ZeroVO: Visual Odometry with Minimal Assumptions

Lei Lai* Zekai Yin* Eshed Ohn-Bar
Boston University
{leilai, zekaiyin, eohnbar}@bu.edu

Abstract

We introduce ZeroVO, a novel visual odometry (VO) algorithm that achieves zero-shot generalization across diverse cameras and environments, overcoming limitations in existing methods that depend on predefined or static camera calibration setups. Our approach incorporates three main innovations. First, we design a calibration-free, geometry-aware network structure capable of handling noise in estimated depth and camera parameters. Second, we introduce a language-based prior that infuses semantic information to enhance robust feature extraction and generalization to previously unseen domains. Third, we develop a flexible, semi-supervised training paradigm that iteratively adapts to new scenes using unlabeled data, further boosting the models' ability to generalize across diverse real-world scenarios. We analyze complex autonomous driving contexts, demonstrating over 30% improvement against prior methods on three standard benchmarks—KITTI, nuScenes, and Argoverse 2—as well as a newly introduced, high-fidelity synthetic dataset derived from Grand Theft Auto (GTA). By not requiring fine-tuning or camera calibration, our work broadens the applicability of VO, providing a versatile solution for real-world deployment at scale.

1. Introduction

For a robot or autonomous vehicle to function reliably in the real world, a *generalized* Visual Odometry (VO) system is essential—one that can robustly estimate the relative camera pose in metric coordinates from a sequence of images under diverse and unforeseen conditions. However, generalization remains a significant challenge for current VO models, which often suffer from lost feature tracks, optimization instability, and drift, particularly when exposed to varying lighting, dynamic scenes, or adverse weather conditions [1, 11, 31, 35, 54, 69].

Due to the inherent difficulty and ambiguity in modeling camera ego-motion, a dynamic 3D world, and real-world scale from 2D images, monocular VO algorithms have traditionally been built on strong assumptions and geometric constraints [3, 7, 10, 13, 16, 19, 20, 51, 52, 57, 84]. While carefully designed camera calibration or evaluation on fixed data distributions can be effective in controlled settings, such approaches can limit adaptability and scalability to real-world scenarios with varying configurations that may not align with such assumptions.

VO techniques have increasingly adopted learning-based components to exploit statistical regularities in scene structure and motion dynamics. However, most learning-based methods rely on privileged ground-truth data (e.g., accurate camera parameters, optical flow) for supervision and often train and evaluate on the same dataset [7, 20, 32, 33, 57, 62, 63, 68, 77]. Although recent studies explore generalization beyond single-dataset settings [37, 42, 62, 63, 70], current models continue to exhibit significant errors in the presence of more complex everyday contexts [1, 11, 31, 35, 54], including harsh conditions such as rainy or snowy nights (e.g., frequent glare, water streaks, reflections, and reduced visibility), lens degradation (e.g., condensation, scratches, dirt), or highly dynamic environments (e.g., dense intersections or aggressive motion). How can we design VO models that generalize across conditions instead of quickly suffering from instability and drift?

In this work, we aim to advance the capabilities of learning-based monocular VO. We introduce **ZeroVO**, a novel transformer-based approach for robustly predicting relative camera motion at real-world scale across variable scenes in a zero-shot manner. By leveraging cross-attention mechanisms [18, 65] to efficiently integrate contextual and geometric priors directly into the network architecture, ZeroVO avoids common limiting assumptions—such as reliance on camera calibration or costly optimization steps. Specifically, we fuse versatile multimodal text [43, 45, 55] and depth-based priors [23, 27, 53, 80] to address inherent scale ambiguity in metric VO. We demonstrate that our proposed model is robust to noisy and uncalibrated setups. We further optimize the model using a novel multimodal semi-supervised training framework that filters noisy

* Equally contributed.

pseudo-labels in a geometry and language-guided process. Our flexible VO framework achieves state-of-the-art, off-the-shelf performance across diverse autonomous driving datasets. To comprehensively assess system generalizability, we also collect and analyze a novel Grand Theft Auto (GTA) dataset featuring challenging scenarios with harsh weather, high-speed motion, complex traffic scenes, and varied camera settings. Our dataset and code are available at https://zvocvpr.github.io/.

2. Related Work

Our framework builds on advances in foundational computer vision models, particularly in metric depth prediction and rich, generalized vision-and-language embeddings.

Learning-Based Monocular Visual Odometry: Learning-based monocular visual odometry tasks can be roughly categorized into two main approaches: neural network models combined with multi-step geometric optimization (e.g., full SLAM [7, 42, 49, 62, 63, 95]) or direct, end-to-end relative pose estimation from two or few consecutive frames [37, 66, 70, 79]. Hybrid methods such as Droid-SLAM [62] have demonstrated strong performance in dense scene reconstruction and pose estimation. In contrast, two-frame pose regression tends to be more robust in short-distance tracking scenarios, while SLAM and other geometry-based approaches typically require continuous, long-frame sequences. These methods often rely on long-term feature matching and global optimization techniques, such as loop closure detection. Although certain methods [30] can aid in initialization, SLAM remains sensitive to environmental features and accurate motion tracking, i.e., can fail to build and update a reliable map in feature-deficient environments (e.g., corridors or repetitive textures) or highly dynamic settings (e.g., crowds). In contrast, two-frame pose regression is less affected by such conditions as it does not rely on maintaining a global representation. However, two-frame pose regression can be prone to drift accumulation, as it lacks the temporal optimization over extended frame sequences needed to correct for drift. Our work improves over two-frame approaches due to inherent efficiency, versatility (i.e., as input to downstream optimization), and minimal assumptions.

