Radiance Fields (NeRF) [MST*21] and Gaussian Splatting (GS) [KKLD23], these problems have been greatly alleviated. These methods simulate light radiance fields and three-dimensional Gaussian distributions, enabling high-fidelity reproduction of complex colour textures and lighting scenes, generating realistic new viewpoint images and demonstrating significant potential. Nevertheless, applying these methods to autonomous driving datasets remains a challenging task. Due to the outward-facing data capture approach used in autonomous driving, these datasets typically feature large-scale scenes with low image overlap rates [XZL*22]. This presents significant challenges in achieving the high level of detail and consistency required for accurate and reliable novel view synthesis.

In this paper, we propose a design scheme for new viewpoint generalisation and scene reconstruction processes for autonomous driving data. This process aims to improve the precision and robustness of raw data by optimising the integration of multimodal autonomous driving data. By combining stereo vision, Structure from Motion, and Novel View Synthesis, we aim to achieve accurate reconstruction of autonomous driving road scenes. Our goal is to enhance geometric and photometric consistency of synthesised new viewpoints, accurately estimate camera poses, and reduce artifacts and blurring issues in the scene. Although this scheme is still in the design phase, we believe it offers new insights and directions for addressing the precision issues of existing systems. The designed process includes the following key steps:

- **Stereo Image Enhancement**: Adopting modular stereo image enhancement algorithms to improve the visual quality and clarity of stereo images, addressing the degradation caused by environmental factors, motion blur, and lighting variations.
- **Stereo-Integrated Structure from Motion**: Making use of the constant relative pose of stereo cameras to provide pose priors and optimise SfM parameters, enhancing accuracy of pose estimation while improving reconstruction precision.
- **Multi-model Depth Completion**: Utilise multi-modal data processed from autonomous driving datasets for depth completion to incorporate complementary information and enhance the robustness and accuracy of the final depth maps.
- **Novel View Synthesis**: Generation of high-fidelity novel views using NeRF with multi-criteria supervision.

Additionally, we have developed a low-cost, open-source solution for building stereo camera systems. This solution can assist researchers and developers in easily constructing and deploying stereo camera setups.

2. Related Work

Our work introduces a new pipeline for NVS in road surface reconstruction datasets. This pipeline incorporates key techniques such as stereo image enhancement, SfM, and NVS. The following section provides a review of the related research in these areas.

2.1. Stereo Image Enhancement

The high dynamics and complexity of autonomous driving scenarios present various challenges, such as motion blur, insufficient resolution of distant objects, and inadequate ambient lighting during image capture. To address these issues, stereo image enhancement techniques are widely applied at the image preprocessing stage to improve image quality. Common techniques include stereo deblurring [ZZZ*19], low-light enhancement [HFX*22], dehazing [NPX*21], and super-resolution [CCY22]. These techniques are crucial in 3D reconstruction and novel view synthesis, as they significantly improve geometric reconstruction accuracy and ensure the consistency and richness of scene details by restoring high-quality scene information.

2.2. Structure From Motion

Structure from Motion (SfM), as a method for estimating spatial geometric information and camera motion from a collection of multi-view images, is widely applied in 3D reconstruction, MVS, and NVS tasks. The primary steps involved in SfM include feature point detection and matching, camera pose estimation, sparse reconstruction, and bundle adjustment [SF16]. In the 3D point cloud reconstruction phase, the structure of objects is recreated using triangulation based on camera poses and feature point positions. Bundle adjustment is a global optimisation process that further refines camera poses and 3D point locations by minimising the reprojection error. Incremental SfM, known for its robustness and scalability, is the most commonly used approach and has led to the development of notable SfM algorithms such as Bundler [Sna08], VisualSfM [W*11], and COLMAP [SF16]. In the pipeline of NVS, COLMAP has established itself as a standard step due to its advanced performance. For autonomous driving datasets, the continuity of image sequences allows incremental SfM to better exploit adjacent frame information, gradually accumulating and refining 3D reconstruction results, thereby enhancing the accuracy.

2.3. Novel View Synthesis With Depth Prior

NVS technology generates images or videos from previously unseen viewpoints based on existing ones. Early NVS methods, such as Local Light Field Fusion (LLFF) [MSOC*19] and Scene Representation Networks (SRN) [SZW19], have demonstrated a certain degree of ability to generalise perspectives by leveraging multiple views of the same scene. However, these approaches encounter significant challenges in complex lighting and reflective environments. Advanced techniques like Neural Radiance Fields (NeRF) [MST*21] and Gaussian Splatting (GS) [KKLD23] have markedly enhanced the fidelity of generated images and improved performance in complex lighting conditions. These advancements, however, depend on the availability of a substantial number of viewpoints to support model training. Consequently, in scenarios with sparse viewpoints, reconstructing 3D model geometry and photometry becomes considerably more challenging.

