OCTOSENSE: SELF-SUPERVISED LEARNING FOR MULTIMODAL ROBOT PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present OctoSense, an open-source sensor platform with stereo RGB and event cameras, LiDAR, a thermal camera, an inertial measurement unit, RTK-corrected global positioning system, and proprioception (CAN bus data from a car). The eponymous OctoSense dataset contains 59 hours of time-synchronized driving data across different types of environments at different times of the day, including situations with highly degraded sensors. We demonstrate multi-modal self-supervised learning using such real-world robotics data, where sensors have different representations, frequencies, latencies and noise. Our approach, a "late-fusion" masked autoencoder, (i) uses modality-specific tokenizers to account for different spatiotemporal characteristics of these sensors, and (ii) caches modality-specific tokens at inference time to process new measurements as they come. This architecture (i) is fast (6.68 ms and 112 ms on NVIDIA 5090 and Orin NX respectively, to compute the representation), (ii) performs better than existing image-only foundation models on tasks such as estimation of optical flow, depth, semantic segmentation, and ego-motion (translation, rotation, and steering angle), and (iii) predicts robustly at nighttime or in situations where sensory data is degraded. Links to the dataset, code, and supplementary videos will be provided with the camera-ready version.

1 INTRODUCTION

Vision-based foundation models trained using self-supervised learning (SSL) such as DINO (Caron et al., 2021), SigLIP (Zhai et al., 2023; Tschannen et al., 2025), and Hiera (Ryali et al., 2023) have transformed robotics. However, in real-world robotic settings, no single sensor suffices (Liu et al., 2024). Global-shutter cameras work poorly in situations with low light, high dynamic range, and rapid motion. LiDAR provides accurate but very sparse depth at low frame rates, with poor semantics. Different sensors have different sensing modalities, angular resolutions, data types and information rates. They respond differently to weather degradation, noise, and packet loss. And they also fail in different ways. Effective robot perception requires representations that are resilient to these failures. This is perhaps why the Global Robotics Technology Roadmap 2025–2035 (Christensen, 2026) notes: "The defining enabling technology across all field applications is multi-sensor fusion robust to environmental degradation—the problem that keeps the most capable field robots indoors". While there is a growing literature on data from depth (Bachmann et al., 2022) or inertial (Girdhar et al., 2023) sensors, we do not have SSL methods that work with all typical robotic sensors. Our contributions towards addressing this gap are as follows.

1. OctoSense hardware and dataset. A central challenge in multi-modal robot representation learning is the lack of large datasets with diverse sensors. To address this gap, we develop an open-source hardware platform called OctoSense that provides time-synchronized data from stereo RGB and event cameras, a thermal camera, LiDAR, an inertial measurement unit (IMU), RTK-corrected GPS, and proprioception information from the CAN bus of a vehicle. We release 59 hours of data from OctoSense. This is one of the largest datasets of its kind. It covers urban, suburban, and rural environments across a mix of highway, residential, and urban driving; recordings span sunrise, daytime, sunset, and nighttime conditions, and include situations where sensors are degraded due to sunflares, dropped packets, etc.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. Multi-modal masked autoencoder. We account for different spatiotemporal characteristics of LiDAR, event cameras and IMUs to develop modality-specific tokenizers which feed a shared transformer-based masked autoencoder (MAE). At inference time, we cache modality-specific features within a temporal window and only process new measurements. This makes it quite fast (6.68 ms and 112 ms on NVIDIA 5090 and Orin NX respectively, for computing the representation).

3. Performance on downstream tasks. On tasks such as estimation of optical flow, depth, semantic segmentation and ego-motion (relative translation and rotation, linear and angular velocity, and steering position), we show that our multi-modal MAE is significantly better than existing image-only SSL approaches. Depth error improves by about 1.65 m and optical flow end-point error by about 7.16 pixels during daytime; segmentation is about on par. Our approach can perform ego-motion estimation very well. Improvements are larger in nighttime data and when sensors are degraded.

2 Related Work

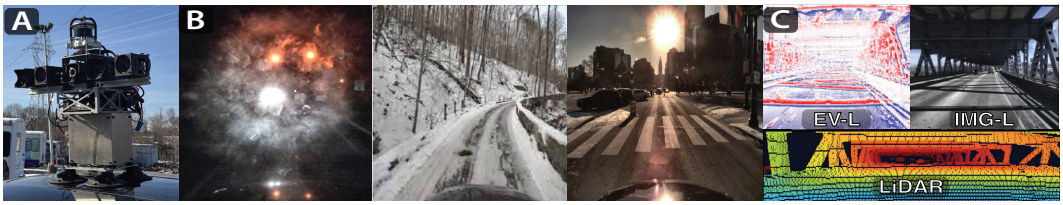
SSL has been instrumental in building highly general representations trained on unlabeled data. These techniques have led to visual foundation models such as DINO (Caron et al., 2021; Oquab et al., 2025; Siméoni et al., 2025), Perception Encoder (Bolya et al., 2025), SigLIP (Zhai et al., 2023; Tschannen et al., 2025), SAM (Ravi et al., 2025) that are the state of the art in computer vision today. There are a number of different pre-training objectives, such as self-distillation (Caron et al., 2021), masked autoencoders (He et al., 2022; Xie et al., 2022; Devlin et al., 2019) and contrastive alignment (van den Oord et al., 2018; Chen et al., 2020). SSL has also been extended beyond image modalities, e.g., to event cameras (Cao et al., 2026; Klenk et al., 2022; Yang et al., 2024; Das et al., 2025), depth data/point clouds (Patel et al., 2026; Pang et al., 2022; Yu et al., 2022; Wang et al., 2021a; Xie et al., 2020), thermal cameras (Munir et al., 2021; Zürn, 2024) and IMUs (Narayanswamy et al., 2025; Xu et al., 2021). These existing efforts have demonstrated strong results on individual sensors. Our work builds upon a number of techniques and best practices in this literature, but our goal is to learn a fused representation from multiple sensors that inherits their individual strengths.