Metric Depth Estimation from Images: We leverage advances in metric depth estimation to address the inherent ambiguity in recovering camera translation at real-world scale. Traditional monocular depth models often rely on scale-invariant losses or sparse supervision, making them unsuitable for tasks such as visual odometry that require consistent metric scale. Recently, models for predicting *metric* depth have demonstrated practical performance [27, 53, 76, 82]. Models such as Depth Anything [76] and UniDepth [53] aim to generalize depth prediction across a wide range of scenes by leveraging large-scale vision foundation models. WordDepth [82] proposes the use of language-guided priors to reduce ambiguity in unconstrained prediction of scale. Metric3Dv2 [27] provides a zero-shot model that was trained across numerous datasets and is capable of predicting real-world scale depth (and surface normals) in diverse settings. By leveraging known camera intrinsics and extrinsics, the model learns to transform inputs into a canonical camera space. While existing models often struggle in challenging real-world scenarios, we adopt Metric3Dv2 to extract real-scale depth features that enable accurate and robust visual odometry. To further increase the flexibility and applicability of our approach, we do not rely on traditional camera calibration or predefined image information [85, 88–90]. Instead, we consider settings where calibration may be unavailable or inaccurate, and incorporate single-image camera parameter estimation techniques such as WildCamera [94] to support inference under uncalibrated conditions.

Rich Vision-and-Language Embeddings: Language-guided models have shown strong generalization capabilities by effectively bridging multiple modalities. Through joint embedding spaces that capture generalized semantic relationships between images and language, Vision-Language Large Models (VLLMs) models have recently achieved state-of-the-art results in diverse tasks such as image captioning [15, 74, 81], visual question answering [2], and cross-modal retrieval [26]. LLaVA [45], for instance, is now being broadly used across contexts and tasks [12, 43, 46, 87]. Preliminary studies in autonomous driving, e.g., Tian et al. [64], have shown VLLMs to be useful for robustness under long-tail events. In our work, we propose to integrate VLLMs to extract high-level semantic descriptions of driving scenes that could serve as language-based priors that guide metric-scale odometry and complement adaptive inference under challenging visual conditions.

Semi-Supervised Learning: Our work aims to develop flexible models that can effectively adapt to new environments, including through the use of unlabeled data. Semi-supervised learning (SSL) is being increasingly used in computer vision and machine learning tasks, particularly in domains where annotated data is scarce, costly, or requires expert supervision [4, 5, 9, 14, 21, 24, 25, 28, 34, 38, 58, 61, 93]. In the context of visual odometry, SSL can potentially enable the use of large-scale, unlabeled video data, such as web videos [37, 86], to expand the diversity of training scenarios and further improve generalization. However, SSL also presents challenges, including noisy pseudo-labels and the risk of propagating errors through repetitive training cycles, which we address in our work through multimodal pseudo-label selection mechanisms.

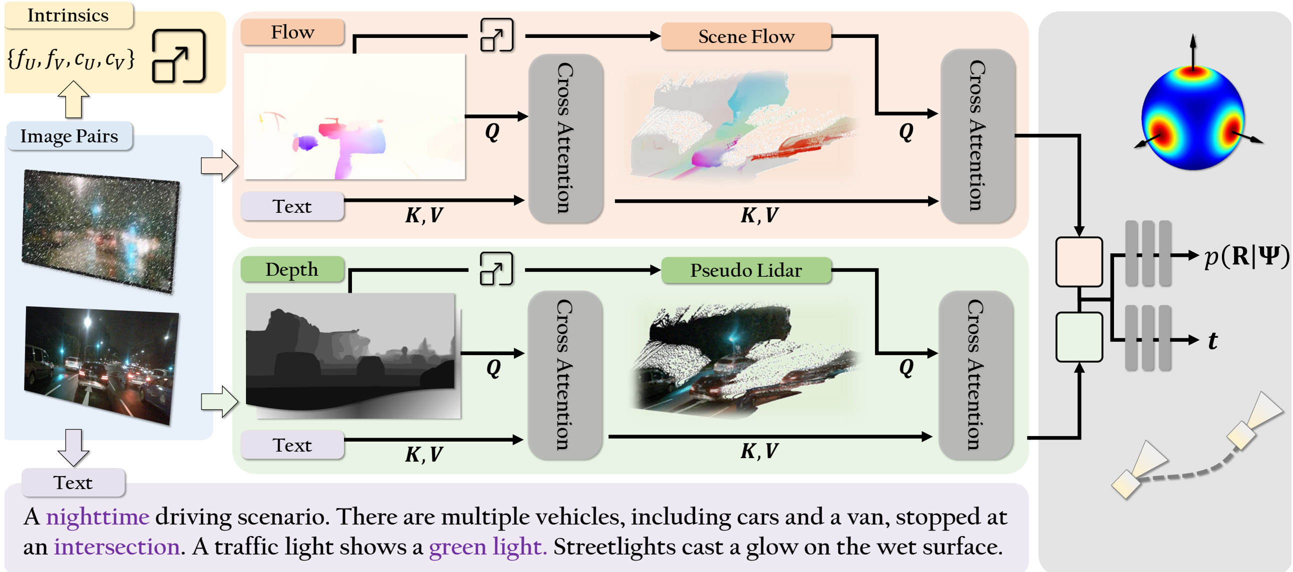

Figure 1. **Multimodal and Geometry-Guided Network Overview.** Given a pair of input images, our model computes a rich multimodal embedding through a transformer-based fusion module. The embedding is then passed to a two-branch decoder MLP that outputs real-world translation and rotation. Our architecture (Sec. 3.1) leverages cross-attention to fuse complementary cues, including flow, depth, camera intrinsics, and language-based features in a geometry-aware manner. The language prior is first used to refine both the depth map and 2D flow estimates. The refined depth is then unprojected into 3D (using estimated parameters) to compute scene flow, which is further enhanced and fused with additional features before decoding. By embedding geometric reasoning and multimodal priors directly into the network structure, our model achieves strong zero-shot generalization across diverse and challenging settings.

3. Method

Our method (Fig. 1) facilitates generalization via minimal and versatile image-based priors, integrated throughout our model structure. In this section, we first formalize our generalized, calibration-free monocular VO task. We then detail the proposed transformer-based geometry and prior-guided network structure in Sec. 3.1 and the semi-supervised training process in Sec. 3.2.