To address this issue, several NeRF-based approaches incorporating depth priors have been proposed. DSNeRF [DLZR22] first utilised sparse point cloud depth information from SfM to supervise light particle distribution. However, the data's sparsity limits model performance improvement. To overcome this, methods to obtain dense depth information have been developed. Dense Depth

Prior NeRF [RBM*22] enhances sparse point cloud depth by introducing depth completion and uncertainty techniques. MonoSDF [YPN*22] uses deep learning-based monocular depth estimation, while StereoNeRF [LJBC24] employs stereo cameras for scene depth estimation.

These methods, however, are mainly limited to indoor scenes. Accurate depth estimation in outdoor environments is more challenging due to greater variations in lighting, weather conditions, and dynamic elements, complicating visual-based depth estimation. In this context, methods such as UrbanNeRF [RLS*22] and StreetNeRF [XZL*22] give up traditional SfM pipelines in favour of LiDAR-based processes to acquire sparse point cloud information and estimate camera poses. These methods employ depth completion techniques to produce denser and more detailed depth maps. While this approach yields more accurate global depth estimations, it introduces a significant dependence on LiDAR data. This reliance not only elevates data collection costs and computational complexity but also limits the integration of multimodal data. For instance, leveraging image texture information to derive prior knowledge of scene geometry can alleviate inaccuracies in sparse depth completion, particularly in local regions where data may be lacking.

3. Aims and Objectives

This work aims to perform high-quality road surface 3D reconstruction and novel view synthesis under varying scenarios with sparse views. The objectives are as follows:

- Design a low-coupling novel view synthesis process by modularising the algorithm steps, aiming to enhance the flexibility of the pipeline and improve its applicability to various tasks and datasets.
- Leverage the multimodal advantages of autonomous driving datasets to jointly integrate and optimise existing data, thereby improving the accuracy and efficiency of synthesis models.
- Improve existing novel view synthesis methods to enhance model performance, focusing on increasing computational efficiency, improving geometric reconstruction accuracy, and ensuring photometric consistency of the generated images.

By achieving these objectives, our proposed pipeline, compared to traditional ones, not only allows for modular and flexible adaptation to different task scenarios but also effectively enhances model accuracy and robustness through comprehensive integration of existing data and advanced novel view synthesis methods.

4. Pipeline Design

To address the challenges of novel view synthesis in autonomous driving road reconstruction datasets, we propose a modular 3D reconstruction and novel view synthesis pipeline that can be flexibly adjusted for different scenarios and data. As illustrated in Figure 2, our designed framework not only optimises the traditional SfM workflow and fully integrates data, but also enhances model performance using novel view synthesis techniques with geometric priors, specifically NeRF. The entire process can be broadly divided into three main components: data acquisition, data preprocessing and fusion, and novel view synthesis with geometric priors. In this section, we will provide a detailed introduction and justification for the design principles of each component.

4.1. Data Acquisition

Due to the limited availability of road reconstruction datasets based on autonomous driving [ZXD*24], we have designed a low-cost, open-source stereo vision device to facilitate data collection for specific task scenarios and model performance testing [Github Page]. Based on the Raspberry Pi platform, it benefits from the extensive community support and flexible hardware expansion capabilities of the Raspberry Pi, enabling users to achieve efficient stereo vision applications at a low cost. It is important to note that the data acquisition phase is not mandatory, as this entire workflow is fully applicable to existing general autonomous driving datasets. To validate the effectiveness of the proposed pipeline, we conducted preliminary experiments on both an autonomous driving-based road reconstruction dataset and a dataset collected using our custom-developed device. The experiments demonstrated that the proposed pipeline can achieve the basic objectives for both types of datasets (see Section 5).

4.2. Data Preprocessing and Fusion

The designed data preprocessing and fusion pipeline is primarily divided into four main modules: a stereo synthesis-based image enhancement module, an SfM-based sparse reconstruction module, a stereo matching module, and a depth completion module. Except for the final depth completion module, the other modules can be flexibly replaced with alternative methods or modified with additional sub-modules. This low-coupling, modular design allows each module to be independently optimised and upgraded, thereby adapting to different datasets and task requirements.

