

# VOCAL: Visual Odometry via ContrAstive Learning

Chi-Yao Huang\* Zeel Bhatt\* Yezhou Yang

Arizona State University

{cy.huang, zbhatt1, yz.yang}@asu.edu

## Abstract

Breakthroughs in visual odometry (VO) have fundamentally reshaped the landscape of robotics, enabling ultra-precise camera state estimation that is crucial for modern autonomous systems. Despite these advances, many learning-based VO techniques rely on rigid geometric assumptions, which often fall short in interpretability and lack a solid theoretical basis within fully data-driven frameworks. To overcome these limitations, we introduce **VOCAL** (Visual Odometry via ContrAstive Learning), a novel framework that reimagines VO as a label ranking challenge. By integrating Bayesian inference with a representation learning framework, VOCAL organizes visual features to mirror camera states. The ranking mechanism compels similar camera states to converge into consistent and spatially coherent representations within the latent space. This strategic alignment not only bolsters the interpretability of the learned features but also ensures compatibility with multimodal data sources. Extensive evaluations on the KITTI dataset highlight VOCAL’s enhanced interpretability and flexibility, pushing VO toward more general and explainable spatial intelligence.

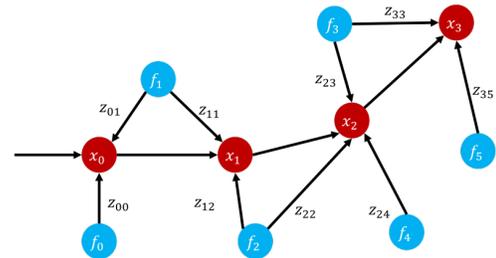
## 1. Introduction

As artificial intelligence advances, the need for multimodal learning methods continues to grow. Visual Odometry (VO)—a key technology for motion estimation in robotics, augmented reality, and autonomous driving—has traditionally relied on geometric constraints, temporal consistency, handcrafted features, and bundle adjustment. While these classical methods are effective, they were developed outside the deep learning paradigm, which instead relies on learning from latent representations.

This divergence creates a notable gap: conventional VO systems struggle to integrate with learning-based frameworks that operate in latent space. Although some learning-based VO models have been proposed, many still depend

\*These authors contributed equally to this work.

(a) Conventional Visual Odometry:



(b) VOCAL:

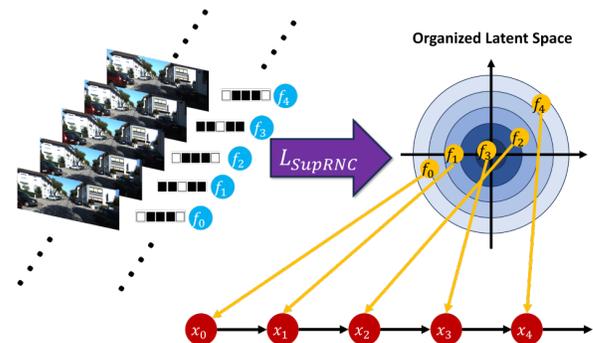


Figure 1. (a) Conventional graph-based visual odometry, where connections between camera states  $x_i$  and features  $f_j$  are modeled using predefined graphs  $z_{ij}$ . This manual design limits both flexibility and interpretability in learning-based VO systems. (b) The VOCAL architecture eliminates the need for handcrafted graph structures by reframing VO as a label-ranking problem. Through contrastive learning, VOCAL organizes features extracted from visual inputs based on their corresponding camera states, ensuring that inputs with similar camera states yield consistent features in the latent space. Our approach improves spatial understanding in visual odometry by establishing a direct correlation between feature representations and 3D camera states.

on geometric constraints that are difficult to interpret and not well aligned with representation learning. As a result, they are hard to adapt to broader multimodal systems such as large language models (LLMs) [21] and vision-language models (VLMs) [19]. Closing this gap is critical for en-

abling VO to collaborate with other learning-based models and to advance spatial intelligence.

In recent years, contrastive learning has emerged as a powerful tool for representation learning. Originally developed as a self-supervised technique, it learns to group similar samples and separate dissimilar ones in the latent space. Methods such as SimCLR [3] and MoCo [11] have demonstrated its effectiveness in image classification. Extensions like RankSim [9] and Rank-N-Contrast [42] further apply ranking mechanisms in latent space to improve regression tasks. Similarly, [39] and [36] show that structured latent representations benefit object pose estimation. Contrastive learning has also become central to multimodal learning. For example, CLIP [23] aligns image and text pairs in a shared latent space using contrastive objectives. Large-scale multimodal systems such as [21], Gemini [32], and LLaVA [19] adopt similar principles to integrate diverse data types.

Despite these successes, the application of contrastive learning to visual odometry remains underexplored. The lack of organized and interpretable latent representations makes it difficult for VO systems to collaborate with other learning-based models.

To address the challenge, we propose VOCAL (Visual Odometry via ContrAstive Learning), a framework that applies contrastive learning to VO to produce structured, interpretable representations (see Fig. 1). VOCAL reframes VO as a label-ranking problem, learning relationships between features and camera states in the latent space without relying on geometric constraints or handcrafted graph structures. This approach leads to a compact, explainable representation that integrates well with other learning-based systems and supports broader goals in spatial intelligence.

