DOI: 10.2195/lj_proc_wei_en_202510_01

Towards Semantically-Aware Few-Shot 3D Reconstruction

Jiahua Wei [1], Juri Zach [2], Hendrik Wilhelm Rose [1], Philipp Braun [1]

[1] *Institute of Logistics Engineering*
Hamburg University of Technology

[2] *Institute of Computer Science*
Hamburg University of Applied Sciences

Acquiring rich object-level information, including shape, texture, and geometry, serves as a fundamental building block across multiple domains. In this context, few-shot reconstruction has become a prominent research field due to the ability to achieve 3D reconstruction from a limited set of input images. By leveraging prior knowledge encoded within a trained neural network, these methods can recover unseen features beyond the information obtained from the recorded sensor data. However, current approaches either model the entire environment without emphasizing specific regions of interest or restrict the process to the target object by completley neglecting the surrounding context in a prepossessing step. One potential approach is to apply object masking in the images and then directly map semantic information from 2D to 3D through deep learning. Nevertheless, this task reflecs highly non-linear properties, and integrating semantic cues remains a significant challenge. In this work-in-progress paper, we explore a pipeline for semantically aware few-shot 3D reconstruction on real-world data.

[Keywords: 3D reconstruction, semantic awareness, deep learning, occlusion handling, scene understanding]

1 Introduction

A central, cross-domain foundational problem lies in the reconstruction of high-fidelity 3D geometries from real-world scenes. Recent advances in radiance fields [1], [2] have significantly improved novel view synthesis (NVS), enabling photorealistic object renderings from multiview angles. By encoding structural, morphological, and material characteristics, 3D modeled objects incorporate a more information-dense understanding of the environment, surpassing the limitations of 2D imagery. Beyond geometric reconstruction, the optimization of radiance fields supports the utilization of sparse image sets by inferring missing content through learned priors. One of the key mechanisms underlying this process is the differentiable rendering formulation, which implements gradient-based optimization and enables the integration into deep learning pipelines. Advanced reconstruction techniques provide accurate localization [3], support real-time decision-making [4], and enhance comprehensive scene understanding [5], all of which are essential for automation in logistics. Prior knowledge of complete 3D object geometry allows autonomous systems to infer occluded structure and plan collision-free trajectories. This is particularly critical in autonomous driving, where access to complete object shapes enhances path planning, allowing the vehicle to predict object boundaries and safely maneuver around them. Similarly, in robotic manipulation, detailed object geometry informs pose and contact estimation, which facilitates reliable grasping and efficient product placement in confined storage spaces.

Although complex scenes can be reconstructed with high realism, the process is typically computationally expensive or demands large-scale image datasets. As a novel approach, feedforward neural networks (FNN) have been used to regress complete 3D structures during training and reconstruct objects from significantly fewer input images [6, 7, 8]. However, previous works are limited by global scene reconstruction [7, 9], the use of generative AI [10], and object-centric training pipelines where the background of the data has already been removed as a prepossessing step [6, 8].

In this work-in-progress paper, a conceptual framework is introduced that leverages deep learning to reconstruct the complete 3D objects from a limited set of target images. A key hypothesis is that a computational model should be able to encode semantic information and exploit learned priors to infer geometric properties even from a single viewpoint. In line with this formulation, the following research question is defined as (**RQ1**):

To what degree can complete and detailed 3D scenes be reconstructed from a limited set of 2D target images?

To mitigate the complexity during initial experiments, the chosen object should possess a well-defined and standardized structure. Since deep learning can capture the semantic features of objects from training data, models can

generalize beyond previously seen instances. To ensure that this generalization is meaningful and reliable, a validation pipeline must be carefully designed to reflect the diversity and complexity of real-world scenes. Therefore, a follow-up research question is defined as (**RQ2**):

Can models trained on a specific object generalize to reconstructing similar objects ?

Object reconstruction from a limited set of input views is inherently difficult, as the geometry can only be derived from the available structural cues. Information outside of the observed image regions remains inaccessible. This results in an underconstrained problem and introduces ambiguity in recovering the complete 3D structure under occlusion. Furthermore, to make the framework practically applicable, the object must be reconstructed in their natural environment. In a deep learning pipeline, the coupling between object and background increases the risk of overfitting to irrelevant background information, potentially leading the reconstruction process to hallucinate non-existent structures. Consequently, the focus should be placed on the target objects, motivating the final research question. (**RQ3**):

What strategies enable semantically-aware reconstruction of objects, while minimizing interference from the surrounding background?

