

VAD-GS: Visibility-Aware Densefication for 3D Gaussian Splatting in Dynamic Urban Scenes

Supplementary Material

1. Multi-Camera Cross-Frame Views

As shown in Fig. 1, the outward-facing multi-camera views have limited overlaps. Prior methods such as [4] typically treat all views indiscriminately during Gaussian training, regardless of their spatial or temporal differences. Nonetheless, structural complexity varies significantly across regions, necessitating a selective reconstruction strategy that prioritizes critical objects over trivial or redundant structures. Object-centric reconstruction strategies generally assume sufficient overlap among views within a bounded range and minimal interference from unrelated perspectives. However, this assumption breaks down in dynamic, unbounded urban scenes. The failure case illustrated in Fig. 1 suggests that observations from the same camera fail to continuously capture a moving target vehicle.

1.1. Visibility Reasoning

Visibility determination, also known as hidden surface removal (HSR) or occlusion culling (OC), which identifies visible surfaces from a given viewpoint, has long been a central topic in computer graphics [2]. Among numerous HSR algorithms, z-buffering is usually the choice due to its simplicity and efficient hardware implementation. In contrast, Gaussian splatting renders pixels by alpha-blending all primitives along each viewing ray rather than explicitly enforcing occlusion. As points cannot occlude one another, no primitive is truly hidden, as illustrated in Fig. 2 in the supplement. With sufficient viewing directions, primitives on visible surfaces may eventually become opaque, which implicitly recovers occlusion. However, when initialization is incomplete and views are limited, ambiguity arises: observed appearance is simply mapped onto whichever primitives are projected to the image plane, regardless of whether the corresponding surface exists. Since incomplete geometry would inevitably mislead optimization, densification strategies have to take visibility reasoning into consideration to assess surface completeness before applying further updates. Most existing Gaussian splatting methods take COLMAP [5] point clouds directly as Gaussian centers and overlook their associated visibility information. Yet COLMAP inherently encodes rich visibility cues. Specifically, the “TRACK” table records the source views in which each 3D point is observed and successfully triangulated. This allows us to exploit the following two real-world visibility observations in a more effective way: (1) Each 3D point lies on the first surface intersected by the pixel rays from all source views, implying that the line of sight be-

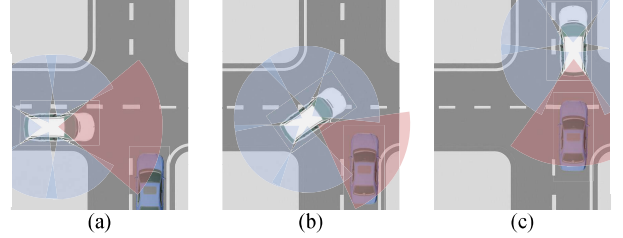


Figure 1. **An illustration of multi-camera, cross-frame views.** For both static and dynamic objects, informative observation views are typically captured by different cameras at different timestamps.

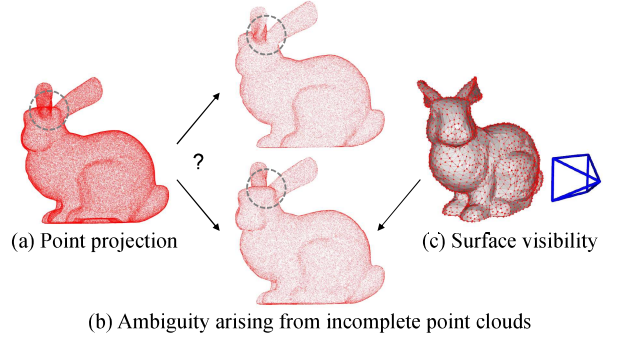


Figure 2. **Is the rabbit facing forward or backward?** (a) Points do not occlude one another, introducing appearance and geometry ambiguity. (b) Incomplete initialization may force the model to map front-side appearance onto back-side primitives, which are then incorrectly optimized and become distorted. (c) Identifying and completing missing structures requires densification with explicit visibility reasoning.

tween the object and each viewpoint is occlusion-free. (2) The local structure surrounding each 3D point is visible in its associated source views, thereby providing reliable supporting-view candidates for object reconstruction.