Multi-modal SSL has employed two key ideas (Zong et al., 2023). The first, contrastive alignment, anchors each modality to a shared embedding, e.g., CLIP (Radford et al., 2021) and SigLIP for vision-language models, ImageBind (Girdhar et al., 2023) for multiple modalities. Recent variants for images and text such as SigLIP 2 (Tschannen et al., 2025) and Perception Encoder (Bolya et al., 2025) observe that these embeddings are not sufficiently sensitive to local image structures (Wang et al., 2021b) and improve this aspect using masking and distillation objectives. Our approach is based on the second key idea, namely masked multi-modal modeling. MultiMAE (Bachmann et al., 2022) uses a Dirichlet masking strategy for multiple input modalities, and 4M (Mizrahi et al., 2023) builds upon this to use modality-specific tokenizers (Bao et al., 2022; Lu et al., 2023; 2024). Some works on driving such as UniM^{2}AE (Zou et al., 2024) and CALICO (Sun et al., 2024) combine cameras and LiDAR to perform tasks such as bird's eye view (BEV) map segmentation. In this paper, we extend masked multi-modal modeling beyond image and text modalities to data from typical robotic sensors, and beyond the camera-LiDAR BEV setting to dense perception and egomotion tasks.

3 OctoSense Hardware and Dataset

OctoSense has eight sensors: stereo RGB and event cameras, LiDAR, a thermal camera, an inertial measurement unit (IMU) that measures accelerations, angular velocities, magnetic fields, and barometric pressure, RTK-corrected global positioning system (GPS), and the proprioception state. See the picture in Fig. 1A. Fig. 1D shows that these sensors have very different data rates, bandwidth requirements and resolutions. OctoSense has a desktop-class CPU for onboard computation. A 24 V battery provides about one hour of operation per charge. Sec. C provides a bill of materials.

Why a new dataset? No existing dataset spans this diversity of sensing modalities, event, RGB, thermal, LiDAR, inertial, GNSS, and proprioception, together with dense labels. The closest, ECMD (Chen et al., 2023), has event, RGB, and thermal cameras, LiDAR, IMU and GPS but no CAN bus and it supports only ego-motion; it also only has 3.4 hrs of data (Fig. 1E, Table S1). The size of our dataset is uncommon but necessary. Our experiments show that multi-modal pretraining performs better than existing vision-only foundation models, beyond about 13 hrs of data (see Sec. 5 and



D

Sensor	Model	Output	Rate	Notes
Event (2×)	SilkyEV VGA	640×480, 15 µm, 120 dB DR	50 MEv/s max	74.6°×59.5°, 390–750 nm filter
RGB (2×)	FLIR Blackfly S	1920×1456, 4.5 µm, 70 dB DR	100 Hz	68.5°×54.7°, 670 nm IR cut
Thermal	FLIR A35	320×256, 17 µm, 7.5–13 µm	50 Hz	45°×35°
LiDAR	Ouster OS1-64 Rev 7	2048×64, 1.3 Mpts/s, 865 nm	10 Hz	360°×42.4°, 170 m range
IMU/AHRS	VectorNav VN-100T-CR	accel., gyro., mag., baro.	400 Hz	hardware-triggered
GNSS/RTK	uBlox ZED-F9P	L1/L2 multi-band RTK	5 Hz	NTRIP via cellular
Vehicle CAN	2021 Mazda CX-5	speed, steer, wheels, pedal, brake	50–100 Hz	FSC CAN

E

Dataset	Duration / Dist.	Sensors	Tasks
KITTI (Geiger et al., 2012)	6 h	RGB, Gray., LiDAR, IMU, GPS	Depth, flow, det., track., seg., egomotion
nuScenes (Caesar et al., 2019)	5 h 33 min	RGB, LiDAR, Radar, IMU, GPS, CAN	3D det., track., seg., egomotion
Waymo Perc. (Sun et al., 2019)	6 h 24 min	RGB, LiDAR	2D/3D det., track., seg., egomotion
NVIDIA PAI (NVIDIA, 2025)	1700 h	RGB, LiDAR, Radar	3D det., track., egomotion, reasoning
DSEC (Gehrig et al., 2021)	53 min	Events, RGB, LiDAR, IMU	Seg., disparity, flow, 2D det.
MVSEC (Zhu et al., 2018)	1 h 08 min	Events, Gray., LiDAR, IMU	Depth, flow, egomotion
M3ED (Chaney et al., 2023)	3 h 46 min	Events, RGB, Gray., LiDAR, IMU, GPS	Seg., depth, egomotion
DDD17/20 (Hu et al., 2020)	51 h	Events, Gray.	Steering angle
OctoSense (Ours)	**59 h / 2474 km**	Events, RGB, IR, LiDAR IMU, GPS, CAN	Depth, flow, segmentation egomotion, steering angle

Figure 1: A) OctoSense hardware mounted on a car. **B)** Degraded sensors: nighttime glare, snow, and sun flare. **C)** Time-synchronized events, RGB, and LiDAR data. **D)** Sensors in OctoSense, which differ widely in their data rate and information structure, full specifications in Table S3. **E)** OctoSense is the largest multi-modal robotics dataset with event cameras (59 h, 2474 km) and, alongside the NVIDIA Physical AI dataset, it is among the largest driving datasets with dense sensory data, full comparison in Table S1.

Fig. S17). This is more than three times larger than the largest existing event-inclusive multi-modal dataset (M3ED). DDD20 (Hu et al., 2020) (51 hr) and Prophesee 1MP (Perot et al., 2020) (15 hr) are also large datasets, but they do not contain other sensor modalities.

Engineering challenges and our solutions. We had to address a large number of engineering issues for recording, processing, and utilizing these robot perception data for building foundation models. For example, we perform online adaptation of exposure and sensor gain for RGB cameras and use a bandpass filter on the event camera to block infrared interference from the LiDAR. We used insulated foam around the event cameras to prevent corrupted timestamps in cold weather and soft-mounting sensors to suppress LiDAR packet loss and IMU noise from vibrations. We spliced the car's camera connector to access the CAN bus. Sec. D gives more details of these modifications.

Many sensor drivers required substantial modifications, which are discussed in Sec. E. Composing drivers and data recorder as ROS 2 Zenoh composable nodes gives zero-copy transfer, which allows us to record those sensors at native rates without dropping any data. We built a web app to monitor system health during data collection, specifically to track per-sensor data rates and timestamps, and to start/stop data recording during real-world experiments. OctoSense produces ∼1.7 GB/s of data. Although this is within the advertised write speeds of modern SSDs, the resulting dataset quickly becomes prohibitively large to store, at roughly 6 TB per hour of recording. We therefore developed drivers that directly record compressed sensor streams, e.g., LiDAR/event packets and H.265 (ITU-T, 2026) video without B frames for RGB and thermal cameras. This reduces the required bandwidth, by 21×, to a much more manageable 78.7 MB/s. For details on spatial calibration see Sec. D and temporal synchronization see Sec. F.