Our contributions are:

- We revisit Bayesian inference—the core principle behind VO—and reinterpret it in a modern latent representation framework.
- We reformulate learning-based VO as a label-ranking problem, introducing a new, explainable way to organize visual features by camera states.
- We provide a detailed analysis of how contrastive learning structures latent space for VO, offering new insights and improving the spatial understanding of learning-based models.

## 2. Related Work

Research in visual odometry (VO) can be broadly categorized into three main paradigms: geometry-based, hybrid, and learning-based approaches.

**Geometry-based Visual Odometry:** These methods rely on geometric principles to estimate 3D structure and determine camera pose. They are typically divided into two categories: direct and feature-based methods. Direct meth-

ods (e.g., [5, 6]) compute camera motion and scene structure directly from pixel intensities by comparing brightness across consecutive frames. Feature-based methods (e.g., [16, 20, 22]) extract and match keypoints or corners across frames to infer motion. Despite their success, these approaches depend on handcrafted features and require manual processes to relate features to camera states. They also lack latent representations, limiting compatibility with learning-based models and hindering progress toward integration with modern AI systems.

**Hybrid Visual Odometry:** Hybrid methods combine learning techniques with traditional geometry-based VO to improve robustness and incorporate semantic understanding. These systems typically consist of two components: a front-end and a back-end. In the front-end, conventional feature pipelines are enhanced with learned features, often extracted via CNNs to estimate depth [31, 40] or via object detectors to provide semantic cues [2, 26–28, 35]. Scene segmentation methods [13, 25] further enrich the map with structural information. In the back-end, traditional optimization techniques like bundle adjustment are refined using learned priors [1, 30]. However, hybrid systems still follow the geometric VO pipeline and require manual efforts to relate visual features to motion, limiting their interpretability and flexibility.

**Learning-based Visual Odometry:** Purely learning-based VO methods aim to estimate camera pose using data-driven approaches. Early work like [14] applied supervised learning to predict camera pose from RGB images, while [37] used recurrent networks to model temporal dynamics. More recent methods such as [18, 44] employ self-supervised learning using stereo image constraints, and [38] introduces an intrinsics layer for improved generalization. [33] demonstrates a fully differentiable architecture that mimics geometric VO. Despite these advances, many learning-based VO methods still incorporate geometric or temporal constraints and lack a clear and interpretable latent space, which makes integration with other learning-based systems difficult.

We draw inspiration from classical graph optimization tools [4, 17] and apply contrastive learning to address the limitations of existing VO methods. By organizing feature representations in a latent space and structuring their relationships with camera states through a label-ranking framework, our method enables a more interpretable and flexible alternative to traditional VO. This also promotes seamless integration with other learning-based systems, supporting broader applications in multimodal and cross-domain tasks.

## 3. Methodology

We propose **VOCAL** (*Visual Odometry via ContrAstive Learning*), a new framework that reformulates visual odom-

etry (VO) by combining Bayesian inference with representation learning. In this section, we first outline the high-level motivation, then present a Bayesian view of VO. We follow this with a label-ranking formulation and describe how contrastive learning captures the relationship between visual features and camera states. This design aligns the latent space with 3D motion, improving spatial understanding and interpretability.

### 3.1. High-Level Idea

Our method is driven by two core goals. First, we aim to establish a consistent relationship between visual observations and camera states in latent space. Second, we seek to create an interpretable latent representation that can integrate seamlessly with other learning-based systems.

Inspired by the human ability to recognize similar motion across diverse visual environments, we hypothesize that **visual inputs associated with similar camera states should be mapped to similar features in latent space—even if they originate from entirely different scenes** (Fig. 2).

To realize this, we adopt contrastive learning, which pulls similar samples closer and pushes dissimilar ones apart in latent space. By encouraging features from similar camera states to cluster, contrastive learning enables a flexible and spatially aligned representation that supports generalization and integration.

### 3.2. Bayesian Inference in Learning-based VO

We begin by examining the VO problem through a Bayesian lens. Traditionally, VO is posed as a conditional probability estimation. Let

$$X = \{x_1, x_2, \dots, x_N\}$$

denote a set of camera states and

$$Z = \{z_1, z_2, \dots, z_N\}$$

represent the corresponding observations, where  $N$  is the batch size. By Bayes' rule, VO can be expressed as:

$$P(X | Z) \propto P(Z | X) P(X). \quad (1)$$

Typically,  $P(\cdot)$  is assumed to follow a Gaussian distribution; however, directly maximizing  $P(X | Z)$  is challenging. As a workaround, many methods approximate  $P(Z | X)$  via reprojection error or pixel intensity constraints under Gaussian assumptions.

