XVO: Generalized Visual Odometry via Cross-Modal Self-Training

Lei Lai* Zhongkai Shangguan* Jimuyang Zhang Eshed Ohn-Bar
Boston University
{leilai, sgzk, zhangjim, eohnbar}@bu.edu

Abstract

We propose XVO, a semi-supervised learning method for training generalized monocular Visual Odometry (VO) models with robust off-the-self operation across diverse datasets and settings. In contrast to standard monocular VO approaches which often study a known calibration within a single dataset, XVO efficiently learns to recover relative pose with real-world scale from visual scene semantics, i.e., without relying on any known camera parameters. We optimize the motion estimation model via self-training from large amounts of unconstrained and heterogeneous dash camera videos available on YouTube. Our key contribution is twofold. First, we empirically demonstrate the benefits of semi-supervised training for learning a general-purpose direct VO regression network. Second, we demonstrate multi-modal supervision, including segmentation, flow, depth, and audio auxiliary prediction tasks, to facilitate generalized representations for the VO task. Specifically, we find audio prediction task to significantly enhance the semi-supervised learning process while alleviating noisy pseudo-labels, particularly in highly dynamic and out-of-domain video data. Our proposed teacher network achieves state-of-the-art performance on the commonly used KITTI benchmark despite no multi-frame optimization or knowledge of camera parameters. Combined with the proposed semi-supervised step, XVO demonstrates off-the-shelf knowledge transfer across diverse conditions on KITTI, nuScenes, and Argoverse without fine-tuning.

1. Introduction

Monocular Visual Odometry (VO) methods for recovering ego-motion from a sequence of images have mostly been studied within a *restricted scope*, where a single dataset, such as KITTI [28], may be used for both training and evaluation under a fixed pre-calibrated camera [37, 45, 54, 77, 109, 112, 116, 124, 126, 128]. However, very few studies have analyzed the task of *generalized VO*, i.e., relative pose estimation with real-world scale across differing scenes and capture setups.

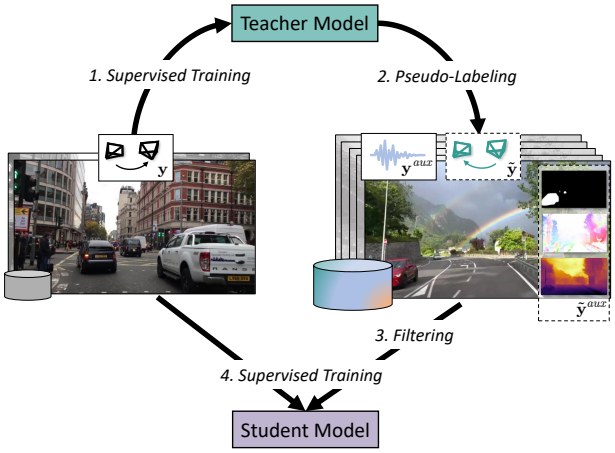

Figure 1: **Learning General-Purpose Monocular Visual Odometry (VO) Models from Multi-Modal and Pseudo-Labeled Videos.** Our proposed XVO framework first trains an ego-motion prediction *teacher model* over a small initial dataset, e.g., nuScenes [7]. We then expand the original dataset through pseudo-labeling of in-the-wild videos. Motivated by how humans learn general representations through observation of large amounts of multi-modal data, we employ multiple auxiliary prediction tasks, including segmentation, flow, depth, and audio, as part of the semi-supervised training process. Finally, we leverage *uncertainty-based filtering* of potentially noisy pseudo-labels and train a robust student model.

For instance, consider an autonomous robot or vehicle deployed at a large scale. The robot is highly likely to encounter environments for which no ground truth ego-motion data was previously collected. In such novel settings, current VO methods will quickly exhibit poor ego-motion estimation and drift [23, 24, 29, 46, 47, 65, 96, 128]. Moreover, our robot may be required to adjust its camera setup over its lifetime (e.g., to a new camera) or leverage data from a fleet of robots with varying or perhaps unknown camera configurations. Yet, existing VO methods generally assume care-

* Equally contributed.

fully calibrated camera parameters during training [23, 24, 65, 77, 96, 109, 128]. Specifically, to simplify the ill-posed monocular pose recovery task, researchers often resort to relying on knowledge of the camera intrinsic parameters to incorporate various geometric or photometric consistency-based mechanisms [37, 45, 54, 89, 112, 116, 124, 126]. In this work, we do not make such an assumption as we are concerned training VO models that can learn from and operate under diverse unconstrained videos in the wild. Specifically, we pursue an orthogonal direction to prior work based on semi-supervised learning and explore more scalable and camera-agnostic deployment settings.

Our key hypothesis is that neural network models can learn to circumvent issues related to pose and scale ambiguity in generalized VO settings through observation of ample amounts of diverse scene and motion video data. This approach is motivated by humans' ability to flexibly estimate motion in arbitrary conditions through a general understanding of salient scene properties (e.g., object sizes) [73]. This general understanding is developed over large amounts of perceptual data, often multi-modal in nature [71, 81, 81]. For instance, cross-modal information processing between audio and video has been shown to play a role in spatial reasoning and proprioception [57, 58, 67, 74, 99]. Indeed, collected online videos often have audio, which can be used as a further source of cross-modal supervision. As further discussed in Sec. 3, we find ambient audio to correlate at times with scenarios where monocular VO tends to fail, such as determining ego-speed when the vehicle is stopped at a dense intersection or as context for the current traffic scenario when estimating translation (e.g., highway driving). Extracting information related to flow, segmentation, or depth can also further guide learning generalized representations. To fully explore the utility of self-training VO models, we analyze a unified multi-modal framework and its impact on guiding semi-supervised VO learning from large amounts of unconstrained sources.

As far as we are aware, we are the first to study the feasibility of self-training for direct, calibration-free, ego-motion pose regression with an absolute real-world scale. Specifically, we find that incorporating additional modalities via simple multi-task learning can significantly enhance model robustness and generalization. When paired with an uncertainty-based filtering module, we achieve state-of-the-art VO performance with a single broadly usable model which we validate for the autonomous driving use-case. Moreover, our training and inference is highly efficient as the auxiliary learning formulation does not alter the two-frame input, i.e., in contrast to methods relying on extracting rich intermediate representations [4, 96, 112, 124]. We demonstrate state-of-the-art results on KITTI using the proposed two-frame VO model structure without requiring elaborate long-term memory, computationally expensive iterative refinement steps, or knowledge of camera parameters. Our code is available at https://github.com/h2xlab/XVO.

