REALM: An RGB- and Event-Aligned Latent Manifold for Cross-Modal Perception

Vincenzo Polizzi[1], David B. Lindell[2], and Jonathan Kelly[1]
{vincenzo.polizzi,jonathan.kelly}@robotics.utias.utoronto.ca
lindell@cs.toronto.edu

[1]University of Toronto, Robotics Institute
[2]University of Toronto, Department of Computer Science

Abstract. Event cameras provide several unique advantages over standard frame-based sensors, including high temporal resolution, low latency, and robustness to extreme lighting. However, existing learning-based approaches for event processing are typically confined to narrow, task-specific silos and lack the ability to generalize across modalities. We address this gap with REALM, a cross-modal framework that learns an **R**GB- and **E**vent-**A**ligned **L**atent **M**anifold by projecting event representations into the pretrained latent space of RGB foundation models. Instead of task-specific training, we leverage low-rank adaptation (LoRA) to bridge the modality gap, effectively unlocking the geometric and semantic priors of frozen RGB backbones for asynchronous event streams. We demonstrate that REALM effectively maps events into the ViT-based foundation latent space. Our method performs downstream tasks, such as depth estimation and semantic segmentation, by simply transferring linear heads trained on the RGB teacher. Most significantly, REALM enables the direct, zero-shot application of complex, frozen image-trained decoders, such as MASt3R, to raw event data. We demonstrate state-of-the-art performance in wide-baseline feature matching, significantly outperforming specialized architectures. Code and models are available at https://papers.starslab.ca/realm/.

1 Introduction

Visual perception is fundamental to modern computer vision and intelligent systems, driving progress in applications ranging from autonomous navigation and augmented reality to computational photography. The vast majority of these systems rely on conventional RGB cameras, which capture dense frames at fixed intervals. While effective in controlled environments, frame-based sensors suffer from fundamental limitations: motion blur, limited dynamic range, and high latency [16]. These constraints degrade performance in challenging conditions, such as high-speed motion [50] or extreme lighting environments [33].

Event cameras are bio-inspired sensors that have found several applications in robotics and computer vision [24,43] and offer a paradigm shift in visual sensing. By asynchronously recording per-pixel brightness changes, event cameras

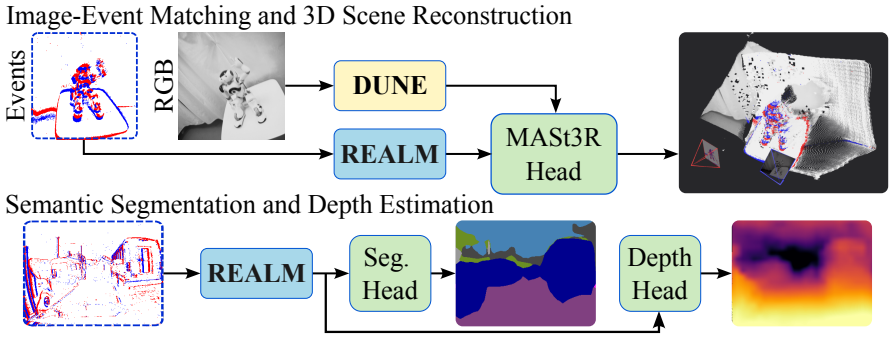

Fig. 1: Overview of REALM's cross-modal versatility. Our model enables diverse downstream applications, which include feature matching using the frozen MASt3R [34] decoder and dense prediction tasks like semantic segmentation and depth estimation using simple linear heads.

produce sparse event streams with microsecond-scale latency and high dynamic range [20]. This sensing principle reduces motion blur and drastically lowers power consumption, making event data distinctively advantageous for analyzing dynamic scenes [39, 55].

Despite these advantages, learning task-agnostic feature representations from event data remains an open challenge. Event streams are sparse, asynchronous, and structurally distinct from images, posing significant challenges for the direct application of standard computer vision techniques. Furthermore, unlike the RGB domain—which benefits from large-scale datasets—labelled event datasets are scarce and small in scale. Consequently, existing event-based models are typically trained from scratch for narrow, specific tasks (e.g., for optical flow computation [51] or classification [12] only), failing to generalize across different downstream applications.

In the RGB domain, the landscape of computer vision techniques has been reshaped by large-scale vision transformers (ViTs). Foundation models such as the DINO family [7, 41, 52] or DUNE [48] have demonstrated that pretraining on massive image collections yields universal encoders that transfer seamlessly to diverse tasks. However, the event vision community has largely been excluded from this revolution due to the significant modality gap between intensity-based frames and temporally-driven event streams.

To bridge this gap, we propose REALM, a cross-modal visual encoder that unifies RGB and event representations within a shared latent space (see Fig. 1). Building upon the DUNE [48] architecture, REALM introduces a lightweight, modality-specific input embedder and leverages Low-Rank Adaptation (LoRA) [29] to align event features with the semantically rich and geometrically consistent RGB manifold. Crucially, this approach allows us to align the modalities without retraining the backbone, preserving the rich semantic and geometric knowledge learned from millions of RGB images. This design enables both intra-modal (event–event) and cross-modal (RGB–event) feature matching, facilitating modality-invariant transfer learning where features remain spatially and seman-

tically consistent regardless of the input sensor. Unlike cross-modal distillation between dense and synchronous sensor modalities such as depth, LiDAR, or thermal, the event modality poses challenges intrinsic to how events are generated. For example, regions of static brightness produce no events, leaving parts of the scene unobserved. The model must therefore learn dense structure without hallucinating content where no signal exists. At the same time, a uniform distillation objective should be avoided as it would penalize the student over regions it cannot observe. Addressing these two properties is what allows a frozen, image-trained decoder to operate zero-shot on event features.

We validate REALM across multiple datasets on three distinct downstream tasks: monocular depth estimation, semantic segmentation, and wide-baseline feature matching. For the latter, we specifically evaluate cross-modal (RGB–event) and intra-modal (event–event) correspondences, demonstrating that our aligned manifold allows a frozen, image-trained matching head to function across disparate sensors without retraining. Our evaluation focuses on two core hypotheses: first, that a shared latent space allows event-based data to leverage knowledge from large-scale RGB foundation models. Second, that this alignment is sufficiently precise to enable the direct, zero-shot[1] application of frozen image-trained decoders to asynchronous event streams. While we demonstrate that the aligned latent space is rich enough to support dense prediction tasks like depth and segmentation using only simple linear heads, our most significant results lie in wide-baseline feature matching. Here, REALM enables the direct use of frozen, RGB-trained geometric heads, such as MASt3R, to perform cross-modal and intra-modal matching. In this challenging domain, our method not only achieves state-of-the-art performance but does so by outperforming specialized event-only baselines.