Monocular VO with Minimal Assumptions: In its most general form, monocular VO assumes two consecutive RGB frames $\mathcal{I} = \{\mathbf{I}_{i-1}, \mathbf{I}_i\}$, $\mathbf{I} \in \mathbb{R}^{W\times H\times 3}$ and learns to predict a real-world relative pose between the two camera views $\mathbf{T}_i = [\mathbf{R}_i|\mathbf{t}_i]$, where $\mathbf{R}_i \in \mathrm{SO}(3), \mathbf{t}_i \in \mathbb{R}^3$ are the relative rotation and translation, respectively. We focus on the efficient two-frame setup as it enables a fair comparison to other baselines methods (e.g.,TartanVO [70]) while quantifying *real-time sequential drift*, i.e., prior to any additional global optimization steps, such as loop closure and bundle adjustment [51, 60, 62]. In Sec. 4, we find ZeroVO to outperform more complex methods that leverage computationally expensive, multi-frame refinement steps. We emphasize that monocular VO methods generally evaluate under up-to-scale settings [51, 63, 70], as estimating a metric-scaled transform from image pairs can be difficult, while reducing the solution space through known camera pose $\mathbf{T}_i^{cam}$ and intrinsics, including the camera's focal length and center, $\{f_U, f_V, c_U, c_V\}$ (these are used in the camera intrinsic matrix, denoted as $\mathbf{K}_i \in \mathbb{R}^{3\times3}$). However, in our formulation, we do not assume any prior knowledge of camera parameters, as it can be limiting and require recalibration in cases of lens issues or different camera setups. Instead, to guide learning and inference, we rely on a set of versatile image-based priors built into the network structure. Specifically, we extract a rich set of modalities, including estimated flow $\hat{\mathbf{F}}_i \in \mathbb{R}^{W\times H\times 2}$, depth map $\hat{\mathbf{D}}_i \in \mathbb{R}^{W\times H}$, camera parameters $\hat{\mathbf{K}}_i$, and rich language-based context features $\mathbf{Z}_i^l \in \mathbb{R}^{W_l\times H_l}$ that provide complementary cues regarding scene semantics, layout characteristics, and scale. Our network structure fuses the estimated cues in a geometrically-guided process, discussed next.

3.1. Geometry and Prior-Guided Network

Our network structure comprises three key components: (1) an *encoding module*, which estimates camera intrinsic parameters and extracts a rich, multimodal set of cues; (2) a text-conditional, *geometry-guided transformer module* that leverages general structural priors to unproject data into 3D space and fuse the different modalities; and (3) a *decoding module* for probabilistically predicting ego-motion.

Intrinsic Parameters Estimation: VO methods generally rely on accurate knowledge of camera extrinsic and intrinsic parameters while training and testing on datasets with fixed camera settings. To enable more generalized VO, we do not rely on such restrictive assumptions. We instead propose to estimate the camera intrinsic parameters leveraging recent advances in in-the-wild, single-image intrinsic parameter estimation [27, 94] (primarily relying on 3D monocular priors). We leverage an off-the-shelf solution [94], as we do not require the estimation to be completely accurate. The intrinsic matrix will also be used to inform the geometry-aware transformer and semi-supervised network training (Sec. 3.2). To align with image-level cues and enable the network to recover from noisy estimates, the intrinsic parameters are encoded into an image-sized array,

$$\mathbf{I}^{\hat{\mathbf{K}}}(u,v) = \frac{|u - c_U|}{f_U} + \frac{|v - c_V|}{f_V} \tag{1}$$

where the intrinsic information is explicitly preserved within each intrinsic map [70]. Encoding parameter information into an image map provides an efficient approach for our transformer module to reason over noisy geometric information, as will be discussed below. We note that $\mathbf{I}^{\hat{\mathbf{K}}}$ uniquely represents a specific camera configuration.

Extracting Multimodal Image Cues: To holistically represent general scene priors, scene dynamics, and camera motion and geometry, we employ a rich and complementary set of image-based features. As in standard VO methods, we extract optical flow [70] from the image pair using a MaskFlownet [91] encoder (We extract the optical flow $\hat{\mathbf{F}}$ as well as a correlation feature $\hat{\mathbf{F}}^c$, which represents 2D correspondences between the images, from the intermediate layer of MaskFlownet). To estimate a metric-scale depth map $\hat{\mathbf{D}}$, we utilize the estimated camera intrinsic parameters with Metric3Dv2 [27]. Finally, although camera information and metric depth can aid in understanding camera projection and motion, estimating these from a single image can be noisy and ill-posed. Thus, in addition to depth-based cues, we propose to leverage complementary text-based cues that can reduce ambiguity by capturing high-level scene semantics and layout characteristics. Specifically, we leverage LLaVA-NeXT [44] to extract rich image descriptions which are encoded using Sentence Transformers [55]. In addition to providing useful context in arbitrary scenes during inference, we leverage the language-based cues to filter noisy pseudo-labels in Sec. 3.2. We fuse modalities in a geometry-guided process, described next.

Unprojection to Pseudo-3D: The estimated depth map can be unprojected into a 3D point cloud $\mathbf{P} \in \mathbb{R}^{W \times H \times 3}$ using the estimated camera matrix [71], i.e., by computing 3D world coordinates $\mathbf{p} = d\hat{\mathbf{K}}^{-1}\mathbf{u}$, where $\mathbf{u} = (u, v)$ is a pixel in homogeneous coordinate and $d = \hat{\mathbf{D}}(\mathbf{u})$. We stack and normalize the resulting unprojection into a 3D array $\hat{\mathbf{D}}^{3\mathrm{D}}$. We unproject the 2D optical flow into 3D to obtain a scene flow $\hat{\mathbf{F}}^{3\mathrm{D}}$ matrix (additional details regarding this step can be found in our supplementary). While these steps integrate physically-coherent camera and 3D information into a consistent representation, we expect the 3D maps to be noisy, particularly in our challenging generalization and adverse settings. Hence, instead of being explicit constraints, the 3D maps are integrated as minimal structures into a transformer-based module.