2. Related Work

Monocular Visual Odometry: Despite recent advances, both geometrical and learning-based VO approaches are still mostly evaluated over limited datasets under similar training and testing conditions [9, 25, 40, 64, 72, 90, 92, 94]. For instance, training and evaluation are both conducted on KITTI [28], which contains a fixed camera setup with limited diversity and density of scenarios [96]. More recently, learning-based approaches leveraging unsupervised learning for VO [45, 47, 75, 112, 115, 126], have shown state-of-the-art performance on KITTI. Notably, UnDeepVO [45] utilizes stereo imagery for training to recover real-world scale without the need for labeled datasets. GeoNet [112] combines depth, optical flow, and camera pose to holistically learn a VO prediction model. TartanVO [96] conditions the VO model on the intrinsic parameters in order to achieve robust generalized performance. However, the aforementioned methods all require known camera intrinsics [45, 112] in inference resulting in a restricted use-case and cannot leverage data with unknown camera parameters. In light of these challenges, our work develops mechanisms to enable VO models to learn from and operate over diverse datasets without known calibration. Specifically, our method leverages semi-supervised and multi-modal learning techniques to learn robust generalized representations for estimating motion and real-world scale. Therefore, our approach is orthogonal to most related methods which emphasize self- and un-supervised learning of models based on warping and consistency tasks which rely on precise camera calibration [37, 45, 112].

Semi-Supervision for Computer Vision Tasks: Semi-supervised learning approaches have been extensively studied within the computer vision and machine learning communities [18, 44, 55, 82, 87, 111]. However, prior works have not yet focused on the monocular VO task, instead emphasizing object detection [8], semantic segmentation [98], 3D reconstruction [102], action recognition [101] or low-level computer vision tasks [6, 10, 17, 33, 38, 59, 61, 70, 84, 107, 110]. Consequently, fundamental research questions related to the impacts of improving VO model generalization remain unanswered, e.g., whether semi-supervised learning can be used to enhance reasoning over real-world scale [43, 76, 97, 120] and long-tail scenarios [119] or how uncertainty mechanisms can contribute to more robust training from heterogeneous video data. Specifically, we explore the role of multi-task and multi-modal learning in order to improve semi-supervised VO model training.

Auxiliary Learning: Our proposed method primarily aims

to enhance the performance of a VO model through auxiliary tasks within a semi-supervised learning framework. Auxiliary learning [49, 51, 122] aims to use auxiliary tasks to enhance the performance of the primary tasks. It has been effectively employed in diverse domains, including computer vision [15, 41, 60, 104], natural language processing [16, 19, 42], and robotics [36, 66, 68, 83, 93, 117, 118]. Xu et al. [104] applied auxiliary image classification and saliency detection to improve the performance of the semantic segmentation. Song et al. [83] leverages an auxiliary task of velocity estimation to enhance the ability to avoid obstacles of an indoor mobile robot. While several related studies employ auxiliary supervision derived from the ground-truth depth and optical flow [88, 89], in this work our goal is to explore the use of such supervision from potentially noisy pseudo-labels, e.g., as regularization for learning robust internal representations for VO.

Cross-Modal Learning: Cross-modal learning is inspired by how biological systems learn by incorporating complementary information from multiple modalities, such as vision, sound, and touch. Prior research in computational cross-modal learning has focused on learning a shared representation space where samples from distinct modalities i.e., image, audio, text, can be aligned [3, 34, 125]. Moreover, the addition multiple tasks and modalities have been shown to benefit generalization for various machine perception and learning tasks [2, 34, 78, 113, 114, 114]. For instance, audio generation [21, 86, 127], image captioning [48, 103], speech recognition [1, 80], navigation [13], and multimedia retrieval [11, 27, 35] have all shown improved performance due to cross-modal training. However, such studies tend to focus on simplified domains, e.g., restricted acoustic or haptic environments, whereas we analyze dense and dynamic scenes in the wild.

3. Method

Our proposed framework comprises three main steps: (1) uncertainty-aware training of an initial (i.e., teacher) VO model (Sec. 3.2); (2) pseudo-labeling with the removal of low-confidence and potentially noisy samples (Sec. 3.3); (3) self-training with pseudo-labeled and auxiliary prediction tasks of a robust VO student model (Sec. 3.4).

3.1. Problem Setting

Direct Pose Regression: Our goal is to learn a general function for mapping two observed image frames $\mathbf{x}_i = \{\mathbf{I}_{i-1}, \mathbf{I}_i\}$, with $\mathbf{I} \in \mathcal{R}^{W \times H \times 3}$, to a relative camera pose with real-world scale $\mathbf{y}_i = [\mathbf{R}_i|\mathbf{t}_i] \in \text{SE}(3)$ with rotation $\mathbf{R}_i \in \text{SO}(3)$ and translation $\mathbf{t}_i \in \mathbb{R}^3$. Given a dataset comprising annotated labels of pose ground-truth, $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, learning-based approaches for VO often optimize for a regression loss [5, 75, 89, 94, 112]. In practice, the direct pose regression task often exhibits drift due to issues with absolute scale ambiguity and compound errors, particularly in cases of dense and dynamic scenes. For instance, small errors in rotation estimation can result in large errors over multiple time steps which impact the evaluation. While we formulate a two-frame regression task, prior methods have relied on longer-term memory in order to improve model robustness [37, 47, 94, 106], however, this comes at a computational and memory cost. Moreover, most monocular methods only produce up-to-scale predictions [46, 96, 112], as will be further discussed in Sec 4. Instead, we rely on a semi-supervised training process to mitigate issues in absolute scale recovery while enabling a simple two-frame model to achieve state-of-the-art results.

Self-Training with Auxiliary Tasks: In addition to a labeled odometry dataset $\mathcal{D}_L$, our framework assumes access to a large dataset that is not annotated with respect to the ego-motion task but potentially other complementary tasks that are auxiliary to the main VO task, i.e., $\mathcal{D}_U = \{(\mathbf{x}_i, \mathbf{y}_i^{aux})\}_{i=1}^M$. Moreover, we assume access to a set of models for generating a *pseudo-labeled* dataset [8, 43, 76, 107, 120], i.e., $\mathcal{D}_{PL} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i, \mathbf{y}_i^{aux}, \tilde{\mathbf{y}}_i^{aux})\}_{i=1}^M$ which can be joined with the original dataset $\hat{\mathcal{D}} = \mathcal{D}_L \cup \mathcal{D}_{PL}$ for supervised training (Sec. 3.4). We note that this is a practical assumption as there are abundant computer vision models for obtaining various pseudo-labels. As will be discussed in Sec. 3.3, these pseudo-labels may be filtered by removing high-uncertainty samples. Overall, the cross-modal self-training objective can be defined as

$$\mathcal{L}_{xvo} = \mathcal{L}_{vo}(\mathbf{y}) + \lambda_u \mathcal{L}_{unc}(\mathbf{y}) + \mathcal{L}_{aux}(\mathbf{y}^{aux}, \tilde{\mathbf{y}}^{aux}) \quad (1)$$

where $\mathcal{L}_{vo}$ is a main VO task loss, $\mathcal{L}_{unc}$ is an uncertainty estimation loss, $\mathcal{L}_{aux}$ is defined over the auxiliary prediction tasks, and λ_u is a scalar hyper-parameter. We demonstrate our semi-supervised formulation to benefit various known issues with VO, e.g., improving scale recovery. Moreover, our formulation is kept efficient during inference as it does not alter the two-frame input $\mathbf{x}$, i.e., in contrast to methods relying on extracting intermediate representations as input, such as flow [96] or depth [124]. Next, we define our network structure and training.