We make the following contributions.

- We introduce REALM, a unified encoder that aligns event and RGB modalities in a shared, semantically rich latent space.
- We demonstrate that our alignment is sufficiently precise to enable the direct, zero-shot application of frozen, image-trained geometric decoders to event streams, establishing a new state-of-the-art in event-based feature matching and pose estimation.
- We show that lightweight LoRA [29] fine-tuning bridges the modality gap without catastrophic forgetting. The resulting frozen encoder generalizes to diverse tasks, from dense depth estimation to semantic segmentation, providing a data-efficient solution to the scarcity of annotated event datasets.
- We provide an open-source implementation of REALM, together with models, at https://papers.starslab.ca/realm/.

[1] Throughout this work, *zero-shot* refers to a head trained exclusively on features from a frozen DUNE [48] encoder and deployed directly onto event features from REALM, with no event-specific supervision, fine-tuning, or adaptation at any stage.

2 Related Work

We first cover event representations, then review the three core tasks our method addresses: semantic segmentation, depth estimation, and feature matching. Finally, we provide an overview of cross-modal knowledge transfer.

2.1 Event Representation

Event cameras asynchronously record changes in scene brightness, producing a sparse stream of events that depends on relative motion between the camera and the scene [24, 43]. Unlike conventional frames, event data encode temporal contrast rather than absolute intensity, which poses challenges for standard deep networks that assume dense, grid-structured inputs.

To bridge the modality gap, several representations have been proposed to convert event streams into tensor formats suitable for deep networks [18]. Common representations include voxel grids [65], which bin events temporally, and Tencode [30], which renders a chunk of events as an RGB image, using the red/blue channels for event polarity and the green channel for the relative timestamp. More recently, ERGO [68] optimizes the event ordering for detection tasks.

2.2 Semantic Segmentation

Semantic segmentation assigns a class label to each pixel in the camera frame, providing dense scene understanding essential for robotics and autonomous driving. In the RGB domain, while early methods relied on CNN-based encoder-decoders like SegNet [2], the field has largely shifted toward Transformer-based architectures that leverage large-scale pretraining for robust pixel-level classification [9, 32, 59]. An in-depth review of visual segmentation methods is in [54].

For event cameras, segmentation is complicated by the sparsity and lack of texture in the data. Early works [1] addressed this by adapting CNN architectures and releasing the first event-segmentation benchmarks. Similarly, ESS [53] proposed a recurrent encoder-decoder to maintain temporal consistency in predictions. More recently, ESEG [63] tackles the sparsity problem by introducing explicit edge-semantic supervision to locate reliable cues. Alternative approaches leverage RGB data to guide learning: EvDistill [57] uses a teacher–student framework to transfer knowledge from frames to events, while HALSIE [11] fuses both modalities to extract rich spatiotemporal features.

However, these methods often require training complex decoders or distillation pipelines from scratch. In contrast, REALM demonstrates that by aligning event representations with the frozen DUNE [48] latent space, we achieve dense semantic maps using only a simple linear head, inheriting strong scene understanding capabilities without the need for extensive task-specific supervision.

2.3 Depth Estimation

Depth estimation predicts per-pixel distances, a critical capability for navigation, mapping, and 3D reconstruction. For RGB images, Eigen et al. [14] originally formulated depth estimation as a continuous regression task using CNNs. However, recent state-of-the-art approaches have recast depth estimation as a classification problem (predicting depth bins), leveraging the global context of Transformers [15,62]. Foundation models like DINOv2 [41], MASt3R [34], VGGT [56], and DepthAnything [60] have further demonstrated that general-purpose features can yield high-quality depth maps without specialized depth architectures.

For event data, monocular depth estimation is complicated by sparsity and motion dependence, as static regions generate no events and create "blind spots" that necessitate temporal aggregation. Previous works have addressed these challenges by employing recurrent U-Nets to aggregate temporal information or by fusing sparse events with frames to densify the resulting predictions [19,27]. Furthermore, self-supervised methods have emerged to bypass the need for ground-truth depth labels by instead leveraging cross-modal consistency between synchronized event streams and intensity frames [65, 67]. A more comprehensive review on depth estimation with event cameras is given by [23].

Unlike these approaches, which often rely on complex recurrent architectures to "fill in" missing data, we adopt a classification-based paradigm and demonstrate that aligning an event encoder with the geometrically rich DUNE [48] latent space is sufficient to produce dense depth maps, effectively bypassing the need for specialized event-depth architectures.

2.4 Feature Matching and Localization

Feature matching under wide-baseline and cross-modal conditions is a cornerstone for tasks such as visual odometry, SLAM, and 3D reconstruction.

The literature on feature extraction and matching for visual cameras is extensive, ranging from classical hand-crafted descriptors [5, 35, 37, 47] to modern learning-based methods [13, 49, 61]. More recent approaches leverage geometric priors to improve feature stability and robustness across viewpoints [34, 56, 58].

In contrast, event-based feature matching remains challenging because the absence of absolute intensity cues makes constructing repeatable descriptors difficult. Existing methods typically rely on spatio-temporal correlations or event accumulation strategies to detect and match features [39,44]. Recent work [30] introduced a novel event representation, Tencode, which learns feature correspondences by leveraging matches from a pre-trained network such as SuperPoint [13]. Other methods, such as MINIMA [46], develop modality-invariant matching by utilizing a generative data engine to synthesize massive multimodal datasets from RGB pairs, enabling cross-modal generalization through fine-tuning.

Our approach differs from these paradigms by leveraging the latent space of a foundation model to incorporate 3D geometric priors. REALM enables the zero-shot use of frozen image-trained heads [34] for matching across both modalities and viewpoint changes. This formulation provides a unified geometric embedding space that facilitates seamless cross-modal correspondence.

2.5 Cross-Modal Knowledge Transfer

Recent successes in event-to-frame reconstruction, such as E2VID [45], have demonstrated that event streams contain sufficient information to recover high-fidelity photometric representations of a scene. While methods like EvDistill [57] leverage this by mapping events into the image domain before utilizing large-scale visual models, this intermediate reconstruction step often introduces computational overhead and potential artifacts. We therefore posit that it is possible to bypass image reconstruction entirely by directly mapping events into the latent representation space of large, pretrained visual encoders.