Language and Geometry-Guided Transformer: We employ transformer [18, 65] to fuse the multimodal priors while reasoning over structure and noisy pseudo-3D information. We process the estimated flow and depth maps to compute two types of language-conditioned descriptors, a depth-based feature $\mathbf{Z}^D$,

$$\mathbf{Z} = \mathrm{CA}(\mathrm{PE}([\hat{\mathbf{D}}, \mathbf{I}^{\hat{\mathbf{K}}}]), \mathbf{Z}^l) \tag{2}$$

$$\mathbf{Z}^{\mathbf{D}} = \mathrm{CA}(\mathrm{PE}(\hat{\mathbf{D}}^{3\mathrm{D}}), \mathbf{Z}) \tag{3}$$

and a flow-based feature $\mathbf{Z}^F$ computed in a similar manner,

$$\mathbf{Z} = \mathrm{CA}(\mathrm{PE}(\hat{\mathbf{F}}^c), \mathbf{Z}^l) \tag{4}$$

$$\mathbf{Z}^{\mathbf{F}} = \mathrm{CA}(\mathrm{PE}(\hat{\mathbf{F}}^{3\mathrm{D}}), \mathbf{Z}) \tag{5}$$

where $\mathrm{CA}(\mathbf{Q}, \mathbf{KV})$ denotes Cross-Attention, with query Q and key-value pair $\mathbf{KV}$, and PE denotes a patch and positional embedding [18]. We note that we concatenate features with the intrinsic image to enable the model to learn coherence under noise, as accurate 3D reasoning is influenced by the focal length [27].

Probabilistic Ego-Motion Decoder: The refined and aligned features, $\mathbf{Z}^{\mathbf{F}}$ and $\mathbf{Z}^{\mathbf{D}}$, are concatenated and decoded into ego-motion. Our decoder consists of two MLP output branches, one predicting translation and the other rotation. For translation, we leverage metric-scale regression [70]. For rotation estimation, we fit a probabilistic distribution, specifically a matrix Fisher distribution (following [37, 48, 50]) to model the rotation distribution in $\mathrm{SO}(3)$.

$$p(\mathbf{R}|\mathbf{\Psi}) = \frac{1}{c(\mathbf{\Psi})} \exp(tr(\mathbf{\Psi}^\top \mathbf{R})) \tag{6}$$

where $\mathbf{R} \in \mathrm{SO}(3)$ is the rotation matrix, $\mathbf{\Psi} \in \mathbb{R}^{3\times3}$ are the parameters of matrix Fisher distribution, and $c(\mathbf{\Psi})$ is a normalization constant [48].

3.2. Model Training via Semi-Supervision

Due to the minimal assumptions employed by our calibration-free VO framework, the model can be effectively trained over in-the-wild, large-scale video collections. Hence, we consider both the standard supervised and a proposed semi-supervised training setup, detailed in this section. We employ the rich priors extracted from Sec. 3.1 in the semi-supervised training to filter noisy pseudo-labeled samples.

Supervised Training: Our model can be trained for a standard VO task, without requiring privileged information, e.g., ground-truth camera parameters, flow, or depth. We optimize the multi-head decoder MLP using Mean Squared Error (MSE) loss over predicted translation $\hat{\mathbf{t}}$ and negative log-likelihood of rotation $\mathbf{R}$ over the predicted distribution parameters $\hat{\mathbf{\Psi}}$,

$$\mathcal{L} = \|\mathbf{t} - \hat{\mathbf{t}}\|_2^2 - \log(p(\mathbf{R}|\hat{\mathbf{\Psi}})) \tag{7}$$

While our supervised model already achieves strong performance, we further explore incorporating an additional training stage using pseudo-labeled samples generated by running the first-stage model on unlabeled data.

Generalization with Semi-Supervised Training: Our goal is to learn effective representations for generalized VO at scale. We thus investigate leveraging semi-supervised training to continue and update the model from unlabeled data. This training involves two stages, first with a supervised (i.e., teacher) model trained using the aforementioned objective function on an annotated dataset. Next, we sample pseudo-labels from the model [9, 39, 56] over a large unconstrained dataset collected from YouTube [75], and re-train the model over the mixed annotated and pseudo-labeled dataset. Thus, the semi-supervised setup enables us to investigate the robustness and flexibility of our model in learning from diverse and challenging data with noisy supervision. While semi-supervised training has become a standard evaluation setup in computer vision [29, 40, 59, 67, 73, 78], as in Sec. 3.1 we explore the benefits of prior-informed mechanisms that can facilitate learning at scale from noisy examples.

Geometry-Guided Pseudo-Label Selection: To robustly learn from potentially noisy pseudo-labels, we employ a geometrical consistency error obtained based on estimated quantities. Specifically, motivated by prior work in unsupervised VO using known camera parameters [41, 47, 47, 83, 92], we warp a frame to the next frame with the estimated intrinsic matrix and ego-motion, $\mathbf{u}_i = \hat{\mathbf{K}}_i(d\hat{\mathbf{R}}_i\hat{\mathbf{K}}_{i-1}^{-1}\mathbf{u}_{i-1} + \hat{\mathbf{t}}_i)$. We then employ a Structural Similarity Index Measure (SSIM) error [6] to quantify the similarity between an observed image $\mathbf{I}_{i+1}$ and $\hat{\mathbf{I}}_{i+1}$. To ensure that we capture diverse patterns of reconstruction challenges, we further normalize by the two-frame SSIM, i.e.,

$$\text{normSSIM} = \frac{\text{SSIM}(\hat{\mathbf{I}}_{i+1}, \mathbf{I}_{i+1})}{\text{SSIM}(\mathbf{I}_i, \mathbf{I}_{i+1})} \tag{8}$$

and exclude samples based on a fixed NormSSIM threshold. We note that SSIM assesses similarity by evaluating structural information, luminance, and contrast, thereby offering a perception-oriented measure of similarity in contrast to traditional measures based on pixel-wise errors.

Language-Guided Pseudo-Label Selection: In addition to the geometry-based consistency pseudo-label check, we leverage our language-based module to filter redundant examples while maintaining an informative and diverse pseudo-labeled dataset. Although distinct text descriptions may not necessarily correspond to distinct pose transformations, we observe that two images characterized by nearly identical text descriptions are likely to be close in the visual space as well. To address sentence sequence variations within a paragraph, rather than serializing all text features into a single vector, we interpret the language feature as a subspace in a higher dimension. We leverage a subspace-based similarity over a short time window H, and compute the text feature similarity between the first image $\mathbf{I}_i$ and the last image $\mathbf{I}_{i+H}$ in the time window [36]. Specifically, we compute similarity as:

$$\text{subspace-sim} = \sin(\arccos(\text{trace}(\Lambda)))^2 \tag{9}$$

where Λ is the eigenvalues matrix obtained via Singular Value Decomposition over $Q_i^\intercal Q_{i+H}$, the orthonormal matrices from the QR decompositions of text features $\mathbf{Z}_i^l$ and $\mathbf{Z}_{i+H}^l$. As in the geometric consistency selection, we remove sequences with low informativeness (i.e., high subspace-sim). The selection mechanism can thus help stabilize learning under the noisy and diverse pseudo-labels.