3.2. Ego-Motion Network Training

Our approach first trains a direct ego-motion teacher model (shown as the main encoder and middle branch in Fig. 1) over the labeled dataset $\mathcal{D}_L$. To enable learning from an unconstrained video, we do not incorporate any dependency on intrinsic parameters, i.e., either as input [23, 24, 96] or for computing a supervisory loss [45, 77, 89, 109, 112]. We find our network design to provide an efficient but surprisingly strong baseline, matching state-of-the-art on the KITTI benchmark despite no elaborate multi-frame optimization.

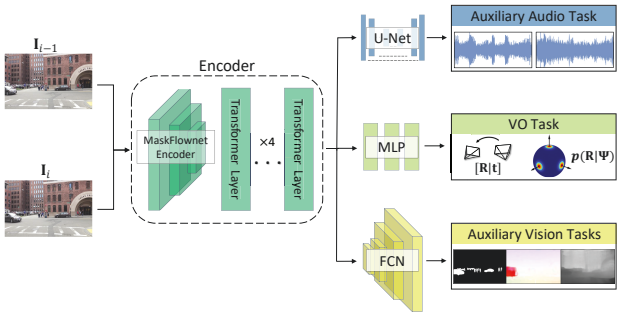

Figure 2: **Network Architecture.** Our initial teacher model (used for pseudo-labeling and filtering) encodes two concatenated image frames and predicts relative camera pose and its uncertainty. The complete cross-modal architecture leverages a similar architecture but with added auxiliary prediction branches with complementary tasks that can further guide self-training, e.g., prediction branches for audio reconstruction, dynamic object segmentation, optical flow, and depth.

Encoder: We employ a high-capacity feature extractor for effectively leveraging the rich multi-task supervision in later stages (Sec. 3.4). The feature extractor is a MaskFlownet encoder [123], which was found to outperform the commonly used PWC-Net [85, 96], followed by four transformer self-attention layers [14, 22]. The patch size is 12×16, with each layer comprising four heads and 256 hidden parameters. The encoder structure for the initial teacher model and cross-modal student is kept the same.

VO Decoder: The VO decoder branch consists of three Fully Connected (FC) layers that regress relative pose $\mathbf{y} = [\mathbf{R}|\mathbf{t}]$ and an uncertainty estimate for the prediction. The VO task optimizes a Mean Squared Error (MSE) loss over predicted translation $\hat{\mathbf{t}} \in \mathcal{R}^3$ and Euler angle rotations $\hat{\boldsymbol{\theta}} \in \mathcal{R}^3$,

$$\mathcal{L}_{vo} = \|\mathbf{t} - \hat{\mathbf{t}}\|_2^2 + \lambda_\theta \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 \tag{2}$$

Uncertainty Estimation: To account for the difficulty in the absolute scale pose regression task, we propose to also model prediction uncertainty. We adopt a matrix Fisher distribution [62], which provides a framework for modeling rotation distribution on SO(3). The probability density function of the matrix Fisher distribution is as follows:

$$p(\mathbf{R}|\boldsymbol{\Psi}) = \frac{1}{c(\boldsymbol{\Psi})} \exp(tr(\boldsymbol{\Psi}^\top \mathbf{R})) \tag{3}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{3\times3}$ are the distribution parameters, $\mathbf{R} \in$ SO(3) is the pose rotation matrix, and $c(\boldsymbol{\Psi})$ is a normalization constant [56]. Given the estimated parameters $\hat{\boldsymbol{\Psi}}$ we use the negative log likelihood of $\mathbf{R}$ in the predicted distribution as a loss, i.e.,

$$\mathcal{L}_{unc} = -\log(p(\mathbf{R}|\hat{\boldsymbol{\Psi}})) \tag{4}$$

As a proxy for prediction (i.e., pseudo-label) quality, we find it is sufficient to model uncertainty in rotation prediction, however more elaborate estimation methods can also be used [39, 76, 108]. The confidence predictions will be used to remove potentially noisy pseudo-labels prior to the self-training process, as discussed next.

3.3. Pseudo-Label Selection

The VO model from Sec. 3.2 can be used to obtain pseudo-labels over an unlabeled (i.e., with respect to the main VO task) data $\mathcal{D}_U$. However, incorrect predictions can introduce noise and heavily degrade model training [76, 82]. Hence, it is crucial to remove low-confidence samples prior to the cross-modal self-training.

In our regression problem, we measure the confidence of a pseudo-label based on the entropy of the predicted matrix Fisher distribution (i.e., a lower entropy represents increased confidence),

$$H(p(\mathbf{R}|\hat{\boldsymbol{\Psi}})) < \tau_u \tag{5}$$

where we set a fixed threshold τ_u to ensure the network prediction is sufficiently certain to be selected. To generate pseudo-labels, the VO model is tested on out-of-domain data with highly diverse and dynamic scenes. Based on our analysis in Sec. 4, we find the uncertainty-aware selection mechanism to be crucial for robust self-training irrespective of the auxiliary training tasks.

3.4. Self-Training with Auxiliary Tasks

To learn effective representations for generalized VO at scale, we propose to incorporate supervision from auxiliary but potentially complementary prediction tasks in addition to the generated VO pseudo-labels on $\mathcal{D}_U$. The introduced auxiliary tasks regularize the self-training process, particularly in cases where VO pseudo-labels may be inaccurate but other modalities may contain relevant information for reducing ambiguity. Our approach is motivated by the success of multi-task frameworks for computer vision tasks [2, 30, 41, 113, 114, 121]. However, we emphasize that related studies often leverage high-quality annotated labels and not noisy pseudo-labels based on model predictions. We sought to incorporate useful auxiliary tasks, as unrelated or noisy supervision can impede the learning process and result in a detrimental effect on the main task model. We set the auxiliary labeled task as an audio prediction task $\mathbf{y}_i^{aux} := \mathbf{A}_i \in \mathcal{R}^{2\times L}$, and the auxiliary pseudo-labeled tasks $\tilde{\mathbf{y}}_i^{aux} := [\tilde{\mathbf{S}}_i, \tilde{\mathbf{D}}_i, \tilde{\mathbf{F}}_i] \in \mathcal{R}^{W\times H\times C}$ as segmentation, depth, and flow prediction, respectively. Subsequently, we leverage multi-task learning (as shown in Fig. 3)

Figure 3: **Illustration of the Importance of Audio.** The frame is consistent with the red arrow marked on the waveform. Left: audio amplitude decreases and maintains a low level when the vehicle is going to wait for traffic lights. Right: audio experiences many ups and downs representing acceleration and brake in a narrow urban area.

and minimize a loss composed of four terms,

$$\mathcal{L}_{aux} = \lambda_a \mathcal{L}_{audio} + \lambda_s \mathcal{L}_{seg} + \lambda_f \mathcal{L}_{flow} + \lambda_d \mathcal{L}_{depth} \quad (6)$$

over the entire dataset $\tilde{\mathcal{D}}$. We note that we drop the explicit label source to avoid clutter. Next, we define each term and corresponding decoder. We empirically observe the additional tasks to improve generalization in evaluation, both within and across VO datasets.