3 Methodology

In this section, we present our approach for training and refining REALM. We first provide background on the DUNE [48] encoder, followed by a formal problem definition for cross-modal alignment. Finally, we detail our architectural modification strategy, focusing on the adaptation of a frozen foundation model for asynchronous event data.

3.1 Background: DUNE

Our work builds upon the DUNE [48] framework, a universal vision transformer (ViT) designed to unify diverse 2D and 3D perception tasks into a single encoder through multi-teacher distillation.

DUNE [48] is trained using a set of N heterogeneous teacher models $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$. Each teacher $\mathcal{T}_i$ is parameterized by its own encoder $t_i(x)$, which maps an input image x into a set of feature vectors $Z \in \mathbb{R}^{(HW+1)\times d}$. These features include HW patch tokens and an optional global CLS token. The goal of distillation is to learn a student encoder $f(x)$ such that its representations, after a teacher-specific projection, match those of the teachers. For each teacher $\mathcal{T}_i$, DUNE [48] employs a teacher-specific transformer projector h_i to map the student's features into the teacher's latent space.

DUNE [48] distills from three distinct teachers: DINOv2 [41], MASt3R [34], and Multi-HMR [3]. After the student encoder f is trained, the task-specific decoders (or heads) are fine-tuned independently while keeping the encoder frozen: $y_{task} = \text{Decoder}_{task}(f(x))$. This decoupled training ensures that the student encoder retains a "universal" latent manifold suitable for diverse downstream applications without needing to store or compute teacher-specific projectors during inference.

In REALM, we leverage the open-source MASt3R [34] head, which was refined on DUNE's universal features. We utilize this head directly in REALM without any event-based refinement, providing a rigorous zero-shot benchmark for our cross-modal alignment strategy.

We emphasize that our alignment framework is not tied to DUNE [48] specifically: any ViT-based encoder with a sufficiently general latent space could serve

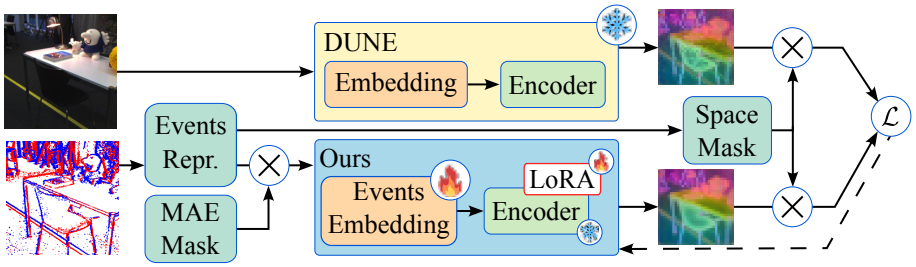

Fig. 2: Overview of the cross-modal distillation framework. Event representations undergo masked autoencoders (MAE)-style dropout before being processed by a trainable embedding layer and a LoRA-adapted student encoder. A progressive spatial mask is applied to the distillation loss ($\mathcal{L}$) to focus the alignment on regions with active event data, preventing background overfitting.

as the teacher. We adopt DUNE [48] because its multi-teacher distillation yields a manifold that already spans semantic, geometric, and human-centric priors, which is well matched to the diverse downstream tasks we target.

3.2 Problem Definition

Our primary objective is to learn a mapping $f : \mathcal{E} \to \mathcal{Z}_E$ that projects an event representation $e \in \mathcal{E}$ into a latent feature vector $\mathbf{Z}_E \in \mathcal{Z}_E$. We optimize f to minimize the distance between $\mathcal{Z}_E$ and $\mathcal{Z}_I$, where $\mathcal{Z}_I$ represents the latent manifold of the DUNE [48] foundation model. As described in Sec. 3.1, for a given image x, the corresponding reference feature is defined as $\mathbf{Z}_I = f_{\text{DUNE}}(x)$, where $\mathbf{Z}_I \in \mathbb{R}^{(HW+1)\times d}$.

To select the optimal input format for e, we evaluated various representations (see Sec. 2) and found that the choice of input has a negligible effect on final alignment quality (see Sec. 4). Consequently, we adopt the voxel grid due to its wide adoption in event-based learning, compatibility with established baselines, and its suitability for the convolutional embedder design we employ, see Appendix Sec. S5.

3.3 Training and Architecture

Standard vision transformer (ViT) architectures, including DUNE, are designed to process RGB images through a patchified embedding stage. Formally, the DUNE [48] encoder can be decomposed as $f_{\text{DUNE}} = f_{\text{back}} \circ f_{\text{emb}}$, where $f_{\text{emb}} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{M\times d}$ projects raw pixels into M patch tokens, and $f_{\text{back}} : \mathbb{R}^{M\times d} \to \mathcal{Z}_I$ represents the transformer backbone.

Event-Specific Embedding. Because our voxel grid representation contains five temporal bins and distinct structural properties compared to RGB frames, the original f_{emb} is incompatible. We substitute the image-based patch module with

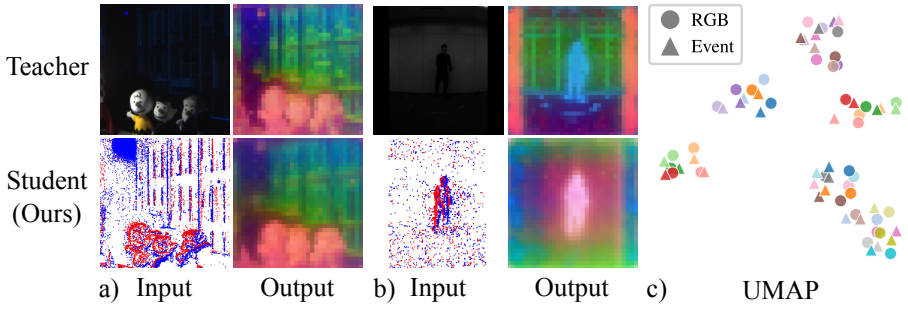

Fig. 3: Qualitative comparison of DUNE and REALM feature manifolds. The spatial feature maps in a) and b), visualized via PCA, show that REALM learns to identify semantic structures similarly to the RGB network. In b), the masking suppresses static background visible to the camera but invisible to the events. In c), the UMAP plot further confirms this by showing event and RGB features form consistent clusters.

f^E_{emb}, a lightweight convolutional stem that maps the voxel grid directly into a set of $M \times d$ tokens. Our goal is not to force the intermediate embeddings to match ($f_{\text{emb}}(x) = f^E_{\text{emb}}(e)$), but rather to ensure that the final latent outputs are aligned within the foundation manifold: $f_{\text{back}}(f_{\text{emb}}(x)) \approx f^E_{\text{back}}(f^E_{\text{emb}}(e))$.