1.2. View Selection

The diversity score s introduced in the main paper quantifies the geometric dissimilarity between a pair of views. However, selecting an informative subset of supporting views for reconstruction requires more than simply maximizing diversity between view pairs, ensuring that the subset is collectively informative and non-redundant. Moreover, as the same reference object may appear repeatedly during training, a deterministic selection based solely on diversity may lead to overfitting or limited generalization. To address this

issue, we propose to sample views via:

$$\max_{\mathcal{V}_s \subset \mathcal{V}_c} \sum_{v_i \in \mathcal{V}_s} s_{iR} \xi_{iR} + \lambda \sum_{\{v_i, v_j\} \subset \mathcal{V}_s} s_{ij} \xi_{ij}, \quad (1)$$

$$|\mathcal{V}_s| = k, \quad \xi \sim \mathcal{N}(1, \epsilon),$$

where \mathcal{V}_c denotes the full set of all candidate views, \mathcal{V}_s represents the selected subset containing k supporting views, s_{iR} denotes the diversity score between each pair of candidate view and the reference view, s_{ij} represent the diversity score among views within the subset, and ϵ represents a noise term introduced to encourage sampling diversity. This randomized selection strategy ensures relevance to the reference view while avoiding deterministic bias, resulting in a diverse yet non-redundant subset of supporting views.

2. Additional Experiments

2.1. Experimental Details

While many 3DGS methods adopt similar train/test splitting strategies, the specific details on these splits remain ambiguous for urban driving scenes. For example, statements such as “randomly select every n -th image of different cameras” can be interpreted in multiple ways: either as discarding specific frames with all associated camera views, or as selectively omitting individual views while retaining the full sequence of frames. Moreover, such random sampling schemes are misaligned with the practical goal of novel view synthesis, which aims to render intermediate views between consecutive video frames captured by multi-camera systems mounted on a moving vehicle.

While both strategies remove the same number of views, randomly selecting individual test views results in more uniform frustum coverage and visually cleaner outputs. However, this approach exploits temporal redundancy and overlooks the realistic constraint that multi-camera views are typically available or missing as a complete observation. In contrast, removing all views at specific timestamps significantly reduces scene coverage and degrades visual quality, particularly when the vehicle is moving rapidly. Despite being more challenging, this setting better reflects real-world deployment constraints and more effectively evaluates the model’s generalizability.

Specifically, we select every fourth frame along with all associated camera views to construct the test set. As a result, spatial observations are entirely unavailable for approximately 25% of the ego vehicle poses. This setting poses significant challenges for models that rely on multi-view consistency or temporal cues, and serves as a rigorous benchmark for evaluating reconstruction robustness under sparse observational conditions.

2.2. Additional Qualitative Comparisons

In this supplement, we provide additional comparative results against recent methods on large-scale driving scenes. Due to the page limitation, qualitative results on the Waymo Open dataset [6] are provided in Fig. 3. For fair comparison, we adopt the validation configuration of StreetGaussians [8] and use only a single forward-facing camera. This setup simplifies view-dependent appearance and geometry consistency constraints, as the forward-facing view undergoes relatively minor temporal changes. However, it inherently limits the acquisition of novel information and significantly reduces overall scene coverage. These minimal inter-frame variations result in highly similar and redundant observations, which can provide limited geometric diversity for triangulation or multi-view spatial-consistency reasoning, thus failing to fully unleash the potential of visibility-aware densification for complete geometry reconstruction. Consequently, high-fidelity rendering quality may not indicate accurate scene geometry recovery, but rather reflect overfitting to specific image observations.

To further demonstrate the high quality of our scene reconstruction, we present an additional example in Fig. 4. This comparison is performed by adopting a multi-camera configuration that utilizes cameras 0, 1, and 2 from the Waymo Open dataset. Although all methods achieve comparable rendering quality, the underlying geometry differs significantly. The traffic sign, highlighted by yellow circles, lies outside the LiDAR scanning range and is only partially visible from a limited number of viewpoints. In OmniRe [1], the sign is reconstructed as a set of scattered and unstructured Gaussians, indicating overfitting to appearance cues in the absence of reliable geometric constraints. As for StreetGaussians, the sign appears fragmented and discontinuous, with Gaussians erroneously updated to positions between the sign and the background trees. These artifacts stem from missing Gaussians caused by incomplete initialization, which in turn lead to erroneous gradient propagation toward trees that should be occluded. The misdirected gradients distort the initial Gaussians representing the leaves, altering their color, position, and shape, and unnaturally pull them toward the sign, ultimately resulting in fragmented and misaligned geometry.