Though these methods perform well in bundle-adjustment-based optimizers, applying a Gaussian model directly to learning-based VO is problematic because optimizing model parameters during training hinders direct parameter modeling in Eq. (1). To address this issue, we reinterpret the training process by drawing parallels with

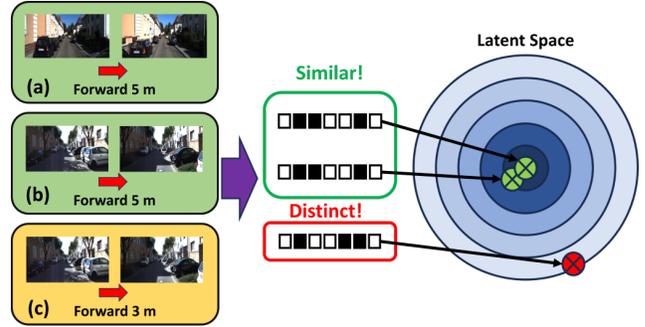


Figure 2. **High-Level Idea:** Panels (a) and (b) show visual inputs from different environments that share the same camera state (“Forward 5 meters”), whereas panels (b) and (c) depict inputs from the same scene but with different camera states (“Forward 5 meters” vs. “Forward 3 meters”). Just as humans can recognize the same motion regardless of environmental differences—and distinguish different motions even in similar scenarios—our approach uses contrastive learning to align features corresponding to similar camera states while separating those corresponding to different states.

conventional localization and mapping: we view backpropagation as inverting the observation model [34] and treat training as a mapping function:

$$P(m_\theta | X, Z) \propto P(Z | X, m_\theta) P(m_\theta | X), \quad (2)$$

where  $m_\theta$  denotes the learned model parameters that define the latent space. During inference, VO is reformulated as:

$$P(X | Z, m_\theta) \propto P(Z | X, m_\theta) P(X | m_\theta). \quad (3)$$

Despite this reformulation, estimating  $P(Z | X, m_\theta)$  remains challenging, as the Gaussian assumption does not always align with the requirements of learning-based VO. This discrepancy has led many methods to rely on geometric loss functions that do not naturally fit within a purely probabilistic framework, underscoring the need for a more theoretically grounded approach.

### 3.3. Label Ranking in Learning-based VO

To overcome the limitation, we reformulate learning-based VO as a *label-ranking* problem. In this formulation, camera states  $X$  serve as query instances, and observations  $Z$  are treated as labels. We adopt the Plackett–Luce model [24] to rank observations according to their corresponding camera states, providing an interpretable framework for organizing feature representations.

We define each ranked feature as:

$$f_{z_i} = m_\theta(X, z_i), \quad (4)$$

where  $f_{z_i}$  denotes the feature associated with camera state  $x_i$ , and  $m_\theta$  is the mapping function. Here, we assume that

$f_{z_i}$  is strictly positive. Under the Plackett–Luce model, the probability of a particular ranking order is given by:

$$P(z_1 > z_2 > \dots > z_N | X) = \prod_{k=1}^N \frac{f_{z_k}}{\sum_{j=k}^N f_{z_j}}. \quad (5)$$

To solve this ranking problem, we employ the *Rank-N-Contrast* (RNC) loss [42], which is specifically designed for continuous label regression tasks. Instead of treating each data point as a distinct class (as in contrastive learning for classification), RNC constructs positive/negative sample relationships based on the ordering of their queries, creating a ranked, continuous latent distribution that facilitates the estimation of the most likely camera states. The probability of each observation in the ranking is defined as:

$$P(f_{z_j} | f_{z_i}, S_{i,j}) = \frac{\exp(\text{sim}(f_{z_i}, f_{z_j})/\tau)}{\sum_{f_{z_k} \in S_{i,j}} \exp(\text{sim}(f_{z_i}, f_{z_k})/\tau)}, \quad (6)$$

where  $\text{sim}(\cdot, \cdot)$  is a similarity function,  $d(\cdot, \cdot)$  denotes the distance between camera states, and

$$S_{i,j} := \{f_{z_k} | k \neq i, d(x_i, x_k) \geq d(x_i, x_j)\}$$

is the set of samples ranked higher than  $f_{z_j}$ , defining the ordering of camera states. The temperature  $\tau$  scales the similarity distribution to ensure stable training.

Through contrastive learning, the model is encouraged to bring feature pairs  $(f_{z_i}, f_{z_j})$  closer when their corresponding camera states  $(x_i, x_j)$  are similar, and to push apart those that are dissimilar (see Fig. 3). Formally,

$$\text{sim}(f_{z_i}, f_{z_j}) \propto \frac{1}{d(x_i, x_j)}. \quad (7)$$

Since  $f_{z_j}$  is obtained through this ranking process, it serves as a maximum likelihood estimate for the corresponding observation:

$$P(f_{z_j} | f_{z_i}, S_{i,j}) \propto P(z_j | X, m_\theta). \quad (8)$$

The ranking-based probability model is linked to learning-based VO, enabling VOCAL to organize observations by their camera states.

### 3.4. Visual Odometry via Contrastive Learning

Having reformulated VO as a label-ranking task (Eq. (4)), we now describe how contrastive learning is incorporated into our framework. In our setup, each observation  $z_i$  consists of optical flow between two consecutive images, and each camera state  $x_i$  is represented by the corresponding relative camera motion. Our objective is to extract discriminative features from these observations and rank them according to their associated camera states.