Low-Rank Adaptation (LoRA). To adapt the backbone f_{back} for the event modality while preserving its pretrained semantic and geometric knowledge, we employ Low-Rank Adaptation (LoRA) [29]. We keep the original weights $W \in \mathbb{R}^{d\times k}$ frozen and model the necessary cross-modal refinement as a low-rank decomposition: $W' = W + AB$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ are trainable matrices with rank $r \ll \min(d, k)$. This allows REALM to bridge the modality gap without the risk of catastrophic forgetting.

To robustly distill knowledge from the DUNE [48] encoder, we propose a dual-masking training strategy described as follows (see Fig. 2).

Progressive Spatial Distillation. To prevent the student model from overfitting to static backgrounds (see Fig. 3b), we constrain the distillation loss using a spatially-aware mask $\mathcal{M}$ derived from the input event activity.

Let $E^i \in \{0,1\}^M$ be a binary occupancy grid for the voxel grid input e_i, where $E^i_j = 1$ if the j-th voxel contains at least one event, and $E^i_j = 0$ otherwise. We define the active distillation mask for e_i at training step s as: $\mathcal{M}^i(s) = \text{Dilation}(E^i, \sigma(s))$, where $\text{Dilation}(\cdot, \sigma)$ denotes a morphological dilation operation with a structuring element of radius σ. We introduce a spatial curriculum by defining $\sigma(s)$ as a monotonically increasing function of the training progress: $\sigma(s) = \min(\sigma_{max}, \lfloor \alpha \cdot s \rfloor)$, where α is the expansion rate and σ_{max} is the radius required to cover the entire spatial grid.

By gradually expanding the penalization region $\mathcal{M}$, we force the model to move beyond trivial edge-matching and learn a holistic scene understanding by leveraging the teacher's static priors in regions initially devoid of events.

Input Patch Masking. To encourage the network to build compressed global representations and discourage reliance on local interpolation, we employ a stochastic patch dropout strategy inspired by Masked Autoencoders (MAE) [25]. Following an initial warmup phase, we uniformly mask a fixed percentage ρ of the input event tokens $\mathcal{X}_E = f^E_{\text{emb}}(e)$ produced by the embedder. By denying the backbone access to the full spatial grid, we force the LoRA-adapted layers to reconstruct dense, cross-modal features from highly degraded inputs. This prevents the model from relying on local interpolation and ensures the learned latent space Z_E captures the geometric and semantic priors of the DUNE teacher.

Loss Function. To ensure feature alignment, we define the total loss as $\mathcal{L}_{\text{total}} = \sum_{i=1}^{B} \mathcal{L}_i$, where $\mathcal{L}_i$ is the loss for the voxel grid i in a batch of size B defined as a weighted combination of ℓ_1 distance, cosine similarity and MSE losses between the DUNE [48] embeddings of the image and event modalities:

$$\mathcal{L}^i = \frac{1}{\sum_{j=1}^{M} \mathcal{M}^i_j(s)} \sum_{j=1}^{M} \mathcal{M}^i_j(s) \cdot \left[\lambda_{\text{MSE}} \mathcal{L}^j_{\text{MSE}} + \lambda_{\cos} \mathcal{L}^j_{\cos} + \lambda_{\ell_1} \mathcal{L}^j_{\ell_1} \right] \tag{1}$$

where λ_{MSE}, $\lambda_{\cos}$, and λ_{L1} control the relative contributions of each term. This objective enforces per-dimension and global directional alignment, promoting cross-modal generalization as shown by the uniform manifold approximation and projection (UMAP) [38] in Fig. 3 c). We show the effect of the proposed masking strategy in Appendix Sec. S1.

4 Results

In this section, we present a comprehensive experimental evaluation of REALM across three primary downstream tasks: monocular depth estimation, semantic segmentation, and wide-baseline feature matching. We report how our model generalizes to other tasks such as image reconstruction in Sec. S6.

To train the event-specific embedder and the LoRA adapters, we utilize a diverse collection of synchronized event-RGB datasets. We selected DSEC [21, 22,33], EventScape [19], EventPointMesh [28], EDS [26], and M3ED [8] to provide a comprehensive overview of diverse geometric and semantic structures, covering a wide range of scenarios. All training input data are resized and cropped to a voxel-grid with a resolution of 448×448 with five temporal bins. For further details on the datasets and our data preparation, refer to the Appendix Sec. S2.

To choose the event representation for our model, we conducted a preliminary evaluation on 5% of our training data comparing voxel grids [65], Tencode [30], and ERGO [68]. All three yielded nearly identical alignment losses (0.0687, 0.0698, and 0.0698, respectively), confirming that the choice of input representation has a negligible effect on cross-modal alignment quality.

We apply LoRA to the attention, projection, and feed-forward layers with a rank of 32 and a learning rate of 1×10^{-3} following a cosine decay schedule.

Dataset	Depth Cut-offs	Avg. Absolute Depth Err. [m] DUNE [48]	E2Depth [27]	Zhu et al. [65]	EMoDepth [67]	REALM (Ours)
outdoor day 1	10 m	1.16	1.85	2.72	**1.40**	1.85
	20 m	1.76	2.64	3.84	**2.07**	2.42
	30 m	2.12	3.13	4.40	**2.65**	2.76
outdoor night 1	10 m	2.13	3.38	3.13	2.18	**2.08**
	20 m	3.10	3.82	4.02	2.70	**2.51**
	30 m	3.51	4.46	4.89	3.64	**3.18**
outdoor night 2	10 m	2.45	**1.67**	2.19	2.06	2.00
	20 m	3.46	2.63	3.15	2.76	**2.31**
	30 m	3.86	3.58	3.92	3.42	**2.98**
outdoor night 3	10 m	2.33	**1.42**	2.86	2.09	1.79
	20 m	3.37	2.33	4.46	2.82	**2.15**
	30 m	3.79	3.18	5.05	3.52	**2.97**

Table 1: Quantitative results on the MVSEC [64] dataset. We report results for DUNE [48] as a reference, and for our event-only method, REALM. Comparisons are provided against a supervised approach [27], as well as unsupervised methods [65, 67].