The remaining paper is structured as follows: In Section 2, the current state of the art is reviewed, focusing on existing approaches to deep learning–based 3D reconstruction and optimization strategies in the context of few-shot targets. Section 3 details the training and validation pipeline, together with the corresponding network architecture. Insights from the literature review and the proposed network structure establish the foundation to address the first research question (**RQ1**). The validation experiments with different input images provide evidence and methodological guidance toward resolving (**RQ2**). Given that extracting robust semantic features remains the most challenging aspect of this work, (**RQ3**) will be addressed progressively across multiple series of future studies. The present contribution lies in identifying the principal difficulties, outlining feasible solution strategies, and setting the stage for further investigation. Finally, Section 4 summarizes the key findings and discusses directions for future research.

2 Related Work

Radiance Field 3D Reconstruction. The field of 3D reconstruction has progressed rapidly, driven by hardware developments increasing the accessibility of LiDAR- and camera-based depth approaches. In contrast to manual modeling, which uses computer-aided design (CAD), 3D reconstruction is induced from sensor-acquired point clouds, images and other spatial input data [11]. Although laser scanning sensors can acquire millions of high-precision points, they perform poorly for transparent or reflective surfaces and often demand costly equipment compared to RGB-cameras. In particular, recent advances in neural rendering have transformed the field of 3D reconstruction by allowing photometric-loss backpropagation through differentiable renderers [7]. One widely used approach is NeRF [2], which represents a scene as a radiance field that maps a 3D location and viewing direction to color and density. These inputs are generated from known camera intrinsic and extrinsic parameters and are processed by a multilayer perceptron (MLP) to implicitly encode the volumetric structure of the scene. However, NeRF-based implicit representations introduce significant computational overhead, as rendering involves expensive per-pixel volumetric sampling [6, 8, 12]. As a real-time capable alternative, 3D Gaussian Splatting (3DGS) was proposed in [1]. The algorithm takes a sparse point cloud as input and represents the scene with 3D Gaussians, each defined by a mean and covariance capturing anisotropic shape. These Gaussians are rendered on the image plane via a customized rasterization technique, which avoids the costly volumetric sampling of NeRF. While vanilla 3DGS requires multiple input views, deep learning priors can infer occluded geometry from limited perspectives by exploiting structural regularities.

Prior-Guided 3D Reconstruction. Inferring object geometries from few-view reconstruction is inherently ill-posed, as the sparsity of the captured images provides insufficient information for precise and unambiguous scene representation. Neural networks mitigate this limitation by enabling the direct generation of 3D points from training images while generalizing across a wide variety of objects [7].

Since NeRF represents scenes as an implicit function through an MLP, the method provides a flexible framework to incorporate learned priors with additional inputs. As one of the earliest approaches, pixelNeRF [13] introduced an FNN that conditions NeRF on image features extracted from a convolutional neural network (CNN) encoder. By training across diverse data, a transferable prior is learned, enabling generalization to previously unseen but related scenes without per-scene optimization. Further works advanced feed-forward NeRF architectures by leveraging pixel-aligned features [13, 14, 15, 16, 17], and later introduced transformer-based representations that employ attention mechanism to aggregate information over multiple viewpoints [18, 19].

Due to the differentiable formulation, 3DGS uses a photometric loss to align projected 3D Gaussian primitives with target images, strongly encouraging the integration of deep learning methods. PixelSplat [7] outlines the fundamental difficulty of optimizing Gaussian positions with gradient

