A Hybrid Framework for Real-Time and Drift-Reduced Camera Pose Estimation

Asrith Pandreka
Arizona State University
apandrek@asu.edu

Deepikaa Anjan Kumar
Arizona State University
danjanku@asu.edu

Jeevan Hebbal Manjunath
Arizona State University
jhebbalm@asu.edu

Sankalp Mucherla Srinath
Arizona State University
smucher1@asu.edu

Yeshwanth Reddy Gurreddy
Arizona State University
ygurredd@asu.edu

Abstract

Camera pose estimation, the task of determining a camera's 6-DoF trajectory, is central to computer vision and robotics. While traditional methods like SLAM and VO offer real-time performance, they suffer from drift over long or dynamic sequences. Neural approaches such as NeRF and iNeRF improve accuracy through photometric refinement but are computationally demanding and unsuitable for real-time use.To address this, we present a hybrid system that combines OCC-VO for real-time semantic mapping and initial pose estimation with a lightweight pose refinement module inspired by UC-NeRF. By leveraging spatiotemporal photometric consistency, this module refines keyframe poses efficiently. A key innovation is our closed-loop design, where refined poses are reinjected into the OCC-VO pipeline to continuously reduce drift and improve map consistency during long-term operation.

1. Introduction

Estimating camera pose in 3D space is a key problem in computer vision with applications in robotics, autonomous driving, AR/VR, and 3D reconstruction. Accurate pose estimation enables systems to localize, plan motion, and interact with dynamic environments.

Traditional methods like SLAM and Visual Odometry (VO) rely on geometric features and bundle adjustment. Systems such as ORB-SLAM and LSD-SLAM perform well in ideal conditions but suffer from drift in dynamic, low-texture, or poorly lit scenes.

Recent learning-based approaches improve robustness by using semantic features. OCC-VO, for instance, replaces sparse keypoints with dense 3D semantic occupancy grids, enhancing performance in challenging environments. However, it still accumulates drift due to its incremental, frame-to-frame nature.

Neural Radiance Fields (NeRF) revolutionize scene representation with volumetric rendering. Extensions like iNeRF and UC-NeRF refine poses using photometric consistency, but their high computational demands hinder real-time use.

To overcome this, we propose a hybrid system combining OCC-VO for real-time odometry with a lightweight, spatiotemporally constrained pose refinement module adapted from UC-NeRF. This module runs intermittently on selected keyframes to correct drift, and refined poses are reintegrated into OCC-VO, forming a closed-loop.

Unlike SLAM systems that rely on loop closure, our method updates poses continuously using local observations, making it suitable for large-scale deployments. The system integrates TPV-Former for semantic extraction, GICP for alignment, and UC-NeRF-based refinement for global consistency.

In essence, our hybrid framework balances speed and precision, enabling real-time, drift-resistant localization for autonomous and immersive systems.

Our key contributions include:

- A hybrid localization framework that integrates OCC-VO with a spatiotemporally constrained UC-NeRF-based pose refinement module.
- A closed-loop reinforcement policy that iteratively improves pose estimates and semantic maps using intermittent NeRF-based corrections.

2. Related Work

2.1. Visual Odometry and Semantic Mapping

Visual odometry (VO) and simultaneous localization and mapping (SLAM) have long served as the foundation for camera pose estimation in robotics and autonomous systems. Classical methods such as ORB-SLAM [9] and

SVO [2] rely on sparse feature matching or semi-direct image alignment to estimate motion across consecutive frames. While these techniques enable real-time performance and work well in structured environments, they suffer from limitations in dynamic scenes, textureless regions, and long-term trajectories without loop closure. In particular, they are prone to drift due to the accumulation of small pose estimation errors over time.

And methods like OCC-VO introduces a learning-based framework that performs dense mapping and robust odometry using only multi-camera RGB inputs. Central to its architecture is TPV-Former, a transformer-based encoder that projects image features into a tri-perspective volumetric representation consisting of top, front, and side views. This design effectively captures 3D spatial context without requiring explicit depth sensing, and enables the generation of semantically rich 3D occupancy grids that describe the free, occupied, and unknown regions of the scene.

Once the semantic occupancy grid is constructed, pose estimation is performed through a semantic-aware Generalized Iterative Closest Point (GICP) algorithm. This alignment process matches the predicted grid of the current frame with the accumulated global map, using semantic filtering to exclude dynamic or non-static objects such as vehicles and pedestrians. Additionally, OCC-VO applies a voxel persistence filter that suppresses transient noise by retaining only frequently observed voxels, thereby increasing map stability.