Audio Decoder: We utilize audio labels, generally available for online videos, as an auxiliary prediction task. We note that prior work often studies such cross-modal reasoning for basic navigation scenarios [12, 13, 26] and not for in-the-wild videos where dense dynamic objects may generate significant ambient noise. In our settings, an audio signal can provide complementary information to visual information regarding the overall traffic scenario as well as ego-speed. This insight will be affirmed by our findings in Sec. 4, where the audio task is shown to provide synergistic supervision, both for the main VO task and when combined with other auxiliary tasks. For instance, Fig. 3 depicts how an idling ego-vehicle may generate lower audio levels, which, in conjunction with the visual scene features, can help disambiguate ego-motion from surrounding motion when stopped at intersections. As drift due to surrounding motion is a common failure mode for VO models, we further incorporate a segmentation task for dynamic objects below.

Our audio decoder is based on a 1D U-Net architecture [79], consisting of a residual 1D convolutional block [32] and an attention block [91], and reconstructs the dual-channel raw audio of the two input frames using the encoder features. We employ a two-term MSE and spectral loss,

$$\mathcal{L}_{audio} = \|\mathbf{A}_i - \hat{\mathbf{A}}_i\|_2^2 + \|\text{FT}(\mathbf{A}_i) - \text{FT}(\hat{\mathbf{A}}_i)\|_2^2 \quad (7)$$

where FT represents the short-time Fourier transform [20].

Segmenting Dynamic Objects: The relative motion caused by dynamic objects can often lead to inaccurate pose predictions, e.g., when stopped at a traffic light with oncoming traffic. To facilitate disambiguating potentially dynamic objects from the static background, we incorporate a segmentation prediction for pedestrian and vehicle classes [4]. As this task involves extensive manual annotation, we intentionally do not assume it is provided as part of the originally labeled dataset $\mathcal{D}_L$ and instead leverage an off-the-shelf model based on Mask R-CNN [31]. The model is pre-trained on the COCO dataset [50]. We use the detector to construct a pseudo-label semantic segmentation $\tilde{\mathbf{S}} \in \mathcal{R}^{W \times H \times 2}$ of foreground and background in the two input frames. We leverage an FCN [52] decoder, consisting of 11 transposed convolutional layers followed by a convolutional layer and a final sigmoid activation function, and minimize a Dice loss,

$$\mathcal{L}_{seg} = 1 - 2\frac{\sum_{j,k,c} \tilde{\mathbf{S}}_i \circ \hat{\mathbf{S}}_i}{\sum_{j,k,c} \tilde{\mathbf{S}}_i^2 + \sum_{j,k,c} \hat{\mathbf{S}}_i^2} \quad (8)$$

where $\hat{\mathbf{S}}_i \in \mathbb{R}^{H \times W \times 2}$ is the decoder predicted segmentation, and $j \in [1, H], k \in [1, W], c \in [1, 2]$. As dynamic objects often cause ego-motion estimation drift, the prediction task can regularize self-training by providing a useful invariant prior (i.e., across datasets and settings) of background and foreground knowledge. Moreover, the segmentation task complements the audio task in many scenarios as dynamic objects may also generate ambient audio.

Depth and Flow Tasks: We explore two additional auxiliary tasks based on depth and optical flow estimation, as they potentially offer valuable information about the structure of the surroundings and the camera motion and are frequently employed in VO tasks [53, 63, 69, 96]. We utilize an MSE as the loss function for both depth and flow tasks,

$$\begin{aligned} \mathcal{L}_{flow} &= \|\tilde{\mathbf{F}}_i - \hat{\mathbf{F}}_i\|_2^2 \\ \mathcal{L}_{depth} &= \|\tilde{\mathbf{D}}_i - \hat{\mathbf{D}}_i\|_2^2 \end{aligned} \quad (9)$$

To simplify the model, we maintain the identical decoder structure used as in the dynamic object segmentation task (see Fig. 2), with the exception of eliminating the final Sigmoid layer.

3.5. Implementation Details

Our models are trained using three NVIDIA RTX 3090 GPUs using a batch size of six. The learning rate is set to 0.001 and with decay 0.99. Given the main VO objective, we set $\lambda_\theta = 1$ and $\lambda_u = 0.1$. Remaining auxiliary loss hyper-parameters, i.e., $\lambda_a, \lambda_s, \lambda_f, \lambda_d$, are set to 0.01. For our semi-supervised training, we obtain a diverse set of 59,000 unlabeled samples across different geographical locations, times of day, and environmental conditions. We split the nuScenes benchmark [7] into training, validation, and evaluation sets, to train an initial teacher model for 15

epochs. The student model is trained for 15 epochs on a mix of labeled nuScenes and pseudo-labeled YouTube data. We note that we do not employ careful ratio optimization [8] when mixing the datasets without and instead solely rely on the uncertainty-based selection mechanism. We leverage data augmentation strategies, including random cropping and resizing, for improving generalization and simulating varying camera intrinsics [96]. During inference, the model runs at 77 FPS on a single NVIDIA GTX 3090 GPU. Additional details regarding the training and experiments can be found in the supplementary.

4. Experiments

In this section, we comprehensively analyze our XVO framework. As our goal is to build generalized VO systems, we emphasize generalization ability across different datasets with various camera setups, specifically in the context of varying autonomous driving settings.

Datasets: To understand the role of cross-modal self-training on model generalization, we evaluate our proposed XVO method using three commonly employed datasets, KITTI [28], nuScenes [7] and Argoverse 2 [100]. Out of the three, KITTI is the most popular VO benchmark, consisting of 11 sequences 00-10 with ground truth. As KITTI is an older benchmark (2012), its camera intrinsics vary significantly from the other two benchmarks. nuScenes consists of about 15 hours of driving data (totaling 197,000 images) from four regions in Boston and Singapore: Boston-Seaport, Singapore-OneNorth, Singapore-Queenstown, and Singapore-HollandVillage. In contrast to KITTI which was captured in sunny driving with mostly static objects, nuScenes incorporates complex real-world driving maneuvers in dense streets and various conditions, e.g., nighttime, difficult illumination conditions with low visibility, as well as artifacts on the camera lens, such as rain droplets or dirt. Finally, Argoverse 2 is a large dataset with 1,000 driving sequences across six US cities. We leverage a test dataset that includes 150 sequences and 48,022 images.