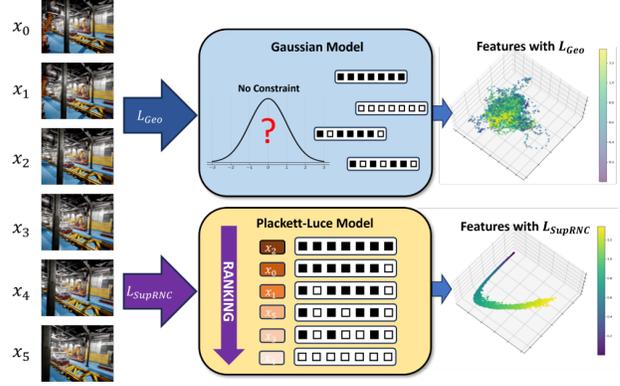


Figure 3. **Gaussian Model vs. Plackett–Luce Model in Learning-based VO:** Most learning-based VO methods rely on geometric loss functions derived from a Gaussian assumption, limiting their alignment with the learning process. In contrast, VOCAL adopts the Plackett–Luce model and employs the *Supervised Rank-N-Contrast* loss ( $L_{SupRNC}$ ) to rank feature representations according to their respective camera states, providing greater flexibility and a clearer interpretation of spatial relationships.

We implement contrastive learning within the mapping function  $m_\theta$ , which takes as input  $z_i$  along with its augmentation  $\text{aug}(z_i)$ . By organizing camera states into a ranking set  $S_{i,j}$ , the label-ranking problem is reformulated as:

$$m_\theta(S_{i,j}, \{z_i, \text{aug}(z_i)\}) \rightarrow f_{z_i}, f_{z_i}^{\text{aug}}. \quad (9)$$

We then apply the *Rank-N-Contrast* (RNC) loss to obtain a ranked feature  $f_{z_i}$ . In practice, we observed that the RNC loss is highly sensitive to the temperature  $\tau$ , sometimes leading to *dimensional collapse* [12], where the learned features become overly sparse. To mitigate this, we introduce an  $L_1$  regularization term weighted by  $\lambda$  and train both the encoder and decoder jointly. The modified loss for a single feature,  $l_{SupRNC}^i$ , is defined as:

$$l_{SupRNC}^i = \quad (10)$$

$$\frac{1}{2N-1} \sum_{\substack{j=1 \\ j \neq i}}^{2N} \left[ -\log \left( \frac{\exp(\text{sim}(f_{z_i}, f_{z_j})/\tau)}{\sum_{f_{z_k} \in S_{i,j}} \exp(\text{sim}(f_{z_i}, f_{z_k})/\tau)} \right) \right] \quad (11)$$

$$+ \lambda L1(x_i, \hat{x}_i). \quad (12)$$

Here,  $N$  is the batch size,  $\tau$  controls the sharpness of the similarity distribution, and  $\hat{x}_i$  denotes the ground-truth camera state. The overall loss is:

$$L_{SupRNC} = \frac{1}{2N} \sum_{i=1}^{2N} l_{SupRNC}^i. \quad (13)$$

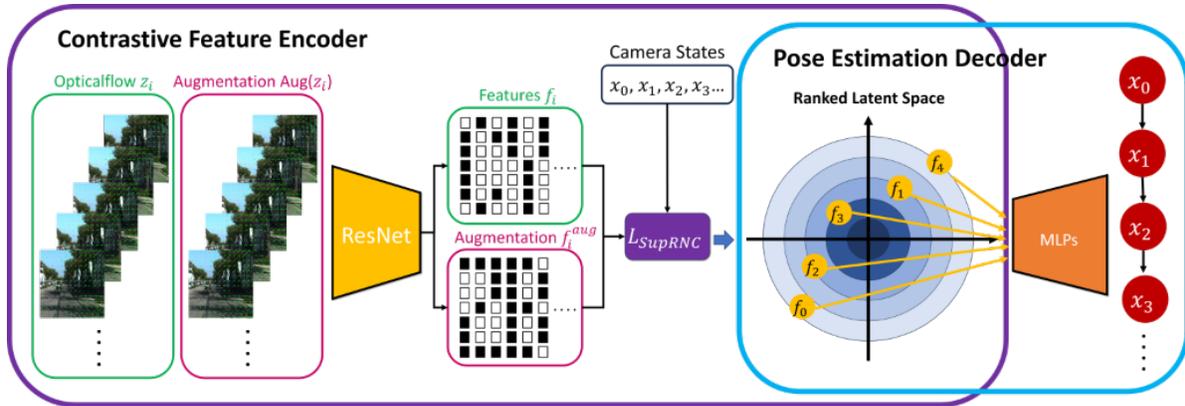


Figure 4. **System Overview:** Our system comprises two main components: a Contrastive Feature Encoder and a Pose Estimation Decoder. The encoder processes optical flow and its augmented variants using a ResNet to generate observation feature vectors. These features are then fed into the Pose Estimation Decoder, which employs Multi-Layer Perceptrons (MLPs) to estimate camera states. During training, the *Supervised Rank-N-Contrast* loss ( $L_{SupRNC}$ ) ranks the features based on camera states, yielding a spatially meaningful and interpretable latent space that facilitates the estimation of the most likely camera states.