Benefiting from visibility reasoning, view selection, and MVS-based reconstruction, VAD-GS densifies Gaussians beyond conventional photometric-based splitting and cloning strategies, greatly alleviating issues related to incomplete or distorted geometry. Notably, VAD-GS accurately recovers the planar structure of the traffic sign, with only minor artifacts at the top border due to limited observations. Additionally, the road surface, highlighted by the white box, demonstrates a more geometrically consistent reconstruction compared to other approaches.

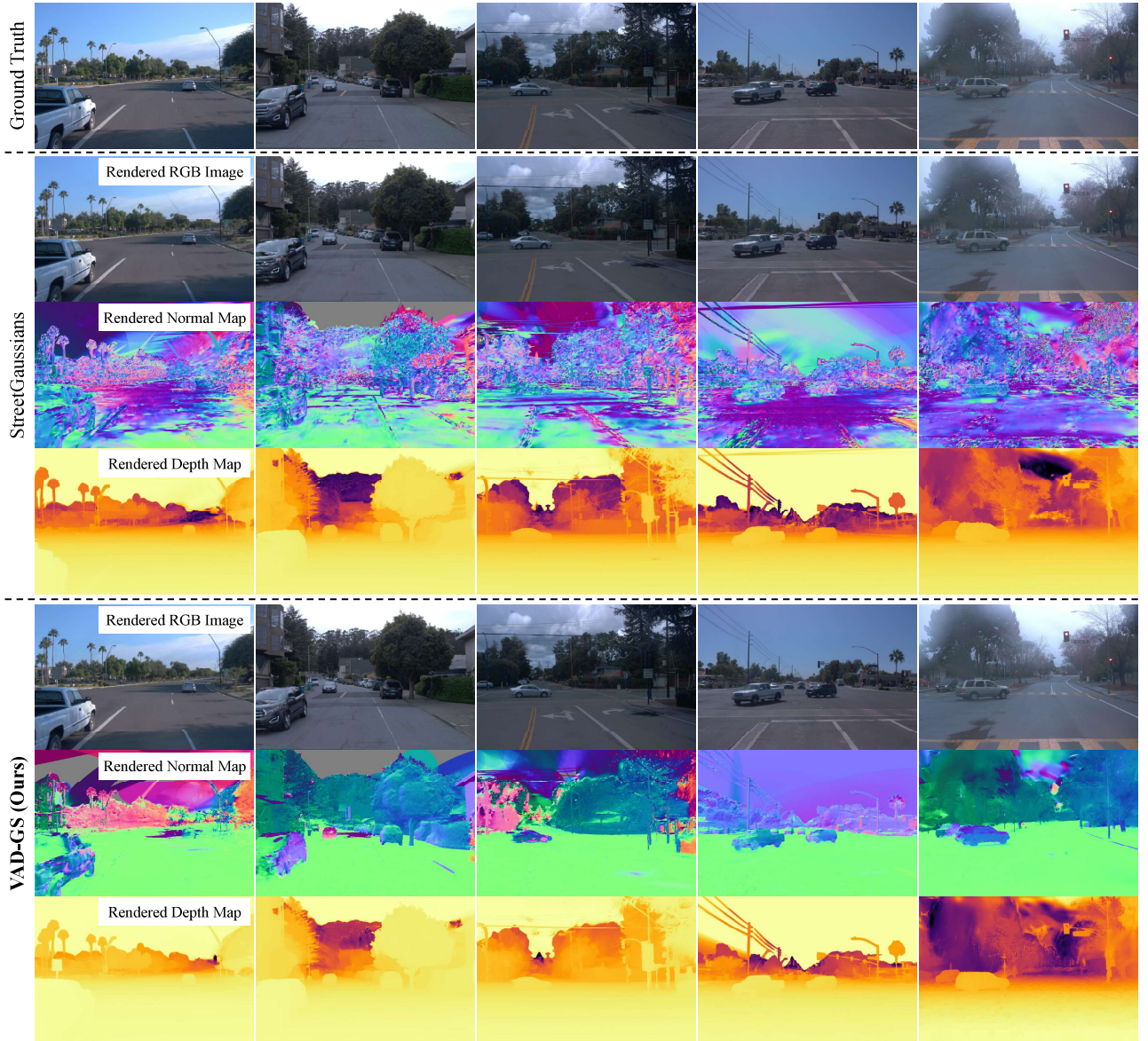


Figure 3. **Additional qualitative results on the Waymo Open dataset.** Due to the single-camera configuration, test views captured by the forward-facing camera exhibit substantial overlap with the training views. While all methods achieve high-fidelity rendering results under this setting, such performance may not reliably indicate the quality of the underlying geometry.

