REPURPOSING A MONOCULAR DEPTH FOUNDATION MODEL FOR SPIKE CAMERAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spike cameras encode incident light as binary streams at high temporal resolution, enabling perception under rapid motion and challenging illumination. Yet monocular depth estimation from spikes is limited by scarce paired spike-depth data and the modality gap between spike measurements and RGB imagery, hindering the reuse of pretrained depth models. We propose to **R**epurpose a m**O**nocular **DE**pth foundation model for **S**pike cameras (RODES) and design a cross-modal distillation framework. We propose a spatio-temporal spike adapter (STSA) to aggregate local temporal dynamics and global context into a foundation-compatible representation. Further, we introduce pretrained-manifold alignment distillation (PMAD) to transfer pretrained RGB priors into shallow spike representations while preserving deeper adaptation to spike-specific cues. To complement this representation-level transfer, we introduce a teacher-guided data expansion strategy that augments the training set with DSEC-derived simulated spike streams and pseudo-depth generated by an RGB teacher. On the DENSE-Spike test set, our method reduces AbsRel from 0.610 to 0.326 over Spike-T. Our ablations validate the complementary benefits of feature distillation and data expansion. Qualitative comparisons on Outdoor-Spike further provide evidence in the real world. **Our code will be available upon publication.**

1 INTRODUCTION

Depth estimation is a fundamental problem in computer vision, providing essential geometric information for 3D scene understanding (Eigen et al., 2014). Most learning-based monocular depth methods operate on images captured by frame-based cameras, whose finite exposure windows can lead to motion blur and over- or under-exposure in highly dynamic scenes (Huang et al., 2023). Inspired by biological vision, spike cameras encode incident-light intensity as continuous binary spike streams at high temporal resolution, making them promising under rapid motion and challenging illumination (Dong et al., 2017; Zhao et al., 2021; Zheng et al., 2026). However, how to jointly model temporal dynamics and spatial scene structure from such high-frequency binary streams remains a key challenge for spike-based depth estimation. An existing method has demonstrated direct depth estimation from spike streams and introduced DENSE-Spike as a benchmark for this task (Zhang et al., 2022). Despite this progress, spike data with depth supervision remain limited in scale and scene diversity. This insufficient supervision and scene coverage may restrict the ability of spike-specific models to generalize to real-world scenes.

Pretrained on large-scale and diverse visual data, Visual Foundation Models (VFMs) have demonstrated strong transferability across visual tasks (Oquab et al., 2024). In particular, RGB depth models such as Depth Anything have demonstrated robust depth estimation across diverse scenes, offering rich visual and geometric priors (Yang et al., 2024b). However, exploiting these priors for spike-based depth estimation requires transferring knowledge learned from spatial images to high-frequency binary spike measurements. In this paper, we investigate how pretrained RGB depth priors can be effectively transferred to the spike domain. Specifically, we focus on three key aspects: **spatio-temporal spike representation learning**, **pretrained feature-manifold alignment**, and **training-data expansion in both scale and scene diversity**.

RGB depth models are pretrained on spatial images, whereas spike measurements distribute intensity information across high-frequency binary sequences. Fixed spike encodings can convert these

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.

sequences into frame-like inputs compatible with pretrained models, but rely on predefined temporal rules that are not optimized for depth estimation (Dong et al., 2019). To bridge this modality gap, we introduce a learnable spatio-temporal spike adapter (STSA). This adapter aggregates task-relevant spatio-temporal cues into a representation compatible with the pretrained backbone, allowing the input mapping to be optimized jointly with the depth model.

Although the spike-aware adapter can learn informative spike representations, these features may not be readily exploited by the RGB-pretrained backbone. Therefore, we expect the proposed adapter not only to extract dense spatio-temporal information from spike streams, but also to organize it into representations that the pretrained backbone can effectively process. However, depth supervision alone is insufficient to satisfy either requirement, as it provides no direct RGB feature reference for the intermediate representations, leaving any mismatch at the modality interface to be resolved indirectly through the final depth objective. We introduce pretrained-manifold alignment distillation (PMAD), in which a frozen RGB teacher provides reference features from paired RGB and spike observations of the same scene. We concentrate this guidance on selected early layers to align spike features with the representation manifold expected by the pretrained backbone, while allowing deeper representations to adapt under the depth supervision.