This first release of OctoSense data covers urban, suburban, and rural environments across a mix of highway, residential, and city driving on Long Island and in Philadelphia, see Fig. S5. Recordings

span sunrise, daytime, sunset, and nighttime conditions, with most sequences lasting for about 10 minutes. Altogether, OctoSense contains 59 hours and 8.4 TB of time-synchronized data. Fig. 1B shows some examples. Fig. 1E shows that OctoSense is one of the largest datasets of its kind; a larger comparison table against many more datasets is provided in Table S1. For exploring and investigating the data using natural language, we used Gemma 4 (Google DeepMind, 2026) to caption three representative frames within each 5-second window, embedded the caption with Qwen3-Embedding-8B (Zhang et al., 2025), and indexed embeddings with FAISS (Douze et al., 2024). See Sec. G for an example.

Ground-truth for odometry, depth, optical flow, and semantic segmentation tasks. We use a LiDAR-inertial odometry method named RKO-LIO (Malladi et al., 2026) to estimate the pose of the platform in $\mathrm{SE}(3)$, along with linear and angular velocities. This method also deskews each 360° LiDAR scan to transform points as if they were captured at a single timestamp, which is very useful for obtaining accurate depth. We accumulate points across 61 consecutive LiDAR scans (a window of 6 sec around a reference time at 10 Hz) to compute relatively dense ground-truth depth. We handle independently moving objects separately. We use YOLO26-medium (Sapkota et al., 2025) on the rectified RGB image closest to each LiDAR scan and remove points that correspond to vehicles or pedestrians. Similarly, if some pixels have multiple depth measurements after projecting the accumulated points onto the rectified RGB image, we keep the minimum depth (this respects disocclusions). In the accumulated point cloud, we prune 3D voxels that are not observed by at least $N = 4$ different scans. We use the odometry pose and LiDAR-camera extrinsics and intrinsics to project the accumulated points into a future RGB frame to compute per-pixel displacement as the optical flow induced by ego-motion. We obtain segmentation pseudo-labels by running a recent segmentation model, EoMT (Kerssies et al., 2025), trained on Cityscapes on the rectified RGB images.

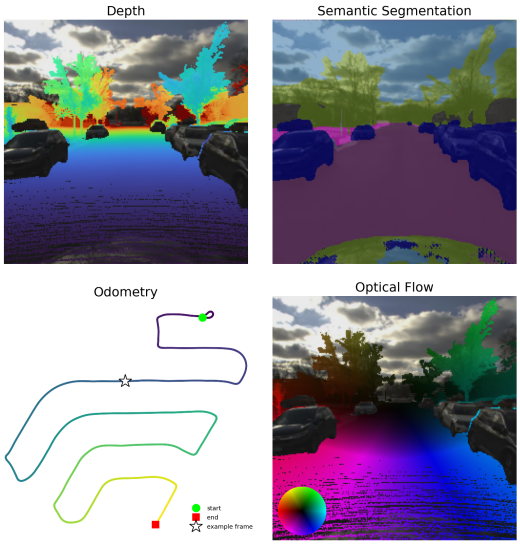


Figure 2: Ground-truth of different downstream tasks in OctoSense

4 METHODS

We next discuss a multi-modal MAE architecture for sensors in OctoSense. These sensors have different spatiotemporal characteristics: dense 2D arrays for RGB, a high-frequency point process for the event camera, an unordered and sparse point cloud for LiDAR and a rapidly varying multi-channel time series for the IMU. This heterogeneity makes it challenging to build a fused representation. Our approach has four key facets: (i) low-level input representations for these diverse sensory modalities, (ii) appropriate per-modality tokenization to ensure effective reconstruction, (iii) an MAE architecture (encoder and decoder) that combines multiple modalities into a shared representation, and (iv) task-specific probes that make predictions from that representation.

Input representation for different sensors We undistort and rectify RGB images before building ViT-like patches (Dosovitskiy et al., 2021). Event cameras produce an event stream $(t_i, u_i, p_i)_{i\in\mathbb{N}}$ for time t_i and pixel coordinates $u_i \in \mathbb{Z}^2$ and polarity p_i indicating the sign of intensity change at the pixel since the last such event (Hawkes & Simonpieri, 1974; Gallego et al., 2022). Events have a very high frequency, bandwidth and spatiotemporal noise, which makes it difficult to incorporate them directly into MAE architectures. After undistortion and rectification, we use a spatiotemporal filter to remove isolated events (Delbruck, 2008). We then use a bank of leaky integrators at each pixel with different bandwidths to obtain a multi-channel image-like representation (Gerstner et al.,

2014; Lagorce et al., 2017; Bisulco et al., 2021). See Sec. H.1 for some more mathematical details.[1] The LiDAR point cloud gives a 64 $\times$ 2048 range image which is cropped to a 90° field of view to get a 64 $\times$ 512 range image. We convolve accelerometer and gyroscope data within a 1.6 sec time window and tokenize it using cross-attention pooling (Jaegle et al., 2021), Sec. H.2 describes the hyper-parameters.

Training modality-specific tokenizers MAEs for RGB data are typically trained to minimize the mean squared error (MSE) between the predicted pixel intensities and those in the original RGB image. This leads to poor training in multi-modal settings because different modalities have different spatiotemporal statistics and token counts. Works such as BEiT (Bao et al., 2022), Unified-IO (Lu et al., 2023) and 4M (Mizrahi et al., 2023) mitigate this issue by training a tokenizer where each patch is represented by a token supported on a finite set. Then, instead of minimizing the MSE on pixel intensities, these methods minimize the discrepancy between the predicted token for masked patches and the ground-truth token of that patch. We follow this approach. We however compute the loss on all patches rather than only the masked ones, which has been found to be important for dense prediction tasks (Mur-Labadia et al., 2026). Each tokenizer uses finite scalar quantization (FSQ) (Mentzer et al., 2024).

While our overall approach above draws on established tokenization methods, each sensory modality requires a tailored implementation. Modality-specific tokenizers are built using a transformer encoder with RoPE (Su et al., 2024) in two dimensions for image-like modalities of RGB, event filter-bank channels, LiDAR range image, and in one dimension for IMU data. The decoder is a hybrid transformer-convolutional architecture which we found to produce sharper reconstructions, see Fig. S9. Tokenizers are trained to minimize the mean ℓ_1 loss with modality-specific modifications. For LiDAR, in addition to ℓ_1 on log-depth, we use a depth-validity mask with binary-cross-entropy loss and a SiLog loss (Eigen et al., 2014). The validity mask accounts for ray drops which produce holes in the range image (Manivasagam et al., 2020). For event cameras where the input is sparse, we minimize a weighted reconstruction loss on activations of the filter bank that balances active (event-containing) and inactive pixels. For the IMU, we regress upon the time series after removing the mean and applying a fixed normalization to the dataset. Sec. H gives more details.