The system's RGB-only input modality makes it particularly suitable for urban driving scenarios where LiDAR or stereo depth may not be available or feasible. However, OCC-VO operates in an incremental frame-to-map manner and lacks global optimization modules such as loop closure or bundle adjustment. As a result, despite its robustness and efficiency, the system remains susceptible to long-term drift, especially in loop-free or large-scale environments. This limitation motivates the integration of a backend refinement module, as presented in our work, to enhance global geometric consistency over time.

2.2. Neural Rendering for Pose Refinement

Neural rendering approaches, particularly Neural Radiance Fields (NeRF) [8], have emerged as powerful tools for 3D scene representation and view synthesis. NeRF models a continuous volumetric field by regressing color and density at any 3D location and viewing direction using a neural network, enabling photorealistic rendering of novel views from sparse observations. While originally designed for high-quality rendering, NeRF and its variants have also been extended for camera pose estimation.

Inverse NeRF (iNeRF) [6] inverts the original NeRF pipeline by treating camera poses as optimization variables, allowing the system to refine pose estimates by minimizing the photometric error between observed and rendered images. However, iNeRF assumes a reasonably accurate initial pose and operates on static scenes, limiting its robustness and applicability in dynamic or large-scale real-world environments.

BARF (Bundle-Adjusting NeRF) [5] introduces a coarse-to-fine positional encoding scheme to address convergence issues in NeRF-based optimization. It jointly optimizes both camera poses and scene geometry using a bundle adjustment-like objective. Although BARF improves convergence stability and performance in structure-from-motion settings, it still requires training a full radiance field and is computationally demanding, making it unsuitable for real-time deployment.

To address these limitations, UC-NeRF [7] proposes a spatiotemporally constrained pose refinement module that separates pose optimization from scene rendering. Instead of rendering synthetic views, UC-NeRF constructs a dense correspondence graph across multiple camera views and time steps. Each edge in this graph connects matching pixels between image pairs and encodes a geometric constraint derived from multi-view observations. The refinement objective minimizes a geometric reprojection loss over this graph, enforcing consistency between matched points across both spatial and temporal domains.

A key strength of UC-NeRF lies in its explicit handling of uncertainty. The method incorporates uncertainty-aware weighting into the optimization, which downweights unreliable correspondences due to occlusions, dynamic objects, or low-texture regions. This enables robust pose refinement even with noisy initializations or imperfect observations. Importantly, UC-NeRF decouples pose refinement from radiance field training, allowing it to be used as a lightweight module that improves global pose accuracy without incurring the full computational cost of NeRF training. This makes it a promising candidate for hybrid VO systems seeking both efficiency and high accuracy.

2.3. Hybrid and Closed-Loop Approaches

Hybrid systems have emerged as a promising direction to reconcile the trade-off between fast visual odometry and the high-accuracy demands of global trajectory consistency. These approaches aim to combine the lightweight front-end estimation of traditional VO pipelines with the dense refinement capabilities of neural rendering techniques.

NeRF-VO [10] exemplifies this class by integrating sparse feature-based visual odometry with a NeRF-based backend that jointly optimizes camera poses and scene geometry using photometric consistency. It uses a sliding window optimization scheme to refine poses in real time without requiring full-sequence access, offering improvements in both loop-free scenarios and geometrically ambiguous environments.

NVINS [3] extends this concept by fusing inertial data with NeRF-augmented supervision in a Bayesian learning framework. The method incorporates synthetic views rendered from NeRF to train a pose regressor and explicitly models the uncertainty of pose predictions. This enhances resilience under dynamic motion and improves trajectory stability by propagating corrections from high-confidence visual segments.

Other systems demonstrate hybridization in different contexts. HybridPose [11] addresses 6D object pose estimation by combining keypoints, edge vectors, and symmetry-aware constraints into a deep network for robustness against occlusions and noise. Although it operates in a different domain, its multi-branch hybrid design exemplifies the value of integrating geometric reasoning with learned priors.

Mix-VIO [13] introduces a hybrid visual-inertial odometry framework that blends handcrafted optical flow tracking with deep learning-based sparse feature matching. Its sliding window backend and dynamic feature weighting enable stable tracking under rapid motion and changing illumination. However, like HybridPose, it focuses on robustness in motion tracking rather than global drift correction.