Method	AbsR ↓	SqR ↓	RMSE ↓	log ↓	SILog ↓	σ_1 ↑	σ_2 ↑	σ_3 ↑	10	20	30	Latency (ms) ↓	Params ↓
DUNE	0.351	0.424	7.828	0.415	0.166	0.559	0.764	0.878	2.02	2.92	3.32	6.02 (A100)	0.39M
EReFormer [36]	**0.275**	0.382	–	–	**0.120**	**0.607**	**0.807**	**0.915**	**1.38**	2.15	**2.73**	35.17 (Tesla V100)	29.9M
D. AnyEv. [4]	0.362	0.697	**6.511**	0.438	0.211	0.494	0.760	0.890	–	–	–	9.20 (A100)	4.06M
AnyEv. Stream [66]	–	–	–	–	–	–	–	–	**1.38**	**2.02**	3.01	22.00 (A100)	10.89M
REALM (Ours)	0.349	**0.312**	8.109	**0.421**	0.169	0.500	0.754	0.875	1.93	2.35	2.97	**6.71 (A100)**	**0.39M**

Table 2: Extended Depth Metrics and Comparison on MVSEC. Specialized transformer decoders reach strong accuracy at substantially higher parameter and latency cost, whereas REALM remains competitive with only a linear head on a frozen backbone. Params are decoder/head parameters only. Values are averaged over scenes.

For each task, we compare REALM against state-of-the-art architectures specifically engineered for event data. We also report the performance of DUNE [48] on synchronized RGB frames to compare REALM to its teacher and show where the event modality helps in perception tasks.

A key strength of our framework is the minimal overhead required for downstream task adaptation: with the segmentation and monocular depth tasks, we demonstrate that REALM's latent space is semantically and geometrically rich enough to support these tasks using only simple linear heads. Notably, these heads are not trained on events, but rather on RGB images on the frozen DUNE [48] encoder. See the Appendix Sec. S5 for further details on the models.

In the same way, for feature matching, we directly utilize the frozen MASt3R [34] head that has been trained on a large dataset of 1.7M images [48]. This requires no retraining or fine-tuning on event data, serving as a strong verification of our cross-modal alignment strategy.

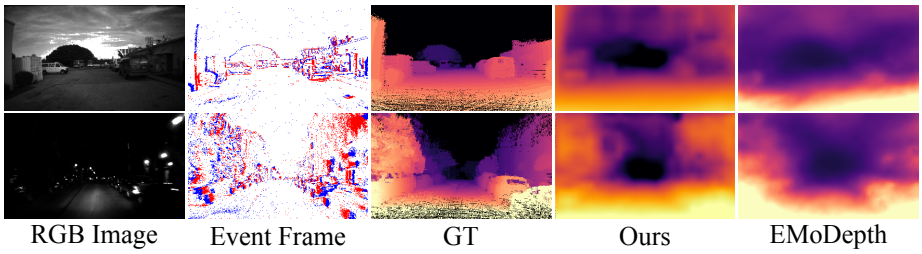

Fig. 4: Qualitative depth estimation results on the MVSEC [64] dataset. From left to right: RGB frame (only for visualization purposes), event frame, ground truth depth, and REALM and EMoDepth [67] predictions. Despite the sparsity of the event stream, REALM preserves the scene structure just using a single linear projector.

4.1 Depth Estimation

We evaluate the dense geometric reasoning capabilities of REALM by performing monocular depth estimation. Following recent trends in foundation model evaluation [41], we aim to demonstrate that a frozen REALM encoder, paired with a minimal prediction head, can compete with specialized event-based depth architectures. Accordingly, our setup is minimal, as we utilize a single linear head trained exclusively on frozen RGB features from DUNE [48], with no event-specific depth supervision at any stage.

Experimental Setup. We benchmark our method on the MVSEC [64] dataset using the evaluation protocol established by E2Depth [27]. We treat depth estimation as a classification task by discretizing depth values into bins and predicting per-pixel probability distributions. To verify the robustness of our latent space, we first train a linear head for 300 epochs on top of the frozen DUNE [48] backbone using real and synthetic RGB images, from MVSEC [64] and DENSE [27], respectively. Then we use the trained head directly on REALM event features with no finetuning. Further details on the training are in the Appendix Sec. S3.

To accommodate the 260×346 resolution of the MVSEC [64] dataset while maintaining the 448×448 input size of the foundation model, we symmetrically pad the original data. We then remove the padding from the prediction. To ensure temporal stability, we implement a memory hold mechanism. That is, in time windows where no event activity is recorded, the model holds the most recent valid depth estimate, preventing noisy outputs during brief sensor inactivity.

Quantitative Results. The results are summarized in Tab. 1 and Fig. 4. Despite this significant gap in both architectural complexity and task-specific supervision, REALM remains competitive across nearly all evaluated sequences. Notably, our approach demonstrates superior robustness in challenging outdoor night scenarios where traditional frame-based sensors degrade. For instance, in the outdoor night 1 sequence at a 20 m cut-off, REALM achieves an error of 2.51 m, significantly outperforming both the unsupervised EMoDepth [67] (2.70 m) and Zhu et al. [65] (4.02 m). Crucially, although REALM is trained only

Metric	DUNE [48]	ESS [53]	EV-SegNet [1]	HALSIE [11]	ESEG [63]	REALM (Ours)
Acc. [%] ↑	93.33	<u>89.25</u>	88.61	89.01	**91.47**	89.23
mIoU [%] ↑	67.64	51.57	51.76	52.43	**57.55**	<u>55.37</u>

Table 3: Quantitative results on the DSEC dataset. We use a linear segmentation head to perform the segmentation task. DUNE shows the upper-bound performance of the shared architecture on intensity frames.

