COMPASS: Contrastive Multimodal Pretraining for Autonomous Systems

Shuang Ma[†], Sai Vemprala[†], Wenshan Wang[⋆], Jayesh K. Gupta[†], Yale Song[†], Daniel McDuff[†] and Ashish Kapoor[†]

***Abstract*— Learning representations that generalize across tasks and domains is challenging yet necessary for autonomous systems. Although task-driven approaches are appealing, designing models specific to each application can be difficult in the face of limited data, especially when dealing with highly variable multimodal input spaces arising from different tasks in different environments. We introduce the first general-purpose pretraining pipeline, COntrastive Multimodal Pretraining for AutonomouS Systems (COMPASS), to overcome the limitations of task-specific models and existing pretraining approaches. COMPASS constructs a multimodal graph by considering the essential information for autonomous systems and the properties of different modalities. Through this graph, multimodal signals are connected and mapped into two factorized spatio-temporal latent spaces: a "motion pattern space" and a "current state space." By learning from multimodal correspondences in each latent space, COMPASS creates state representations that models necessary information such as temporal dynamics, geometry, and semantics. We pretrain COMPASS on a large-scale multimodal simulation dataset TartanAir [1] and evaluate it on drone navigation, vehicle racing, and visual odometry tasks. The experiments indicate that COMPASS can tackle all three scenarios and can also generalize to unseen environments and real-world data. [1].**

I. Introduction

A fundamental facet of human intelligence is the ability to perceive the environment and encode multimodal sensory signals into complex neural representations [2], [3], which are then used to complete a wide variety of tasks. Similarly, learning representations that capture the underlying state of an environment from different sensors, while taking into account an agent's dynamic capabilities is crucial for autonomous systems. Such concise, jointly learned representations have the potential to effectively transfer knowledge across tasks and enable learning with fewer environmental interactions. The ability to perceive and act is crucial for any embodied autonomous agent and is required in many situations involving different form factors and scenarios. For example, localization (or being able to answer "Where am I?") is a fundamental question that needs to be answered by any autonomous agent prior to navigation, this is often achieved via visual odometry. Highly dynamic tasks, such as vehicle racing, necessitate collision avoidance and require precise understanding for planning a trajectory and meeting objectives. In both cases learning geometric and semantic information from the environment is crucial. Task-specific approaches produce promising results, but they involve learning only the part of information tailored for the intended tasks, which can be limiting in utility by failing to generalize to new scenarios. We investigate whether it is possible to build a general-purpose pretrained models in a task-agnostic fashion, which can be useful in solving various downstream tasks relevant to the perception-action loops in autonomous systems.

Although pretrained models have shown strong performance in domains such as NLP [4], [5] and computer vision [6], [7], building such models for autonomous systems brings unique challenges. First, the environments are usually perceived through multimodal sensors, so the model needs the ability to make sense of multimodal data. Existing multimodal learning approaches primarily focus on mapping multimodal data into joint latent spaces [8], [9], [10]. These approaches are suboptimal for autonomous systems as they do not address aspects such as differing sampling rate, temporal dynamics, and geo-centric or object-centric spatial factors. These are crucial factors in our scenario due to variations that arise from sensor and actuator configurations in autonomous systems. Secondly, autonomous systems deal with a complex interplay between perception and action. The target learning space is highly variable due to a large variety of environmental factors, application scenarios, and system dynamics. This is in stark contrast to language models that focus on underlying linguistic representations, or visual models centered on object semantics. Finally, unlike NLP and computer vision, there is a scarcity of multimodal data that can be used to train large pretrained representations for autonomous systems.

In this work, we introduce COntrastive Multimodal Pretraining for AutonomouS Systems (COMPASS), a multimodal pretraining approach for *perception-action loops*. COMPASS builds a general-purpose representation that generalizes to different environments and tasks. Unlike the prevalent approaches, COMPASS aims to learn a generic representation by exploiting underlying properties across multiple modalities, while appropriately considering the dynamics of the autonomous system. Self-supervised learning using a large corpus of multimodal data collected from various environments allows the model to be completely agnostic to downstream tasks.

Our design choices are informed by seeking answers to two questions: 1) *What information would be essential to solve common tasks in autonomous systems?* 2) *How can*

[†] Microsoft Redmond, WA {shuama,sai.vemprala,jayesh.gupta,yalesong,damcduff,akapoor}@microsoft.com

[⋆] Carnegie Mellon University Pittsburgh, PA wenshanwang@cmu.edu

[1]Our code implementation can be found at https://github.com/microsoft/COMPASS

we represent such information by learning from multisensory multimodal data captured by autonomous agents? First, we posit that information essential for autonomous systems lies in a spatio-temporal space that models motion (ego-motion or environmental), geometry and semantic cues. We also observe that such information is typically perceived by an autonomous agent through multimodal sensors. Consequently, we propose a multimodal graph as a core building block that models such spatio-temporal relationships and statistical characteristics of different modalities (Fig. 1). Intuitively, the graph maps all modalities into a factorized spatio-temporal latent space comprising of two subspaces: a *motion pattern space* and a *current state space*. The first subspace explicitly models and handles the temporal and system dynamics of autonomous systems, while the latter is designed to encode geometric and semantic information coming from modalities representing the states at certain local time points, e.g. a single RGB frame. Training COMPASS is then geared towards learning to associate multimodal data from a large training corpus. Such a factorized representation captures important spatio-temporal structure important for various downstream tasks while allowing different sensors to use the same pretrained model. By evaluating the pretrained COMPASS on three downstream tasks, i.e. *Vehicle Racing*, *Visual Odometry*, and *Drone Navigation*, with variations across environments, dynamics, and application scenarios, we observe that COMPASS generalizes well to different tasks, unseen environments and real-world challenges even in the low-data regimes.