2.3. Additional Ablation Studies

In this supplementary material, we also present additional qualitative ablation study results, including rendered RGB images, depth maps, and normal maps, to further demonstrate the effectiveness of each module in VAD-GS. As shown in Fig. 6, the sparse point clouds provide limited surface coverage. Each LiDAR scan line in the ground-truth point typically contributes only two or three points to thin structures such as tree trunks or utility poles. Additionally, due to the limited scanning angle and sparse sampling in-

tervals, the resulting point cloud distribution exhibits substantial gaps and covers only a narrow field of view. These limitations pose significant challenges for capturing complete geometry, particularly for large and distant surfaces such as buildings and walls.

Furthermore, we select several challenging test views to more clearly demonstrate the contribution of each component. A common issue during densification is the emergence of floaters, where Gaussians become misaligned with the actual scene geometry. While most floaters are of-

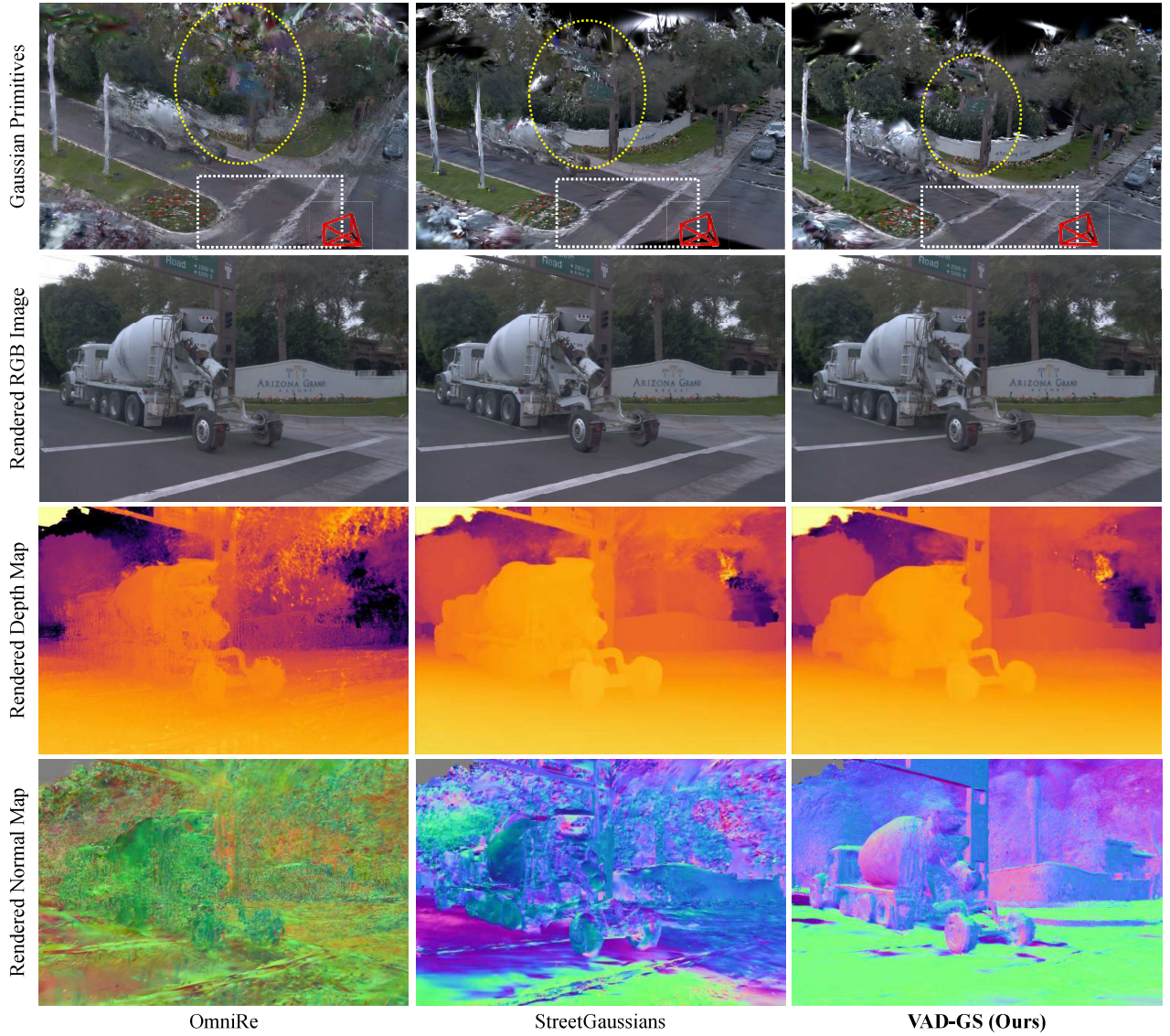


Figure 4. **Qualitative comparison between VAD-GS and other SoTA methods on the Waymo Open dataset when a multi-camera configuration is used.**

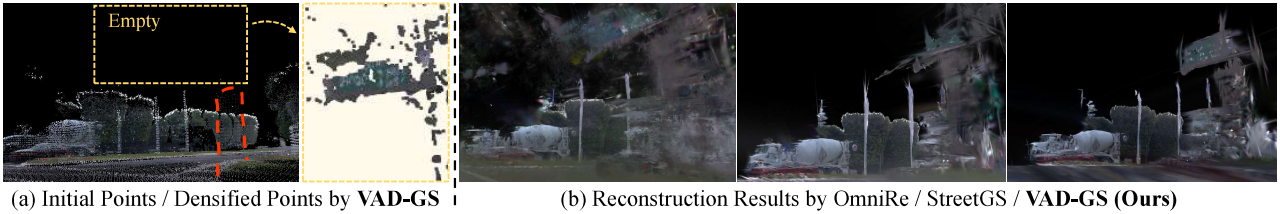


Figure 5. **Qualitative comparison of reconstructed scene geometry.** Revisiting the Waymo traffic sign example given in the paper, the LiDAR and SfM points are not only noisy but **missing**, with only a few points available within the **red** dashed area. In this extremely challenging case, the exact contribution of our densification lies in recovering points in the **yellow** box.