Despite their innovations, most existing hybrid systems lack mechanisms for closed-loop reintegration of refined pose estimates back into the VO pipeline. This limits their ability to perform long-term drift correction or to incrementally improve mapping over time. Our framework addresses this gap by incorporating a spatiotemporally constrained refinement module and reinjecting refined poses into the OCC-VO trajectory, enabling persistent correction and enhanced map consistency across large-scale deployments.

2.4. Our Position

Our work builds upon recent progress in visual odometry and neural pose refinement by proposing a closed-loop hybrid architecture that addresses both efficiency and accuracy in real-time pose estimation. Specifically, we combine the real-time semantic odometry capabilities of OCC-VO [4] with the spatiotemporally constrained pose optimization module introduced in UC-NeRF [7]. Rather than adopting the full NeRF-based radiance field optimization pipeline, which is computationally prohibitive for real-time use, we isolate the pose refinement component from UC-NeRF and adapt it for keyframe-level corrections in our pipeline.

This selective refinement strategy allows our system to periodically correct accumulated drift by enforcing geometric consistency across time and viewpoints—without incurring the full computational load of neural rendering. The corrected poses are then reinjected into the OCC-VO mapping loop, which improves subsequent pose predictions. This closed-loop feedback mechanism effectively bridges the gap between low-latency odometry and high-fidelity mapping, enabling long-term trajectory stability even in dynamic or unstructured environments.

Our approach maintains modularity and scalability, allowing intermittent neural optimization to complement fast visual odometry without breaking real-time constraints. As such, it presents a practical and deployable solution for robust camera localization in autonomous driving and robotic navigation scenarios.

3. Approach

We propose a hybrid pose estimation framework that combines real-time semantic visual odometry with drift correction via spatiotemporally constrained pose refinement. The system integrates three key components: an OCC-VO-based frontend for frame-to-frame tracking, a selective UC-NeRF-inspired refinement module for keyframe optimization, and a closed-loop mechanism to reinject corrected poses.

OCC-VO leverages TPV-Former to convert multi-view RGB images into dense 3D semantic occupancy grids. These grids are aligned incrementally using semantic-aware GICP, enhanced by dynamic object and voxel persistence filtering for robustness in dynamic environments.

To address long-term drift, selected keyframes are refined through photometric optimization over a local sliding window, guided by uncertainty modeling to focus on reliable correspondences. This refinement improves pose coherence across space and time without requiring full NeRF rendering.

Refined poses are asynchronously fed back into the OCC-VO pipeline, incrementally improving map alignment and trajectory accuracy. The resulting system maintains real-time responsiveness while achieving long-term consistency in complex environments.

3.1. Real-Time Visual Odometry with OCC-VO

The front-end of our system is built on the OCC-VO framework, which provides real-time camera pose estimation by projecting multi-view RGB images into a unified semantic 3D representation. This is achieved through TPV-Former, a transformer-based encoder-decoder architecture that extracts features from multi-camera images and projects them into a tri-perspective volume (TPV)—consisting of top, front, and side views. These volumetric features are then fused and decoded to produce dense 3D semantic occupancy grids without the need for explicit depth estimation.

These occupancy grids serve as a structured and semantically meaningful representation of the scene, which is leveraged for frame-to-map motion estimation. The pose of each incoming frame is estimated by aligning the newly predicted occupancy grid with the global accumulated map using a semantic-aware Generalized Iterative Closest Point

(GICP) algorithm. To increase robustness, the alignment is guided by semantic confidence weights and excludes dynamic objects that are likely to change over time, such as vehicles or pedestrians. This filtering is performed using semantic segmentation labels and temporal consistency checks to identify non-static regions.

OCC-VO is designed to operate efficiently in real-world settings with urban complexity. It avoids reliance on sparse keypoints or handcrafted features, and instead relies on dense semantic geometry, which allows it to remain robust in low-texture, cluttered, or dynamic environments. However, because it performs pose estimation in an incremental manner—without global bundle adjustment or loop closure—it is susceptible to long-term drift, especially in scenarios where large-scale consistency is required. This motivates the integration of a global refinement module in our overall framework.