Feature-level guidance enriches supervision on the available samples, but it cannot broaden the scene coverage and data scale for cross-modal adaptation. With DENSE-Spike alone, cross-modal adaptation remains confined to a limited set of synthetic scenes. To address this issue, we introduce a teacher-guided data expansion strategy. Specifically, we derive simulated spike streams from real-world multimodal sequences in DSEC and use a pretrained depth teacher to generate pseudo-depth labels from the corresponding RGB observations, enabling training on additional scenes without new spike-depth annotations (Gehrig et al., 2021).

Experimental results on the DENSE-Spike test set show that our method reduces AbsRel from 0.610 to 0.326 compared with Spike-T, a relative reduction of 46.6% (Zhang et al., 2022). Qualitative comparisons on Outdoor-Spike further show improved depth-map quality on unseen real spike scenes, suggesting better real-world generalization. In summary, our contributions are as follows:

- We present RODES, which repurposes a pretrained RGB depth model for spike-only depth estimation by combining a learnable spike-to-VFM input adapter (STSA) with shallow feature alignment to a frozen RGB teacher (PMAD).
- We introduce a teacher-guided data expansion strategy that pairs DSEC-derived simulated spike streams with relative pseudo-depth from a frozen RGB depth teacher.
- Experiments on DENSE-Spike, DSEC-Spike, and Outdoor-Spike show consistent improvements over spike-specific and foundation-model-based baselines, reducing AbsRel by 14.9% over the strongest baseline on DENSE-Spike.

2 Related Work

Spike-Based Vision. Unlike event cameras, which record logarithmic intensity changes (Lichtsteiner et al., 2008), spike cameras integrate incident light and fire upon reaching a threshold, retaining intensity cues in both static and dynamic regions (Dong et al., 2017; 2019). Spike-specific models have been developed for image reconstruction (Zhu et al., 2020; Zhao et al., 2021; 2023; 2024), optical flow (Hu et al., 2022; Zhao et al., 2022; 2026), deblurring (Chen et al., 2023), and 3D reconstruction (Zhang et al., 2024). For depth, Spike Transformer (Zhang et al., 2022) and SpikeStereoNet (Gao et al., 2026) address monocular and stereo estimation with spike-native architectures trained solely under spike-domain supervision, leaving the transfer of pretrained RGB depth priors largely unexplored.

Monocular Depth Foundation Models. Monocular depth estimation has progressed from supervised regression (Eigen et al., 2014) to models trained across mixed datasets (Ranftl et al., 2022; 2021). Built on DINOv2 features (Oquab et al., 2024), Depth Anything and its successor (Yang et al., 2024a;b) achieve robust zero-shot depth through large-scale pseudo-labeling, but assume conventional images and cannot directly consume spike streams.

Cross-Modal Distillation. Knowledge distillation transfers supervision through outputs or intermediate features (Hinton et al., 2015; Romero et al., 2015) and has been extended across modalities (Gupta et al., 2016). For event cameras, image models have been distilled via reconstruction guidance or embedding alignment (Wang et al., 2021; Sun et al., 2022), and more recently via VFM depth targets (Bartolomei et al., 2025), sparsity-aware feature distillation (Zhu et al., 2025), or representation pretraining (Chen et al., 2026). Close to ours, Bartolomei et al. (2025) likewise use VFM pseudo-depth, but supervise only the outputs of event models built on a fixed encoding; we instead learn a spike-to-VFM input interface and align shallow features to the pretrained manifold, which also benefits a fixed reconstruction-based interface.