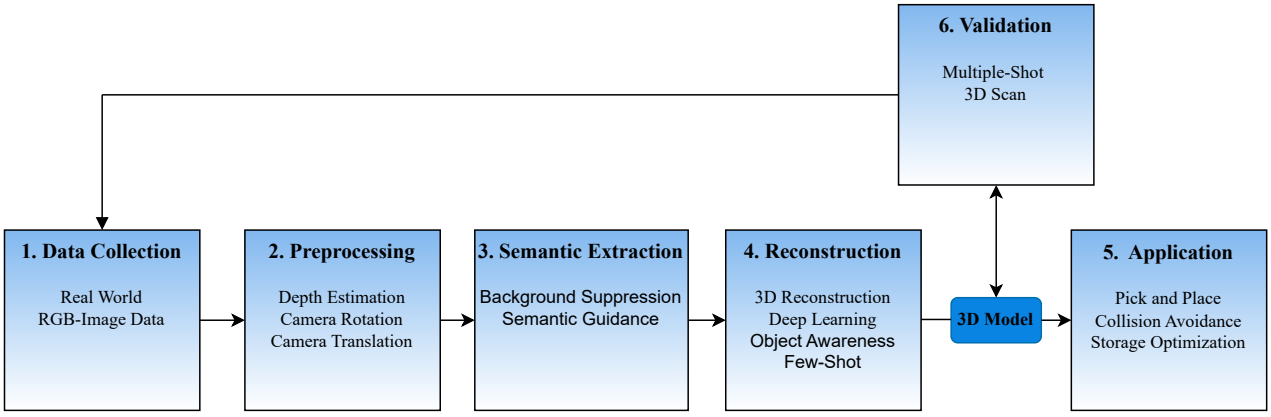

Figure 1: Workflow of the few-shot 3D reconstruction framework.

descent, which is a highly non-convex problem. Poor initialization or misleading gradients caused by occlusions can easily trap optimization in local minima. To address this, PixelSplat employs a feed-forward architecture that predicts a probability distribution over Gaussian means, enabling sampling to adaptively adjust the positions. Subsequent methods have increasingly adopted 3DGS in conjunction with feedforward reconstruction, extending the application from large-scale scene optimization [3, 8, 9] to single-object reconstruction [6, 10, 12]. The highlighted techniques consistently demonstrate real-time rendering performance exceeding 35 FPS. Although, recent few-shot Gaussian splatting methods can reconstruct single objects, they do not inherently address semantic object focus, since their training datasets typically consist of already centered or background-free objects.

Semantically-Aware 3D Reconstruction. Integrating deep learning into 3DGS not only enables the prediction of Gaussian positions, but also allows for the incorporation of semantic object information. Such semantic cues can be explicitly introduced through the use of segmentation models to map 2D masks into 3D scenes [20, 21, 22]. Although any segmentation model can be integrated, recent approaches highlight the advantages of the Segment Anything Model (SAM) [23], which offers strong zero-shot generalization and reliable mask predictions across a wide range of different object categories. Alternatively, few-shot FNN architectures demonstrate strong generalization capabilities under extremely sparse input [6, 7], indicating an implicit learning ability that allows the network to semantically infer and complete object representations.

From the literature review, it can be inferred that few-shot reconstruction methods already exist which are capable of detecting objects in real time, providing an initial answer to (**RQ1**). However, these approaches are typically limited to synthetic datasets or on background-removed objects. Moreover, current methods can lift 2D segmentations into 3D, which is particularly applicable in multi-view scenarios. Nevertheless, the integration of FNNs substantially increases the overall complexity to include semantic awareness in the mapping process. Consequently, (**RQ2**) represents a novel and meaningful contribution when explored in conjunction with few-shot reconstruction.

3 Method

The conceptual workflow of the proposed framework is illustrated in Figure 1. The process starts with the collection of RGB image data from real-world objects. During training, multiple views of each object are captured to ensure sufficient geometric and appearance coverage. In contrast, at inference time, the model is expected to generalize from limited observations. Despite recent advances in acceleration, NeRF-based methods remain computationally expensive due to the need for dense point sampling across the entire scene space [12]. To achieve a better trade-off between efficiency and reconstruction fidelity 3DGS is used. However, this approach requires additional supervisory signals in the form of a sparse or dense depth map and positional information. To ensure object-centric reconstruction, a semantic extraction stage is employed, which leverages guidance to distinguish the target object from the background and suppress irrelevant scene content. The resulting 3D models obtained from the few-shot reconstruction are evaluated by comparison against either high-fidelity 3D scans or objects generated from multi-view input data. In cases where the reconstructed models exhibit significant artifacts or deviations from the reference, additional data must be collected to improve the quality.