Given that stereo cameras have become standard equipment in autonomous vehicles, stereo vision technology has been integrated into the data preprocessing workflow. Due to the high dynamic nature of autonomous driving scenarios, the quality of data collected by image sensors is often compromised. To address this, a stereo image enhancement module has been introduced to improve the quality of raw images. This module utilises advanced image processing techniques such as dehazing [NPX*21], deblurring [ZZZ*19], and low light enhancement [HFX*22] to improve image clarity and detail fidelity, ensuring that subsequent processing steps can operate on higher-quality data for analysis. Additionally, the stereo matching module performs initial dense depth estimation by utilising image disparity and texture information.

For the SfM-based sparse reconstruction process (such as using COLMAP), road scenes sometimes lack effective reference features, and road textures degrade with increasing distance, leading to potential drift or instability in pose estimation. To mitigate this issue, we plan to introduce rigid relative pose constraints from the stereo cameras to optimise the camera pose estimation. Additionally, we have adjusted and optimised the default parameters in the COLMAP process based on the characteristics of autonomous driving scenarios to improve the convergence speed and accuracy of sparse reconstruction.

The final step in data preprocessing and fusion is depth completion. Traditional depth completion methods often use only one or two modalities of data as input. Our design, however, fully processes existing data and extracts information from multiple modalities, increasing the dimensionality of the guidance data for depth

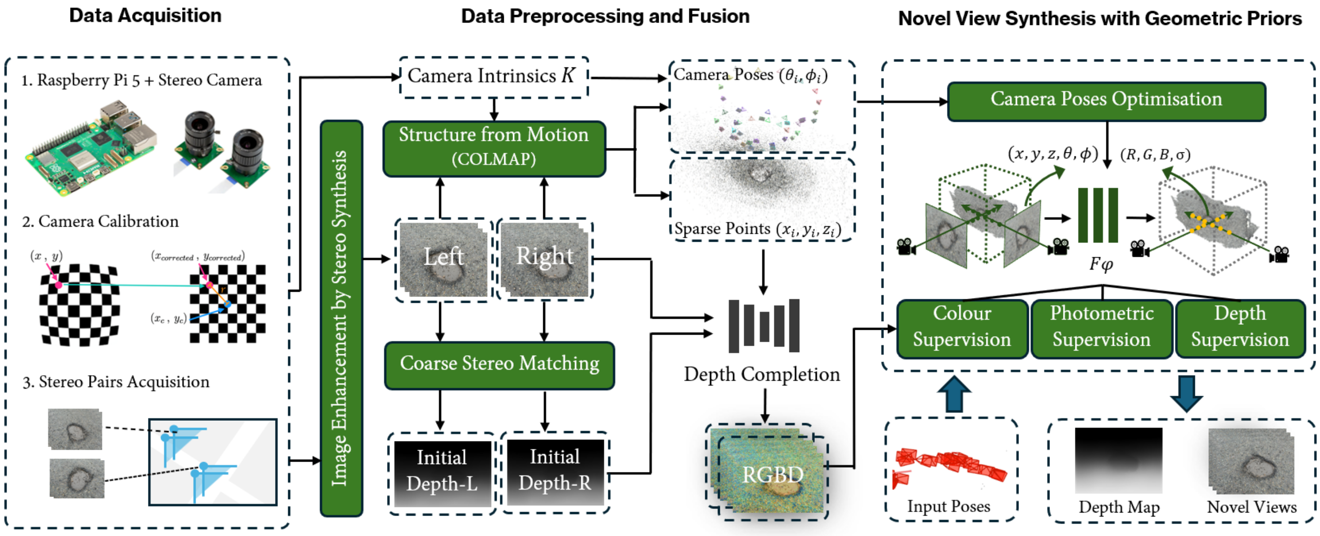

Figure 2: *The proposed pipeline can be divided into three main stages: data acquisition, data preprocessing, and novel view synthesis. First, data is collected using a stereo camera (this step can be skipped if using an existing dataset). During data preprocessing, image quality is enhanced through stereo synthesis techniques. COLMAP is then used for SfM to estimate camera poses and generate sparse point clouds, followed by coarse stereo matching and depth completion to create dense depth maps. In the novel view synthesis stage, camera pose optimisation, colour supervision, photometric supervision, and geometric supervision are applied to improve 3D reconstruction accuracy and photometric consistency. Finally, by inputting new camera poses, corresponding depth images and synthesised novel views are generated.*

completion to three dimensions. These three-dimensional data include initial dense depth maps from stereo matching, texture-based geometric contours from RGB images, and sparse point clouds from 3D reconstruction. Utilising the modular nature of our workflow, the sparse point clouds can be replaced with LiDAR point clouds as needed.