3.2. Spatiotemporally Constrained Pose Refinement

To mitigate drift and enforce global geometric consistency, we incorporate a spatiotemporally constrained pose refinement strategy derived from the UC-NeRF framework. Unlike standard NeRF-based methods that typically optimize per-frame poses independently or rely on photometric consistency alone, this module jointly refines camera poses by leveraging dense geometric correspondences across both time and viewpoint. This enables the refinement process to integrate multi-view constraints without requiring full radiance field supervision or synthetic view rendering.

The core idea behind this module is to construct a correspondence graph E, where each edge connects a pair of matched pixels across space (i.e., different camera views) and time (i.e., different frames). Each correspondence implicitly encodes 3D scene structure through consistent observations from multiple viewpoints, without requiring explicit depth. These correspondences are discovered using dense optical flow and keypoint detection between image pairs using algorithms such as SIFT, and are maintained over a sliding temporal window to ensure computational efficiency.

Given this correspondence graph, the pose refinement objective is formulated as a geometric reprojection loss. For any matched pixels $\mathbf{q}_i^k$ and $\mathbf{p}_j^l$, observed from camera k at time i and camera l at time j, respectively, the loss penalizes the reprojection error induced by transforming the source pixel into the target view:

$$\mathcal{L}_{\text{rpj}} = \sum_{((i,k),(j,l))\in E} \left\| \mathbf{p}_j^l - \Pi_l \left((T_j \Delta T_l)^{-1} T_i \Delta T_k \, \Pi_k^{-1}(\mathbf{q}_i^k) \right) \right\|^2 \tag{1}$$

Here, Π_k^{-1} denotes the back-projection of a 2D pixel $\mathbf{q}_i^k$ into a normalized 3D ray in the coordinate frame of camera k, while Π_l represents the projection function of camera l. The global poses T_i and T_j describe the ego-motion of the rig at times i and j, and ΔT_k, ΔT_l denote the fixed extrinsics from the rig to the respective cameras. This formulation effectively transports a pixel from one view to another using the estimated global and relative poses, and compares it against the observed correspondence in the target frame.

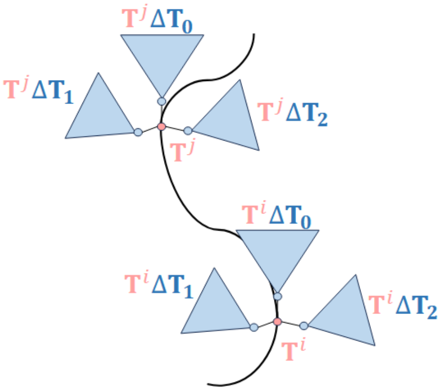
$\mathbf{T}^j\Delta\mathbf{T_0}$
$\mathbf{T}^j\Delta\mathbf{T_1}$
$\mathbf{T}^j\Delta\mathbf{T_2}$
$\mathbf{T}^j$
$\mathbf{T}^i\Delta\mathbf{T_0}$
$\mathbf{T}^i\Delta\mathbf{T_1}$
$\mathbf{T}^i\Delta\mathbf{T_2}$
$\mathbf{T}^i$

Figure 1. Visualization of the spatiotemporal reprojection geometry used in the pose refinement loss. A 3D point observed from camera k at time i is transformed and projected into the frame of camera l at time j, enforcing geometric consistency across space and time.

By minimizing this loss, the optimization aligns all matched keypoints across views, thereby enforcing consistent geometry in the estimated trajectory. In our implementation, we do not optimize for scene radiance or density fields, but rather isolate the pose graph refinement component to significantly reduce computational burden while preserving the drift-correcting benefits of the correspondence formulation.

Importantly, this pose refinement operates over a fixed-size window, allowing the method to scale efficiently while still propagating corrections across multiple frames. The use of spatiotemporal constraints enables the system to leverage redundant viewpoints in real-world sequences, such as those captured by surround-view camera rigs, and improves robustness under partial occlusions or textureless regions. This selective, correspondence-driven optimization is invoked intermittently on keyframes identified by motion and uncertainty heuristics, making it an efficient and effective component of our overall closed-loop system.

3.3. Closed-Loop Reintegration

Following the spatiotemporally constrained pose refinement stage, the corrected camera poses are systematically reintegrated into the OCC-VO pipeline. Although OCC-VO itself performs frame-to-map registration in an incremental manner [4], it does not incorporate a global feedback mechanism to revise past estimates. In our system, however, the refined poses generated by the UC-NeRF module are injected back into the ongoing trajectory and used to update the global semantic occupancy grid.