Fusing individual modalities in the MAE The multi-modal MAE (Fig. 3) takes as input tokens from each sensor over a 1.4 sec window, with the exception of the IMU, which is tokenized over a slightly wider 1.6 sec window.[2] We subsample and downsample the OctoSense data: each datum consists of 8 uniformly spaced LiDAR scans within this window, 8 192 $\times$ 192 RGB images from the left-camera with timestamps closest to these scans, 8 192 $\times$ 192 left-event camera filter bank responses at these timestamps, along with accelerometer and gyroscope time-series.

Late-fusion architecture. Building a representation of this data requires reasoning over spatial structure within each modality, temporal structure across the time window, and cross-modal structure across modalities. It is possible to take the input representations in Sec. 4 and process them "directly" using a transformer architecture—we call this an early fusion architecture. This architecture can exploit synergistic or complementary information across sensors early. The downside however is that it puts severe demands on GPU memory and bandwidth, as we will see in the experiments the early fusion architecture has a large latency on embedded platforms. We therefore propose a late-fusion architecture with modality-specific spatial and temporal tubelet encoders that feed into a cross-modal encoder to create the representation.

Masking strategy. Consider a sensor $s \in \{1, 2, 3\}$ which could be RGB/event camera or the LiDAR. The masking ratio is $r_s = 1 - \min(1, \frac{3}{5}\, p_s)$ where $p \sim \text{Dirichlet}(\gamma, \gamma, \gamma)$ with γ drawn uniformly randomly from $\{0.5, 1, 2\}$. Masks are spatio-temporal tubes that span the entire time window for each modality. For one-dimensional IMU data, tokens are kept (left unmasked) independently with Bernoulli parameter $r_s = 0.2 + \frac{1}{2} r_0$ where $r_0 \sim \text{Beta}(2, 5)$.

[1] We normalize each channel individually to account for the fact that filters with different bandwidths result in different statistics at each channel.

[2] Ingesting all sensors at their raw throughput and resolution to train the multi-modal MAE is well beyond limits of GPU memory and memory bandwidth. This is an interesting avenue for future work. For the same reason, most experiments use four sensors (left RGB, left event, IMU, and LiDAR), and we describe this four-sensor variant in the main text. Details of the seven-sensor variant are provided in Sec. H, and experiments using it are noted explicitly.

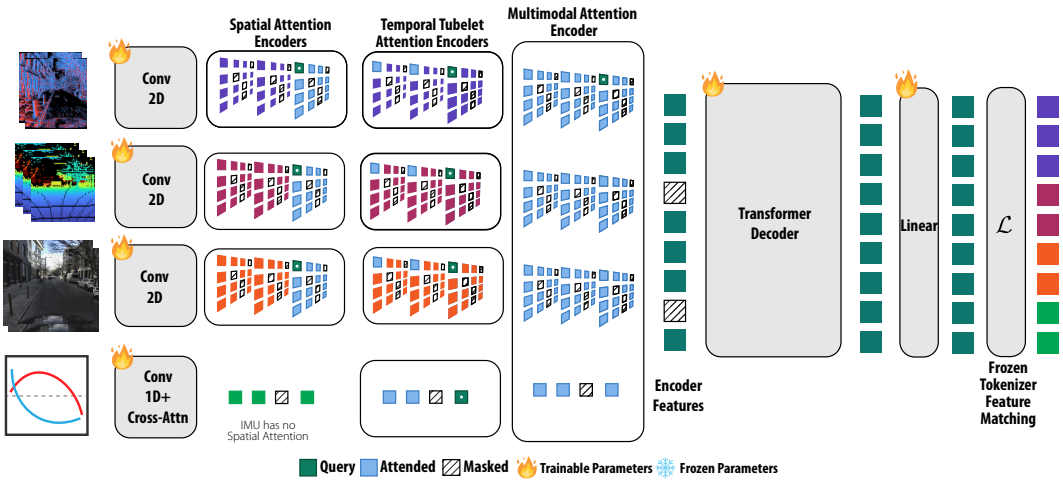


Figure 3: Late-fusion MAE. Per-modality tokens are masked in spatiotemporal tubes. The encoder applies attention in three stages: spatial, within each frame; temporal, along per-patch tubes across time; and multimodal, jointly over all tokens. Mask tokens are inserted after encoding, and a shared transformer decoder with per-modality linear heads predicts the codes of the frozen FSQ tokenizers. Fig. S10 in Sec. H shows the masking strategy and further details.

Architecture. For the fusion MAE, we use a four-dimensional variant of RoPE corresponding to coordinates (t, u_1, u_2, s) where $t \in \{0, 1, \ldots, 7\}$ indicates position within the temporal window, (u_1, u_2) indicates patch coordinate for image-like inputs (set to zero for IMU) and s indicates the sensor index. The encoder in our late-fusion MAE applies a per-modality spatial encoder at each time instant, followed by a per-modality temporal encoder across each unmasked tube (let us call this "tubelet attention").[3] Both encoders are self-attention based transformers. A cross-modal encoder processes features from all these modalities further to create the fused representation. The decoder receives the encoder output together with a learned mask token inserted at every masked patch. See Fig. 3 for rough schematic of the architecture. The multi-modal MAE is trained to regress the (frozen) FSQ tokenizer's values using an ℓ_1 loss. Sec. H gives more details of the architecture and Sec. K.1 discusses some examples of reconstruction of original sensory data.

Task-specific probes After pretraining the encoder, we can use the representation to train task-specific probes. For dense prediction tasks such as segmentation, optical flow and depth estimation, we train a dense prediction transformer (DPT)-based (Ranftl et al., 2021) fusion head using features from four encoder layers uniformly separated in depth. Before feeding into the DPT, for each of these layers, we (i) process features using a self-attention block to benefit from temporal information within the time window of each modality, and (ii) select a fixed number of features (as many as the number of patches in one RGB frame) as input to the DPT.[4] The same DPT network is used to predict multiple tasks using task-specific heads, trained with SiLog loss (Eigen et al., 2014) for depth, Charbonnier loss (Charbonnier et al., 1997) for flow, and cross-entropy loss for segmentation. Fig. S10 provides a schematic of the DPT while Sec. I provides more details of the task prediction heads and the evaluation protocol. For ego-motion prediction, we train three independent cross-attention modules whose queries attend to the last layer of the encoder. These probes predict linear and angular velocity, relative translation and rotation across time instants within the window, car speed and steering position. For the rotation heads, we minimize the squared Frobenius norm between the predicted and ground-truth rotation matrices; at inference-time predictions are projected onto $\mathrm{SO}(3)$ (Geist et al., 2024; Levinson et al., 2020). The other task-heads are trained with a Huber loss on Z-scored targets.