By employing the *Supervised Rank-N-Contrast* loss ( $L_{SupRNC}$ ), our model learns features that are both highly continuous and systematically ranked by their underlying camera states. This structured organization ensures that visual inputs with similar camera states are closely aligned in the latent space—even when these states originate from different scenes—improving the model’s spatial understanding of the relationship between observations and camera states. Moreover, by regulating the feature distribution, VOCAL generates an interpretable latent space that serves as a critical bridge for integrating with other learning-based models, ultimately increasing the flexibility of visual odometry.

## 4. Experiments

**Implementation Details:** We represent each camera state using six degrees of freedom (6-DoF):  $\{x, y, z, \text{roll}, \text{pitch}, \text{yaw}\}$ . To handle these dimensions, we employ six separate encoder-decoder networks (see Fig. 4), each responsible for regressing a specific pose component. For the encoder, we use ResNet-18 [10] to process the optical flow between consecutive image pairs and generate a 512-dimensional feature vector  $\mathbf{f} \in \mathbb{R}^{512}$ . Each decoder is a three-layer Multi-Layer Perceptron (MLP) that takes  $\mathbf{f}$  as input and predicts one of the six pose parameters.

Prior to training, we compute the optical flow for each pair of consecutive images using the Gunnar-Farneback method. We set the parameters in the OpenCV function as follows:  $\text{pyr\_scale} = 0.5$ ,  $\text{levels} = 3$ ,  $\text{winsize} = 15$ ,  $\text{poly\_n} = 5$ , and  $\text{poly\_sigma} = 1.2$ . The resulting flow maps are then cropped to  $224 \times 224$ . We apply Gaussian noise augmentation with mean  $\mu = 0.0$  and standard deviation  $\sigma = 0.05$ .

During training, the *Supervised Rank-N-Contrast* loss

( $L_{SupRNC}$ ) is used to rank encoder-generated features based on their corresponding camera states, resulting in a ranked, interpretable, and continuous latent feature distribution. The decoder then regresses the final pose from this latent representation. For the ranking policy in  $L_{SupRNC}$ , we use the negative  $L_2$ -norm as the similarity function  $\text{sim}(\cdot, \cdot)$ , the  $L_1$ -norm as the distance function  $d(\cdot, \cdot)$ , a temperature factor  $\tau = 2.0$ , and a regularization weight  $\lambda = 2.0$ .

**Datasets:** We conduct our experiments on the KITTI dataset [8], a standard benchmark for visual odometry and simultaneous localization and mapping. Following [37], we train our model on sequences 00, 02, 08, and 09, and evaluate it on sequences 03, 04, 05, 06, 07, and 10.

**Evaluation Metrics:** We evaluate VOCAL’s performance using three primary assessments:

- **Feature Ranking:** We measure how well the ranked features align with their corresponding camera states by computing Spearman’s rank correlation [29] and Kendall’s rank correlation [15]. Additionally, we visualize the latent-space feature distribution to verify that  $L_{SupRNC}$  effectively ranks features according to camera states.
- **Generalization Capability:** We assess the model’s generalization by training on different fractions of the training set. Specifically, we partition the training data into sizes of 0.2, 0.4, 0.6, 0.8, and 1.0 of the total data, and then measure the Spearman correlation coefficients and VO metrics on the test set.
- **VO Performance:** We follow standard VO evaluation protocols by measuring the average translational RMSE drift ( $t_{rel}$ , in %) and the average rotational RMSE drift ( $r_{rel}$ , in  $^\circ/100$  m) on trajectory segments of 100–800 m.

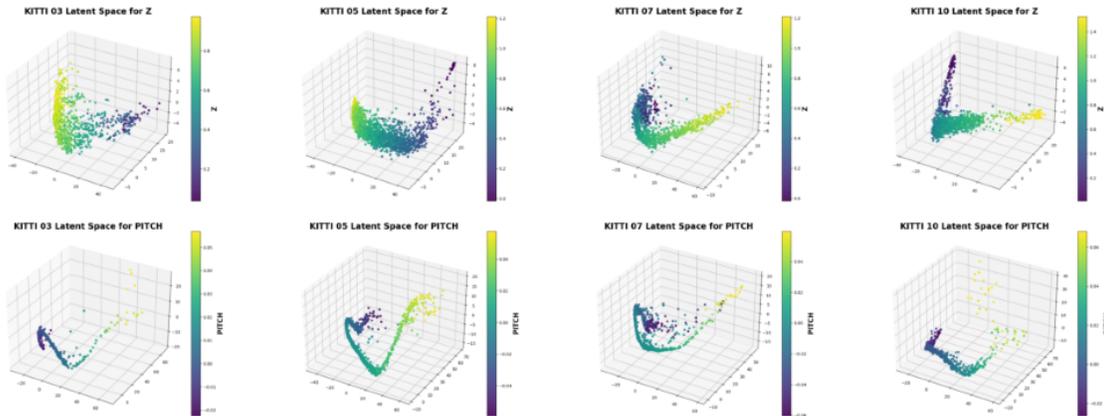


Figure 5. **Feature Distribution in Latent Space:** Lighter features (yellow) correspond to larger camera motions, while darker features (purple) denote smaller motions. The results, based on KITTI sequences 03, 05, 07, and 10, reveal a continuous gradient from lighter to darker features, highlighting the effective ranking of features according to their camera states.