This reinjection is realized by modifying the pose graph maintained by OCC-VO. Specifically, when refined poses are available for previously visited keyframes, they overwrite the original estimates, allowing the semantic map to be realigned with greater geometric consistency. Because the occupancy grid is accumulated over time, this correction helps eliminate artifacts or drift-induced misalignments that would otherwise propagate through the mapping process. In effect, the refined poses act as loop closure constraints that implicitly regularize the global structure of the trajectory, without requiring explicit re-observation of past locations.

Moreover, the corrected poses serve as improved priors for subsequent frame-to-map registrations. By continuously injecting globally consistent estimates into the local visual odometry loop, we reduce sensitivity to transient noise and improve the stability of short-term motion estimation. This closed-loop architecture, though not present in the original OCC-VO design, is inspired by the modular separation of pose optimization in UC-NeRF [7], where pose updates are decoupled from radiance modeling and can be applied independently. Leveraging this modularity, we maintain the efficiency of the OCC-VO frontend while augmenting it with drift-resilient corrections that accumulate over time.

The synergy between fast visual odometry and delayed, high-accuracy refinement yields a system capable of real-time operation with long-term robustness. This is particularly valuable in large-scale or dynamic environments, where traditional frame-to-frame VO systems may accumulate substantial drift and lack the ability to correct past errors. By integrating refinement as an asynchronous feedback loop, our framework achieves improved localization accuracy and semantic map consistency without compromising runtime performance.

4. Experimental Setup

4.1. Dataset

We conduct our experiments on the nuScenes dataset [1], a large-scale, multimodal benchmark specifically designed for autonomous driving research. The dataset includes 1,000 driving scenes collected in urban environments such as Boston and Singapore, each lasting 20 seconds and annotated at 2 Hz. For the purposes of this study, we focus on the image-based components of the dataset, utilizing only the multi-view RGB streams.

nuScenes provides synchronized images from six surround-view cameras—front, front-left, front-right, back, back-left, and back-right—capturing 360° coverage of the vehicle's surroundings. This configuration enables robust evaluation of spatiotemporal pose consistency in complex urban scenarios, including dynamic objects, occlusions, and varying lighting conditions. Although the full dataset includes LiDAR, radar, and GPS-IMU streams, we restrict our analysis to the visual modality and the corresponding ego-pose annotations provided as ground truth.

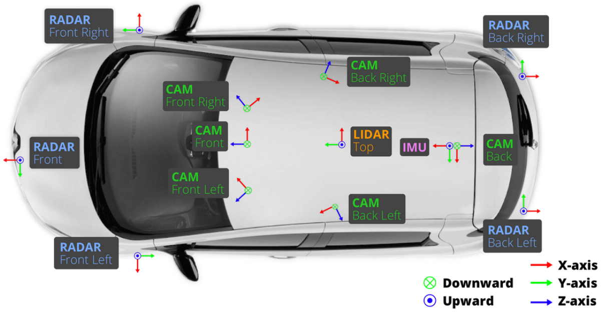

Figure 2. Sensor configuration in the nuScenes dataset, including camera, LiDAR, IMU, and radar placements.

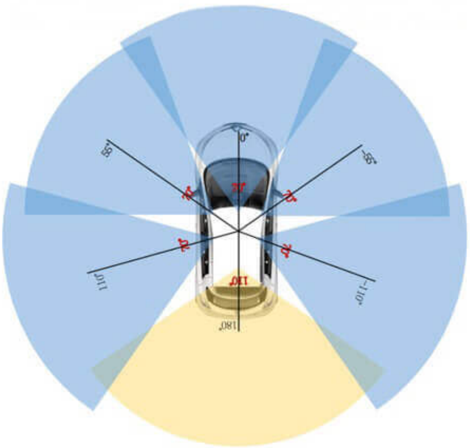

Figure 3. Field-of-view coverage of nuScenes camera rig. Cameras are positioned to cover overlapping 360° perspectives.

This setup allows us to test the generalizability and robustness of our hybrid pose estimation framework under realistic conditions, where accurate visual odometry and consistent global pose tracking are critical for autonomous navigation.

4.2. Evaluation Strategy

To evaluate the effectiveness of our hybrid pose estimation framework, we adopt standard trajectory evaluation metrics widely used in the visual odometry and SLAM literature. These metrics allow for a rigorous comparison of estimated camera trajectories against ground truth and are particularly effective in quantifying both local and global consistency.