[3] In contrast to tubelet attention, "full attention" attends to all spatio-temporal tokens within a modality.

[4] There are two things to note here. First, when we run comparisons against baseline methods, say DINO, we encode each of the 8 RGB images within the time window and then run the self-attention block in step (i); this helps in tasks such as optical flow which require temporal context. Second, we select a fixed number of tokens in step (ii) to ensure that the DPT receives a fixed-size grid as input.

Table 1: Late-fusion MAE architecture predicts multiple tasks robustly across varied conditions. We report root mean square error of depth prediction at pixels with depth within [3, 200] m. For flow, we report the average end-point error in pixels. For segmentation, we report mean intersection-over-union (mIoU) over 19 Cityscapes classes. Ground-truth segmentation labels are unavailable at night. For ego-motion we report RMSE on translation (m), rotation angle (°) and steering position (°). All metrics are reported on data where speed is at least 5 mph. These metrics are a very heavily summarized evaluation of the performance, e.g., RMSE depth could be measured in terms of SiLog loss, mean absolute error, errors on objects at different distances from the platform etc. We report such detailed metrics in Sec. K.2. Our approach also performs well on these detailed metrics. For V-JEPA 2.1 (stereo) *, we ran the original model on both the left and right RGB images.

Encoder	Depth (m) ↓		Flow (px) ↓		Seg. (mIoU ↑)	Trans. (m) ↓		Rotation (°) ↓		Steering (°) ↓	
	day	night	day	night	day	day	night	day	night	day	night
Existing approaches											
DINOv2	6.82	9.95	19.46	18.36	0.410	0.96	0.84	0.80	0.75	22.62	23.11
DINOv3	6.93	10.03	19.43	16.97	0.382	0.93	0.79	0.79	0.76	25.24	21.93
SigLIP 2	7.22	10.36	20.22	18.12	0.377	0.98	0.78	0.95	0.94	36.12	36.00
Perception Enc.	7.12	10.20	19.90	17.84	0.388	0.96	0.91	0.94	0.93	35.62	34.54
V-JEPA 2.1	6.38	9.51	9.13	10.90	0.402	0.77	0.76	0.47	0.47	5.98	6.83
Our baselines											
V-JEPA 2.1 (stereo)*	6.40	9.51	9.03	10.55	0.405	0.77	0.71	0.48	0.47	5.69	6.47
RGB-only image MAE	7.78	10.58	20.88	18.98	0.296	0.92	1.05	0.88	0.77	31.57	21.39
RGB-only video MAE	7.41	10.20	10.11	11.56	0.327	0.76	0.72	0.50	0.50	12.96	12.45
Early-fusion MAE	4.98	6.30	2.01	2.45	0.379	0.14	0.15	0.56	0.53	4.58	4.48
Late-fusion MAE	**4.73**	6.08	1.97	2.39	**0.411**	0.06	0.06	0.24	0.23	2.51	2.27
w/ full attention	**4.73**	**6.07**	**1.96**	**2.38**	0.405	**0.05**	**0.05**	**0.17**	**0.17**	**2.46**	**2.25**
w/ all 7 sensors	**4.73**	6.13	1.98	2.40	0.404	**0.05**	0.06	0.22	0.20	2.77	2.45
w/ pixel targets (no tokenizers)	4.90	6.33	2.11	2.51	0.390	**0.05**	0.06	0.20	0.21	2.94	3.06
Leave one modality out/in at test time in late-fusion MAE											
w/o RGB	4.70	5.99	2.02	2.43	0.366	0.06	0.06	0.24	0.23	2.65	2.34
w/o EV	4.88	6.32	2.00	2.42	0.403	0.05	0.05	0.21	0.20	2.59	2.42
w/o LiDAR	6.74	9.55	7.19	8.05	0.351	0.73	0.57	0.46	0.45	3.97	4.00
w/o IMU	4.67	6.00	2.05	2.49	0.402	0.06	0.06	0.24	0.22	2.78	2.61
w/ only RGB	7.20	10.09	11.31	12.81	0.310	0.83	0.77	0.47	0.47	8.29	7.75

5 RESULTS

This section provides experimental evidence and analysis to support our claim that multimodal sensors provide complementary information that vision-only foundation models cannot recover. We are particularly interested in understanding performance when sensors are degraded or missing.

Baselines. We compare our approach to existing vision-only methods (DINOv2, DINOv3, SigLIP 2, Perception Encoder and V-JEPA 2.1) across five tasks (depth, flow, segmentation, egomotion and steering position). We also develop six other baselines that are based on our approach: (i) RGB-only MAE (with only the left camera) trained on FSQ targets for single frames which only has access to spatial information, (ii) RGB-only video MAE (with only the left camera) trained on FSQ targets for all 8 frames within the temporal window and thereby also receives some temporal information, (iii) early fusion MAE as described in Sec. 4 which uses a monolithic transformer architecture and has access to multi-modal information, (iv) late-fusion MAE (w/ full attention) that performs temporal self-attention over all tokens in a modality, as opposed to our proposed architecture which only performs self-attention over spatiotemporal tubes independently, (v) late-fusion MAE (w/ all 7 sensors), which adds the right RGB, right event and infrared cameras to the default four-sensor input, and (vi) late-fusion MAE (w/ pixel targets), which regresses raw sensor values instead of the frozen FSQ codes. To evaluate vision-only baselines, we encode each of the 8 RGB images within the time window and then proceed to the self-attention-block implemented before the DPT, as discussed in Sec. 4. Sec. I provides more details.

Training and evaluation methodology. We split OctoSense data into 293 training and 78 test sequences, each about 10 mins long. Test set contains data from both daytime and nighttime

conditions, with sensor-degraded sequences reserved as a separate held-out set in Sec. K.2. During training we split data into 1.4 sec windows that have an overlap of 1 sec with each other; temporal windows for test data do not have any overlap. We train our late-fusion MAE (ViT-B scale) for 150 epochs with a batch size of 512 on four NVIDIA B200 GPUs; hyper-parameters are provided in Sec. H. The experiments in this paper required ∼7,000 GPU-hours. All task probes are trained for 60 epochs. Secondary metrics are reported in Sec. K.2.