3 Preliminaries: Spike Generation Mechanism

A spike camera independently integrates incident light at each pixel and fires when the accumulated photoelectric response reaches a threshold (Dong et al., 2017; Zhao et al., 2021). For intensity $I(t)$, the k-th firing time satisfies:

$$\int_{t_{k-1}}^{t_k} \alpha I(t)\,\mathrm{d}t = \theta, \tag{1}$$

where α and θ are the photoelectric conversion coefficient and firing threshold. Inter-spike intervals therefore encode intensity, with brighter pixels producing shorter intervals. Given a polling interval Δt, $s(x, y, n) \in \{0, 1\}$ indicates whether pixel (x, y) fires a spike during the n-th interval $((n-1)\Delta t, n\Delta t]$ (Zhang et al., 2022). A sequence of T consecutive intervals is represented as:

$$\mathbf{S} = [s(x, y, n)]_{n=1}^{T} \in \{0, 1\}^{T\times H\times W}.$$

The simulator in Section A.3 follows the same integrate-and-fire principle and carries the accumulator across adjacent intervals.

4 Method

Our goal is to leverage the knowledge of vision foundation models to improve depth estimation and generalization capabilities in the spike modality. We propose **RODES**, a cross-modal distillation framework that transfers knowledge from depth foundation model to the spike modality. This transfer is realized through two complementary pathways. On the model side, the **Spatio-Temporal Spike Adapter (STSA)** maps binary spike sequences to a latent representation compatible with the pretrained image encoder, while the **Pretrained-Manifold Alignment Distillation (PMAD)** provides direct representation-level guidance from an RGB teacher. On the data side, we **expand the training dataset** with simulated spike streams derived from real-world driving sequences and supervise them using pseudo-depth generated by a frozen pretrained RGB depth teacher. Together, these two pathways enable the spike student to exploit pretrained depth priors while requiring only spike input at inference. Figure 1 summarizes the complete training and inference framework.

4.1 Model-Side Prior Transfer

4.1.1 Spatio-Temporal Spike Adapter

Let $\mathbf{S} \in \{0, 1\}^{T\times H\times W}$ denote a spike sequence recorded over T time steps. Directly feeding $\mathbf{S}$ into an image-pretrained encoder is infeasible because its temporal channels do not conform to the three-channel input interface, whereas naively collapsing the temporal dimension may discard dynamic cues useful for depth estimation. To address this issue, we introduce the STSA to transform the raw spike sequence into a compact, foundation-compatible latent representation. As detailed in the lower-left inset of Figure 1, STSA consists of three stages: chunk-wise spatio-temporal encoding, cross-chunk interaction and adaptive aggregation, and contextual feature fusion.

To preserve short-term temporal variations, we partition $\mathbf{S}$ into K contiguous, non-overlapping chunks and apply a shared lightweight encoder. The encoded chunks are stacked and passed through a temporal mixer, whose depth-wise temporal operation exchanges information across neighboring chunks and point-wise transformation mixes channels. A parallel global branch processes the complete window, complementing these local dynamics with long-range scene context (Zhang et al.,