4.3. Novel View Synthesis with Geometric Prior

Due to the use of discrete 3D Gaussian points to represent scene information, Gaussian splatting makes it challenging to accurately supervise training results using dense depth information. In contrast, NeRF utilises ray sampling techniques, enabling geometric supervision of image pixels based on the density distribution of colour particles. Therefore, the NeRF are selected for preliminary experiments in the pipeline. Based on the derivations in DSNeRF [DLZR22], the depth estimation formula derived from the volumetric rendering equation is as follows:

$$\hat{D}(\mathbf{r}) = \sum_{i=1}^{N} w_i t_i, \quad w_i = T(t_i)\left(1 - \exp(-\sigma_i \Delta t_i)\right) \tag{1}$$

where N is the total number of samples, t_i is the position of the i-th sample point, w_i is the weight of each sample point and $T(t_i)$ is the accumulated transmittance from the origin to t_i.

Given that the dense depth information estimated in the previous steps is derived from basic stereo matching and the fusion of sparse point clouds, to handle local depth anomalies more robustly and smoothly, the depth supervision will be based on the following Huber Loss formula. This formula combines the advantages of Mean Squared Error (MSE) and Mean Absolute Error (MAE), providing robustness while simplifying computations:

$$\mathcal{L}_\delta(z, z_{\text{sensor}}) = \begin{cases} \frac{1}{2}(z - z_{\text{sensor}})^2 & \text{for } |z - z_{\text{sensor}}| \leq \delta, \\ \delta|z - z_{\text{sensor}}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \tag{2}$$

where δ denotes the threshold at which the loss function transitions between quadratic and linear loss. When the absolute error between the predicted value z and the sensor measurement z_{sensor} is less than or equal to δ, the loss function uses the quadratic form $\frac{1}{2}(z - z_{\text{sensor}})^2$. Conversely, when the absolute error $|z - z_{\text{sensor}}|$ exceeds δ, the loss function switches to the linear form $\delta|z - z_{\text{sensor}}| - \frac{1}{2}\delta^2$. This linear loss reduces the impact of large errors, improving the model's robustness to outliers.

Additionally, in the complex and uncontrollable lighting conditions of autonomous driving scenarios, the quality of data captured by onboard image sensors can be significantly affected. To address this issue, extended photometric loss [MBRS*21] will be introduced during model training. By optimising photometric loss, colour and brightness consistency can be maintained under varying lighting conditions, thereby improving the accuracy of 3D reconstruction and the realism of novel view synthesis.

5. Preliminary Experiment Results

This section presents the preliminary experiment results of our proposed novel view synthesis pipeline for road reconstruction in autonomous driving scenarios. The experiments were conducted to evaluate the effectiveness of incorporating camera pose constraints in SfM, the performance of the pipeline on different datasets, and the efficiency of depth supervision methods.

SfM with Camera Pose Constraints

In our preliminary experiments, we introduced additional camera pose constraints in SfM and optimised the parameters of SfM to improve the accuracy of pose estimation. By incorporating rigid relative pose constraints from stereo cameras, we observed a reduction in pose drift and increased stability in dynamic scenes. In Figure 3, we selected two scenes to demonstrate the changes in camera poses before and after applying pose constraints and parameter optimisation during sparse reconstruction.

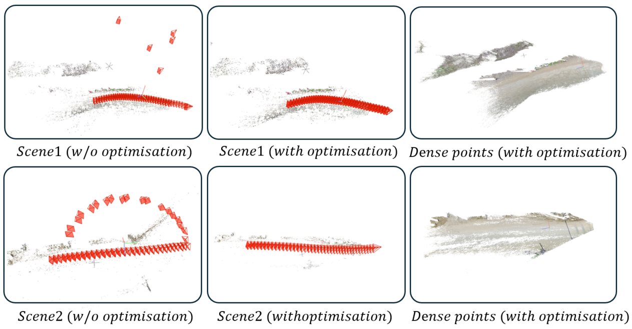

Figure 3: *The first column of images shows the results of sparse reconstruction without pose constraints and parameter tuning, where noticeable camera pose drift is evident. The second column illustrates the results after introducing rigid relative pose constraints and parameter tuning, demonstrating a significant reduction in pose drift. The final column presents the optimised dense point cloud reconstruction, showing a high quality of dense reconstruction.*

Generalisation of processes across different datasets

To demonstrate the versatility of the proposed pipeline in road surface reconstruction, we evaluated the pipeline on various datasets, including the existing autonomous driving road reconstruction dataset RSRD [ZXD*24] and datasets collected using our custom-developed stereo camera. As a preliminary experiment, we implemented the backbone network of the pipeline using Instant Neural Graphics Primitives (Instant NGP) [MESK22] embedded with depth supervision, aiming to accelerate the training process. In terms of metrics, we chose the Peak Signal-to-noise Ratio (PSNR) and Structural Similarity Index (SSIM) [WBSS04].