3.1 Camera Odometry

Training Gaussian splatting on real-world data requires multiple images of the object captured from different viewpoints, together with accurate information about the relative

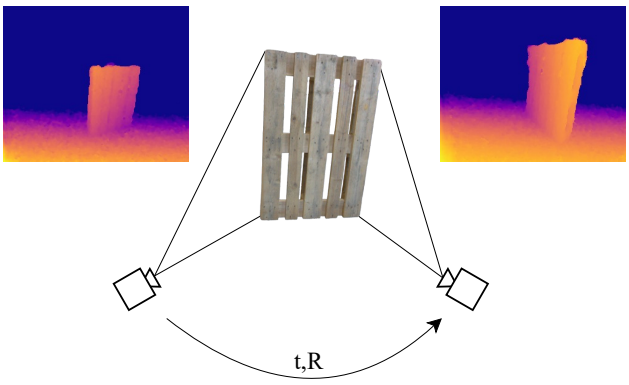

Figure 2: Estimation of depth and camera positions from multi-view RGB inputs, providing initialization for the 3D reconstruction framework

position and orientation of the camera. Knowledge of the camera pose enables self-supervised training, as Gaussian splats can be projected into the corresponding 2D image plane. This projection allows the captured images to serve as ground-truth supervision for optimizing the 3D representation. High-quality reconstructions further depend on precise depth estimates and accurate camera poses. While external sensors can provide such measurements, they often suffer from sparsity, limited accuracy, high cost, and synchronization issues. In contrast, computer vision methods can directly infer depth and camera motion from the images, removing the need for additional hardware or sensor fusion.

Visual odometry estimates the translational and rotational motion of the camera between consecutive frames, enabling reconstruction of the complete camera trajectory. Although generally precise, it is a dead-reckoning approach, and cumulative drift can introduce substantial errors over long sequences. This limitation is negligible for short trajectories required by the proposed framework. Benchmark datasets such as KITTI [24, 25] further show that state-of-the-art depth prediction is dominated by deep neural network–based methods. Despite these advances, existing methods are not universally suitable for the proposed framework. Supervised methods cannot be assumed to generalize reliably to newly collected data of unseen objects and constructing a new labeled dataset lies beyond the scope of this work. Monocular methods can produce accurate relative estimates but inherently suffer from scale ambiguity, necessitating external reference measurements. A well-established strategy to overcome this limitation is the use of stereo cameras.

Deep stereo visual odometry addresses these challenges by leveraging photometric consistency to train neural networks for depth and motion estimation in a self-supervised manner. Methods such as [26] and [27] demonstrate state-of-the-art performance in both depth prediction and camera motion estimation. A simple example of the data prepossessing pipline is shown in Figure 2.

3.2 Few-Shot Gaussian Splatting

In 3DGS, the radiance field is expressed through a set of N Gaussian primitives, whose starting configuration is derived by the depth estimator. Each primitive is parameterized as $g_n = \{\mu_n, \Sigma_n, \alpha_n, S_n\}$, where the mean μ_n is initialized at the corresponding coordinates of the point cloud, Σ_n represents the covariance matrix, α_n the opacity and S_n the spherical harmonic coefficients for the view-dependent color [1, 7]. The 3D Gaussians are rendered into a 2D image $I(x, y)$ by applying the camera extrinsics (W, P) followed by the projection matrix K. In this process, 3DGS emplyos a non-linear reparameterization of the camera space coordinates to obtain the projection of the covariance Σ^{proj} by utilizing the Jacobian J locally linearized at the mean:

$$\Sigma^{proj} = JW\Sigma W^T J^T. \tag{1}$$

The rendering $\mathcal{R}$ is performed through a fast rasterization algorithm, enabling high-speed performance [1].

In a few-shot setting, the 3D reconstruction pipeline has only access to a small set of input images. As a result, the recovered point cloud is sparse and incomplete, providing limited cues for matching and surface reconstruction. To address this problem, an FNN is proposed [7, 6, 12] to be trained on multi-view datasets and capture object priors. At inference, the network acts as an inverse mapping $\mathcal{G} = \{g_n\}_{n=1}^{M} = F(I(x, y))$, which generates missing primitives M and creates the geometry of the whole scene from the target images. A lightweight and efficient few-shot approach providing initial insights into the first research question (**RQ1**) was recently introduced in [6], serving as a promising foundation for further investigations. In this approach, a SongUNet [28] is employed to directly predict Gaussian Primitives $\mathcal{G}$ at every image pixel. The mean μ is calculated by backprojecting the pixel coordinates $u = [x_i, y_i, 1]^T$ with the associated depth d and augmenting the result with a learned offset δ to increase the flexibility of the positioning:

$$\mu = du + \delta. \tag{2}$$

For each pixel, the network predicts $12 + kc$ parameters by means of a 1×1 convolutional output layer, which are mapped to the Gaussian attributes opacity $\alpha \in \mathbb{R}_+$, offset $\delta \in \mathbb{R}^3$, depth $d \in \mathbb{R}_+$, scale $s \in \mathbb{R}^3$, rotation quaternion $q \in \mathbb{R}^4$ and $c \in \mathbb{R}^{k_c}$. Since, the depth and the external rotation are already given by the camera odometry, the network output is reduced to $7 + kc$ parameters. An illustration of the networks output is shown in Figure 3. During training, a source image $I(x, y)$ is passed through the network to produce the Gaussian primitives. These are rendered via the differentiable renderer $\mathcal{R}(F(I(x, y)), \pi)$ into the image plane of a novel viewpoint π and the result is compared against