Procedure and Baselines: We generally train within one region on nuScenes (HollandVillage) and evaluate the remaining regions and datasets. This is in contrast to prior evaluation procedures where models can learn to memorize the scale and camera setup without generalization through training and testing on the same camera setup and similar environments. We also directly compare with prior state-of-the-art using the standard KITTI protocol [45, 112]. As our approach does not leverage known intrinsics, we separate approaches that do assume such knowledge in their pipeline to ensure meaningful analysis. We further indicate whether methods predict pose with absolute scale, as some methods output up-to-scale estimates and use the ground-truth scale to align and evaluate their model, e.g., TartanVO [96]. Nonetheless, TartanVO is one of the few approaches that have shown generalization across datasets without the need to fine-tune or perform online adaptation strategies and is therefore our main baseline.

Metrics: We follow standard evaluation metrics of average translation t_{rel} (in %) and rotation r_{rel} (in degrees per 100 meters) errors, computed over all possible subsequences within a test sequence of lengths $(100, ..., 800)$ meters [28, 96]. We refer the reader to the KITTI leaderboard for more details regarding the metric. However, we observe prior measures to only provide a proxy evaluation of real-world scale predictions as the errors could potentially vary along the trajectory independently of trajectory-level measures (our supplementary contains additional details). To explore the benefits of semi-supervised training on real-world scale estimation, we sought to directly quantify scale recovery within consecutive frames in a single metric. We therefore also report the *average scale error (se)* over predicted and ground-truth translation,

$$se = 1 - \min\left(\frac{\|\hat{\mathbf{t}}\|_2}{\max(\|\mathbf{t}\|_2, \epsilon)}, \frac{\|\mathbf{t}\|_2}{\max(\|\hat{\mathbf{t}}\|_2, \epsilon)}\right) \quad (10)$$

where ϵ prevents dividing by zero.

4.1. Results

We examine the role of the main components in the proposed framework below. Complete ablation, e.g., across modality combinations and training settings, can be found in the supplementary.

Teacher Model Performance: Table 1 compares our proposed encoder architecture for the main VO task with prior methods. When trained in a supervised learning manner on KITTI, our teacher model achieves the lowest translation error of 3.4% even without access to camera intrinsics or multi-step optimization. This suggests that basic modifications to underlying network structure, e.g., through an improved encoder and attention-based mechanism, can result in significant gains for the monocular VO task. Given the effective network structure, we now turn to analyzing the benefits of the proposed semi-supervised framework.

Semi-Supervised VO Training: Table 1 also analyzes the generalization performance of the proposed semi-supervised learning framework on KITTI. Specifically, we show our initial teacher model that is trained on the nuScenes (HollandVillage) dataset to not generalize well to the KITTI testing set (25.27% translation and 8.17° rotation error) due to domain shift and differing camera settings. However, after semi-supervised training, the errors for the student model are reduced by 40% and 50% in translation and rotation errors, respectively. The best self-trained model with auxiliary tasks (complete ablation can be found

Table 1: **Analysis on the KITTI Benchmark.** We abbreviate 'intrinsics-free' as I (i.e., a method which does not assume the intrinsics) and 'real-world scale' as S (i.e., a method is able to recover real-world scale). To ensure meaningful comparison, we categorize models based on supervision type. Firstly, we present unsupervised learning methods, followed by supervised learning methods, then generalized VO methods, and finally our XVO ablation. In the case of TartanVO, we analyze robustness to noise applied to the intrinsics. We train two teacher models: one based on KITTI (as shown in supervised learning approaches) and the other on nuScenes (as displayed at the end of the Table with ablations).

Method	I	S	**Seq 03**		**Seq 04**		**Seq 05**		**Seq 06**		**Seq 07**		**Seq 10**		**Avg**	
			t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
Unsupervised Methods:																
SfMLearner [126]	✗	✗	10.78	3.92	4.49	5.24	18.67	4.10	25.88	4.80	21.33	6.65	14.33	3.30	15.91	4.67
GeoNet [112]	✗	✗	19.21	9.78	9.09	7.55	20.12	7.67	9.28	4.34	8.27	5.93	20.73	9.10	14.45	7.40
Zhan *et al.* [115]	✗	✓	15.76	10.62	3.14	2.02	4.94	2.34	5.80	2.06	6.49	3.56	12.82	3.40	8.16	4.00
UnDeepVO [45]	✗	✓	5.00	6.17	4.49	2.13	3.40	1.50	6.20	1.98	3.15	2.48	10.63	4.65	5.48	3.15
Supervised Methods:																
DeepVO [94]	✓	✓	8.49	6.89	7.19	4.97	2.62	3.61	5.42	5.82	3.91	4.60	8.11	8.83	5.96	5.79
ESP-VO [95]	✓	✓	6.72	6.46	6.33	6.08	3.35	4.93	7.24	7.29	3.52	5.02	9.77	10.2	6.16	6.66
GFS-VO [105]	✓	✓	5.44	3.32	2.91	1.30	3.27	1.62	8.50	2.74	3.37	2.25	6.32	2.33	4.97	2.26
Xue et al. [106]	✓	✓	**3.32**	2.10	2.96	1.76	2.59	1.25	4.93	1.90	**3.07**	**1.76**	3.94	1.72	3.47	**1.75**
Our Teacher (KITTI)	✓	✓	3.46	**2.00**	**1.67**	**0.70**	**2.12**	**0.92**	**3.92**	**1.46**	5.93	3.96	**3.31**	**1.52**	**3.40**	1.76
Baseline Generalized VO Methods:																
TartanVO (TartanAir) [96]	✗	✗	4.20	2.80	6.19	4.35	5.84	3.24	4.21	2.51	7.11	4.96	8.00	3.21	5.93	3.51
TartanVO (10% Noise)	✗	✗	9.33	3.12	10.88	4.71	11.77	5.39	11.88	4.52	14.70	10.74	11.76	3.61	11.72	5.35
TartanVO (20% Noise)	✗	✗	17.79	4.42	21.58	5.04	20.12	8.54	18.80	6.26	21.34	16.27	17.45	5.03	19.51	7.59
TartanVO (30% Noise)	✗	✗	25.89	7.06	34.91	4.54	22.48	10.17	19.32	5.23	19.40	13.33	25.06	8.43	24.51	8.13
Proposed Generalized VO Methods:																
Our Teacher (nuScenes)	✓	✓	26.78	4.92	26.02	2.42	23.65	8.85	23.97	6.47	30.66	20.32	20.57	6.01	25.27	8.17
Student w/o Filter	✓	✓	26.98	9.68	22.56	2.15	14.77	5.83	**11.38**	**1.62**	16.45	9.35	20.23	8.99	18.73	6.27
Student	✓	✓	20.30	3.97	16.33	1.57	11.12	4.19	15.60	5.69	7.77	3.48	19.91	5.59	15.17	4.08
XVO	✓	✓	**14.53**	**3.93**	**16.29**	**0.96**	**8.31**	**2.76**	15.31	5.49	**5.86**	**3.00**	**12.17**	**3.45**	**12.08**	**3.27**

Table 2: **Average Quantitative Results across Datasets.** We test on KITTI (sequences 00-10), Argoverse 2, and the unseen regions in nuScenes. All results are the average over all scenes. We present translation error, rotation error and scale error. Approaches such as TartanVO do not estimate real-world scale but may be aligned with ground truth (GT) scale in evaluation. A, S, F, D are the abbreviation of Audio, Seg, Flow, Depth.