Overall results. Table 1 summarizes the performance of our proposed approach as well as some of the baselines. Overall, on all tasks, the late-fusion MAE performs much better than other approaches. Depth RMSE is better by 1.65 m compared to the best existing method (V-JEPA 2.1). This difference is more than twice as large in nighttime testing, which demonstrates the utility of having multiple modalities as well as the robustness of our approach. Trends are very similar for flow estimation, more than 7.16 and 8.51 pixel improvement over V-JEPA 2.1 in daytime and nighttime respectively. On segmentation, the late-fusion MAE marginally exceeds the best existing model (0.411 vs. 0.410). The narrow margin is likely because our pseudo-labels in OctoSense are generated by a DINOv2-based model. Our late-fusion MAE matches DINOv2 on frequent categories, but the former does poorly on rare categories (poles, buses, people), see Table S17. Predicting ego-motion from RGB-only models is quite difficult, and our approach performs significantly better because it has access to LiDAR as well as inertial information. We achieve a prediction error of 0.06 m in translation and 0.23–0.24° in rotation. The relative advantage of having multiple sensory modalities is higher at nighttime when visual sensors are degraded (Table 2). Using all 7 sensors (stereo RGB, stereo event, infrared, IMU, LiDAR) performs on par with the default model, suggesting the extra stereo views and infrared are largely redundant given LiDAR. Replacing tokenizer targets with raw pixels hurts dense prediction (e.g., depth 4.90 vs. 4.73 m, segmentation 0.390 vs. 0.411 mIoU) but remains competitive on ego-motion.

Comparison to baselines. RGB-only image MAE performs on par with existing methods, although the former is trained on our data while the latter are being evaluated zero-shot on our data. This is a positive datapoint to suggest that one can train useful self-supervised learning methods using data at the OctoSense scale. The video MAE's gain over the image MAE (∼11 px on daytime flow) matches V-JEPA 2.1's gain over the image-based existing methods, both reflect pretraining with temporal information. Late-fusion MAE performs about as well as the early-fusion MAE. This is somewhat remarkable because it indicates that a monolithic transformer can adequately represent information across these modalities. As Table 3 shows, the late-fusion MAE has about 60% (Orin) smaller latency and is much more amenable to implementation on NVIDIA Orin. Late-fusion MAE (w/ full attention) which uses full temporal self-attention performs on-par with our proposed tubelet attention.

Table 2: Performance of different encoders on degraded data from OctoSense. We report RMSE (m) for depth, end-point error (px) for flow, translation (m) and rotation (°) for ego-motion estimation.

Encoder	Depth (m) ↓	Flow (px) ↓	Trans.(m) ↓	Rot. (°) ↓
DINOv2	7.41	19.97	0.65	0.58
DINOv3	7.49	18.82	0.65	0.61
SigLIP 2	7.55	19.02	0.71	0.70
Perception Enc.	7.56	19.52	0.72	0.71
V-JEPA 2.1	7.40	10.85	0.51	0.43
Late-fusion MAE	**4.77**	**2.12**	**0.06**	**0.23**

Table 3: Latency of the forward pass for late-fusion MAE with similar number of parameters as that of the early-fusion MAE (batch-size of 1). Tables S9 and S10 elaborate upon this.

Fusion strategy	Nvidia GPU	# weights (M)	Latency (ms)
Early	5090	85	11.35
Late	5090	83.7	6.68
Early	Orin NX	85	288
Late	Orin NX	83.7	112

Table 4: Transfer to M3ED (Chaney et al., 2023), a different vehicle with a closely matched sensor suite. All encoders are frozen.

Encoder	Depth (m) ↓	Flow (px) ↓	Trans.(m) ↓	Rot. (°) ↓
DINOv2	9.53	5.52	0.46	1.13
DINOv3	9.29	5.09	0.55	1.24
SigLIP 2	9.69	5.43	0.70	1.24
Perception Enc.	9.75	5.26	0.53	1.32
V-JEPA 2.1	9.72	**1.59**	**0.22**	**0.64**
Late-fusion MAE	**9.05**	1.60	0.36	0.82

Leaving out modalities at test time. Leaving out a specific modality at test time in Table 1 helps understand the relative contribution from

that modality. LiDAR is a dominant sensor for ego-motion estimation tasks, without it the error in translation, rotation and steering angle is very high compared to late-fusion MAE. Dropping event images or the IMU has a smaller effect on the performance compared to dropping LiDAR. LiDAR is also the most important sensor for dense prediction: dropping it sharply degrades both depth and flow (RMSE $4.73 \rightarrow 6.74$ m, EPE $1.97 \rightarrow 7.19$ px), whereas dropping RGB leaves them essentially unchanged. The RGB camera instead matters most for segmentation, where leaving it out is what hurts performance the most. Event cameras and IMU contribute marginally on most tasks. It is remarkable that well-trained foundation models like V-JEPA 2.1 that are trained only on RGB data perform comparably to our late-fusion MAE that does not have access to LiDAR data (w/o LiDAR row in Table 1). In view of the fact that OctoSense is a much smaller dataset than those used to train industry-scale foundation models, this suggests that having LiDAR information can go a long way.

Indeed, LiDAR provides a direct geometric signal for ego-motion while RGB images serve to spatially anchor the tokens of other modalities for pixel-aligned semantic prediction. The fact that we do not see event cameras and IMU contributing to the performance much is perhaps due to the fact that we evaluate these tasks at 5 Hz. This is far below the native temporal resolution of these sensors (microsecond-scale for event cameras and $\sim$400 Hz for IMU). These sensors will be useful for high-speed tracking and low-latency maneuvers; we leave this for future work.

Transfer to a closely matched sensor platform. Foundation models like DINO or V-JEPA generalize well to new settings, in part because they are trained on large datasets. To check whether a representation learned from OctoSense generalizes to a similar platform, we evaluate on the M3ED dataset (Chaney et al., 2023). M3ED is the only other large existing dataset with event camera data, a similar sensor suite and ego-motion ground truth.[5] As Table 4 shows, our late-fusion architecture is competitive on dense perception tasks compared to existing foundation models, and second best on ego-motion tasks.

Novelty of our approach. MAEs, learned tokenizers, and multi-modal fusion are established approaches. In this paper, our key contribution is to develop techniques to make these approaches work on heterogeneous robot sensors; as we showed, a straightforward combination of these existing architectures to diverse modalities does not quite work as well. Concretely: (i) modality-specific tokenizers with modality-specific losses, and cross-attention pooling for the IMU (Sec. 4); (ii) a late-fusion factorization into spatial and tubelet attention with 4D RoPE, which matches full temporal attention at a fraction of the cost (Tables 1 and S9); (iii) streaming token caching, uniquely enabled by this factorization (Table S9); and (iv) the resulting architecture, whose dependence on multi-modality we establish by comparing against RGB-only encoders and by ablating each sensor at test time (Table 1). The early-fusion baseline is the control: a monolithic transformer over the same inputs is both more expensive and worse on every dense task (Tables 1 and S9).