II. Related Work

Representation learning has been an area of great interest in machine learning as well as in robotics. Self-supervised learning has been shown to be effective in vision particularly through the use of contrastive objectives [11], [12], [6], [7]. Recently, there is growing interest in learning "object-centric" representations of visual scenes [13], [14]. Contrastive learning has also been applied to reinforcement learning to match data augmentations with raw observations [15].

Learning multimodal representations has been examined in several domains such as vision-language [16], [17], vision-audio [18], [19], image registration [20], [21], and video understanding [9], [22]. Tsai et al. [23] present a framework learning intra-modal and cross-modal interactions from input, and Alayrac et al. [24] present a multimodal learning approach for text, audio and video. Inspired by the success of large-scale pretraining in the text domain [4], [5], pretrained models have also been developed for vision-language tasks [25], [26], [27]. A natural extension of multimodal learning algorithms has been applied to the multi-task learning setting [28], [29], [30]. Numerous surveys on multimodal learning are also largely focused on vision, text and speech [31], [32], [33]. Baltrusaitis et al. [34] point out the opportunity for co-learning with multimodal data where knowledge from one (resource rich) modality can be exploited in modeling another (resource poor) modality.

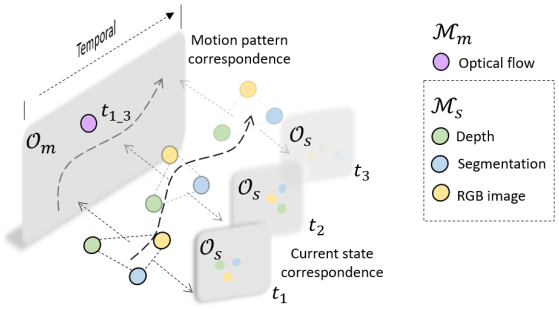


Fig. 1: We introduce COntrastive Multimodal Pretraining for AutonomouS Systems (COMPASS). Given spatial and temporal modalities $\mathcal{M}_s$ and $\mathcal{M}_m$. COMPASS learns two factorized latent spaces, i.e., a motion pattern space $\mathcal{O}_m$ and a current state space $\mathcal{O}_s$, using multimodal correspondence as the self-supervisory signal.

Autonomous systems require rich, well-grounded representations and benefit from the existence of multiple sensors of different modalities. Robotics tasks such as manipulation have been shown to benefit from object-centric representations [35], [36], and combining geometry has been shown to be effective for navigation tasks. Multimodal representation learning has been applied to robotic manipulation and grasping in the form of visuo-tactile representations [37], as well as scene understanding and exploration by combining RGB and depth [38], and human robot interaction [39]. Cross-modal representation learning has been combined with imitation learning to result in drone navigation policies [40]. Multi-task learning has been examined for self-driving under different modes [41] and visual odometry/semantic segmentation [42]. Instead of leveraging specific designs which are tailored for each of these tasks, here we propose a general-purpose representation learning approach in the context of the perception stack of an embodied autonomous agent. With the single unified pretraining model, different tasks can hence be easily achieved with a fast finetuning on a small amount of data.

III. Approach

We set out to create a model that can be pretrained on simulated environments providing data from different modalities. Our goal is to learn a model that can produce *general-purpose* representations and be further adapted to various autonomous tasks. We posit that autonomous systems need representations that encode both the current state of the environment as well as the temporal dynamics. As such, we propose a *multimodal graph* that respects the underlying spatio-temporal relationship between modalities (Section III-A), and use it as the basis for designing contrastive learning objectives (Section III-B) to learn two factorized latent spaces that model them, respectively.

A. Multimodal Graph Construction

Given a set of multimodal data $\{\mathcal{M}\}_N$ with N modalities, existing multimodal learning approaches mainly lie in two folds: 1) learning a **joint common multimodal latent space**, which maps all modalities into a common latent space. It benefits from the simple design, while suffers from the issue that different complementary properties across various

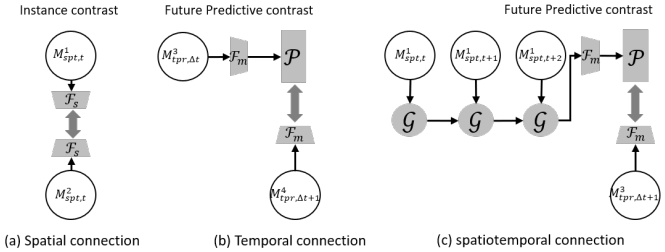


Fig. 2: Three types of connections. $\mathcal{M}^i_{s,t}$: i-th spatial modality at time step t. $\mathcal{M}^i_{m,\delta t}$: i-th temporal modality in a time window δt. z^i_t: extracted representation from modality i at time t, z'^i_t: the latent code mapped by spatial projection head $\mathcal{F}_s$. $z'^i_{t\to t+j}$: the latent code projected by temporal projection head $\mathcal{F}_m$. $c^i_{t\to t+j}$: context vector induced by aggregation network $\mathcal{G}$. $\hat{z}$: predicted latent code. E_i: modality encoder for modality i. Different color signifies different modality. The modules in shaded gray share weights among all modalities, i.e. $\mathcal{F}_s$, $\mathcal{F}_m$, $\mathcal{G}$ and $\mathcal{P}$.