Dimension	$x$		$y$		$z$	
Correlation	$r_s \uparrow$	$r_k \uparrow$	$r_s \uparrow$	$r_k \uparrow$	$r_s \uparrow$	$r_k \uparrow$
Seq 03	0.460	0.323	0.012	0.008	0.878	0.697
Seq 04	0.058	0.039	0.174	0.117	0.454	0.308
Seq 05	0.759	0.575	0.370	0.252	0.924	0.764
Seq 06	0.586	0.423	0.424	0.290	0.873	0.687
Seq 07	0.734	0.546	0.232	0.156	0.840	0.655
Seq 10	0.599	0.429	0.292	0.197	0.901	0.731

Dimension	$roll$		$pitch$		$yaw$	
Correlation	$r_s \uparrow$	$r_k \uparrow$	$r_s \uparrow$	$r_k \uparrow$	$r_s \uparrow$	$r_k \uparrow$
Seq 03	0.886	0.719	0.990	0.919	0.626	0.450
Seq 04	0.804	0.644	0.378	0.296	0.487	0.345
Seq 05	0.888	0.723	0.975	0.901	0.545	0.387
Seq 06	0.860	0.687	0.936	0.816	0.467	0.339
Seq 07	0.819	0.657	0.942	0.839	0.423	0.296
Seq 10	0.886	0.717	0.989	0.915	0.526	0.369

Table 1. **Correlations between Features and Camera States:** High correlation scores for  $z$  translation and  $pitch$  rotation reflect the dominant motion in the KITTI dataset. In contrast, the  $y$  and  $yaw$  dimensions—typically regarded as noise in this dataset—exhibit lower correlation scores for both  $r_s$  and  $r_k$ .

We compare our results with those of current state-of-the-art learning-based VO methods.

#### 4.1. Feature Ranking

We analyze the feature distribution produced by our contrastive feature encoder in the latent space. By leveraging the *Supervised Rank-N-Contrast* loss ( $L_{SupRNC}$ ), the learned features are expected to closely correlate with the ground-truth camera states. Table 1 reports the Spearman ( $r_s$ ) and Kendall ( $r_k$ ) rank correlation coefficients, where values approaching 1.0 indicate more effective ranking. Notably, the  $z$ -translation and  $pitch$ -rotation achieve high cor-

relation scores, reflecting the dominant motions along the  $z$ -axis and changes in  $pitch$  in the KITTI dataset. In contrast, the  $y$  and  $yaw$  dimensions, generally considered noise, exhibit lower correlation scores.

Fig. 5 illustrates the feature distribution for  $z$  and  $pitch$  (distributions for the remaining dimensions are provided in the supplementary material Sec. 7). In the figure, lighter colors (yellow) represent larger camera state values, while darker hues (purple) indicate smaller values. The continuous gradient from light to dark underscores the effective ranking of features by camera state, yielding an interpretable and flexible latent representation that also supports the estimation of the most likely camera states.

#### 4.2. Generalization Capability

We assess VOCAL’s generalization capability by evaluating its performance with varying amounts of training data. Specifically, we partition the training set (sequences 00, 02, 08, and 09) into fractions of the total data (0.2, 0.4, 0.6, 0.8, and 1.0) and evaluate the resulting feature ranking performance on the test set using Spearman’s rank correlation coefficient ( $r_s$ ) as well as the VO metrics  $t_{rel}$  and  $r_{rel}$ .

Fig. 6 plots the rank correlation for all six degrees of freedom (DoF) as a function of the training data proportion. Notably, the  $z$  and  $pitch$  dimensions—representing the dominant motions in the KITTI dataset—exhibit gradually increasing  $r_s$  values as more training data becomes available, indicating that our method ranks features more effectively with additional data. In contrast, the  $y$  and  $yaw$  dimensions, which are more susceptible to noise in KITTI, do not converge regardless of the training set size. Additionally, Fig. 6 presents the performance of  $t_{rel}$  and  $r_{rel}$  on different test datasets, showing that our model performs better as the training data scale increases. Even with a rela-

Method	Training dataset	KITTI Sequences ( $t_{rel}/r_{rel}$ )					
		03	04	05	06	07	10
VISO2-M [7]	-	8.47/8.82	4.69/4.49	19.22/17.58	7.30/6.14	23.61/29.11	41.56/32.99
SfMLearner [44]	KITTI 00-08	12.56*/4.52*	<b>4.32*/3.28*</b>	12.99*/4.66*	15.55*/5.58*	12.61*/6.31*	15.25/4.06
GeoNet [41]	KITTI 00-08	19.41*/9.80*	10.81*/7.00*	22.68*/7.70*	9.90*/4.30*	9.82*/5.90*	23.90/9.04
DeepVO [37]	KITTI 00, 02, 08, 09	8.49/6.89	7.19/6.97	<b>2.62/3.61</b>	5.42/5.82	<b>3.91/4.60</b>	8.11/8.83
TartanVO [38]	TartanAir (~40k)	-	-	-	<b>4.72/2.96</b>	<b>4.32/3.41</b>	<b>6.89/2.73</b>
VOCAL (ours)	KITTI 00, 02, 08, 09	<b>4.60/1.99</b>	5.18/ <b>1.51</b>	<u>5.45/2.51</u>	<u>5.11/6.06</u>	12.13/4.79	<u>7.72/7.61</u>

Table 2. **Visual Odometry Results ( $t_{rel}/r_{rel}$ ) on KITTI Sequences:**  $t_{rel}$  denotes the average translational RMSE drift (in %) over trajectory segments of 100–800 m, and  $r_{rel}$  denotes the average rotational RMSE drift (in  $^{\circ}/100$  m) over trajectory segments of 100–800 m. An asterisk (\*) marks training data (which may be overfitted).