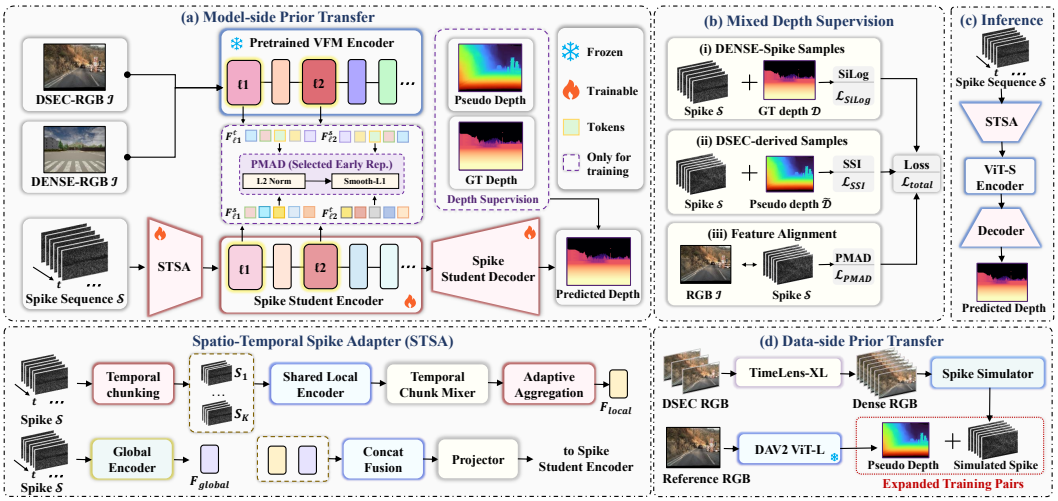

Figure 1: **Overview of RODES.** (a) Model-side prior transfer with STSA and PMAD, including the STSA architecture. (b) Mixed depth supervision. (c) Spike-only inference. (d) Teacher-guided DSEC data expansion. Snowflakes and flames denote frozen and trainable modules, respectively. Metric depth and relative pseudo-depth use distinct colormaps to distinguish scale conventions.

2022). Given the temporally mixed chunk features $\{\widetilde{\mathbf{U}}_k\}_{k=1}^{K}$, STSA adaptively aggregates them according to their input-dependent relevance:

$$\alpha_k = \frac{\exp\Big(q\Big(\mathrm{GAP}(\widetilde{\mathbf{U}}_k)\Big)\Big)}{\sum_{j=1}^{K}\exp\Big(q\Big(\mathrm{GAP}(\widetilde{\mathbf{U}}_j)\Big)\Big)}, \qquad \mathbf{F}_{\mathrm{local}} = \sum_{k=1}^{K}\alpha_k\widetilde{\mathbf{U}}_k, \tag{2}$$

where $q(\cdot)$ scores globally pooled chunk features. Each sample-dependent scalar α_k changes a chunk's contribution while preserving its spatial organization. The resulting local summary is concatenated with the global feature, fused, and projected to $\mathbf{X}^s \in \mathbb{R}^{3\times H\times W}$.

The three-channel projection is deliberate, allowing $\mathbf{X}^s$ to pass through the original patch embedding and enter the encoder through the same interface as pretrained RGB inputs. Rather than reconstructing an image, $\mathbf{X}^s$ serves as a learned latent interface, jointly optimized by the depth and distillation objectives to produce patch tokens compatible with the pretrained representation space.

4.1.2 Pretrained-Manifold Alignment Distillation

STSA resolves the input-interface mismatch, but input compatibility alone does not ensure that its tokens occupy a feature space that the RGB-pretrained encoder can exploit. With only final depth supervision, this representation gap must be corrected indirectly through the entire network, which is difficult under limited paired spike-depth data. We therefore introduce PMAD, shown in Figure 1(a), which uses a frozen RGB teacher to anchor the spike student's representations at strategically selected shallow stages.

We focus alignment on shallow representations because the modality discrepancy is introduced at the input interface and during early feature formation, while constraining deeper features may interfere with their adaptation to spike-specific cues. PMAD therefore aligns the patch-embedding output and a selected early Transformer block, leaving subsequent representations to be shaped primarily by the depth objective. This design is further examined in Table 5.

Let $\mathbf{F}_{\ell}^{m} \in \mathbb{R}^{B\times N_\ell\times C_\ell}$ denote the features of branch $m \in \{student, teacher\}$ at a selected location ℓ. PMAD applies token-wise normalization and minimizes a Smooth-L_1 discrepancy at each selected location, where ρ_1 denotes the Smooth-L_1 function:

$$\overline{\mathbf{F}}_{\ell,bn}^{m} = \frac{\mathbf{F}_{\ell,bn}^{m}}{\|\mathbf{F}_{\ell,bn}^{m}\|_2}, \qquad \mathcal{L}_{\mathrm{NSL1}}^{(\ell)} = \frac{1}{BN_\ell C_\ell}\sum_{b=1}^{B}\sum_{n=1}^{N_\ell}\sum_{c=1}^{C_\ell}\rho_1\Big(\overline{F}_{\ell,bnc}^{s} - \overline{F}_{\ell,bnc}^{t}\Big) \tag{3}$$

Here, B, N_ℓ, and C_ℓ denote batch size, token count, and channel dimension.

Token-wise normalization reduces sensitivity to modality-dependent feature magnitudes, making the alignment focus on the relative channel pattern of each patch token. The normalized teacher tokens thus act as local anchors in the RGB-pretrained representation space, guiding spike features toward the pretrained manifold without constraining modality-specific scales.