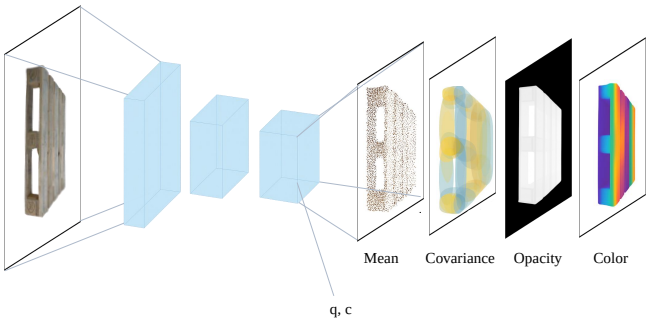

Figure 3: Neural network output of the few-shot gaussian splatting.

the corresponding target image $T(x, y)$. This is done over the complete dataset $\mathcal{D}$ for training to minimize the average reconstruction loss $\mathcal{L}$:

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum \|T(x, y) - \mathcal{R}(F(I(x, y)), \pi)\|^2 \qquad (3)$$

The proposed framework was originally designed for single-image input. For more than one source image, each image contributes primitives defined by the corresponding viewpoint π. Therefore, overlapping Gaussians may be generated in regions that are observed by more than one view. The optimization process addresses this redundancy by reducing the predicted opacity value α in regions with multiple overlapping Gaussians, minimizing the influence during training [6].

3.3 Semantic Extraction

The third research question (**RQ3**) concerns the problem of semantic object extraction, for which an initial solution can be outlined. Addressing this problem requires progressive scaling and represents a central challenge within the broader long-term research agenda.

When reconstructing objects from real-world data, images are captured against diverse and cluttered backgrounds, introducing unwanted contextual biases into the learning process. To mitigate this, semantic awareness of the objects is necessary, as models are prone to hallucinating undesired environment details inherited from the training data. One straightforward approach for semantic feature extraction is given by pre-filtering the background using a segmentation model. In this setting, SAM can serve as a baseline tool, allowing manual specification of the target object and ensuring that feature extraction is guided by the generated masks $\mathcal{M} = (x_1, y_1, ...x_i, y_i)$. During FNN training, background regions are treated as empty pixels that correspond to non-informative Gaussians, ensuring that they do not contribute to the reconstruction objective.

Although segmentation preprocessing methods can provide a short-term solution, they are dependent on manually annotated labels and often fail to generalize across diverse environments. Directly projecting 2D segmentation masks into 3D Gaussians has recently attracted attention [20, 21, 22]. However, the FNN in a few-shot setting must learn a complex non-linear mapping from image pixels to 3D primitives. As the geometry of occluded regions cannot be directly observed, the network is forced to infer their structure by hallucinating unseen content or reallocating representational capacity from noisy Gaussians. This highlights a fundamental limitation, since lifting 2D masks into 3D space is insufficient to classify objects in regions when solely relying on visible image features. Furthermore, the covariances of multiple Gaussians may overlap during the rendering process, making the overlapping regions difficult to unambiguously associate with a single segmentation identity.

One potential solution is to integrate the semantic extraction process directly into the FNN. During training, the network captures the structural dependencies among similar objects, allowing the usage of semantic information as prior knowledge that becomes implicitly encoded in the network weights through inductive bias. This effect is reinforced by the higher mutual information of real-world objects present in 3D models compared to 2D images, enabling the model to exploit geometric and contextual consistency for richer and more discriminative feature extraction.

3.4 Adaptation in the Logistics Domain

The presented few-shot reconstruction framework is a domain-agnostic module and can be integrated across different fields as a core building block. In this section, a specific example use case from the logistics domain is introduced.