modalities are not fully utilized. Projection to a shared single latent space results in a loss of information. 2) learning **disjoint cross-modal latent spaces**, which learns multimodal data by capturing the cross-modal relation between each pair of modalities. Though, it enables the specificity of different modality, scaling it to a more complex setting, e.g., more than two modalities, presents several computational challenges.

Spatio-temporal Multimodal latent space. The model we propose enables us to model the complementary information across the multimodal signals in a scalable manner via a novel multimodal graph design. The key insight we build upon is that the information essential for autonomous systems lies in a spatio-temporal space that can be partitioned to model temporal statistics and spatial and/or semantic aspects. Consequently, we start by first categorizing our multimodal data streams into two classes: *spatial* modalities and *temporal* modalities. Given a set of N multimodal data streams $\{\mathcal{M}\}_N$ with n spatial modalities $\{\mathcal{M}_s\}_n$ and l temporal modalities $\{\mathcal{M}_m\}_l$, we then jointly learn two latent spaces, a "motion pattern space" $\mathcal{O}_m$ and a "current state space" $\mathcal{O}_s$.

Let $\{\mathcal{M}^1_m, \mathcal{M}^2_m, ..., \mathcal{M}^l_m\}_{\delta T}$ denote a data sequence from temporal modalities that are time-synchronized within a window δT, which share a certain motion pattern arising from the agent or the environment. As shown in Fig. 2, we construct temporal connections by mapping them to the common motion pattern space $\mathcal{O}_m$ through a projection head $\mathcal{F}_m$, i.e. $\mathcal{F}_m(\{\mathcal{M}^i_m\}^{i\in[1,l]}_{\delta T}) \to \mathcal{O}_m$. Similarly, we construct spatial connections by mapping data from spatial modalities at each discrete time step t, i.e., $\{\mathcal{M}^1_s, \mathcal{M}^2_s, ..., \mathcal{M}^n_s\}_t$, to the common current state space $\mathcal{O}_s$ through a projection head $\mathcal{F}_s$, i.e. $\mathcal{F}_s(\{\mathcal{M}^i_s\}^{i\in[1,n]}_t) \to \mathcal{O}_s$. Furthermore, to better associate spatial modalities with the temporal ones, we rely on the intuition that multiple instances of one/more spatial modalities from a window of time can be associated with observations from a temporal modality from that same window of time. Thus, a spatio-temporal connection is added by aggregating sequential data from spatial modalities using an aggregation head $\mathcal{G}$ and mapping them to the common motion pattern space $\mathcal{O}_m$ using the projection head $\mathcal{F}_m$, i.e. $\mathcal{F}_m(\mathcal{G}(\{\mathcal{M}^i_{s,t}, \mathcal{M}^i_{s,t+1}, ..., \mathcal{M}^i_{s,t+\delta t}\}^{i\in[1,n]})) \to \mathcal{O}_m$.

B. Training Objective

Contrastive Objective for Temporal Connections. To encode temporal information in the motion pattern space $\mathcal{O}_m$, we solve a contrastive learning objective that associates pairs of time-synced data from different modalities. Intuitively, if a model successfully captures temporal information from one modality, it should have the predictive capacity to model a few future time steps for itself as well as the other modalities. We formulate this intuition into a contrastive learning objective.

Given a set of m time-synced temporal modalities, we define positive pairs as a sequence of embeddings for the true observations $[z_{t+1}, z_{t+2}, ..., z_{end}]_{\mathcal{M}^{i\in[1,m]}}$ and a sequence of embeddings predicted recursively from an anchor modality $[\hat{z}_{t+1}, \hat{z}_{t+2}, ..., \hat{z}_{end}]_{\mathcal{M}^a}$, where $\mathcal{M}^a$ is the anchor modality. The positive pairs from m modalities include the true future observations of anchor modality, its own, and the remaining $m-1$ modalities. Thus, the comparison is performed both in an intra-modal and a cross-modal fashion. As shown in Fig. 3, the modality-specific encoders E, extract embeddings from each modality. These are then mapped to the common motion pattern space $\mathcal{O}_m$ through the motion pattern projection head $\mathcal{F}_m$. A prediction head $\mathcal{P}$ is added on top to perform future prediction. The contrastive loss is computed between the predicted future representations and their corresponding encoded true representations. Our contrastive objective is then:

$$\mathcal{L}_m = -\sum_{t,i} \log \frac{\exp(\hat{z}^T_{t,a} z_{t,i})}{\exp(\hat{z}^T_{t,a} z_{t,i}) + \sum_{j\neq t} \exp(\hat{z}^T_{t,a} z_{j,i})} \quad (1)$$

where t, i, a denote the time, modality, and anchor indices.