Absolute Trajectory Error (ATE) [12] measures the global consistency of a trajectory by computing the root mean square error between the estimated camera positions and the ground truth after aligning the two trajectories. This metric is crucial for assessing long-term drift, as it reflects the cumulative deviation from the true path over the entire sequence.

Relative Pose Error (RPE) [12] quantifies the accuracy of motion estimation between consecutive frames. It calculates the deviation in relative transformations over short temporal intervals and serves as a good indicator of the system's frame-to-frame consistency. RPE is particularly useful for evaluating the short-term motion smoothness and the responsiveness of the odometry frontend.

Reprojection Error [14] evaluates the fidelity of pose estimates by projecting 3D scene points into multiple views and measuring the pixel-wise difference between the predicted projections and the actual observed image points. This metric is especially sensitive to pose inconsistencies and is well-suited for assessing spatiotemporal alignment in multi-view refinement modules, such as our UC-NeRF-based backend.

Although our system is currently in development and undergoing iterative refinement, these metrics provide a principled and well-established framework for assessing performance. As our refinement module matures, we aim to minimize these error terms to ensure both global accuracy and local smoothness in estimated trajectories.

4.3. Implementation Details

Our hybrid framework integrates components from both geometric and neural methods, built atop publicly available codebases with extensive custom adaptations for compatibility and efficiency. The front-end visual odometry module is based on a modified version of the official OCC-VO repository [4], optimized to process RGB inputs from the six synchronized cameras provided in the nuScenes dataset. TPV-Former is used as the semantic encoder to project multi-view images into tri-perspective volumes (top, front, side), which are fused into dense 3D occupancy grids representing the semantic layout of the environment.

To estimate camera pose in real-time, we perform frame-to-map registration using a semantic-aware Generalized Iterative Closest Point (GICP) algorithm. This implementation incorporates dynamic object filtering, which masks out moving classes like vehicles and pedestrians based on semantic segmentation labels, and employs a voxel persistence filter to retain only stable structures for alignment. These additions improve robustness to noise and transient scene elements. All front-end operations run at interactive frame rates on an NVIDIA A100 GPU.

For the backend, we isolate the spatiotemporally constrained pose refinement module from UC-NeRF [7], adapting it to function without radiance field rendering. Instead of learning a full volumetric scene representation, we construct a sparse correspondence graph by extracting keypoints using SuperPoint and associating matches across time and camera views via SuperGlue. These correspondences form the basis for geometric reprojection constraints.

Pose refinement is formulated as a non-linear least squares optimization problem, implemented using the Ceres Solver. The objective minimizes reprojection error over a sliding temporal window while preserving pose consistency across spatially distributed views. Uncertainty weighting is applied to each correspondence, allowing the system to down-weight low-confidence or ambiguous matches during optimization. This design enables scalable and accurate refinement without incurring the full computational overhead of a neural rendering pipeline.

4.4. Current Status

The current implementation of the system is operational and capable of executing the full hybrid pipeline, including real-time visual odometry using OCC-VO and offline spatiotemporal pose refinement. Preliminary outputs have demonstrated that the pipeline successfully tracks camera motion and generates 3D semantic occupancy grids with reasonable fidelity. However, the system's absolute pose accuracy and long-term trajectory consistency are still suboptimal, particularly in scenes with high dynamic content or low-texture regions.

Several areas remain under active development. These include tuning the voxel filtering thresholds within OCC-VO to reduce noise, improving the robustness of keypoint matching across time and camera views, and optimizing the refinement loss function parameters to ensure better convergence. Additionally, we are investigating strategies to automate keyframe selection for refinement and to accelerate correspondence graph construction.

While the current setup provides a stable foundation for testing, more extensive quantitative and qualitative evaluations are planned to assess generalization across diverse driving environments. The pipeline is modular and designed for iterative refinement, and ongoing improvements will target both accuracy and computational efficiency as the system matures.