Method	**KITTI 00-10**			**Argoverse 2**			**nuScenes**		
	t_{err}	r_{err}	se	t_{err}	r_{err}	se	t_{err}	r_{err}	se
Baseline Generalized VO Methods:									
TartanVO w/ GT Align	6.37	3.32	/	8.55	5.77	/	9.61	6.83	/
TartanVO w/o GT Align	21.67	3.33	0.29	41.11	5.77	0.40	28.23	6.83	0.29
Proposed Generalized VO Methods:									
Teacher (nuScenes)	26.16	6.84	0.25	10.89	3.40	0.16	15.93	6.73	0.20
Student w/o Filter	20.64	5.68	0.21	10.80	7.33	0.14	9.32	4.60	0.14
Student	17.04	4.02	0.16	9.16	3.40	0.14	10.54	3.94	0.13
Student+Seg	16.31	3.77	0.16	9.17	**3.18**	0.13	11.35	4.05	0.14
Student+Flow	15.60	3.19	0.19	9.04	4.45	0.13	**9.13**	4.06	**0.13**
Student+Depth	17.49	3.89	0.20	9.25	4.11	0.13	11.86	6.46	0.15
Student+Audio	**14.37**	**3.06**	**0.16**	**8.00**	**3.08**	**0.12**	**9.26**	**3.20**	**0.12**
Student+Audio+Seg	**14.20**	**3.02**	**0.16**	8.67	3.63	0.13	11.29	3.70	0.14
Student+S+F+D	18.23	3.88	0.21	8.79	4.89	0.13	8.93	**3.44**	0.13
Student+A+S+F+D	16.74	4.40	0.18	**7.89**	3.54	**0.12**	9.98	4.36	0.15

in the supplementary) results in further student performance gains, e.g., a further reduction in translation error by 20%.

We also compare with the most related TartanVO [96] baseline which utilizes the ground-truth for scale alignment and has access to camera intrinsics. However, even with the ground-truth alignment, TartanVO exhibits quick degradation with minimal noise in the intrinsics (enabling a more fair comparison as our method is not provided these as input). Moreover, we explore the generalization of our training framework by evaluating on various datasets in Table 2. We emphasize that none of the trained models have access to samples from Argoverse 2 or KITTI dataset during training. By predicting real-world scale, our student model with all auxiliary tasks outperforms the baseline TartanVO in all three datasets, e.g., by 80% in translation and 70% scale error on Argoverse 2, without any ground-truth alignment. This indicates the proposed method to improve reasoning over scale and scene semantics across arbitrary conditions.

Impact of Uncertainty-Aware Sample Selection: When inspecting the various pseudo-labels, we observed many cases of drift and incorrect predictions due to the harsh generalization settings. Hence, the uncertainty-aware pseudo-label selection mechanism plays a crucial role in the semi-supervised learning process. As shown in Table 1 and Table 2, discarding pseudo-labels with low confidence consistently improves performance, both with and without multi-modal supervision. We notice how a student model without the uncertainty-aware sample removal (i.e., 'Student w/o

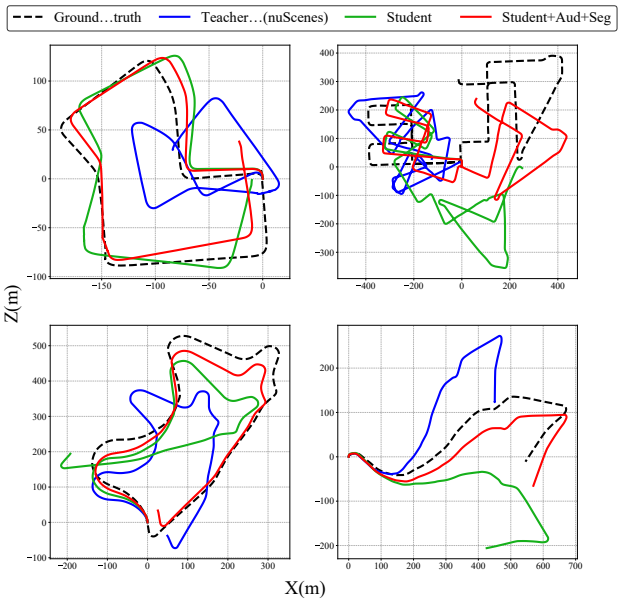

Figure 4: **Qualitative Analysis on KITTI.** We find that incorporating audio and segmentation tasks as part of the semi-supervised learning process significantly improves ego-pose estimation on KITTI.

Filter') provides only mild improvements compared to the teacher. Once noisy samples are filtered out of the dataset, the performance on KITTI and nuScenes improve significantly, e.g., from 26.16 to 17.04 and 15.93 to 10.54 translation error respectively.

Ablation on Auxiliary Tasks: We sought to understand the role of the various explored auxiliary tasks, i.e., audio, segmentation, depth, and flow. We first analyze the impact of adding an audio reconstruction task for the VO problem. Although extracted audio includes some ambient noise, we can see that XVO consistently benefits from the proposed audio supervision across the evaluation datasets. This can be explained by the consistent quality of the ground-truth audio labels, i.e., when compared to the noise in pseudo-labels generated by the auxiliary prediction models on our unconstrained videos. In general, we find that audio, segmentation, and flow tasks result in better performance when compared to the depth prediction task. While prior research often leverages monocular depth prediction for improving VO on KITTI, this is a significantly challenging task in more general settings which results in noisier pseudo-labels. We also investigate various combinations of auxiliary branches and find the combination of segmentation and audio branch performs better than a single auxiliary task on KITTI. While this is encouraging, KITTI contains simpler scenarios with relatively few dynamic traffic participants. In such simpler settings, our segmentation branch can be used to obtain reliable pseudo-labels and learn efficient generalized features. However, this finding does not extend to nuScenes and Argoverse 2 which frequently contain dense and dynamic scenes. We also find that simply adding prediction tasks does not provide further gains due to the pseudo-label noise and a more brittle and difficult optimization process. Complete ablations on auxiliary tasks can be seen in our supplementary.