Contrastive Objective for Spatial Connections. The current state space $\mathcal{O}_s$ is expected to encode geometric and semantic information by associating the data from spatial modalities $\mathcal{M}_s$ together. We again utilize a contrastive objective to formulate this idea.

Given n spatial modalities, we define the positive pairs as the representation from an anchor modality at time step t, i.e. $z_{t,a}$, and the representations induced by all the spatial modalities at the same time step t, i.e. $\{z_{t,i}\}_{i\in[1,n]}$. The negative pairs are sampled from representations induced by spatial modalities at different time steps. We formulate this instance level contrastive loss as:

$$\mathcal{L}_s = -\sum_{t,i} \log \frac{\exp(z^T_{t,a} z_{t,i})}{\exp(z^T_{t,a} z_{t,i}) + \sum_{j\neq t} \exp(z^T_{t,a} z_{j,i})} \quad (2)$$

where t, i, a denote time, modality and anchor, respectively. Note that the anchor representation is sampled from $\mathcal{O}_s$, i.e. $\mathcal{F}_s(E(\mathcal{M}_a)) \to z_a$. This is different from Eq. (1) where the anchor is the estimated representation induced by the prediction head, i.e. $\mathcal{P}(\mathcal{F}_m(E(\mathcal{M}_a))) \to \hat{z}_a$. Therefore, while Eq. (1) computes contrastive loss through future prediction, Eq. (2) computes it at an instance level.

Objective for Spatio-temporal Connections. The spatio-temporal connections encode motion patterns from consecutive observations of spatial modalities. Given a sequence $[M_{a,t}, M_{a,t+1}, ..., M_{a,t+\delta t}]$ from an anchor modality $\mathcal{M}_a \in \{\mathcal{M}_s\}_n$, we obtain embeddings using the

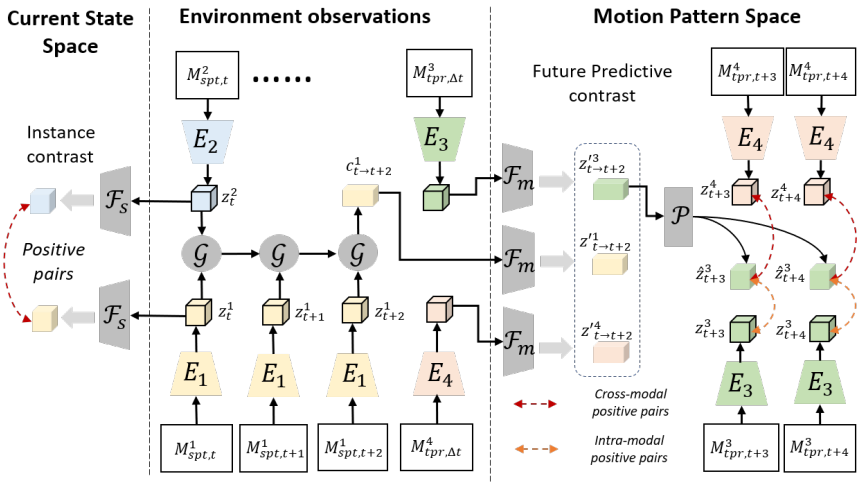


Fig. 3: Pretraining pipeline of COMPASS. $\mathcal{M}^i_{s,t}$ denotes the i-th spatial modality at time step t. $\mathcal{M}^i_{m,\delta t}$ denotes the i-th temporal modality in a time window δt. z^i_t denotes the extracted representation from modality i at time t, z'^i_t is the latent code mapped by spatial projection head $\mathcal{F}_s$. $z'^i_{t\to t+j}$ is the latent code projected by temporal projection head $\mathcal{F}_m$. $c^i_{t\to t+j}$ denotes the context vector induced by aggregation network $\mathcal{G}$. $\hat{z}$ represents the predicted latent code. E_i is the modality encoder for modality i. Different color signifies different modality. The modules in shaded gray share weights among all modalities, i.e. $\mathcal{F}_s$, $\mathcal{F}_m$, $\mathcal{G}$ and $\mathcal{P}$.

modality encoder, i.e. $E([M_{a,t}, M_{a,t+1}, ..., M_{a,t+\delta t}]) \to [z_{a,t}, z_{a,t+1}, ..., z_{a,t+\delta t}]$. We then use the aggregation network $\mathcal{G}$ to project them to $\mathcal{O}_m$ and produce an aggregated context vector c_a, i.e. $\mathcal{G}([z_{a,t}, z_{a,t+1}, ..., z_{a,t+\delta t}]) \to c_a$. Given this context vector, we can compute future predictions similar to the way the motion pattern loss was computed, i.e., by inputting c_a to $\mathcal{P}$ for future prediction as $\mathcal{P}(c_a) \to \hat{z}_a$. To this end, we again utilize Eq. (1) to minimize the contrastive objective of $\mathcal{L}_{sm}$. Our learning objective is: $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_s + \mathcal{L}_{sm}$.

IV. Experiments

The experiments aim to demonstrate the effectiveness of COMPASS as a general-purpose pretraining approach. We tackle three downstream scenarios that are representative of autonomous system tasks: vehicle racing (Section IV-A), visual odometry (Section IV-B), and drone navigation (Section IV-C), for all of which we finetune a single pretrained COMPASS model.