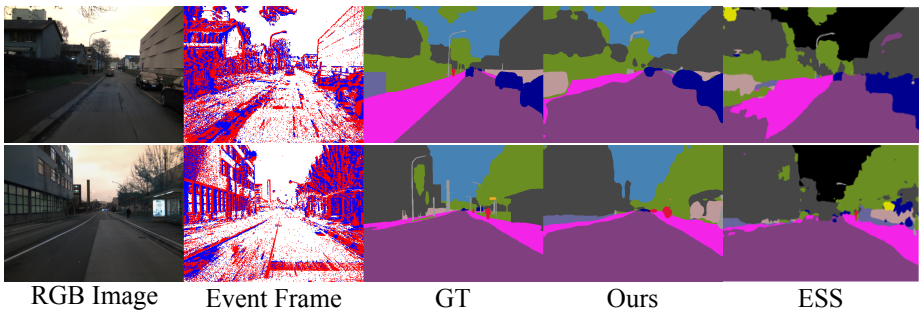

Fig. 5: Qualitative semantic segmentation on the DSEC [21] driving dataset. REALM effectively identifies key classes such as road, vehicles, and sidewalk. Our model achieves these results using a frozen backbone and a single linear head, demonstrating the semantic richness of the aligned latent space.

on RGB data, distilled from the frame-based teacher DUNE [48], it surpasses the teacher model in low-illumination and high-dynamic-range conditions, that is, in the regime where events are more informative than frames, which lose detail in under- and over-exposed regions. The student can thus exceed the source of its own supervision: our aligned manifold not only inherits DUNE's representation but also lets the event modality contribute cues unavailable in RGB.

In Tab. 2 we report additional depth metrics alongside recent transformer-based methods [4, 36, 66]. While these specialized decoders achieve strong accuracy on depth estimation, they do so with 10–77× more parameters and 1.4–5× higher latency than REALM, which attains competitive results using only a linear head on a frozen backbone.

4.2 Semantic Segmentation

To evaluate the semantic qualities of the aligned latent space, we benchmark REALM on the task of semantic segmentation. This experiment serves to demonstrate that the high-level scene understanding present in RGB foundation models can be seamlessly transferred to the event modality.

Experimental Setup. We utilize the DSEC [21] dataset, which provides semantic labels across eleven distinct classes. Following the evaluation protocol of ESS [53], we compute standard confusion-matrix-based metrics, including Pixel Accuracy

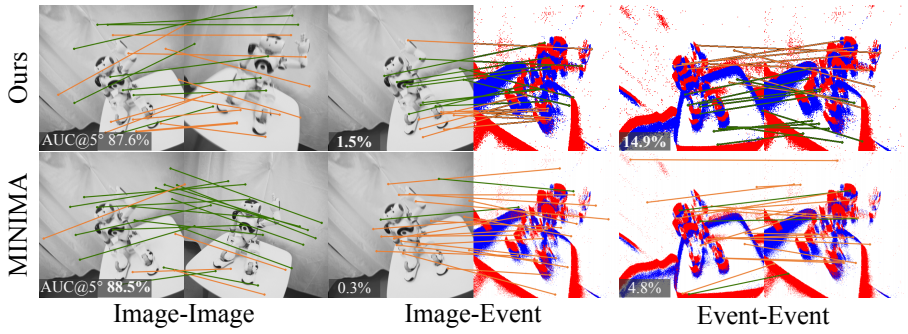

Fig. 6: Cross-modal and intra-modal feature matching. We report the AUC at 5° per scene and visualize the correspondence estimation results for inliers and outliers in green and orange, respectively.

and Mean Intersection over Union (mIoU). To highlight the off-the-shelf utility of our features, we employ a minimal single-layer linear head. The head is trained using the frozen DUNE [48] encoder on RGB images and subsequently used as is, with no further finetuning, on REALM event embeddings.

To accommodate the 480×640 resolution of the DSEC [21] dataset while maintaining the 448×448 input size of the foundation model, we implement an overlapping tiling strategy. During inference, we extract four 448×448 tiles from the corners of the input grid. The final per-pixel prediction is determined by accumulating the raw model outputs across tiles and normalizing the result by a count mask to average overlapping regions. To ensure temporal stability, we implement a memory hold mechanism as described in Sec. 4.1.

Quantitative Results. The results are summarized in Tab. 3 and Fig. 5. REALM achieves a Pixel Accuracy of 89.23% and an mIoU of 55.37%. These results are highly competitive with specialized architectures such as ESS [53] (51.57% mIoU) and EV-SegNet [1] (51.76% mIoU). While ESEG [63], which fine-tunes a SegFormer [59] decoder with edge supervision, maintains a slight lead in mIoU (57.55%), REALM attains comparable performance without requiring specialized edge-guidance modules or complex recurrent decoders. We report a detailed per-class IoU and the corresponding confusion matrix in the Appendix Sec. S4.

4.3 Feature Matching

We evaluate the quality of the learned latent space by benchmarking REALM on wide-baseline feature matching and relative pose estimation. Unlike prior works that rely on task-specific descriptors, we equip REALM with the MASt3R [34] head. Crucially, this geometric head remains frozen and was trained exclusively on RGB data. This setup serves as the ultimate test for our cross-modal alignment: if REALM correctly maps events into the DUNE manifold, the MASt3R [34] head should treat event features as native RGB representations.

Dataset	Metric AUC	LLAK [10]	RATE [31]	EventPoint [30]	SuperEvent [6]	REALM (Ours)
ECD	@5° ↑	0.7	3.3	1.6	22.7	**26.2**
	@10° ↑	1.4	8.4	3.0	35.8	**46.8**
	@20° ↑	2.1	18.0	5.4	46.7	**63.3**
EDS	@5° ↑	0.5	2.1	1.6	15.2	**18.3**
	@10° ↑	0.7	5.1	2.8	26.4	**34.1**
	@20° ↑	1.0	10.3	5.2	40.1	**55.3**

Table 4: Quantitative results on the ECD [40] and EDS [26] datasets. We report the Area Under Curve (AUC) in % at different thresholds.

Modality	Metric	MINIMA [46]	SuperEvent [6]	REALM (Ours)
Event-Event	AUC@5° ↑	39.0	5.7	**47.7**
	AUC@10° ↑	53.9	11.4	**64.9**
	AUC@20° ↑	68.0	19.6	**79.1**
	Med. Err (°) ↓	3.93	46.33	**2.75**
	S. Rate (%) ↑	**100.0**	99.9	99.7
Image-Event	AUC@5° ↑	23.8	-	**26.0**
	AUC@10° ↑	39.4	-	**46.3**
	AUC@20° ↑	55.1	-	**64.6**
	Med. Err (°) ↓	6.46	-	**5.45**
	S. Rate (%) ↑	**100.0**	-	99.6

Table 5: Pose estimation performance grouped by sensing modality. Higher AUC in % and success rate are better, while a lower median angular error is better. The experiments are conducted on the VECtor [17] dataset.