4.2. Qualitative Results

Fig. 4 depicts the prediction of driving trajectories on KITTI sequences 7, 8, 9, and 10. The trajectory predicted using the teacher model that is trained on nuScenes is not able to recover scale accurately. Due to the semi-supervised training process, the student model is shown to have better scale recovery and generalization despite the lack of calibration knowledge. Nonetheless, the student model fails to estimate accurate rotation in more challenging scenes on KITTI, e.g., top right and bottom left scenarios. Finally, the cross-modal trained model is shown to robustly estimate translation, rotation, and scale, even in the most complicated route in Fig. 4-top right. Additional qualitative examples are provided in the supplementary.

5. Conclusion

In this paper, we present XVO, a novel method for generalized visual odometry estimation via cross-modal self-training. Our efficient network structure achieves state-of-the-art results on KITTI, despite having no knowledge of camera parameters or multi-frame optimization as in prior methods. Moreover, our framework leverages a mixed-label semi-supervised setting over a large dataset of internet videos to further enhance generalization performance. Specifically, we show that additional auxiliary segmentation and audio reconstruction tasks can significantly impact cross-dataset generalization. Our trained VO models can be used across platforms and settings without fine-tuning, i.e., due to general reasoning over semantic visual characteristics of scenes. Moreover, our training settings of improving the performance of a model that is initially trained on a small and restricted dataset are broadly applicable to various robotics use-cases. We hope our work can inspire future researchers to explore scalable VO models that can benefit a broad range of applications. Given the limited utility of combining multiple auxiliary tasks in our settings, a future direction would be to study better methods for learning from noisy and diverse auxiliary pseudo-labels. Moreover, while we achieved state-of-the-art results with a two-frame approach, multi-frame optimization could provide further benefits by alleviating drift.

Acknowledgments: We thank the Red Hat Collaboratory and National Science Foundation (IIS-2152077) for supporting this research.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. In *PAMI*, 2018. 3

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 3, 4

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. In *PAMI*, 2018. 3

[4] Aseem Behl, Kashyap Chitta, Aditya Prakash, Eshed Ohn-Bar, and Andreas Geiger. Label-efficient visual abstractions for autonomous driving. In *IROS*, 2020. 2, 5

[5] Aseem Behl, Despoina Paschalidou, Simon Donné, and Andreas Geiger. Pointflownet: Learning representations for rigid motion estimation from point clouds. In *CVPR*, 2019. 3

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

[7] 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. 1, 5, 6

[8] Benjamin Caine, Rebecca Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Zhifeng Chen, and Jonathon Shlens. Pseudo-labeling for scalable 3d object detection. *arXiv preprint arXiv:2103.02093*, 2021. 2, 3, 6

[9] 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*, 37(6):1874–1890, 2021. 2

[10] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019. 2

[11] Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D Manning. Text to 3d scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*, 2015. 3

[12] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021. 5

[13] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 3, 5

[14] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 4

[15] Amirhossein Dadashzadeh, Alan Whone, and Majid Mirmehdi. Auxiliary learning for self-supervised video representation via similarity-based knowledge distillation. In *CVPR*, 2022. 3

[16] Keqi Deng, Songjun Cao, Yike Zhang, Long Ma, Gaofeng Cheng, Ji Xu, and Pengyuan Zhang. Improving ctc-based speech recognition via knowledge transferring from pre-trained language models. In *ICASSP*, 2022. 3

[17] Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. In *AISTAT*, 2021. 2

[18] Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *J. Nonparametric Stat.*, 32(1):42–72, 2020. 2

[19] Lucio M Dery, Paul Michel, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Aang: Automating auxiliary learning. *arXiv preprint arXiv:2205.14082*, 2022. 3

[20] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 5

[21] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *ACMMM*, 2021. 3

[22] 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. In *ICLR*, 2021. 4

[23] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. In *PAMI*, 2017. 1, 2, 3

[24] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 1, 2, 3

[25] Mahdi Abolfazli Esfahani, Han Wang, Keyu Wu, and Shenghai Yuan. Aboldeepio: A novel deep inertial odometry network for autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.*, 21(5):1941–1950, 2019. 2

[26] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 5

[27] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. Multimodal multimedia retrieval with vitrivr. In *ICMR*, 2019. 3

[28] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 2, 6

[29] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *CVPR*, 2019. 1

[30] Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *ICCV*, 2019. 4

[31] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 5

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[33] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019. 2

[34] Zanming Huang, Zhongkai Shangguan, Jimuyang Zhang, Gilad Bar, Matthew Boyd, and Eshed Ohn-Bar. ASSISTER: Assistive navigation via conditional instruction gen-

eration. In *ECCV*, 2022. 3

[35] Bogdan Ionescu, Henning Müller, Renaud Péteri, Johannes Rückert, Asma Ben Abacha, Alba G Seco de Herrera, Christoph M Friedrich, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, et al. Overview of the imageclef 2022: Multimedia retrieval in medical, social media and nature applications. In *CLEF*, 2022. 3

[36] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *RA-L*, 4(2):2188–2195, 2019. 3

[37] Ganesh Iyer, J Krishna Murthy, Gunshi Gupta, Madhava Krishna, and Liam Paull. Geometric consistency for self-supervised end-to-end visual odometry. In *CVPRW*, 2018. 1, 2, 3

[38] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 2

[39] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 4

[40] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 2

[41] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 3, 4

[42] Po-Nien Kung, Sheng-Siang Yin, Yi-Cheng Chen, Tse-Hsuan Yang, and Yun-Nung Chen. Efficient multi-task auxiliary learning: selecting auxiliary data by feature similarity. In *EMNLP*, 2021. 3

[43] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 2, 3

[44] Qimai Li, Xiao-Ming Wu, Han Liu, Xiaotong Zhang, and Zhichao Guan. Label efficient semi-supervised learning via graph filtering. In *CVPR*, 2019. 2

[45] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *ICRA*, 2018. 1, 2, 3, 6, 7

[46] Shunkai Li, Xin Wu, Yingdian Cao, and Hongbin Zha. Generalizing to the open world: Deep visual odometry with online adaptation. In *CVPR*, 2021. 1, 3

[47] Shunkai Li, Fei Xue, Xin Wang, Zike Yan, and Hongbin Zha. Sequential adversarial learning for self-supervised deep visual odometry. In *ICCV*, 2019. 1, 2, 3

[48] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 3

[49] Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*, 2018. 3

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[51] Shikun Liu, Andrew Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. In *NeurIPS*, 2019. 3

[52] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 5

[53] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. Cnn-svo: Improving the mapping in semi-direct visual odometry using single-image depth prediction. In *ICRA*, 2019. 5

[54] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018. 1, 2

[55] Gideon S Mann and Andrew McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, 2007. 2

[56] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000. 4

[57] Viorica Marian, Sayuri Hayakawa, and Scott R Schroeder. Cross-modal interaction between auditory and visual input impacts memory retrieval. *Front. Neurosci.*, 15:661477, 2021. 2