3.3. Implementation Details

In our implementation, we leverage the pre-trained WildCamera [94] model to estimate camera intrinsics. We utilize the MaskFlowNet encoder [91] and Metric3Dv2 [27], as flow and depth backbones, respectively. Sentence Transformers [55] is used to extract a 15×768 language-based feature matrix. For semi-supervised training, we follow prior work and collect a large-scale, unconstrained web video dataset for additional training [37, 75]. In our analysis, we present three model variants: ZeroVO, ZeroVO+, and LiteZeroVO+. ZeroVO serves as the default model in our experiments, while ZeroVO+ is further trained on the web video dataset with the proposed multimodal pseudo-label selection mechanism. LiteZeroVO+ shows a resource-constrained variant that omits the language-conditioned input modules by replacing the cross-attention module (for conditioning on the language cues and refining the estimated flow and depth maps) with self-attention. The training protocol remains consistent with that of the standard ZeroVO+. We train our network architecture using NVIDIA RTX 4090 GPU with a batch size of 16. ZeroVO+ achieves an inference speed of approximately 0.6 FPS, primarily constrained by the slower Lava-Next module (0.7 FPS), and LiteZeroVO+ obtains an inference speed of 5 FPS. Complete implementation and training details can be found in our supplementary.

4. Experiments

4.1. Experimental Setup

Real-World Datasets: To study the generalization ability of our model, we conduct experiments using five datasets including three widely adopted datasets for autonomous driving: nuScenes [8], KITTI [22], and Argoverse 2 [72], as well as an introduced Grand Theft Auto V (GTA) simulated dataset with challenging environmental and lens conditions. nuScenes covers four distinct regions across Boston and Singapore: Boston-Seaport, Singapore-OneNorth, Singapore-Queenstown, and Singapore-Holland Village. It encompasses various challenging conditions, such as heavy traffic, nighttime driving, and scenarios involving strong light reflections, making nuScenes particularly valuable for assessing the robustness of models under diverse and complex real-world conditions. In our evaluation, we train on a subset of nuScenes, and test on other benchmarks in a zero-shot manner. KITTI is the most widely evaluated dataset in the VO task. Specifically, the camera intrinsics in KITTI differ significantly from those of the other three benchmarks, making it an important dataset for evaluating a model's ability to adapt to varying camera configurations. Argoverse 2 collects data from six distinct U.S. cities and encompasses a wide range of weather conditions and driving scenarios. Notably, the dataset includes grayscale images captured by the stereo front camera, which provides another generalization stress-test for the model. We also follow Lai et al. [37] and leverage online driving videos from YouTube, encompassing footage across multiple cities, including urban areas, villages, national parks, mountainous regions, and coastal areas, under a wide range of weather conditions. This dataset enables us to study the benefits of diverse unlabeled data while providing an ideal environment for the model to self-learn numerous variations induced by camera motions.

GTA Dataset: Besides the three public datasets, we introduce a newly generated simulated dataset derived from the high-fidelity, GTA simulation. Our GTA dataset consists of 922 driving sequences captured within a simulated city environment, encompassing a range of diverse weather conditions, driving speeds (particularly high-speed maneuvers not found in other public datasets), traffic scenarios, and times of day. Compared to other commonly used open-source simulation platforms such as CARLA [17], GTA offers several key advantages: (1) enhanced image realism through the application of the reshade graphic settings that support higher quality rendering, and (2) a wider variety of road conditions across various weather scenarios. For on-road driving, these conditions include significant uphill and downhill gradients, tunnels, and underground parking facilities; for off-road driving, the environment features mountains, deserts, snow-covered terrains, and forests, thereby enabling more precise and complex rotational dynamics throughout the map.

Experimental Setting: Similar to XVO [37], our framework is trained on data from a single city in the nuScenes dataset. Unlike XVO, we observed that Boston-Seaport, Singapore-Queenstown, and Singapore-Holland Village contain the majority of challenging conditions, such as rain, nighttime driving, light reflections, and heavy traffic. Therefore, we use Singapore-OneNorth as our supervised training dataset and the remaining regions, KITTI, Argovere 2, and GTA, as test datasets. It is important to note the main evaluation is done on datasets that were unseen by our model during training and without assumed camera parameters.

Baselines: We compared the four most related baselines that demonstrate generalization across datasets without requiring additional fine-tuning: TartanVO [70], XVO [37], DPVO [63], and Metric3D+Droid-SLAM (M+DS) [27, 62]. TartanVO employs effective random cropping and resizing techniques to simulate diverse camera configurations, thereby enhancing the generalization of rotation estimation across unseen datasets. XVO leverages a multi-modality architecture to implicitly extract richer spatial features and integrates self-training to achieve robust generalization performance in both rotation estimation and real-world scale recovery. DPVO employs a recurrent update operator for patch-based correspondence, complemented by differentiable bundle adjustment, demonstrating strong zero-shot performance in rotation estimation. M+DS utilizes the generalization capabilities of Metric3D v2 and Droid-SLAM to accurately estimate metric depth and rotation, effectively recovering the motion trajectory at a real-world scale. Our main baseline is M+DS which achieves state-of-the-art generalization results across dataset.