Experimental Setup. We evaluate on the ECD [40], EDS [26], and VECtor [17] datasets. Following the protocol in [30], we report the Area Under Curve (AUC) for relative rotation error at thresholds of 5°, 10°, and 20°, as well as the median angular error. We compare against specialized event-based baselines, including LLAK [10], RATE [31], EventPoint [30], SuperEvent [6], and the recent cross-modal MINIMA [46].

Quantitative Results. As summarized in Tab. 4 and Fig. 6, REALM outperforms all existing methods by a significant margin. On the ECD [40] dataset, our method achieves an AUC@20° of 63.3%, representing a substantial improvement over the previous state-of-the-art SuperEvent [6] (46.7%). These results demonstrate that leveraging the geometric priors of an RGB foundation model is more effective than training specialized event descriptors from scratch.

Wide-Baseline Robustness. To further probe the robustness of our alignment, we conduct a wide-viewpoint evaluation following the protocol in [42]. We match a query event stream against targets at increasing angular distances. Tab. 5 shows that while MINIMA [46] degrades as the viewpoint change increases, REALM is capable of producing more accurate feature matches, leading to lower orientation

errors. REALM, leveraging the MASt3R [34] decoder, which provides strong geometric cues, outperforms the other methods in both event-event and image-event matching tasks. We refer the reader to Sec. S4 for further analysis.

Notably, REALM achieves these results in a zero-shot manner, whereas MINIMA [46] requires a specialized generative data engine and fine-tuning to reach generalization. This highlights that our joint LoRA-based optimization effectively "unlocks" universal geometric reasoning for the event modality. We report in the Appendix a breakdown of per-scene evaluation in Sec. S4.

5 Limitations and Future Work

Our approach has two main limitations that also point to directions for future work. First, our convolutional embedder operates on fixed-size voxel grids, limiting its flexibility across sensors of differing resolution; we currently bridge this gap with padding and overlapping tiling (Secs. 4.1 and 4.2). A resolution-agnostic embedder operating directly on the asynchronous stream would remove this constraint. Second, the embedder does not explicitly model long-term temporal dynamics. We handle low-event periods with a memory-holding mechanism, which is effective but remains a practical workaround; asynchronous or recurrent embedders are a principled next step, maintaining temporal consistency through sparse periods without frame-level heuristics.

6 Conclusion

We have presented REALM, a cross-modal visual encoder that leverages large-scale RGB pretraining to achieve state-of-the-art performance in event-based vision. By employing a cross-modality alignment strategy, REALM effectively bridges the gap between asynchronous event streams and intensity frames, enabling the direct use of foundation models originally designed for image data. Our results demonstrate that REALM successfully inherits the deep geometric and semantic priors of the DUNE [48] teacher. Most notably, this alignment enables the direct, zero-shot application of complex image-trained decoders, such as MASt3R [34], to raw event data—outperforming specialized architectures. Finally, by showing that event data can inhabit the same manifold as RGB foundation models, REALM provides a solution to the scarcity of annotated event datasets and unlocks a broad ecosystem of RGB-based techniques, from calibration-free SLAM to cross-modal localization, for the event domain.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the RGPIN program and the Canada Research Chairs program, the Canada Foundation for Innovation, and the Ontario Research Fund. This research was enabled in part by support provided by Compute Ontario and the Digital Research Alliance of Canada.

References

1. Alonso, I., Murillo, A.C.: EV-SegNet: Semantic segmentation for event-based cameras. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 1624–1633 (2019)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
3. Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., Lucas, T.: Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In: Eur. Conf. Comput. Vis. (ECCV). pp. 202–218. Springer (2024)
4. Bartolomei, L., Mannocci, E., Tosi, F., Poggi, M., Mattoccia, S.: Depth AnyEvent: A cross-modal distillation paradigm for event-based monocular depth estimation. In: Int. Conf. Comput. Vis. (ICCV). pp. 19669–19678 (2025)
5. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Eur. Conf. Comput. Vis. (ECCV). pp. 404–417 (2006)
6. Burkhardt, Y., Schaefer, S., Leutenegger, S.: SuperEvent: Cross-modal learning of event-based keypoint detection for SLAM. In: Int. Conf. Comput. Vis. (ICCV). pp. 8918–8928 (2025)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Int. Conf. Comput. Vis. (ICCV) (2021)
8. Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K.: M3ED: Multi-robot, multi-sensor, multi-environment event dataset. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 4016–4023 (2023)
9. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1290–1299 (2022)
10. Chiberre, P., Perot, E., Sironi, A., Lepetit, V.: Long-lived accurate keypoints in event streams. arXiv e-prints (2022)
11. Das Biswas, S., Kosta, A., Liyanagedera, C., Apolinario, M., Roy, K.: HALSIE: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In: IEEE Winter Conf. Appl. Comput. Vis. (WACV). pp. 5964–5974 (2024)
12. Deng, Y., Chen, H., Liu, H., Li, Y.: A voxel graph CNN for object classification with event cameras. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1172–1181 (2022)
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW). pp. 224–236 (2018)
14. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Adv. Neural Inf. Process. Syst. **27** (2014)
15. Farooq Bhat, S., Alhashim, I., Wonka, P.: AdaBins: Depth estimation using adaptive bins. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4009–4018 (2021)
16. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (2022)