Through our experiments, we explore the following questions:

1) Can COMPASS adapt to unseen environments and real-world scenarios? COMPASS is pretrained on simulation data (TartanAIR [1]) and we demonstrate experiments on a real-world benchmark (KITTI [43]) to understand sim2real performance (Section IV-B). Similarly, our experiments with the vehicle racing task investigate how well we can generalize to completely unseen environments (Section IV-A).

2) What are the benefits of COMPASS when compared to other representation learning approaches? We compare COMPASS with task-specific approaches and representative pretraining/multimodal learning approaches in Section IV-A).

3) Can COMPASS improve data efficiency? We compare models finetuned over COMPASS representations to task specific models trained from scratch as we vary the data set size, as we analyze the learning performance (section IV-C).

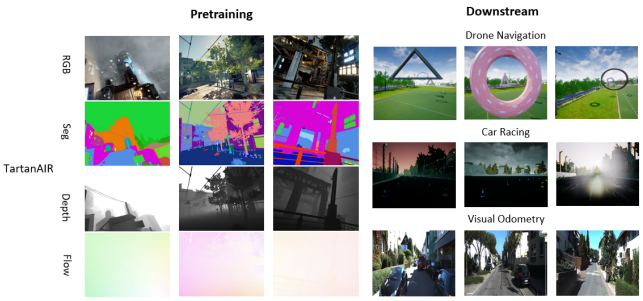


Fig. 4: Samples from TartanAIR and the downstream task datasets. Notice the difference in the visual scene: a soccer field (drone navigation), a racing track with varying backgrounds (vehicle-racing), and a real world scene (VO).

TABLE I
Various datasets used in our experiments.

Dataset	Usage	Scale	Env.
TartanAIR	Pretrain	1M	16
AirSim-Car	Vehicle Racing	17k	9
KITTI	Visual Odometry	23K	11
Drone-Gate	Drone Navigation	3k	1

Pretraining. We use a 3D-ResNet18 [44] architecture for the encoder for each modality E_m, a two-layer CNN for the future prediction head P, and a bidirectional ConvGRU [45] for the aggregation head G (shared across modalities) We use the TartanAIR [1] dataset for pretraining that contains 400K sensor samples from diverse environments including indoor, outdoor, urban, nature, and sci-fi scenes. The dataset is generated with a simulated pinhole camera, and provides multimodal signals. We pretrain COMPASS on 16 environments of TartanAIR with data from three modalities: RGB, depth, and optical flow. Sample data from the pretraining dataset can be seen in Fig. 4 along with the downstream task datasets. In Table I, we list some details about the extent of data used for pretraining and task-specific finetuning.

TABLE II
RESULTS ON VEHICLE RACING EXPERIMENTS. THE NUMBERS ARE MEAN/STD OF L1 ERRORS ON STEERING ANGLE PREDICTION OVER 5 RUNS.

Model	Seen environment	Unseen environment
SCRATCH	0.085 ± 0.019	0.120 ± 0.009
CPC	**0.037** ± 0.012	0.101 ± 0.017
CMC	0.039 ± 0.013	0.102 ± 0.012
JOINT	0.055 ± 0.016	0.388 ± 0.018
DISJOINT	0.039 ± 0.017	0.131 ± 0.016
COMPASS	0.041 ± 0.021	**0.071** ± 0.023

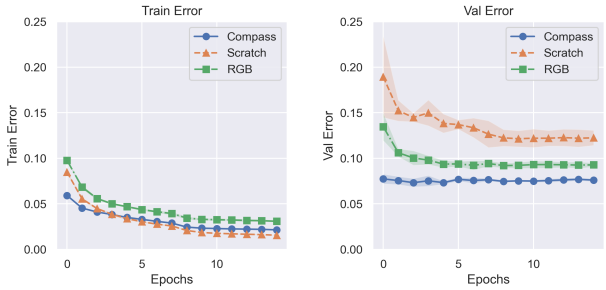
A. Vehicle Racing

Task and setting. The goal here is to enable autonomous vehicles to drive in a competitive Formula racing environment. We use the AirSim-Car dataset [46] that provides 9 simulated racetrack environments in the AirSim simulator, each with 2 lanes separated with different colored traffic cones. The environment contains visual distractors such as ad signs, tires, grandstands, and fences, which help add realism and increase task difficulty. The control module is expected to predict the steering angle such that the car can successfully maneuver through the tracks and avoid obstacles. We construct a perception module with the RGB encoder from COMPASS pretrained on TartanAIR and define a control module as a two-layer MLP with a prediction head that outputs the steering wheel angle (normalized to $[0, 1]$). We finetune the model on the AirSim-Car dataset with $L1$ loss measuring the per-frame angle discrepancy.

Baselines. We compare COMPASS with a model trained from scratch (Scratch), 2 pretraining approaches (CPC [11] and CMC [10]), and 2 multimodal learning approaches (JOINT and DISJOINT). Scratch is directly trained on the AirSim-Car dataset without pretraining, whereas the pretraining and multimodal learning approaches are pretrained on TartanAIR before finetuning on the AirSim-Car dataset. More details on the baselines:

- Scratch trains a randomly initialized network (the same architecture as ours) from scratch.
- CPC [11] is a contrastive learning approach that learns representations by predicting the future representations in the latent space.
- CMC [10] is a contrastive learning approach that captures information shared across modalities. Unlike CPC, it learns from multiple views and the contrastive loss is defined at an instance level rather than in a predictive manner.
- JOINT learns multimodal data from a single joint latent space by mapping modalities with a single projection head.
- DISJOINT learns multimodal data from disjoint latent spaces. Other than using a single projection head, it creates a cross-modal latent space for each modality pairs, and all of the latent spaces are disjoint.