4.2 Data-Side Prior Transfer

PMAD transfers representations for available RGB-spike pairs but does not broaden the limited spike-depth scene distribution. We complement model-side alignment with a **teacher-guided data expansion strategy**, illustrated in Figure 1(d), which introduces additional driving scenes without requiring spike-depth annotations.

To broaden the training distribution, we derive additional samples from DSEC driving sequences (Gehrig et al., 2021). The image sequences are temporally densified using TimeLens-XL (Ma et al., 2024) and converted into spike streams through the integrate-and-fire simulation described in Section 3. Each spike window is paired with its reference RGB image, from which a frozen DAv2 ViT-L teacher generates relative pseudo-depth (Yang et al., 2024b). The paired RGB observation can additionally provide the feature targets used by PMAD through its separate ViT-S feature teacher. The two teachers therefore provide data-level and representation-level guidance, respectively.

The original and expanded data provide complementary supervision in different scale spaces, as summarized in Figure 1(b). DENSE-Spike provides metric ground truth to anchor the student to physical depth scale, whereas DSEC-derived samples provide relative pseudo-depth to transfer geometric structure across a broader scene distribution. Rather than fine-tuning ViT-L on the limited DENSE-Spike set, which may over-specialize its pretrained prior, we keep it frozen in its original relative-depth regime and apply source-appropriate objectives to the shared spike student.

Metric ground-truth supervision. For a DENSE-Spike sample, let Ω_{D} denote its pixel domain, and define $e_p = \log \widehat{D}_p - \log D_p^{\star}$, where $\widehat{D}_p$ and $D_p^{\star}$ are the predicted and ground-truth metric depths, respectively. We adopt the scale-invariant logarithmic loss (Eigen et al., 2014):

$$\mathcal{L}_{\mathrm{SiLog}} = \sqrt{\frac{1}{|\Omega_{\mathrm{D}}|}\sum_{p\in\Omega_{\mathrm{D}}} e_p^2 - \lambda_{\mathrm{var}}\left(\frac{1}{|\Omega_{\mathrm{D}}|}\sum_{p\in\Omega_{\mathrm{D}}} e_p\right)^2}. \tag{4}$$

We set $\lambda_{\mathrm{var}} = 0.5$ in all experiments. This objective anchors the shared student to metric scene geometry while retaining the absolute-depth supervision available in DENSE-Spike.

Relative pseudo-depth supervision. Because the pseudo-label is defined up to an unknown scale and shift, we compare it with the student's inverse-depth prediction after robust per-sample normalization. Following Ranftl et al. (2022), let $\mathcal{R}_{\Omega}(\cdot)$ denote median centering followed by scaling with the mean absolute deviation from the median. For each relative-supervision sample, we construct:

$$\overline{\mathbf{Q}}_j = \mathcal{R}_{\Omega_j}\left(\frac{1}{\widehat{\mathbf{D}}_j}\right), \qquad \overline{\mathbf{R}}_j = \mathcal{R}_{\Omega_j}\left(\widetilde{\mathbf{R}}_j\right). \tag{5}$$

The relative-depth objective is then:

$$\mathcal{L}_{\mathrm{SSI}} = \frac{1}{|\mathcal{B}_{\mathrm{R}}|}\sum_{j\in\mathcal{B}_{\mathrm{R}}} \operatorname*{mean}_{u\in\Omega_j} \rho_1\left(\overline{Q}_j(u) - \overline{R}_j(u)\right). \tag{6}$$

Here, ρ_1 is the Smooth-L_1 penalty. This objective removes the incompatible scale and offset while preserving the teacher's relative geometric structure.

Joint Training Objective. The model-side and data-side transfer pathways are optimized jointly in a single training stage. The complete PMAD and training objectives are:

$$\mathcal{L}_{\mathrm{PMAD}} = \sum_{\ell\in\mathcal{P}} \lambda_\ell \mathcal{L}_{\mathrm{NSL1}}^{(\ell)}, \qquad \mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SiLog}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{SSI}} + \mathcal{L}_{\mathrm{PMAD}}. \tag{7}$$

Here, $\mathcal{P}$ denotes the selected PMAD alignment locations. In our method, we select the patch-embedding output and the third Transformer block. The weights λ_ℓ control the contribution of each alignment location, while λ_{rel} weights the relative-depth supervision. At inference, both RGB teachers and all RGB observations are removed. As shown in Figure 1(c), depth is predicted from the spike sequence using only STSA and the trained student encoder-decoder.