17. Gao, L., Liang, Y., Yang, J., Wu, S., Wang, C., Chen, J., Kneip, L.: VECtor: A versatile event-centric benchmark for multi-sensor SLAM. IEEE Robot. Autom. Lett. **7**(3), 8217–8224 (2022)
18. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Int. Conf. Comput. Vis. (ICCV) (October 2019)
19. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robot. Autom. Lett. **6**(2), 2822–2829 (2021)
20. Gehrig, D., Scaramuzza, D.: Low-latency automotive vision with event cameras. Nature (2024)
21. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: DSEC: A stereo event camera dataset for driving scenarios. IEEE Robot. Autom. Lett. (2021)
22. Gehrig, M., Millhäusler, M., Gehrig, D., Scaramuzza, D.: E-RAFT: Dense optical flow from event cameras. In: Int. Conf. 3D Vis. (3DV). pp. 197–206 (2021)
23. Ghosh, S., Gallego, G.: Event-based stereo depth estimation: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **47**(10), 9130–9149 (2025)
24. He, B., Wang, Z., Zhou, Y., Chen, J., Singh, C.D., Li, H., Gao, Y., Shen, S., Wang, K., Cao, Y., Xu, C., Aloimonos, Y., Gao, F., Fermuller, C.: Microsaccade-inspired event camera for robotics. Sci. Robot. (2024)
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 16000–16009 (2022)
26. Hidalgo-Carrió, J., Gallego, G., Scaramuzza, D.: Event-aided direct sparse odometry. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5781–5790 (2022)
27. Hidalgo-Carrió, J., Gehrig, D., Scaramuzza, D.: Learning monocular dense depth from events. In: Int. Conf. 3D Vis. (3DV). pp. 534–542 (2020)
28. Hori, R., Isogawa, M., Mikami, D., Saito, H.: EventPointMesh: Human mesh recovery solely from event point clouds. IEEE Trans. Vis. Comput. Graph. **31**(09), 5593–5610 (Sep 2025)
29. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (ICLR) (2022)
30. Huang, Z., Sun, L., Zhao, C., Li, S., Su, S.: EventPoint: Self-supervised interest point detection and description for event-based camera. In: IEEE Winter Conf. Appl. Comput. Vis. (WACV). pp. 5396–5405 (2023)
31. Ikura, M., Le Gentil, C., Müller, M.G., Schuler, F., Yamashita, A., Stürzl, W.: RATE: Real-time asynchronous feature tracking with event cameras. In: IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS). pp. 11662–11669 (2024)
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment Anything. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4015–4026 (2023)
33. Kong, L., Liu, Y., Ng, L.X., Cottereau, B.R., Ooi, W.T.: OpenESS: Event-based semantic scene understanding with open vocabularies. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2024)
34. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: Eur. Conf. Comput. Vis. (ECCV). pp. 71–91. Springer (2024)
35. Leutenegger, S., Chli, M., Siegwart, R.Y.: BRISK: Binary robust invariant scalable keypoints. In: Int. Conf. Comput. Vis. (ICCV). pp. 2548–2555. IEEE (2011)

36. Liu, X., Li, J., Shi, J., Fan, X., Tian, Y., Zhao, D.: Event-based monocular depth estimation with recurrent transformers. IEEE Trans. Circuits Syst. Video Technol. **34**(8), 7417–7429 (2024)
37. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
38. McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: Uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018)
39. Messikommer, N., Fang, C., Gehrig, M., Scaramuzza, D.: Data-driven feature tracking for event cameras. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2023)
40. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. Int. J. Rob. Res. (2017)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024)
42. Polizzi, V., Cannici, M., Scaramuzza, D., Kelly, J.: FaVoR: Features via voxel rendering for camera relocalization. In: IEEE Winter Conf. Appl. Comput. Vis. (WACV). pp. 44–53 (February 2025)
43. Polizzi, V., Yang, S., Clark, Q., Kelly, J., Gilitschenski, I., Lindell, D.B.: VibES: Induced vibration for persistent event-based sensing. In: Int. Conf. 3D Vis. (3DV). pp. 1–10 (2026)
44. Rebecq, H., Horstschaefer, T., Scaramuzza, D.: Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In: Proc. Br. Mach. Vis. Conf. (BMVC). vol. 2, p. 7 (2017)
45. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 1964–1980 (2021)
46. Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X.: MINIMA: Modality invariant image matching. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2025)
47. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: Int. Conf. Comput. Vis. (ICCV). pp. 2564–2571. IEEE (2011)
48. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., de Jorge, P., Larlus, D., Kalantidis, Y.: DUNE: Distilling a universal encoder from heterogeneous 2D and 3D teachers. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 30084–30094 (June 2025)
49. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4938–4947 (2020)
50. Shariff, W., Dilmaghani, M.S., Kielty, P., Moustafa, M., Lemley, J., Corcoran, P.: Event cameras in automotive sensing: A review. IEEE Access **12**, 51275–51306 (2024)
51. Shiba, S., Aoki, Y., Gallego, G.: Secrets of event-based optical flow. In: Eur. Conf. Comput. Vis. (ECCV). pp. 628–645 (2022)
52. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt,

L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv:2508.10104 (2025)
53. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: ESS: Learning event-based semantic segmentation from still images. In: Eur. Conf. Comput. Vis. (ECCV). pp. 341–357. Springer (2022)
54. Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., Herath, D.: Semantic segmentation using vision transformers: A survey. Eng. Appl. Artif. Intell. **126**, 106669 (2023)
55. Vidal, A.R., Rebecq, H., Horstschaefer, T., Scaramuzza, D.: Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. IEEE Robot. Autom. Lett. **3**(2), 994–1001 (2018)
56. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2025)
57. Wang, L., Chae, Y., Yoon, S.H., Kim, T.K., Yoon, K.J.: EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 608–619 (June 2021)
58. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2024)
59. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. **34**, 12077–12090 (2021)
60. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the power of large-scale unlabeled data. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2024)
61. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: Learned invariant feature transform. In: Eur. Conf. Comput. Vis. (ECCV). pp. 467–483 (2016)
62. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected CRFs for monocular depth estimation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3916–3925 (2022)
63. Zhao, Y., Lyu, G., Li, K., Wang, Z., Chen, H., Yang, Z., Deng, Y.: ESEG: Event-based segmentation boosted by explicit edge-semantic guidance. In: AAAI Proc. AAAI Conf. Artif. Intell. vol. 39, pp. 10510–10518 (2025)
64. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. IEEE Robot. Autom. Lett. **3**(3), 2032–2039 (2018)
65. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 989–997 (2019)
66. Zhu, J., Pan, T., Cao, Z., Liu, Y., Kwok, J.T., Xiong, H.: Depth Any Event Stream: Enhancing event-based monocular depth estimation via dense-to-sparse distillation. In: Int. Conf. Comput. Vis. (ICCV). pp. 5146–5155 (2025)
67. Zhu, J., Liu, L., Jiang, B., Wen, F., Zhang, H., Li, W., Liu, Y.: Self-supervised event-based monocular depth estimation using cross-modal consistency. In: IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS). pp. 7704–7710 (2023)
68. Zubić, N., Gehrig, D., Gehrig, M., Scaramuzza, D.: From chaos comes order: Ordering event representations for object recognition and detection. In: Int. Conf. Comput. Vis. (ICCV). pp. 12846–12856 (October 2023)