5. Results (Preliminary)

5.1. Qualitative Observations

We conducted qualitative evaluations on selected nuScenes sequences to assess the practical behavior of our hybrid pose estimation framework. The OCC-VO front-end

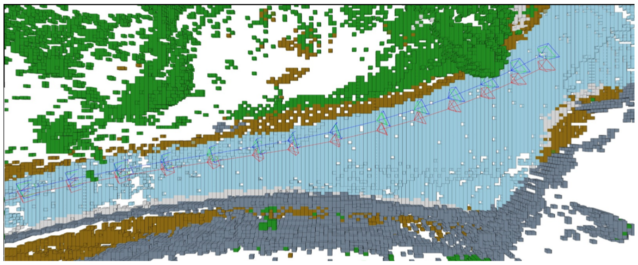

Figure 4. Result 1

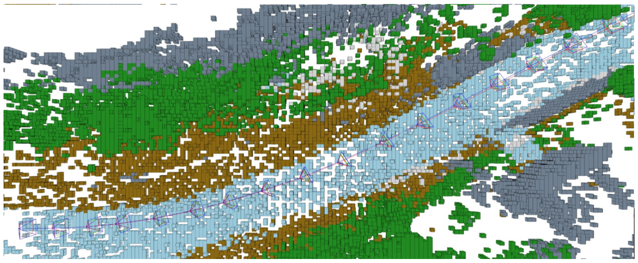

Figure 5. Result 2

successfully produced coarse but coherent camera trajectories and semantic 3D occupancy maps in real time, consistent with the findings reported in [4]. These occupancy maps offered a high-level geometric abstraction of the driving environment, capturing both free and occupied spaces from monocular RGB input. However, due to its incremental frame-to-frame pose estimation strategy, OCC-VO exhibited noticeable drift accumulation over longer trajectories, particularly in scenes with extended motion or few revisited landmarks.

To address this, we applied our UC-NeRF-inspired spatiotemporal pose refinement module to selected keyframes. Following the methodology outlined in [7], we constructed a correspondence graph across cameras and time steps and optimized pose estimates via reprojection loss minimization. The refined trajectories demonstrated improved local alignment, especially in sequences involving curved trajectories or sudden vehicle turns. We observed enhanced frame-to-frame consistency in the vicinity of refined keyframes, suggesting that the joint optimization across space and time effectively compensates for the accumulated drift.

Furthermore, the reinjection of these refined poses into the OCC-VO pipeline led to partial correction of map artifacts—such as stretched structures or voxel discontinuities—in the semantic occupancy grids. While these improvements were not always uniform due to occasional correspondence mismatches and residual motion blur in keyframes, the trend points toward better geometric consistency and map reliability. As future work, we aim to integrate more robust matching strategies and extend the refinement window to further enhance global accuracy.

5.2. Planned Evaluation Metrics

To evaluate the effectiveness of our hybrid pose estimation pipeline, we define a structured set of metrics that capture both global accuracy and local consistency. These include Absolute Trajectory Error (ATE), which quantifies cumulative drift over time by measuring the Euclidean distance between estimated and ground-truth poses; Relative Pose Error (RPE), which assesses short-term consistency between consecutive frames; and 2D Reprojection Error, which evaluates alignment quality at the image level by projecting reconstructed 3D points into camera views and comparing with observed keypoints. This combination of spatial and photometric metrics is consistent with evaluation protocols adopted in prior pose refinement and visual odometry works [4, 7].

ATE is particularly sensitive to global drift and is useful for benchmarking the benefits of pose reinjection. RPE captures improvements from local pose smoothness after refinement, and 2D reprojection error reflects the photometric alignment driven by the spatiotemporal optimization module. These metrics will be evaluated on standard sequences from the nuScenes dataset using ground-truth ego-pose annotations.

Table 1 provides a placeholder structure for our planned results. Final values will be updated after the full system—including keyframe selection, refinement, and closed-loop reinjection—is integrated and tuned.

Table 1. Planned Evaluation Metrics (Placeholder)

Method	**ATE (m) ↓**	**RPE (deg) ↓**
OCC-VO Only	–	–
+ Pose Refinement (UC-NeRF)	–	–
+ Reinjection into OCC-VO	–	–

Note: Metrics are defined but values are placeholders. Final evaluation will follow full integration.

6. Discussion

This project focused on building a hybrid pose estimation system that balances real-time odometry with global consistency through pose refinement. Our goal was not only to integrate these components but to also evaluate their feasibility under practical constraints such as limited compute, sparse supervision, and incomplete calibration.

6.1. What's Working

The OCC-VO front-end, supported by TPV-Former, successfully generated 3D semantic occupancy grids and estimated trajectories from multi-camera RGB input. The pipeline is robust enough to run on a consumer-grade GPU and yields useful intermediate outputs for pose tracking.