5 EXPERIMENTS

We evaluate RODES with four objectives in mind: establishing its overall performance on the synthetic spike-depth benchmark, examining its transfer to real-world scenes and physical recordings, identifying the contribution of STSA and PMAD, and determining whether DSEC-derived supervision provides an additional data-level signal. Unless otherwise stated, all variants share the same backbone, data split, training schedule, and evaluation protocol, and are evaluated on DENSE-Spike.

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate on synthetic, real-world-derived, and physical-camera spike data. DENSE-Spike (Zhang et al., 2022) uses the standard sequence-level split of 4,995/1,998/999 training/validation/test samples. Each sample contains 128 spike frames at 346×260 with synchronized RGB and metric depth. DSEC-Spike, constructed from DSEC driving sequences (Gehrig et al., 2021), provides 6,150 paired 128-frame windows with relative pseudo-depth for auxiliary training and a sequence-disjoint 1,800-sample test split; the latter measures agreement with generated relative references rather than metric accuracy. Outdoor-Spike (Zhang et al., 2022) contains physical-camera recordings at 400×250 and 40,000 Hz; because it lacks depth ground truth, it is excluded from training and model selection and used only for zero-shot qualitative evaluation. All modalities are center-cropped to 224×224. Further details are provided in Appendix B.

Metrics. On DENSE-Spike, we report the standard metric-depth errors AbsRel, SqRel, $\mathrm{RMSE}_{\log}$, and SILog, together with threshold accuracies δ_k for $k \in \{1, 2, 3\}$, evaluated without per-image scale or shift alignment. On DSEC-Spike, which provides only relative references, we report scale-and-shift-invariant Smooth-L1, Spearman's ρ, and ordinal accuracy. Protocol-specific preprocessing is detailed in Appendix B.

Baselines. We compare against the U-Net baseline (Ronneberger et al., 2015) used by Spike Transformer, the recurrent E2Depth model (Hidalgo-Carrió et al., 2020), the spike-specific Spike-T architecture (Zhang et al., 2022), and the foundation-model-based DepthAnyEvent (DAE) and DepthAnyEvent-R (DAE-R) variants (Bartolomei et al., 2025). All baselines are retrained under the same protocol with 128-frame inputs, 224×224 crops, and a common valid-depth mask. RODES combines STSA and PMAD using only DENSE-Spike, whereas RODES+ additionally uses DSEC-derived mixed supervision. All models use spike-only input at inference.

5.2 IMPLEMENTATION DETAILS

We instantiate RODES with Depth Anything V2 (Yang et al., 2024b): the student and online feature teacher use the DAv2 ViT-S/14 backbone with a DPT decoder, and a frozen DAv2 ViT-L supplies offline relative pseudo-depth. PMAD is applied at the patch-embedding output and block 3. We train for 20 epochs with AdamW. Remaining hyperparameters (learning rates, alignment/loss weights, DSEC sampling ratio, and schedule) are listed in Appendix B. RGB observations, frozen teachers, and pseudo-depth labels are required only during training.

5.3 EXPERIMENTAL RESULTS

Experiments on the Synthetic Dataset We first evaluate all methods on the DENSE-Spike test set. As shown in Table 1, Spike-T is the strongest native spike baseline, while spike-adapted DepthAnyEvent variants further benefit from pretrained visual priors. RODES achieves the best SqRel, whereas RODES+ ranks first on the remaining six metrics. Compared with Spike-T, RODES+ reduces AbsRel from 0.610 to 0.326 and improves δ_1 from 0.682 to 0.769. RODES+ further improves over RODES, demonstrating the additional benefit of data-side transfer.