Can COMPASS generalize to unseen environments? We explore the hypothesis pretraining can help with generalization to unseen environments. Consequently, we compare COMPASS with Scratch (no pretraining) and the other pretraining approaches: CPC, CMC, JOINT, and DISJOINT. We evaluate these models in two settings: 1) trained and evaluated on the same environments ("seen"); 2) trained and evaluated on different environments ("unseen"). Table II shows that overall, there is a gap between Scratch and COMPASS.



(a) Train error profile (b) Validation error profile

Fig. 5: Comparison of train and test error profiles for vehicle racing between COMPASS, a model pretrained only with RGB modality, and a model trained from scratch for the task.

Can COMPASS benefit from multimodal pretraining regime? We investigate the effectiveness of pretraining on multimodal data by analyzing loss curves from different pretrained models on the same 'unseen' environments. Fig. 5 compares the train/validation loss curves of COMPASS, RGB, and Scratch over five random seeds. Here RGB is the model pretrained by using the same backbone and training objectives with COMPASS, but only pretrained with RGB modality. By comparing COMPASS with RGB, we observe that pretraining on multimodal data helps COMPASS achieve the best performance overall. Also, both of these pretraining models show large gaps when compared to a model trained from scratch (Scratch). When comparing Fig. 5a to Fig. 5b, we also see that Scratch suffers more from an overfitting issue than the other two models.

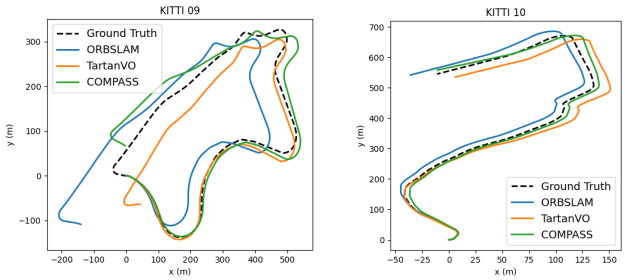
B. Visual Odometry

Task and setting. Visual odometry (VO) aims to estimate camera motion from consecutive image frames. It is a fundamental component in visual SLAM and is a widely used for localization in robotics. Note that we focus on a simple visual odometry task, which only takes in two consecutive images as inputs. This is different from full-fledged SLAM systems [47], [48], [49], which utilize key-frame optimization in the back-end - our task can be considered as the pose tracking module in the SLAM front-end.

We evaluate COMPASS for the VO task using a real-world dataset KITTI [43] which is one of the most widely used benchmarks in the VO/SLAM literature. It contains 11 labeled sequences including 23,201 image frames in a driving scenario. On this dataset, we examine the generalization ability of COMPASS that was pretrained purely on simulation data to the real-world data. We attempt the VO task using a model designed as a two-stage structure, following TartanVO [50]. In the first stage, an off-the-shelf pretrained optical flow estimation network, PWC-Net [51], is utilized to convert consecutive RGB image pairs into optical flow images by extracting the dense correspondence information. In the second stage, a pose network estimates the camera motion

TABLE III
COMPARISON OF TRANSLATION AND ROTATION ERRORS ON KITTI DATASET. VISO2-M AND ORB-SLAM ARE GEOMETRY-BASED, WHILE THE OTHERS ARE LEARNING-BASED APPROACHES. METRICS FOR COMPASS AND SCRATCH AVERAGED OVER FIVE RANDOM SEEDS.

Methods[2]	Seq. #09		Seq. #10	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}
VISO2-M [52]	**4.04**	1.43	25.2	3.8
ORB-SLAM† [47]	15.3	**0.26**	**3.71**	**0.3**
DeepVO*† [53]	N/A	N/A	8.11	8.83
Wang et al.*† [54]	8.04	1.51	6.23	0.97
TartanVO‡ [50]	6.00	3.11	6.89	2.73
UnDeepVO* [55]	N/A	N/A	10.63	4.65
GeoNet* [56]	26.93	9.54	20.73	9.04
SCRATCH*	1.88	0.74	3.05	1.10
COMPASS*	1.90	0.78	3.14	1.11



(a) (b)

Fig. 6: Comparison of the KITTI 9 and 10 trajectories predicted by different approaches. TartanVO [50] is a learning-based VO (only relies on two frames, same as ours), and ORBSLAM2 [47] is a geometry-based SLAM system (includes multi-frame optimization).

based on the optical flow. In our case, we utilize the pretrained optical flow encoder from COMPASS coupled with a 2-layer MLP prediction head as the second-stage pose network, so that we can evaluate the effectiveness of COMPASS for the flow modality. The model is asked to estimate the camera translation and rotation.