6 Discussion

Our paper demonstrates the utility and viability of multi-modal SSL across a variety of robot perception tasks and experimental settings. But it is a research prototype. Our dataset is orders-of-magnitude smaller than the datasets used to train foundation models such as DINO and V-JEPA, by our estimate about $100\times$. So while it is remarkable that the multi-modal approach leads to such dramatic performance gains over existing foundation models, training on more data can improve the generalizability of our learned representation. In effect it will allow us to build a "foundation" multi-modal robot perception model. For our approach to replace existing foundation models, we will need a community-wide effort to pool data.

Our computing budget restricted us from training with larger networks, which could improve the performance further. One of the key avenues for future work is to use temporal windows longer than 1.4 sec with recurrent architectures. This can enable new tasks such as object tracking and 3D instance detection (after appropriate pseudo annotation), and also further reduce the computational complexity of our approach.

[5] Cross-dataset generalization requires a dataset with a sensor suite close to ours and reliable ground-truth odometry. To our knowledge, M3ED is the only publicly available dataset that meets both criteria. Evaluating generalization across substantially different sensor configurations e.g., 16/32/64/128-beam LiDARs, is itself an open research problem that we leave to future work.

AI USE STATEMENT

Generative AI Claude was used during this investigation. For our purposes, we used it to perform copy editing for the paper and help with stylization of figures, including creating larger font, better ticks, and smaller grid areas. Claude was used during the coding of this investigation for uses including UI design our web application/semantic search, refactoring/modularization of functions into class/subclasses for scalable development, code-generation for various utilities to help visualize, inspect and monitor our data collection process, and debugging various issues encountered.

ETHICS STATEMENT

We do not believe these results raise any ethical concerns.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure that our results can be reproduced. **Hardware and data collection:** the full sensor specifications are in Table S3, a complete bill of materials in Sec. C, per-sensor configuration, calibration and mechanical design in Sec. D, the modifications we made to open-source sensor drivers in Sec. E, and our time-synchronization hardware and algorithm in Sec. F. **Data processing and ground truth:** the generation of ground truth for odometry, depth, optical flow and segmentation is described in Sec. 3, and the training/test split in Sec. 5. **Models:** input representations for event cameras and IMU are given in Secs. H.1 and H.2, and the architectures and hyper-parameters of the tokenizers, the late-fusion MAE and all baselines in Tables S5 to S7. **Evaluation:** the probe architectures and training recipe are in Table S8, the loss functions, metrics, and M3ED protocol in Sec. I. Upon publication, we will release the OctoSense dataset, all code for data collection, processing and model training, and the hardware schematics and CAD files.

REFERENCES

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: End-to-end davis driving dataset. *ArXiv*, abs/1711.01458, 2017.

Anthony Bisulco, Fernando Cladera Ojeda, Volkan Isler, and Daniel D. Lee. Fast Motion Understanding with Spatiotemporal Neural Networks and Dynamic Vision Sensors. In *IEEE International Conference on Robotics and Automation*, 2021.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Doll'ar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, 2025.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2019.

Jianwen Cao, Jiaxu Xing, Nico Messikommer, and Davide Scaramuzza. Generative event pretraining with foundation model alignment. *arXiv preprint 2603.23032*, 2026.

Nicholas Carlevaris-Bianco, Arash K. Ushani, and Ryan M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35:1023 – 1035, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ED: Multi-robot, multi-sensor, multi-environment event dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11305–11315, 2022.

P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2):298–311, 1997.

Peiyu Chen, Weipeng Guan, Feng Huang, Yihan Zhong, Weisong Weisong Wen, Li-Ta Hsu, and Peng Lu. ECMD: An event-centric multisensory driving dataset for slam. *IEEE Transactions on Intelligent Vehicles*, 9: 407–416, 2023. URL https://api.semanticscholar.org/CorpusID:265033288.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

Henrik I. Christensen. Global robotics technology roadmap 2025–2035: A multi-regional, cross-domain strategic perspective for europe, asia, and the united states. Technology roadmap, University of California San Diego, April 2026. Version 1.02.

Richeek Das, Kostas Daniilidis, and Pratik Chaudhari. Fast feature field (F3): A predictive representation of events. *arXiv preprint 2509.25146*, 2025.

Tobi Delbruck. Frame-free dynamic digital vision. In *Proceedings of the International Symposium on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pp. 21–26, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Carlos Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, Wei-Lun Chao, Bharath Hariharan, Kilian Q. Weinberger, and Mark E. Campbell. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. 2024.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

Jonas Frey, Turcan Tuna, Frank Fu, Katharine Patterson, Tianao Xu, Maurice Fallon, Cesar Cadena, and Marco Hutter. Grandtour: A legged robotics dataset in the wild for multi-modal perception and state estimation, 2026. URL https://arxiv.org/abs/2602.18164. *Equal contribution (Turcan Tuna and Jonas Frey).

Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1280–1286, 2013.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxun Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. VECtor: A versatile event-centric benchmark for multi-sensor slam. *IEEE Robot. and Autom. Lett.*, 7(3):8217–8224, June 2022.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robot. and Autom. Lett.*, March 2021.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

Andreas Rene Geist, Jonas Frey, Mikel Zobro, Anna Levina, and Georg Martius. Learning with 3D rotations, a hitchhiker's guide to SO(3). *International Conference on Machine Learning*, 2024.

Wulfram Gerstner, Werner M. Kistler, Richard Naud, and Liam Paninski. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind one embedding space to bind them all. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2026. Open model release.

Benjamin R. Hamilton, Xiaoli Ma, Qi Zhao, and Jun Xu. ACES: adaptive clock estimation and synchronization using kalman filtering. In *ACM/IEEE International Conference on Mobile Computing and Networking*, 2008.

T. Hawkes and P. Simonpieri. Signal coding using asynchronous delta modulation. *IEEE Trans. on Comm.*, 22 (5):729–731, March 1974.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD20 End-to-End Event Camera Driving Dataset: Fusing frames and events with deep learning for improved steering prediction. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, 2020.