The implementation of spatiotemporally constrained pose refinement—adapted from UC-NeRF—has shown promise in correcting localized drift. When tested on selected keyframes, refined poses demonstrated improved consistency across views and time steps. The design of the correspondence graph and reprojection-based optimization aligns well with the constraints and data modalities available in the nuScenes dataset.

6.2. Development Path and Lessons Learned

Initially, we explored the use of iNeRF for pose estimation by inverting a NeRF model trained on RGB images. However, we quickly discovered that NeRF pipelines—especially iNeRF—are highly sensitive to initial pose estimates. Without accurate priors, optimization diverged or produced meaningless reconstructions.

We then attempted to use BARF (Bundle-Adjusting NeRF), which jointly optimizes camera poses and scene geometry using positional encoding. While BARF produced better convergence, the training process was computationally expensive and time-consuming, making it impractical for real-time use or large-scale data.

Eventually, we identified the pose refinement module in UC-NeRF, which provided a more modular and computationally feasible solution. By isolating this component and avoiding full radiance field learning, we were able to focus on the geometric aspects of pose correction—matching the real-time constraints of our system.

6.3. Current Challenges

Our refinement results remain limited by several unresolved challenges:

- Sparse and noisy correspondences between views reduce the stability of the optimization.

- Pose refinement currently runs offline and is not yet integrated into a continuous update loop.

- Lack of loop closure or global landmarks in OCC-VO constrains the ability to correct long-term drift.

6.4. Expected Outcomes and Future Work

We expect that with improved keypoint filtering, dynamic keyframe selection, and tighter integration between OCC-VO and the refinement module, the system will achieve significant drift reduction over longer sequences. Our future work includes automating reinjection of refined poses, extending the optimization to a sliding window, and scaling to more sequences in the nuScenes dataset.

Ultimately, we believe this hybrid approach—leveraging learning-based front-ends with geometry-aware refinement—is a promising direction for scalable and accurate pose estimation.

7. Conclusion

We presented a hybrid camera pose estimation system that combines the real-time capabilities of OCC-VO with the accuracy of a spatiotemporally constrained pose refinement module adapted from UC-NeRF. By isolating pose optimization from full NeRF rendering, our method offers an efficient, correspondence-based approach to correcting drift over long trajectories.

Leveraging multi-camera RGB input from the nuScenes dataset, the system generates dense 3D semantic maps using TPV-Former and incrementally estimates poses via semantic-aware GICP. To improve global consistency, selected keyframes undergo pose refinement through dense reprojection loss optimization over a spatiotemporal correspondence graph. This corrects accumulated drift and improves trajectory coherence, particularly in visually challenging scenes.

Early results show improved local smoothness, though current limitations include manual reinjection and sensitivity to correspondence quality. The modular design enables future integration of real-time updates and additional sensing modalities.

In summary, our approach delivers accurate, drift-resilient localization by uniting efficient VO and lightweight global optimization. Future efforts will focus on automating refinement, enhancing robustness, and scaling evaluation across larger datasets.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[2] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 15–22, 2014. 2

[3] Juyeop Han, Lukas Lao Beyer, Guilherme V. Cavalheiro, and Sertac Karaman. Nvins: Robust visual inertial navigation fused with nerf-augmented camera pose regressor and uncertainty quantification. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12601–12608, 2024. 3

[4] Heng Li, Yifan Duan, Xinran Zhang, Haiyi Liu, Jianmin Ji, and Yanyong Zhang. Occ-vo: Dense mapping via 3d occupancy-based visual odometry for autonomous driving, 2024. 3, 5, 6, 7

[5] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields, 2021. 2

[6] Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021. 2

[7] Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. Uc-nerf: Uncertainty-conditioned neural radiance fields for camera pose optimization, 2023. 2, 3, 5, 6, 7

[8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 2

[9] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[10] Jens Naumann, Binbin Xu, Stefan Leutenegger, and Xingxing Zuo. Nerf-vo: Real-time sparse visual odometry with neural radiance fields, 2024. 2

[11] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations, 2020. 3

[12] Jürgen Sturm, Niko Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 573–580. IEEE, 2012. 6

[13] Huayu Yuan, Ke Han, and Boyang Lou. Mix-vio: A visual inertial odometry based on a hybrid tracking strategy. *Sensors*, 24(16):5218, 2024. 3

[14] Zheng Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251, 2018. 6