Metrics: To provide a comprehensive analysis of the results, we utilize Translation Error (t_{err}), Rotation Error (r_{err}), Absolute Trajectory Error (ATE), and Scale Error (s_{err}) [22, 37]. t_{err} and r_{err} compute the average translation error (%) and rotation error ($^{\circ}$/100 m) across all possible subsequences within a test sequence with lengths ranging from 100 to 800 meters. ATE measures the deviation between the estimated trajectory and the ground-truth trajectory by comparing the positions of corresponding poses, making it an effective metric for measuring drift over time. The scale error (s_{err}) measures the average discrepancy between the predicted translation and the ground truth translation. Combined with rotation error (r_{err}) and Absolute Trajectory Error (ATE), it allows us to effectively determine whether accumulated drift is attributed to scale inaccuracies or rotational deviations.

Table 1. **Comparative Analysis Across Datasets.** We compare ZeroVO variants with existing baselines using standard metrics of translation, rotation, absolute trajectory, and scale errors. All methods are provided with estimated camera intrinsics and metric depth. ZeroVO+ is our model trained with further data using semi-supervision, and LiteZeroVO+ is a smaller model variant for resource-constrained settings. Our models demonstrate strong performance across metrics and datasets, particularly in metric translation estimation. As highlighted by the scale error, GTA and nuScenes contain challenging evaluation settings, including nighttime, weather variations, haze, and reflections. We note that TartanVO and DPVO baselines (in gray) only predict up-to-scale motion and use privileged information, i.e., ground-truth scale alignment in evaluation.

Method	KITTI 00-10				nuScenes				Argoverse				GTA			
	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}
XVO [37]	16.82	3.84	168.43	0.17	12.75	5.11	8.30	0.16	9.13	4.86	5.70	0.12	25.56	12.64	28.02	0.21
M+DS [27]	14.22	2.72	154.77	0.09	17.08	**1.46**	10.46	0.18	16.67	**1.79**	8.51	0.13	23.53	10.38	12.96	0.26
ZeroVO	7.69	2.72	105.07	0.07	10.98	4.48	6.79	0.14	6.83	3.13	4.10	0.11	14.74	10.63	8.55	0.17
ZeroVO+	**6.81**	**2.69**	**104.69**	**0.06**	**9.74**	4.37	**6.03**	**0.12**	**4.64**	2.83	**3.05**	**0.09**	**13.42**	**7.99**	**8.24**	**0.17**
LiteZeroVO+	8.85	2.90	118.54	0.08	11.57	4.44	6.87	0.13	7.65	3.82	5.28	0.11	15.93	12.16	11.26	0.18
TartanVO [70]	13.85	3.27	103.07	-	10.27	6.35	6.26	-	11.17	5.30	7.03	-	10.56	9.35	3.82	-
DPVO [63]	8.31	2.37	78.53	-	4.34	2.85	2.66	-	2.66	1.25	1.59	-	12.65	10.67	4.33	-

Table 2. **Ablation Analysis for Model and Training Components.** We analyze various model components: Flow module (**F**), Depth module (**D**), Language prior (**L**), Semi-supervised training (**S**), and Pseudo-label Selection (**P**). Flow, depth, and language correspond to the proposed supervised ZeroVO model. Results with additional semi-supervised training are shown as ZeroVO+ (showing state-of-the-art performance by integrating all of our proposed components).

F	D	L	S	P	KITTI 00-10				nuScenes				Argoverse				GTA			
					t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}	t_{err}	r_{err}	ATE	s_{err}
✓					18.76	5.49	174.24	0.18	19.40	7.42	12.54	0.22	12.23	6.34	9.42	0.20	25.68	15.52	25.38	0.25
✓	✓				8.99	2.92	123.42	0.08	12.26	5.23	8.40	0.15	8.62	4.11	5.71	0.11	16.76	12.75	12.37	0.19
✓	✓	✓			7.69	2.72	105.07	0.07	10.98	4.48	6.79	0.14	6.83	3.13	4.10	0.11	14.74	10.63	8.55	0.17
✓	✓	✓	✓		9.11	2.88	117.49	0.08	12.25	5.39	7.53	0.14	7.98	3.95	5.13	0.11	16.49	11.95	10.27	0.18
✓	✓	✓	✓	✓	**6.81**	**2.69**	**104.69**	**0.06**	**9.74**	**4.37**	**6.03**	**0.12**	**4.64**	**2.83**	**3.05**	**0.09**	**13.42**	**7.99**	**8.24**	**0.17**

4.2. Results

Generalization Performance: To examine the generalization ability of our model, we evaluate it on entire sequences on KITTI, the unseen regions in nuScenes, and the simulated dataset GTA. Table 1 compares ZeroVO+ with prior baselines in a zero-shot setting. For a fair comparison of the zero-shot performance, all models are provided with the same estimated camera intrinsics and metric depth (if required). TartanVO and DPVO can only estimate rotation and require scale alignment with ground-truth translation to reconstruct the trajectory at a real-world scale. From the results in Table 1, our model achieves superior performance across nearly all metrics on the four datasets. It is important to note that sequences on KITTI are significantly longer compared to those in other datasets, making them more prone to accumulating large drift (i.e., high ATE). Our method accurately predicts rotation and translation scale on KITTI, resulting in the lowest ATE among all baselines, even without incorporating multi-frame temporal optimization. The results on the GTA dataset further demonstrate the strong generalization capability of our model, achieving ATE results comparable to scale-aligned DPVO, which leverages privileged evaluation. In Table 3, we divide the remaining regions in nuScenes into different subsets based on various weather conditions: day, night, rain, and light. The strong light scenario is caused by severe light reflections. We find that night and strong light conditions present the most challenging scenarios, as it is difficult for the model to detect and extract valuable information. We demonstrate that our model achieves the best performance across all conditions, highlighting its robustness against external noise.