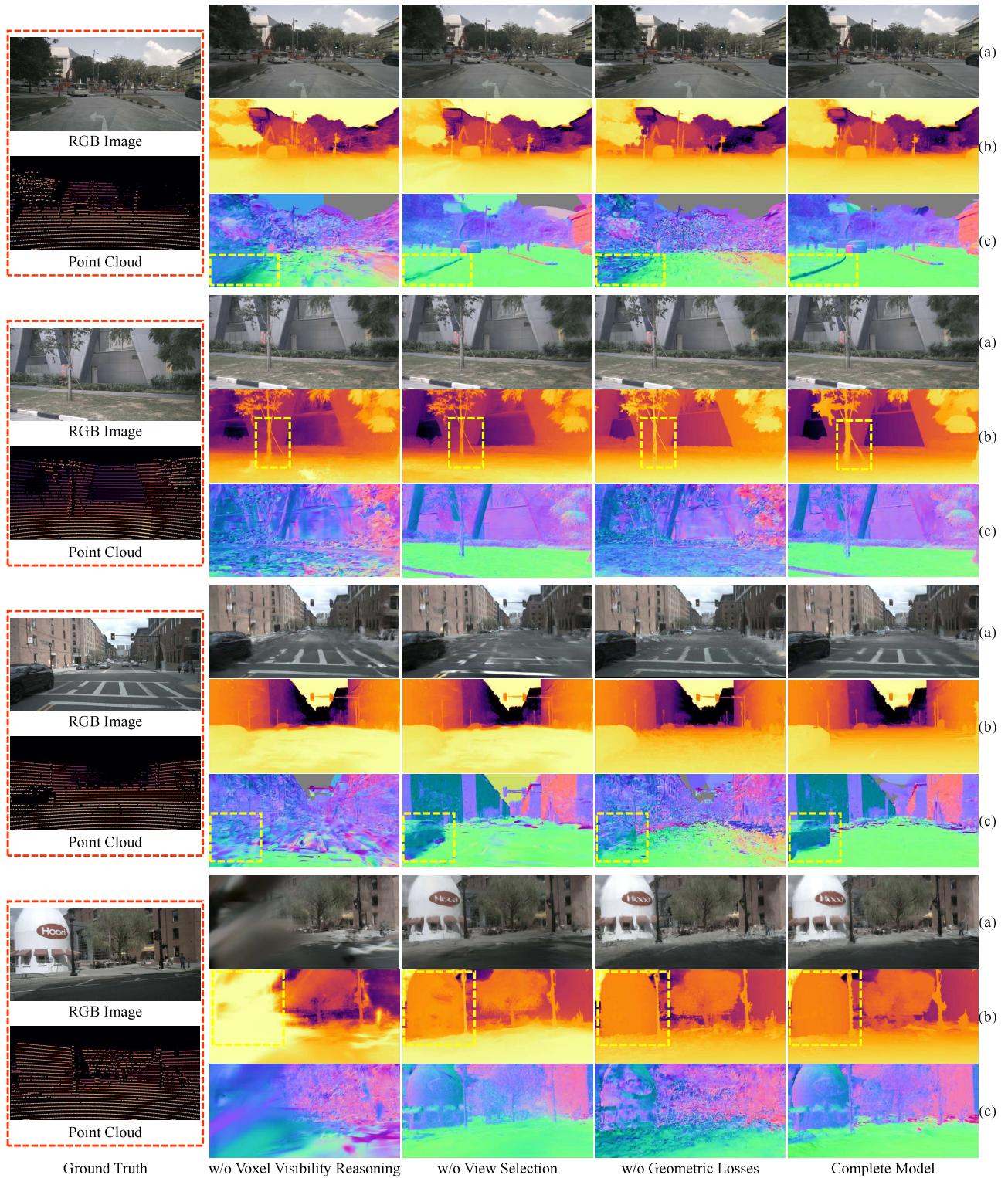


Figure 6. **Additional qualitative ablation study results on the nuScenes dataset.** The rendered RGB images, depth maps, and normal maps are visualized in (a), (b), and (c), respectively.

ten naturally pruned or corrected when they appear in regions well-covered by training views, they tend to persist in sparsely observed areas. In selected test views where these floaters are prominent, our geometric loss effectively penalizes them, encouraging alignment with the correct underlying surfaces. This process significantly improves the final surface quality and substantially reduces visual artifacts. Moreover, objects that are only transiently visible, such as moving vehicles or structures primarily observed from side views, often suffer from sparse observations. Our view selection and MVS-based reconstruction modules improve the instance-level fidelity in these challenging regions, including dynamic vehicles, small trees, and complex landmarks such as the bottle-shaped building.

3. Failure Cases and Limitations

Despite achieving high-fidelity performance, VAD-GS still exhibits several known limitations. The primary challenge lies in its inability to effectively handle deformable objects, such as pedestrians. Given that our objective is to recover geometry in complex urban scenes, the presence of walking pedestrians is inevitable. Nonetheless, these non-rigid objects violate the rigidity assumption required by MVS-based reconstruction. Future work will explore the integration of state-of-the-art Gaussian-based deformable object modeling approaches, such as 4DGS [7] and SC-GS [3], to address this issue.