In intralogistics, autonomous warehouse management is emerging as an increasingly advanced topic, driven by the integration of adaptive control and intelligent scheduling methods. Among the core tasks are reliable grasping operations and the efficient placement of objects within storage systems. The availability of known 3D positions can reduce the complexity of robotic pose estimation by providing the controller with spatial information about the object. Moreover, few-shot reconstruction based on Gaussian Splatting requires only camera sensors, making the approach hardware efficient and real-time capable.

As an initial experiment, pallets are selected as the primary focus, providing a structured test case to validate the fundamental capabilities of the proposed pipeline. Furthermore, pallets are standardized in their geometry, which enables straightforward validation of the reconstructed models. The image acquisition is carried out using two Basler ace 2 Pro industrial cameras in a stereo setup, each featuring a resolution of 24.4 MP. Before data collection, the camera parameters must be calibrated to obtain both intrinsic and extrinsic properties. For the experiments, the stereo setup is calibrated in a warehouse environment using a PuzzleBoard

patterns that include positional encoding. For the training of the FNN multiple viewpoints of the pallets are required. To construct a sufficient dataset, five representative pallets were selected, and images were freely captured over a 360° trajectory with tightly spaced viewpoints around each object. The image resolution is set to 1152×1332, while retaining the flexibility for post-collection adjustment through resizing or data augmentation. In addition, the dataset $\mathcal{D}$ can be redefined after acquisition to a sparse subset of $\{2, 4, 6, 8\}$ images, allowing controlled investigations of the training. The remaining images are employed as target views to assess the accuracy of the 3D reconstruction via novel view synthesis, with performance optimized against benchmark metrics such as PSNR, LPIPS, and DSSIM [6, 7]. However, while these metrics are widely used, they provide shallow approximations that miss important perceptual details. Deep feature–based measurements resulting from the latent space could offer a stronger alternative for future research, providing more reliable visual fidelity assessments [29]. To provide another baseline for comparison, the performance can be further evaluated against Gaussian splats generated from multiple input images.

Lastly, to address the research question (**RQ2**), additional data from different objects must be collected to validate the generalizability of the network structure. An ablation study is employed to evaluate the reconstruction algorithm trained on standard Euro pallets, assessing the performance of the model to partially missing pallet structures and on structurally related variants such as plastic and collar pallets. At the same time, the training process should be extended to include varying backgrounds, lighting conditions and object positions within the images to reduce the risk of overfitting.

4 Conclusion and Outlook

Few-shot 3D reconstruction with semantic awareness represents an important advancement, enabling both object recognition and the instantaneous estimation of the objects dimensions. Together, these capabilities yield a more comprehensive representation of the physical environment, which in turn supports robust and reliable scene interpretation. This is particularly valuable in domains such as autonomous driving and robotic manipulation, where additional knowledge about the geometric shape is critical to effectively interact with the environment. The aim of this work-in-progress paper is to identify the initial methodological components and key challenges involved in creating a prio-based reconstruciton pipeline for real world objects.

In this context, a strategy for few-shot 3D reconstruction using only image data was demonstrated to address (**RQ1**). The literature review revealed that the current state-of-the-art frameworks lack semantic awareness. This limitation hinders background-invariant reconstruction. As a result, the proposed pipeline in this work is composed of four principal building blocks: a camera odometry algorithm, a 3D modeling method, a prior-informed FNN, and a semantic extraction module. Furthermore, (**RQ2**) emphasizes the need for generalization beyond a single object instance, ensuring that the approach remains scalable. Without this capability, the approach would only be applicable to objects encountered during training. For this reason, a structured data acquisition protocol was established to guide the collection of training data. The protocol is designed to ensure sufficient variation across object classes, background conditions, and camera distances, allowing a systematic ablation study on the influence of these factors. Finally, the long-term research question (**RQ3**) concerns the explicit integration of the semantic extraction approach. A central challenge is that current neural architectures are not inherently suited to transforming 2D segmentation masks into 3D representations via prio-based 3DGS. Therefore, this work presents initial suggestions and provides first steps in defining the problem.

To enable the implementation in the intralogistic domain, an image dataset of pallets and semantically similar objects was collected to evaluate the performance of the algorithmic components. An important direction for future research will be the integration of the semantic extraction process directly into the FNN. Although a simple U-Net architecture from SplatterImage [6] may serve as a baseline for initial experiments, the gradient flow and loss function must be adapted to place emphasis on the semantic object features. Moreover, since validation and training losses based on low-level or color characteristics are often insufficient, future investigations should incorporate comparisons using deep feature representations in the latent space [29].