Ablation Study: In Table 2, we study the roles of each module in our model structure. We begin by analyzing the impact of our depth module. When the model is equipped with only the flow module, the model struggles to generalize to unseen scenarios, particularly in terms of scale estimation. This outcome is expected, as predicting scale from a single image without any additional context is an ill-posed problem. By incorporating the depth module, the model demonstrates improvements across all metrics, particularly in scale estimation. This improvement indicates that by concatenating the estimated metric depth with the intrinsic image, the model can effectively learn coherent

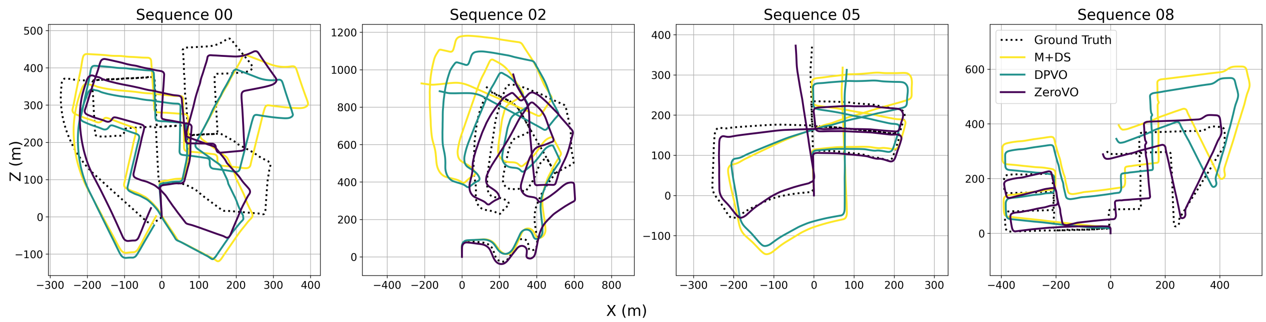

Figure 2. **Qualitative Results on KITTI.** We show trajectory prediction results across the four most complex driving sequences (00, 02, 05, and 08) from the KITTI dataset. Each subplot illustrates the trajectories generated by our proposed model and the baseline models alongside the ground truth trajectory. The qualitative results demonstrate that our approach achieves the highest alignment with the ground truth, particularly in challenging turns and extended straight paths. These findings highlight the robustness of our method in handling complex and diverse driving scenarios.

3D spatial information, even in the presence of noise, and accurately estimate scale. It is also noteworthy that the depth module improves rotation estimation performance. This demonstrates that leveraging both depth and optical flow to unproject 3D scene flow provides crucial 3D correspondence information that leads to improved rotation estimation. The experiment with textual information further demonstrates the model's robustness against noise. Under challenging driving conditions, such as numerous dynamic objects, darkness, strong light reflections, rain, and fog, the estimated camera intrinsics and metric depth are highly susceptible to becoming unreliable. The general text description is able to provide extra 3D information, such as object layouts and movements, which helps the model maintain robustness in highly noisy environments. At last, we demonstrate the effectiveness of our semi-supervision approach using pseudo-label selection. Without pseudo-label selection, we observe a drop in the model's performance compared to the supervised trained model. This decline is due to the introduction of excessive pseudo-labeled examples with redundancy and uncertain label quality, which hinders model training. Our pseudo-label selection process effectively filters out highly redundant and low-quality pseudo-labeled examples, achieving the best performance among all zero-shot metric-scale models. Further ablations and analysis can be found in our supplementary.

Qualitative Analysis: Fig. 2 depicts the most complex and longest trajectories on KITTI, compared with the two best-performing baselines. The trajectory of DPVO is aligned with the ground-truth translation after scale adjustment. Therefore, it is straightforward to see how inaccurate rotation estimation results in drift accumulation. A comparison between the results of DPVO and M+DS reveals how inaccuracies in translation estimation further exacerbate drift accumulation. By leveraging general textual information and unprojecting 2D data into 3D space, our model effectively extracts more accurate and inherent correspondence features, which enhance robustness even when the estimated depth or camera intrinsics are noisy.

Table 3. **Condition Breakdown on nuScenes.** We show results breakdown (ATE) over scenes categorized by weather and lens settings. We sample from nuScenes the Day, Night, and Rainy scenes, along with particularly challenging frames that include severe light reflections. Our ZeroVO+ model performs best overall. We note that TartanVO and DPVO baselines only predict up-to-scale motion and use ground-truth scale alignment in inference.

Method	Day	Night	Rainy	Light
XVO [37]	6.61	14.41	15.99	15.73
M+DS [27]	6.08	17.19	17.49	18.54
ZeroVO	3.90	10.33	12.63	13.33
ZeroVO+	**3.60**	**10.26**	**10.10**	**11.15**

5. Conclusion

We introduced ZeroVO, a novel transformer-based framework designed to tackle the challenge of visual odometry generalization under adverse and unseen conditions. ZeroVO integrates rich multimodal cues—spanning geometry, language, and vision—within a unified architecture to enhance robustness and adaptability in complex environments. Its camera-agnostic design, combined with a semi-supervised training paradigm, enables effective handling of noisy data and seamless adaptation to novel scenarios. Extensive evaluation across diverse and challenging benchmarks demonstrates that ZeroVO establishes a new standard for zero-shot VO performance, underscoring its promise for real-world deployment without the need for camera recalibration or domain-specific tuning.

6. Acknowledgments

We thank the Red Hat Collaboratory (awards 2024-01-RH02, 2024-01-RH07) and National Science Foundation (IIS-2152077) for supporting this research.

References

[1] Aditya Agarwal, Daniel Maturana, and Sebastian Scherer. Visual odometry in smoke occluded environments. *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-15-07*, 2014. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, 2015. 2

[3] Ali Azarbayejani and Alex P Pentland. Recursive estimation of motion, structure, and focal length. *PAMI*, 1995. 1

[4] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 2

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019. 2

[6] Dominique Brunet, Edward R Vrscay, and Zhou Wang. On the mathematical properties of the structural similarity index. *T-IP*, 2011. 5

[7] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *T-RO*, 2016. 1, 2

[8] 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. 6

[9] Benjamin Caine, Rebecca Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Zhifeng Chen, and Jonathon Shlens. Pseudo-labeling for scalable 3d object detection. In *arXiv preprint arXiv:2103.02093*, 2021. 2, 5

[10] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *T-RO*, 2021. 1

[11] Andrea Ceccarelli and Francesco Secci. Rgb cameras failures and their effects in autonomous driving applications. *T-DSC*, 2022. 1

[12] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*, 2023. 2

[13] Alessandro Chiuso, Paolo Favaro, Hailin Jin, and Stefano Soatto. Structure from motion causally integrated over time. *PAMI*, 2002. 1

[14] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*, 2018. 2

[15] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *NeurIPS*, 30, 2017. 2

[16] Ernst Dieter Dickmanns. *Dynamic vision for perception and control of motion*. Springer, 2007. 1