ITU-T. Series h: Audiovisual and multimedia systems: Infrastructure of audiovisual services - coding of moving video: High efficiency video coding. Technical Report ITU-T H.265, International Telecommunication Union, 2026. Version 01/2026.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32:922–923, 1976.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is Secretly an Image Segmentation Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Simone Klenk, David Bonello, Lukas Koestler, and Daniel Cremers. Masked Event Modeling: Self-supervised pretraining for event cameras. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

Lingdong Kong, Dongyue Lu, Xiang Xu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. Eventfly: Event camera perception from ground to the sky. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1472–1484, 2025.

Xavier Lagorce, G. Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1346–1359, 2017.

Alex Junho Lee, Younggun Cho, Young sik Shin, Ayoung Kim, and Hyun Myung. ViViD++ : Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7:6282–6289, 2022.

Kenneth Levenberg. A method for the solution of certain non – linear problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An Analysis of SVD for Deep Rotation Estimation. *Advances in Neural Information Processing Systems*, 2020.

Daniil Lisus, Katya M. Papais, Cédric Le Gentil, Elliot Preston-Krebs, Andrew Lambert, Keith Y.K. Leung, and Timothy D. Barfoot. Boreas Road Trip: A multi-sensor autonomous driving dataset on challenging roads. *ArXiv*, abs/2602.16870, 2026.

Yu Liu, Shuting Wang, Yuanlong Xie, Tifan Xiong, and Mingyuan Wu. A review of sensing technologies for indoor autonomous mobile robots. *Sensors*, 24, 2024.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations*, 2023.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

William P. Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36:15 – 3, 2017.

M.V.R. Malladi, T. Guadagnino, L. Lobefaro, and C. Stachniss. A robust approach for lidar-inertial odometry without sensor-specific modeling. *IEEE Robotics and Automation Letters*, 11(6):7420–7427, 2026.

Sivabalan Manivasagam, Shenlong Wang, K. Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Binh Yang, Wei-Chiu Ma, and Raquel Urtasun. LiDARsim: Realistic lidar simulation by leveraging the real world. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

David Mizrahi, Roman Bachmann, Ouguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In *Advances in Neural Information Processing Systems*, 2023.

Farzeen Munir, Shoaib Azam, and Moongu Jeon. Sstn: Self-supervised domain adaptation thermal object detection for autonomous driving. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint 2603.14482*, 2026.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, S. Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak N. Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations*, 2025.

Haoyi Niu, Jianming Hu, Guyue Zhou, and Xianyuan Zhan. A comprehensive survey of cross-domain policy transfer for embodied agents. *ArXiv*, abs/2402.04580, 2024. URL https://api.semanticscholar.org/CorpusID:267522875.

NVIDIA. PhysicalAI-Autonomous-Vehicles dataset. https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles, 2025.

Edwin Olson. AprilTag: A robust and flexible visual fiducial system. *IEEE International Conference on Robotics and Automation*, 2011.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. In *International Conference on Learning Representations*, 2025.

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*, 2022.

Manthan Patel, Jonas Frey, Mayank Mittal, Fan Yang, Alexander Hansson, Amir Bar, César Cadena, and Marco Hutter. DeFM: Learning foundation representations from depth for robotics. *arXiv preprint 2601.18923*, abs/2601.18923, 2026.

Etienne Perot, Pierre de Tournemire, Davide Oscar Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Neural Information Processing Systems*, 2020.

Bernd Pfrommer. Frequency cam: Imaging periodic signals in real-time. *arXiv preprint 2211.00198*, 2022.

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. In *International Conference on Machine Learning*, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *IEEE/CVF International Conference on Computer Vision*, 2021.

Harry Ernest Rauch, Fang-Wu Tung, and Charlotte T. Striebel. On the maximum likelihood estimates for linear dynamic systems. In *AIAA Journal*, 1965.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao Wu, Ross B. Girshick, Piotr Doll'ar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Joern Rehder, Janosch Nikolic, Thomas Schneider, Timo Hinzmann, and Roland Siegwart. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes. In *IEEE International Conference on Robotics and Automation*.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, 2023.

Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint 2509.25164*, 2025.

Harald Schafer, Eder Santana, Andy Haden, and Riccardo Biasini. A commute in data: The comma2k19 dataset. *ArXiv*, abs/1812.05752, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint 2508.10104*, 2025.

Matthew Sivaprakasam, Parv Maheshwari, Mateo Guaman Castro, Samuel Triest, Micah Nye, Steven Willits, Andrew Saba, Wenshan Wang, and Sebastian A. Scherer. TartanDrive 2.0: More modalities and better infrastructure to further self-supervised learning research in off-road driving tasks. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12606–12606, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568, 2024. ISSN 0925-2312.

Jiachen Sun, Haizhong Zheng, Qingzhao Zhang, Atul Prakash, Zhuoqing Morley Mao, and Chaowei Xiao. CALICO: Self-supervised camera-lidar contrastive pre-training for bev perception. In *International Conference on Learning Representations*, 2024.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M. Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2019.

Samuel Triest, Matthew Sivaprakasam, Sean J. Wang, Wenshan Wang, Aaron M. Johnson, and Sebastian A. Scherer. TartanDrive: A large-scale dataset for learning off-road dynamics models. *IEEE International Conference on Robotics and Automation*, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim M. Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiao-Qi Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint 2502.14786*, 2025.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:376–380, 1991.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint 1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *IEEE/CVF International Conference on Computer Vision*, 2021a.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *ArXiv*, abs/2301.00493, 2023.

Saining Xie, Jiatao Gu, Demi Guo, C. Qi, Leonidas J. Guibas, and Or Litany. PointContrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision*, 2020.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: a simple framework for masked image modeling. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 2020.

Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. LIMU-BERT: Unleashing the potential of unlabeled data for imu sensing applications. In *ACM Conference on Embedded Networked Sensor Systems*, 2021.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Benjamin Sapp, Mingxing Tan, Jyh-Jing Hwang, and Drago Anguelov. WOD-E2E: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint 2510.26125*, 2025.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision*, 2024.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint 2506.05176*, 2025.

Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robt. and Autom. Lett.*, 3:2032–2039, Feb. 2018.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M. Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:5299–5318, 2023.

Jian Zou, Tianyu Huang, Guanglei Yang, Zhenhua Guo, and Wangmeng Zuo. UniM2AE: Multi-modal masked autoencoders with unified 3d representation for 3d perception in autonomous driving. In *European Conference on Computer Vision*, 2024.

Jannik Zürn. *Self-supervised and Multi-modal Learning for Perception in Mobile Robots and Autonomous Driving*. PhD thesis, University of Freiburg, 2024.