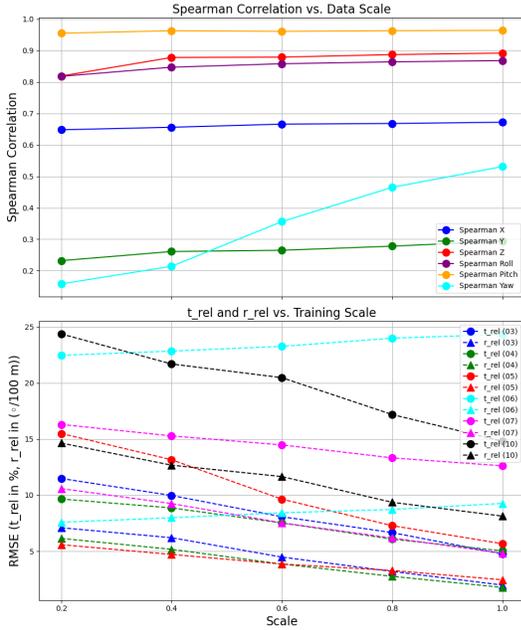


Figure 6. **Correlation and VO Metrics vs. Data Scale:** While most learning-based VO methods require extensive training data to achieve high performance, VOCAL converges and delivers competitive results even with limited datasets. As the amount of training data increases, our method achieves higher Spearman rank correlation and lower  $t_{rel}$  and  $r_{rel}$  values, indicating that the visual features are effectively ranked according to camera states. This outcome demonstrates VOCAL’s flexibility and provides a clearer interpretation of spatial relationships.

tively small training set, VOCAL achieves competitive performance compared with other methods (as will be detailed in the next section). This early-stage performance is due to our method’s ability to directly capture the relationship between visual features and camera states through label ranking.

In contrast, many existing learning-based VO models struggle during early training stages and require extensive training to converge due to a lack of interpretability. By comparison, VOCAL not only benefits from larger datasets

but also maintains strong performance with limited data, underscoring its potential for effective generalization.

### 4.3. VO Performance

We evaluate VOCAL’s visual odometry performance on the KITTI dataset by training on sequences 00, 02, 08, and 09 and testing on sequences 03–07 and 10. Table 2 compares translation and rotation accuracy against other VO methods, with the best results in **bold**, the second-best in underline, and an asterisk (\*) marking the training data (potentially overfitted). References to the original papers, [43], and [45] are provided for these comparisons.

VISO2-M [7], a classic geometry-based method, is efficient and low-cost but falls short of modern learning-based approaches in accuracy. SfMLearner [44] is an end-to-end model for single-view depth and pose estimation. However, it relies heavily on geometric and temporal constraints and performs worse than VOCAL even on its own training data (trained on KITTI sequences 00–08). GeoNet [41], also trained unsupervised on sequences 00–08, depends on pretrained depth and flow models; despite these additional resources, it underperforms VOCAL and introduces extra complexity. DeepVO [37] uses a similar training protocol but depends on an RNN architecture that requires sequential constraints and multiple-frame optimization. In contrast, VOCAL processes a single optical flow input (only two frames) while achieving comparable or superior results. TartanVO shows strong performance on sequences 06 and 07 but demands a large training corpus—hindering reproducibility—and offers limited flexibility for broader integration with learning-based models.

Our experimental results show that VOCAL achieves the best translation performance on sequence 03 and ranks second on sequences 05, 06, and 10. For rotation accuracy, VOCAL attains the best performance on sequences 03, 04, and 05. Notably, although our training data is relatively smaller than that used in [41, 44], VOCAL still performs competitively, even outperforming those methods on training sequences 03–05. These findings demonstrate that VOCAL not only provides an interpretable latent representation but also delivers competitive VO performance.

## 5. Conclusion

We presented **VOCAL** (Visual Odometry via ContrAstive Learning), a novel framework that reinterprets visual odometry as a label-ranking problem by integrating Bayesian inference with contrastive representation learning. By organizing visual features in a latent space according to underlying camera states, VOCAL produces an interpretable, continuous representation that effectively captures spatial relationships.

Our experiments on the KITTI dataset show that VOCAL achieves competitive translation and rotation accuracy compared to other learning-based VO methods, while avoiding reliance on geometric or temporal constraints. Unlike conventional approaches that depend on handcrafted features or sequential inputs, VOCAL operates on optical flow from only two frames, enabling a flexible and efficient model that generalizes well across different scenarios.

Moreover, VOCAL serves as a bridge between visual odometry and other learning-based models. Traditional Bundle-Adjustment-based methods offer little control over the latent space, and their optimization logic rarely generalizes to other learning-based frameworks. In contrast, our approach produces a visualized, interpretable, and ranked latent representation that can be seamlessly integrated with additional sensor modalities and learning-based systems, thereby enhancing performance in multimodal applications.