[17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017. 6

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 4

[19] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. In *PAMI*, 2017. 1

[20] Friedrich Fraundorfer and Davide Scaramuzza. Visual odometry: Part i: The first 30 years and fundamentals. *RAM*, 2011. 1

[21] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *CVPR*, 2019. 2

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6

[23] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *CVPR*, 2023. 1

[24] Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*, 2019. 2

[25] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020. 2

[26] Noriaki Hirose, Fei Xia, Roberto Martín-Martín, Amir Sadeghian, and Silvio Savarese. Deep visual mpc-policy learning for navigation. *RA-L*, 4(4), 2019. 2

[27] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *PAMI*, 2024. 1, 2, 4, 5, 6, 7, 8

[28] Young Kyun Jang and Nam Ik Cho. Generalized product quantization network for semi-supervised image retrieval. In *CVPR*, 2020. 2

[29] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 5

[30] Takayuki Kanai, Igor Vasiljevic, Vitor Guizilini, and Kazuhiro Shintani. Self-supervised geometry-guided initialization for robust monocular visual odometry. *arXiv preprint arXiv:2406.00929*, 2024. 2

[31] Nimet Kaygusuz, Oscar Mendez, and Richard Bowden. Mdn-vo: Estimating visual odometry with confidence. In *IROS*, 2021. 1

[32] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 1

[33] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1

[34] Hee Jae Kim and Eshed Ohn-Bar. Motion diversification networks. In *CVPR*, 2024. 2

[35] Pyojin Kim, Hyon Lim, and H Jin Kim. Robust visual odometry to irregular illumination changes with rgb-d camera. In *IROS*, 2015. 1

[36] Andrew V Knyazev and Merico E Argentati. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 2002. 5

[37] Lei Lai, Zhongkai Shangguan, Jimuyang Zhang, and Eshed Ohn-Bar. XVO: Generalized visual odometry via cross-modal self-training. In *ICCV*, 2023. 1, 2, 4, 5, 6, 7, 8

[38] Lei Lai, Eshed Ohn-Bar, Sanjay Arora, and John Seon Keun Yi. Uncertainty-guided never-ending learning to drive. In *CVPR*, 2024. 2

[39] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 5

[40] Qimai Li, Xiao-Ming Wu, Han Liu, Xiaotong Zhang, and Zhichao Guan. Label efficient semi-supervised learning via graph filtering. In *CVPR*, 2019. 5

[41] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *ICRA*, 2018. 5

[42] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *ECCV*, 2024. 1, 2

[43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1, 2

[44] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1, 2

[46] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023. 2

[47] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018. 5

[48] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*. 2000. 4

[49] Nico Messikommer, Giovanni Cioffi, Mathias Gehrig, and Davide Scaramuzza. Reinforcement learning meets visual odometry. *ECCV*, 2024. 2

[50] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. In *NeurIPS*, 2020. 4

[51] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *T-RO*, 2017. 1, 3

[52] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *T-RO*, 2015. 1

[53] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 1, 2

[54] Alberto Pretto, Emanuele Menegatti, Maren Bennewitz, Wolfram Burgard, and Enrico Pagello. A visual odometry framework robust to motion blur. In *ICRA*, 2009. 1

[55] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*. 1, 4, 5

[56] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 5

[57] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *3DV*. IEEE, 2022. 1

[58] Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. Rethinking semi-supervised learning with language models. *arXiv preprint arXiv:2305.13002*, 2023. 2

[59] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017. 5

[60] Chengzhou Tang and Ping Tan. BA-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018. 3

[61] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 2021. 2

[62] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. *NeurIPS*, 2021. 1, 2, 3, 6

[63] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *NeurIPS*. 1, 2, 3, 6, 7

[64] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv preprint arXiv:2407.00959*, 2024. 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 4

[66] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Learning of structure and motion from video. In *CVPR*, 2017. 2

[67] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3DIoUMatch: Leveraging IoU prediction for semi-supervised 3D object detection. In *CVPR*, 2021. 5

[68] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *ICRA*, 2017. 1

[69] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020. 1

[70] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. In *CoRL*, 2021. 1, 2, 3, 4, 6, 7

[71] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 4

[72] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting, 2023. 6

[73] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 5

[74] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022. 2

[75] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *CVPR*, 2024. 5

[76] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[77] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *CVPR*, 2020. 1

[78] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021. 5

[79] Weicai Ye, Xinyue Lan, Shuo Chen, Yuhang Ming, Xingyuan Yu, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Pvo: Panoptic visual odometry. In *CVPR*, 2023. 2

[80] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *CVPR*, 2023. 1

[81] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2

[82] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *CVPR*, 2024. 2

[83] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018. 5

[84] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, Ravi Garg, and Ian Reid. Df-vo: What should be learnt for visual odometry? *arXiv preprint arXiv:2103.00933*, 2021. 1

[85] Hui Zhang, K Wong Kwan-yee, and Guoqiang Zhang. Camera calibration from images of spheres. *PAMI*, 2007. 2

[86] Jimuyang Zhang, Ruizhao Zhu, and Eshed Ohn-Bar. Selfd: Self-learning large-scale driving policies from the web. In *CVPR*, 2022. 2

[87] Jimuyang Zhang, Zanming Huang, Arijit Ray, and Eshed Ohn-Bar. Feedback-guided autonomous driving. In *CVPR*, 2024. 2

[88] Yueqiang Zhang, Langming Zhou, Haibo Liu, and Yang Shang. A flexible online camera calibration using line segments. *Journal of Sensors*, 2016. 2

[89] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 2000.

[90] Zhengyou Zhang. Camera calibration with one-dimensional objects. *PAMI*, 2004. 2

[91] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *CVPR*, 2020. 4, 5

[92] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 5

[93] Ruizhao Zhu, Peng Huang, Eshed Ohn-Bar, and Venkatesh Saligrama. Learning to drive anywhere. In *CoRL*, 2023. 2

[94] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: in-the-wild monocular camera calibration. *NeurIPS*, 2024. 2, 4, 5

[95] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *3DV*, 2024. 2