Can COMPASS be adapted to real-world scenarios? In this experiment, we finetune the model on sequences 00-08 of KITTI and test it on sequence 09 and 10. For comprehensive investigation, we compare COMPASS with both SLAM methods and visual odometry methods. The results are shown in Table III using the relative pose errors (RPE) on rotation and translation, which is the same metric used on KITTI benchmark. The first three [47], [48], [49] are SLAM methods that perform optimization on multiple frames to optimize the trajectory. Other baselines are VO methods that do not contain backend optimization. We compared against both geometry-based VO (VISO2-M [52], ORB-SLAM2 [47]) and learning-based VO, including supervised methods (DeepVO [55], Wang et. al [53], TartanVO [50]) and unsupervised methods (UnDeepVO [55], GeoNet [56]). We note here that the baseline methods apart from TartanVO use RGB images directly as inputs to their models whereas COMPASS and TartanVO use an optical flow network and perform pose estimation from flow data. TartanVO is a VO-specific pretrained model, which unlike the other learning based approaches is not finetuned on KITTI data.

Using the pretrained flow encoder from COMPASS within our VO pipeline achieves comparable results when comparing with other VO-specific approaches including SLAM methods. Fig. 6 shows the predicted trajectories of sequences 09 and 10 compared to ground truth. For clarity, we also select one representative model from the geometry-based and learning-based approaches each. We can see that, when pretrained purely on simulation data, COMPASS adapts well to real-world scenarios. We also show the results where the pretrained flow encoder from COMPASS is intialized with random parameters (depicted as Scratch). As shown in in Table III, we observe that the performance of COMPASS and Scratch are fairly close to each other. We hypothesize that the gains we are seeing on the VO task using the COMPASS architecture might primarily be due to the network architecture. For instance the existing pre-trained component in the VO pipeline to estimate optical flow might be helping much more than the fine-tuning. Another possible reason is our current reliance on RGB as the anchor modality. It might be possible to further the performance by training the network without relying solely on the RGB modality as an anchor. Finally, using regular contrastive loss has been shown to have limitations in capturing dense geometry information [57].

C. Drone Navigation

Task and setting. The goal of this task is to enable a quadrotor drone to navigate through a series of gates whose locations are unknown to it a priori. It takes as input monocular RGB images and predicts the velocity command. We construct the module by adding a policy network on top of the RGB encoder from pretrained COMPASS. We use a dataset from a simulated environment in AirSim Drone Racing Lab [58] where the environment contains a diverse set of gates varying in shape, sizes, color, and texture, through which a drone is flown to obtain the velocity commands.

In Fig. 7, we compare the validation errors of the x, y, z, and yaw velocity predictions from a finetuned COMPASS model with those of a model trained from scratch. We can observe that finetuning COMPASS for this velocity prediction task results in better performance than training a model from scratch, and reaches low errors quicker than the scratch model.

Can COMPASS improve data efficiency? Furthermore, we observe that finetuning pretrained COMPASS models exhibits more data efficient learning than training models from scratch. Fig. 8 compares finetuning performance with different amounts of data to training from scratch, over five different seeds. As we reduce the amount of training samples, model finetuned on `COMPASS` generalizes better than the model trained from scratch. Also, to match the performance of a `Scratch` model trained on 100% data, the model finetuned on `COMPASS` only required around 50% of the data.

[2] t_{rel}: average translational RMSE drift (%) on a length of 100-800 m. r_{rel}: average rotational RMSE drift ($^\circ/100m$) on a length of 100-800 m. $*$: the methods are trained or finetuned on the KITTI dataset. †: the methods use multiple frames to optimize the trajectory after the VO process. ‡: the method is trained on large amount of labeled data.

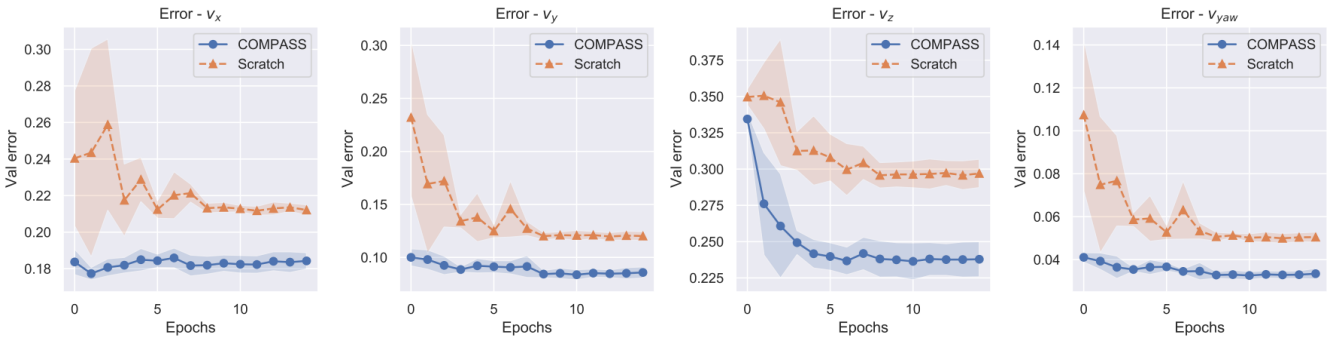


Fig. 7: Comparison of validation error profiles of velocity predictions between COMPASS and a model trained from scratch.