Second, our method assumes locally consistent visibility among neighboring points. While this assumption enables effective occlusion modeling and supports continuous surface reconstruction, it may fail in extreme cases involving complex structures such as wire fences or glass surfaces. These structures often reflect LiDAR beams, producing dense point clouds that resemble those from regular surfaces. Nevertheless, the simultaneously captured images may reveal background objects without occlusion, leading to discrepancies between geometric and visual observations. Accurately and efficiently modeling occlusion relationships in such challenging and visually ambiguous regions remains an important direction for future research.

References

- [1] Ziyu Chen et al. OmniRe: Omni urban scene reconstruction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 85508–85527, 2025. [2](#)
- [2] Daniel Cohen-Or et al. A survey of visibility for walkthrough applications. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):412–431, 2003. [1](#)
- [3] Yi-Hua Huang et al. SC-GS: Sparse-controlled Gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4220–4230, 2024. [6](#)
- [4] Bernhard Kerbl et al. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4): 1–14, 2023. [1](#)
- [5] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. [1](#)
- [6] Pei Sun et al. Scalability in perception for autonomous driving: Waymo Open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020. [2](#)
- [7] Guanjun Wu et al. 4D Gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. [6](#)
- [8] Yunzhi Yan et al. Street Gaussians: Modeling dynamic urban scenes with Gaussian splatting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 156–173. Springer, 2024. [2](#)