As Figure 4 shows, our proposed pipeline exhibits certain robustness in novel view synthesis with road reconstruction datasets of varying qualities and scenarios. By incorporating depth supervision, the performance of the pipeline significantly surpass traditional algorithms (COLMAP) in scenarios with relatively sparse images. Three datasets were used in preliminary experiments: an 80-frame dense road reconstruction dataset, a 50-frame self-collected sparse road dataset, and a 42-frame self-collected pothole dataset. The image sizes are downsampled by factors of 8, enabling the networks to converge in under 15 minutes. For the road reconstruction datasets, significant differences highlighted in red boxes demonstrated our method's superiority over COLMAP. Comparative experiments with the RSRD dataset underscored the importance of depth supervision, showing that methods with depth supervision achieve better results in SSIM and PSNR metrics and qualitative image comparisons with the same number of training iterations (30,000). While preliminary, these results highlight our methodology's potential. Future work will involve comprehensive experiments to validate the pipeline's effectiveness and generalisability across various datasets and conditions.

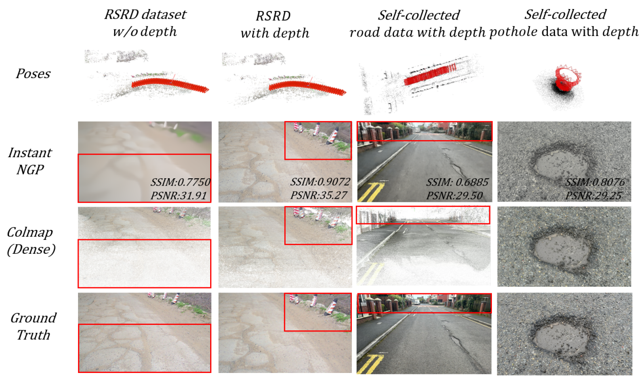

Figure 4: *Preliminary experimental results. To provide a clearer visualisation of the scenes, we included reconstruction results with camera poses in the top row.*

6. Conclusion

In this paper, we discuss the design idea of a novel view synthesis pipeline developed for road reconstruction tasks. This approach employs virtual camera technology to synthesise images from unvisited camera poses, thereby enhancing and expanding the existing road reconstruction dataset. The pipeline primarily consists of three steps: data acquisition, data preprocessing and fusion, and novel view synthesis with geometric priors. Furthermore, the modular design allows each component of the pipeline to be independently optimised and upgraded, enabling flexibility to adapt to different datasets and task requirements. Through this proposed pipeline, we aim to enhance the robustness, realism, and photometric consistency of difficult view synthesis for autonomous driving road reconstruction. This method is capable of handling complex dynamic scenes and maintaining high-quality image synthesis across varying lighting conditions.

Additionally, we plan to develop our proposed modular process into software and release it under an open source licence. We will also integrate existing software and algorithms as submodules into this process, providing users with more options. Furthermore, we will optimise and open source our current stereo camera hardware and software system. By achieving these goals, we aim to make significant contributions to the fields of autonomous driving and 3D reconstruction, offering new insights and tools for future research and development.

References

[BHLL23] Bao Z., Hossain S., Lang H., Lin X.: A review of high-definition map creation methods for autonomous driving. *Engineering Applications of Artificial Intelligence 122* (2023), 106125. 1

[CBL*20] Caesar H., Bankiti V., Lang A. H., Vora S., Liong V. E., Xu Q., Krishnan A., Pan Y., Baldan G., Beijbom O.: nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 11621–11631. 1

[CCY22] Chu X., Chen L., Yu W.: Nafssr: stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 1239–1248. 2

[COR*16] Cordts M., Omran M., Ramos S., Rehfeld T., Enzweiler M., Benenson R., Franke U., Roth S., Schiele B.: The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 3213–3223. 1

[DLZR22] Deng K., Liu A., Zhu J.-Y., Ramanan D.: Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12882–12891. 2, 4