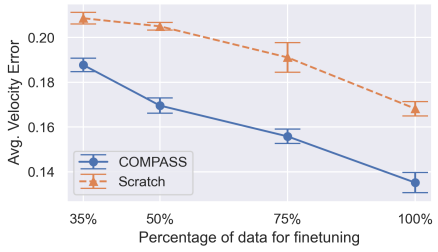


Fig. 8: Comparison of average velocity errors between COMPASS and Scratch with varying sizes of the finetuning dataset. We see that COMPASS consistently reaches a lower error even in the low-data regime.

V. Conclusion

We introduced COntrastive Multimodal Pretraining for AutonomouS Systems (COMPASS), a general purpose pre-training approach that learns multimodal representations for various downstream autonomous system tasks. In contrast to existing task-specific approaches in autonomous systems, COMPASS is trained entirely agnostic to any downstream tasks, with the primary goal of extracting information that is common to multiple tasks. COMPASS also learns modality specific properties, allowing it to encode the spatio-temporal nature of data commonly observed in autonomous systems. We demonstrate that COMPASS generalizes well to different downstream tasks – vehicle racing, visual odometry and drone navigation – even in unseen environments, real-world environments and in the low-data regime. More work is required for improving the local geometric information present in the learned representations, and to ensure encoders corresponding to all modalities can learn sufficiently strong features. Extending the ideas presented here to even more data modalities, especially with different sampling rates and understanding scaling laws [59] for such models is also important future work.

References

[1] 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. 2020.

[2] Richard A. Andersen, Lawrence H. Snyder, David C. Bradley, and Jing Xing. Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, 20(1):303–330, March 1997.

[3] Simon Lacey and Krish Sathian. Crossmodal and multisensory interactions between vision and touch. In *Scholarpedia of Touch*, pages 301–315. Springer, 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[8] Shuang Ma, Daniel Mcduff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7597–7606, 2019.

[9] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[10] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[11] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[12] Wenshan Wang, Aayush Ahuja, Yanfu Zhang, Rogerio Bonatti, and Sebastian Scherer. Improved generalization of heading direction estimation for aerial filming using semi-supervised regression. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5901–5907. IEEE, 2019.

[13] Daniel M Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B Tenenbaum, et al. Learning physical graph representations from visual scenes. *arXiv preprint arXiv:2006.12373*, 2020.

[14] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020.

[15] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

[16] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[17] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020.

[18] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019.

[19] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[20] Alexis Roche, Grégoire Malandain, Xavier Pennec, and Nicholas Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1115–1124. Springer, 1998.

[21] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13, 2018.

[22] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

[23] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[24] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *arXiv preprint arXiv:2006.16228*, 2020.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[26] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

[27] 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 *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[28] Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. Omninet: A unified architecture for multi-modal multi-task learning. *arXiv preprint arXiv:1907.07804*, 2019.

[29] Ryohei Kuga, Asako Kanezaki, Masaki Samejima, Yusuke Sugano, and Yasuyuki Matsushita. Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 403–411, 2017.

[30] Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, and Dhruv Batra. Embodied multimodal multitask learning. *arXiv preprint arXiv:1902.01385*, 2019.

[31] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[32] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[33] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[34] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[35] Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep object-centric representations for generalizable robot learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7111–7118. IEEE, 2018.

[36] Peter R Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv preprint arXiv:1806.08756*, 2018.

[37] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[38] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33, 2020.

[39] Joseph Campbell, Simon Stepputtis, and Heni Ben Amor. Probabilistic multimodal modeling for human-robot interaction tasks. *arXiv preprint arXiv:1908.04955*, 2019.

[40] Rogerio Bonatti, Ratnesh Madaan, Vibhav Vineet, Sebastian Scherer, and Ashish Kapoor. Learning visuomotor policies for aerial navigation using cross-modal representations. *arXiv preprint arXiv:1909.06993*, 2020.

[41] Sauhaarda Chowdhuri, Tushar Pankaj, and Karl Zipser. Multinet: Multi-modal multi-task learning for autonomous driving. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1496–1504. IEEE, 2019.

[42] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414, 2018.

[43] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[44] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[45] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

[46] Dean Zadok, Tom Hirshberg, Amir Biran, Kira Radinsky, and Ashish Kapoor. Explorations and lessons learned in building an autonomous formula sae car from simulations. *arXiv preprint arXiv:1905.05940*, 2019.

[47] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[48] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[49] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1281–1292, 2020.

[50] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. 2020.

[51] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[52] S. Song, M. Chandraker, and C. C. Guest. High accuracy monocular sfm and scale correction for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):730–743, 2016.

[53] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050. IEEE, 2017.

[54] X. Wang, D. Maturana, S. Yang, W. Wang, Q. Chen, and S. Scherer. Improving learning-based ego-motion estimation with homomorphism-based losses and drift correction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 970–976, 2019.

[55] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7286–7291. IEEE, 2018.

[56] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018.

[57] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3023–3032, 2021.

[58] Ratnesh Madaan, Nicholas Gyde, Sai Vemprala, Matthew Brown, Keiko Nagami, Tim Taubner, Eric Cristofalo, Davide Scaramuzza, Mac Schwager, and Ashish Kapoor. Airsim drone racing lab. *arXiv preprint arXiv:2003.05654*, 2020.

[59] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.