Beyond visual odometry, VOCAL offers a new perspective on learning spatial and conceptual relationships across modalities. The label-ranking principle in latent space can be extended to various space-to-space or concept-to-concept tasks, such as aligning visual inputs with camera states, associating 2D images with text or 3D reconstructions, or linking sensor data (e.g., depth, IMU, GPS) with high-level representations (e.g., language, multi-agent states, spatial concepts). Compared to classical graph-based approaches, our framework provides new insights and broader potential for advancing spatial intelligence.

In summary, VOCAL advances the state of visual odometry, provides a foundation for flexible and interpretable multimodal systems, and opens new directions for solving broader space-to-space problems. Future work will focus on incorporating additional sensor modalities, refining latent representations through advanced learning strategies, and validating the approach in more diverse real-world and conceptual domains. We believe the ideas behind VOCAL have the potential to drive significant progress in spatial intelligence—and beyond.

## References

[1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro,

Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 2

[2] Sean L. Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J. Pappas. Probabilistic data association for semantic slam. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1722–1729, 2017. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 2

[4] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. *Georgia Institute of Technology, Tech. Rep*, 2:4, 2012. 2

[5] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsdslam: Large-scale direct monocular slam. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13*, pages 834–849. Springer, 2014. 2

[6] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *CoRR*, abs/1607.02565, 2016. 2

[7] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968, 2011. 7

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[9] Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression, 2022. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019. 2

[12] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *CoRR*, abs/2110.09348, 2021. 4

[13] Masaya Kaneko, Kazuya Iwami, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Mask-slam: Robust feature-based monocular slam by masking using semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 2

[14] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 2

[15] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 5

[16] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007. 2

- [17] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613. IEEE, 2011. 2
- [18] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. *CoRR*, abs/1709.06841, 2017. 2
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2
- [20] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. Orbslam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2
- [21] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Rei-ichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Felipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Valone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 1, 2
- [22] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *CoRR*, abs/1708.03852, 2017. 2
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2
- [24] Label Ranking. Label ranking methods based on the plackett-luce model. *Poster presented at ICML*, page 27, 2010. 3
- [25] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696. IEEE, 2020. 2
- [26] Martin Rünz and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. *CoRR*, abs/1804.09194, 2018. 2
- [27] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, and Richard Newcombe. Frodo: From detections to 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

- [28] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, and Andrew J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2
- [29] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987. 5
- [30] Tetsuya Tanaka, Yukihiro Sasagawa, and Takayuki Okatani. Learning to bundle-adjust: A graph network approach to faster optimization of bundle adjustment for vehicular slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6250–6259, 2021. 2
- [31] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction. *CoRR*, abs/1704.03489, 2017. 2
- [32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gura, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Meray, Martin Bauml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anas White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maroon, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruiho Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomnech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sbastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogoziska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Gimnez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yin Chang, Paul Komarek, Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphal Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller,

Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimentko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussebot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tshilas, Arpi Vezar, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chandra, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezedegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinta Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna,

Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bølle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjit Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Raga Kotikalapudi, Chalance Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Pawel Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal,

Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajt Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaime Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Lu-wei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikolaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHosseini Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong

Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uribe, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Lohar, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Ram-mohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman,

- Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulic, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listfk, Mathias Carlen, Jan van de Kerkhof, Marcin Píkus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanian, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. [2](#)
- [33] Zachary Teed and Jia Deng. DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras. *CoRR*, abs/2108.10869, 2021. [2](#)
- [34] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005. [3](#)
- [35] Jingwen Wang, Martin Rünz, and Lourdes Agapito. Dpslam: Object oriented slam with deep shape priors. In *2021 International Conference on 3D Vision (3DV)*, pages 1362–1371, 2021. [2](#)
- [36] Jiayun Wang, Stella X. Yu, and Yubei Chen. Trajectory regularization enhances self-supervised geometric representation. *CoRR*, abs/2403.14973, 2024. [2](#)
- [37] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. *CoRR*, abs/1709.08429, 2017. [2](#), [5](#), [7](#)
- [38] Wenshan Wang, Yaoyu Hu, and Sebastian A. Scherer. Tartanvo: A generalizable learning-based VO. *CoRR*, abs/2011.00359, 2020. [2](#), [7](#)
- [39] Yang Xiao, Yuming Du, and Renaud Marlet. Posecontrast: Class-agnostic object viewpoint estimation in the wild with pose-aware contrastive learning. *CoRR*, abs/2105.05643, 2021. [2](#)
- [40] 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 (CVPR)*, 2020. [2](#)
- [41] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. *CoRR*, abs/1803.02276, 2018. [7](#)
- [42] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression, 2023. [2](#), [4](#)
- [43] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, Ravi Garg, and Ian Reid. Df-vo: What should be learnt for visual odometry?, 2021. [7](#)
- [44] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. *CoRR*, abs/1704.07813, 2017. [2](#), [7](#)
- [45] Yuliang Zou, Pan Ji, , Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. In *European Conference on Computer Vision*, 2020. [7](#)