Figure 2 provides qualitative comparisons on representative samples. Consistent with the quantitative results, RODES+ produces more coherent scene geometry and cleaner depth boundaries across rural and urban scenes, particularly for building structures and thin objects such as poles and trunks. In contrast, convolutional and recurrent baselines often exhibit local structural gaps and incomplete depth predictions, whereas DAE-R* tends to oversmooth object boundaries.

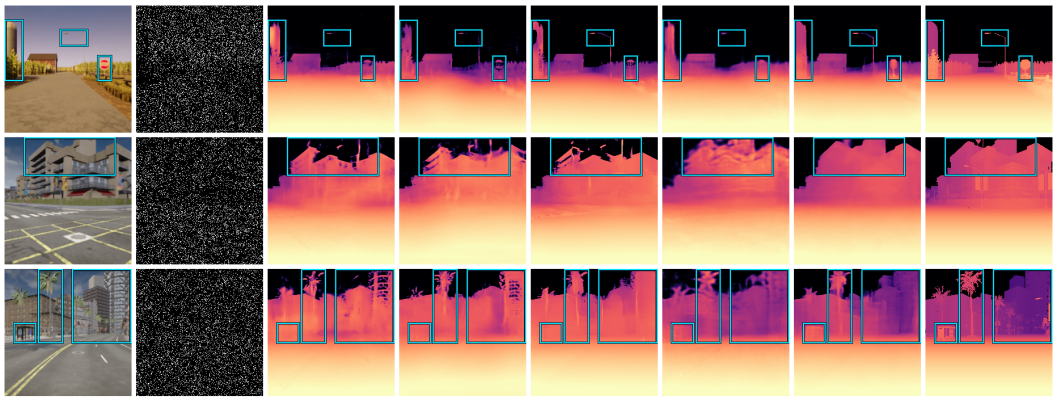

Figure 2: **Qualitative comparison on the DENSE-Spike dataset.** From left to right: RGB scene reference, spike input, U-Net, E2Depth, Spike-T, DAE-R*, RODES+, and Ground-Truth depth.

Method	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
U-Net (Ronneberger et al., 2015)	0.961	35.527	0.736	0.435	0.695	0.762	0.811
E2Depth (Hidalgo-Carrió et al., 2020)	0.761	23.446	0.784	0.459	0.643	0.722	0.786
Spike-T (Zhang et al., 2022)	0.610	18.664	0.707	0.396	0.682	0.762	0.813
DAE* (Bartolomei et al., 2025)	0.423	12.451	0.458	0.173	0.734	<u>0.859</u>	<u>0.929</u>
DAE-R* (Bartolomei et al., 2025)	0.383	9.503	0.521	0.223	0.699	0.813	0.893
RODES (Ours)	<u>0.345</u>	**6.983**	<u>0.445</u>	<u>0.165</u>	<u>0.735</u>	0.848	0.919
RODES+ (Ours)	**0.326**	<u>7.585</u>	**0.421**	**0.158**	**0.769**	**0.883**	**0.941**

Table 1: **Quantitative comparison on the DENSE-Spike test set.** All methods are retrained and evaluated under the same protocol. Best in **bold**, second-best <u>underlined</u>.

Method	SSI Smooth-L1↓	Spearman's ρ ↑	Ordinal accuracy↑
U-Net	0.081	0.916	0.925
E2Depth	0.074	0.933	0.924
Spike-T	0.050	0.941	0.941
DAE-R*	<u>0.021</u>	0.972	0.968
RODES	0.026	<u>0.976</u>	<u>0.973</u>
RODES+	**0.006**	**0.990**	**0.989**

Table 2: **Relative-depth evaluation on the DSEC-Spike test split.** RODES is zero-shot; RODES+ is trained with DSEC-derived relative supervision. Best in **bold**, second-best <u>underlined</u>.

Experiments on Real-World Data We evaluate real-world transfer in two complementary settings. DSEC-Spike contains simulated spike streams derived from real-world DSEC scenes and provides generated relative-depth references for quantitative evaluation. Outdoor-Spike consists of measurements captured directly by a physical spike camera and is entirely absent from training. It therefore provides the most direct evidence of generalization to real spike-camera observations.