[58] Sarah Maslen. "hearing" ahead of the sound: How musicians listen via proprioception and seen gestures in performance. *Senses Soc.*, 2022. 2

[59] Alexander Mey and Marco Loog. Improved generalization in semi-supervised learning: A survey of theoretical results. In *PAMI*, 2022. 2

[60] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 3

[61] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016. 2

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

[63] Peter Muller and Andreas Savakis. Flowdometry: An optical flow and deep learning based approach to visual odometry. In *WACV*, 2017. 5

[64] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *T-RO*, 31(5):1147–1163, 2015. 2

[65] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *T-RO*, 33(5):1255–1262, 2017. 1, 2

[66] Ishan Nigam. Learning from auxiliary supervision. Master's thesis, Carnegie Mellon University, 2018. 3

[67] Casey O'Callaghan. Seeing what you hear: Cross-modal illusions and perception. *Philos.*, 18:316–338, 2008. 2

[68] Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, and Andreas Geiger. Learning situational driving. In *CVPR*, 2020. 3

[69] Tejas Pandey, Dexmont Pena, Jonathan Byrne, and David Moloney. Leveraging deep learning for visual odometry using optical flow. *J. Sens.*, 21(4):1313, 2021. 5

[70] Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L Martel. A cluster-then-label semi-supervised learning approach for pathology image classification. *Sci. Rep.*, 8(1):1–13, 2018. 2

[71] Jean Piaget, Margaret Cook, et al. *The origins of intelli-*

gence in children, volume 8. IUP Press NY, 1952. 2

[72] Jin-Chun Piao and Shin-Dug Kim. Real-time visual–inertial slam based on adaptive keyframe selection for mobile ar applications. *TMM*, 21(11):2827–2836, 2019. 2

[73] Sabrina Pitzalis, Stefano Sdoia, Alessandro Bultrini, Giorgia Committeri, Francesco Di Russo, Patrizia Fattori, Claudio Galletti, and Gaspare Galati. Selectivity to translational egomotion in human brain motion areas. *PloS one*, 8(4), 2013. 2

[74] Monique Radeau. Auditory-visual spatial interaction and modularity. *CPC*, 13(1):3–51, 1994. 2

[75] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019. 2, 3

[76] 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. 2, 3, 4

[77] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. *arXiv preprint arXiv:2208.08988*, 2022. 1, 2, 3

[78] Alexander Sax, Bradley Emi, Amir R Zamir, Leonidas Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning active tasks. In *CoRL*, 2019. 3

[79] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023. 5

[80] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth $32 \times 32 \times 8$ voxels. In *ASRU*, 2021. 3

[81] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artif. Life*, 11(1-2):13–29, 2005. 2

[82] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2, 4

[83] Hailuo Song, Ao Li, Tong Wang, and Minghui Wang. Multimodal deep reinforcement learning with auxiliary task for obstacle avoidance of indoor mobile robot. *J. Sens.*, 21(4):1363, 2021. 3

[84] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017. 2

[85] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 4

[86] Xiaodong Tan, Mathis Antony, and H Kong. Automated music generation for visual art through emotion. In *ICCC*, 2020. 3

[87] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[88] Zachary Teed and Jia Deng. Droid-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. *NeurIPS*, 2021. 3

[89] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *arXiv preprint arXiv:2208.04726*, 2022. 2, 3

[90] Alexander Vakhitov, Victor Lempitsky, and Yinqiang Zheng. Stereo relative pose from line and point feature triplets. In *ECCV*, 2018. 2

[91] 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. 5

[92] Ke Wang, Siyuan Zhang, Junlan Chen, Fan Ren, and Lei Xiao. A feature-supervised generative adversarial network for environmental monitoring during hazy days. *Sci. Total Environ.*, 748:141445, 2020. 2

[93] Lirui Wang, Yu Xiang, Wei Yang, Arsalan Mousavian, and Dieter Fox. Goal-auxiliary actor-critic for 6d robotic grasping with point clouds. In *CoRL*, 2022. 3

[94] 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. 2, 3, 7

[95] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *IJRR*, 37(4-5):513–542, 2018. 7

[96] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. In *CoRL*, 2021. 1, 2, 3, 4, 5, 6, 7

[97] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *ICCV*, 2021. 2

[98] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, 2022. 2

[99] Shams Watkins, Ladan Shams, Sachiyo Tanaka, J-D Haynes, and Geraint Rees. Sound alters activity in human v1 in association with illusory visual perception. *Neuroimage*, 31(3):1247–1256, 2006. 2

[100] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2021. 6

[101] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *CVPR*, 2023. 2

[102] Zhen Xing, Hengduo Li, Zuxuan Wu, and Yu-Gang Jiang. Semi-supervised single-view 3d reconstruction via prototype shape priors. In *ECCV*, 2022. 2

[103] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 3

[104] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging aux-

iliary tasks with affinity learning for weakly supervised semantic segmentation. In *ICCV*, 2021. 3

[105] Fei Xue, Qiuyuan Wang, Xin Wang, Wei Dong, Junqiu Wang, and Hongbin Zha. Guided feature selection for deep visual odometry. In *ACCV*, 2019. 7

[106] Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In *CVPR*, pages 8575–8583, 2019. 3, 7

[107] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *ICCV*, 2019. 2, 3

[108] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*, 2021. 4

[109] 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, 2, 3

[110] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021. 2

[111] Yingda Yin, Yingcheng Cai, He Wang, and Baoquan Chen. Fishermatch: Semi-supervised rotation regression via entropy-based filtering. In *CVPR*, 2022. 2

[112] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 1, 2, 3, 6, 7

[113] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020. 3, 4

[114] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 3, 4

[115] 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. 2, 7

[116] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, and Ian Reid. Visual odometry revisited: What should be learnt? In *ICRA*, 2020. 1, 2

[117] Jimuyang Zhang, Zanming Huang, and Eshed Ohn-Bar. Coaching a teachable student. In *CVPR*, 2023. 3

[118] Jimuyang Zhang and Eshed Ohn-Bar. Learning by watching. In *CVPR*, 2021. 3

[119] Jimuyang Zhang, Minglan Zheng, Matthew Boyd, and Eshed Ohn-Bar. X-World: Accessibility, vision, and autonomy meet. In *ICCV*, 2021. 2

[120] Jimuyang Zhang, Ruizhao Zhu, and Eshed Ohn-Bar. Selfd: self-learning large-scale driving policies from the web. In *CVPR*, 2022. 2, 3

[121] Yu Zhang and Qiang Yang. A survey on multi-task learning. *TKDE*, 34(12):5586–5609, 2021. 4

[122] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 3

[123] 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

[124] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *CVPR*, 2020. 1, 2, 3

[125] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *CVPR*, 2019. 3

[126] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1, 2, 7

[127] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 3

[128] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018. 1, 2