[DTP21] Dey J., Taylor W., Pasricha S.: Vespa: A framework for optimizing heterogeneous sensor placement and orientation for autonomous vehicles. *IEEE Consumer Electronics Magazine 10* (2021), 16–26. doi:10.1109/MCE.2020.3002489. 1

[GLSU13] Geiger A., Lenz P., Stiller C., Urtasun R.: Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research 32*, 11 (2013), 1231–1237. 1

[GUGK17] Gupta K., Upadhyay S., Gandhi V., Krishna K. M.: Small obstacle detection using stereo vision for autonomous ground vehicle. *Proceedings of the 2017 3rd International Conference on Advances in Robotics* (2017). doi:10.1145/3132446.3134889. 1

[HFX*22] Huang J., Fu X., Xiao Z., Zhao F., Xiong Z.: Low-light stereo image enhancement. *IEEE Transactions on Multimedia* (2022). 2, 3

[KKLD23] Kerbl B., Kopanas G., Leimkühler T., Drettakis G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics 42*, 4 (July 2023). URL: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/. 2

[LHH16] Ley A., Hänsch R., Hellwich O.: Syb3r: A realistic synthetic benchmark for 3d reconstruction from images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (2016), Springer, pp. 236–251. 1

[LJBC24] Lee H., Jin W., Baek S.-H., Cho S.: Generalizable novel-view synthesis using a stereo camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 4939–4948. 3

[MBRS*21] Martin-Brualla R., Radwan N., Sajjadi M. S. M., Barron J. T., Dosovitskiy A., Duckworth D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR* (2021). 4

[MESK22] Müller T., Evans A., Schied C., Keller A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. 41*, 4 (July 2022), 102:1–102:15. URL: https://doi.org/10.1145/3528223.3530127, doi:10.1145/3528223.3530127. 5

[MSOC*19] Mildenhall B., Srinivasan P. P., Ortiz-Cayon R., Kalantari N. K., Ramamoorthi R., Ng R., Kar A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–14. 2

[MST*21] Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., Ng R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM 65*, 1 (2021), 99–106. 2

[NPX*21] Nie J., Pang Y., Xie J., Pan J., Han J.: Stereo refinement dehazing network. *IEEE Transactions on Circuits and Systems for Video Technology 32*, 6 (2021), 3334–3345. 2, 3

[PNC*22] Platinsky L., Naseer T., Chen H., Haines B. A., Zhu H., Grimmett H., Pero L. D.: Quantity over quality: Training an av motion planner with large scale commodity vision data. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), 5752–5759. doi:10.1109/IROS47612.2022.9982116. 1

[pol] Explore | polycam. https://poly.cam/explore. Accessed: 2024-06-14. 1

[RBM*22] Roessle B., Barron J. T., Mildenhall B., Srinivasan P. P., Niessner M.: Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12892–12901. 3

[RLS*22] Rematas K., Liu A., Srinivasan P. P., Barron J. T., Tagliasacchi A., Funkhouser T., Ferrari V.: Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 12932–12942. 3

[SF16] Schönberger J. L., Frahm J.-M.: Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). 2

[Sna08] Snavely N.: Bundler: Structure from motion (sfm) for unordered image collections. *http://phototour. cs. washington. edu/bundler/* (2008). 2

[SZW19] Sitzmann V., Zollhöfer M., Wetzstein G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems 32* (2019). 2

[W*11] Wu C., et al.: Visualsfm: A visual structure from motion system. 2

[WBSS04] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612. 5

[XZL*22] Xie Z., Zhang J., Li W., Zhang F., Zhang L.: S-nerf: Neural radiance fields for street views. In *The Eleventh International Conference on Learning Representations* (2022). 2, 3

[YLL*20] Yao Y., Luo Z., Li S., Zhang J., Ren Y., Zhou L., Fang T., Quan L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 1790–1799. 1

[YPN*22] Yu Z., Peng S., Niemeyer M., Sattler T., Geiger A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems 35* (2022), 25018–25032. 3

[ZXD*24] Zhao T., Xie Y., Ding M., Yang L., Tomizuka M., Wei Y.: A road surface reconstruction dataset for autonomous driving. *Scientific Data 11* (2024). doi:10.1038/s41597-024-03261-9. 3, 5

[ZZZ*19] Zhou S., Zhang J., Zuo W., Xie H., Pan J., Ren J. S.: Davanet: Stereo deblurring with view aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10996–11005. 2, 3