REFERENCES

[1] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, Jul. 2023.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, p. 99–106, Dec. 2021.

[3] H. Zhai, X. Zhang, B. Zhao, H. Li, Y. He, Z. Cui, H. Bao, and G. Zhang, "Splatloc: 3d gaussian splatting-based visual localization for augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 5, pp. 3591–3601, 2025.

[4] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting fornbsp;multi-task robotic manipulation," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXXV*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 349–366.

[5] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 162–179.

[6] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, "Splatter image: Ultra-fast single-view 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 10 208–10 217.

[7] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 19 457–19 467.

[8] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 370–386.

[9] K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu, "Gs-lrm: Large reconstruction model for 3d gaussian splatting," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXII*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 1–19.

[10] C. Wewer, K. Raj, E. Ilg, B. Schiele, and J. E. Lenssen, "Latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 456–473.

[11] L. Zhou, G. Wu, Y. Zuo, X. Chen, and H. Hu, "A comprehensive review of vision-based 3d reconstruction methods," *Sensors*, vol. 24, no. 7, 2024.

[12] S. Zheng, B. Zhou, R. Shao, B. Liu, S. Zhang, L. Nie, and Y. Liu, "Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 19 680–19 690.

[13] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4576–4585.

[14] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14 124–14 133.

[15] H. Lin, S. Peng, Z. Xu, Y. Yan, Q. Shuai, H. Bao, and X. Zhou, "Efficient neural radiance fields for interactive free-viewpoint video," in *SIGGRAPH Asia 2022 Conference Papers*, ser. SA '22. New York, NY, USA: Association for Computing Machinery, 2022.

[16] A. Jain, M. Tancik, and P. Abbeel, "Putting nerf on a diet: Semantically consistent few-shot view synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5885–5894.

[17] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "Ibrnet: Learning multi-view image-based rendering," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4688–4697.

[18] Y. Du, C. Smith, A. Tewari, and V. Sitzmann, "Learning to render novel views from wide-baseline stereo pairs," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4970–4980.

[19] M. M. Johari, Y. Lepoittevin, and F. Fleuret, "Geonerf: Generalizing nerf with geometry priors," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 344–18 347.

[20] Q. Shen, X. Yang, and X. Wang, "Flashsplat: 2d tonbsp;3d gaussian splatting segmentation solved optimally," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXII*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 456–472.

[21] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 162–179.

[22] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Segment any 3d gaussians," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025.

[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[24] M. Menze, C. Heipke, and A. Geiger, "Object scene flow," *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018.

[25] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *International Conference on 3D Vision (3DV)*, 2017.

[26] N. Yang, L. von Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1278–1289.

[27] J. Zach and P. Stelldinger, "Self-supervised deep visual stereo odometry with 3d-geometric constraints," in *Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 336–342.

[28] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jiahua Wei (M.Sc.), is a Research Associate at the Institute of Logistics Engineering. Prior, he obtained his Mechatronics Master's degree at Hamburg University of Technology. Address: Theodor-Yorck-Straße 8, 21079 Hamburg, Germany, Phone: +49 40 42878-4893, E-Mail: jiahua.wei@tuhh.de

Juri Zach, (M.Sc.), is a Research Associate at the Institute of Computer Science. Prior, he obtained his Computer Science Master's degree at Hamburg University of Applied Sciences. Address: Berliner Tor 5, 20099 Hamburg, Germany, E-Mail: juri.zach@haw-hamburg.de

Hendrik Rose, (M.Sc.), studied Mechanical and Industrial Engineering at Hamburg University of Technology. Since 2021, he has been working as a Research Associate and Chief Engineer (since 2023) at the Institute of Logistics Engineering. Address: Theodor-Yorck-Straße 8, 21079 Hamburg, Germany, Phone: +49 40 42878-3668, E-Mail: hendrik.wilhelm.rose@tuhh.de

Philipp Braun, (M.Sc.), studied Logistics & Mobility as well as International Industrial Engineering and Management at the Hamburg University of Technology. Since 2019, he has been working at the Institute of Logistics Engineering, initially as a Research Associate and, since 2023, as Chief Engineerg. Address: Theodor-Yorck-Straße 8, 21079 Hamburg, Germany, Phone: +494042878-3556,E-Mail: philipp.braun@tuhh.de.