Table 2 reports relative-depth evaluation on the DSEC-Spike test split. Without using DSEC-derived samples during training, RODES is evaluated zero-shot and reduces SSI Smooth-L1 from 0.050 to 0.026 relative to Spike-T, while achieving the strongest two depth-ordering scores among methods without DSEC-derived supervision. With additional DSEC-derived relative supervision, RODES+ ranks first under all three criteria and further improves over RODES across all metrics. These results show that model-side transfer already generalizes to unseen DSEC scenes, with data-side transfer further improving relative-depth consistency.

Additional qualitative comparisons in Appendix D.2 further show that RODES preserves coherent scene geometry under zero-shot transfer, while RODES+ produces sharper structures after incorporating DSEC-derived supervision.

Finally, Fig. 3 evaluates the methods on unseen Outdoor-Spike recordings. Despite the shift from simulated training data to physical spike-camera measurements, RODES+ recovers coherent scene geometry across diverse environments, with more complete structures and stable foreground-background transitions than the baselines. These observations provide qualitative evidence that the learned representation transfers to previously unseen recordings from a physical spike camera.

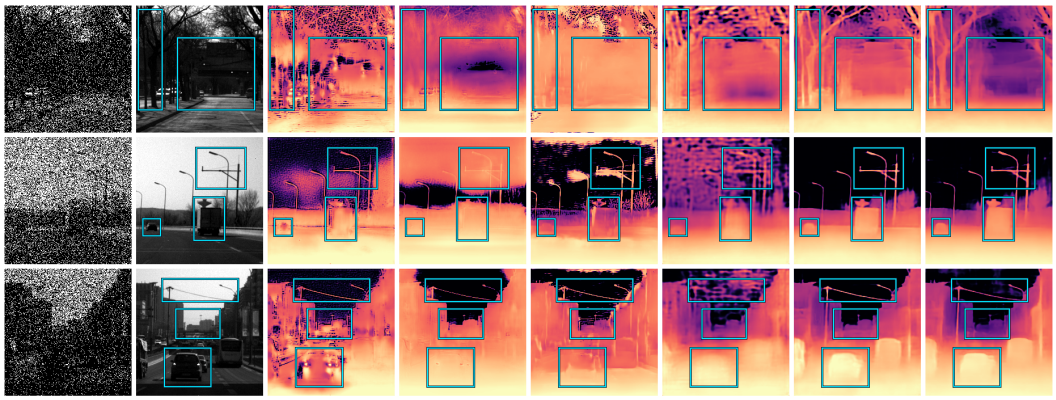

Figure 3: **Qualitative comparison on the unseen Outdoor-Spike dataset.** From left to right: raw spike input, scene reconstructions, U-Net, E2Depth, Spike-T, DAE-R*, RODES and RODES+.

Input strategy	AbsRel↓	SqRel↓	$\text{RMSE}_{\log}$↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Direct input	1.538	54.439	0.776	0.519	0.711	0.797	0.849
Convolutional projection	1.255	46.847	0.722	0.460	**0.725**	0.799	0.850
Residual projection	1.070	39.404	0.698	0.412	0.711	0.786	0.846
Spike-T embedding	0.947	33.898	0.678	0.385	0.707	0.785	0.847
Reconstruction-based input	0.669	27.339	**0.565**	**0.277**	**0.725**	**0.806**	**0.875**
STSA	**0.665**	**24.791**	0.637	0.327	0.706	0.777	0.840

Table 3: **Comparison of spike adaptation strategies on DENSE-Spike test set.** Feature distillation and DSEC-derived supervision are disabled. The best result for each metric is shown in bold.

5.4 Ablation Studies

Comparison of Spike-to-VFM Input Interfaces We first compare STSA with several representative strategies for mapping spike streams to the input space of the pretrained depth encoder. As shown in Table 3, direct input yields the largest errors, while structured input mappings substantially improve performance. Reconstruction-based input remains competitive across several metrics, while STSA achieves the lowest AbsRel and SqRel, reducing them by 56.8% and 54.5% over direct input. This trade-off suggests that input adaptation alone does not establish a uniformly superior interface for prior transfer. Under the same PMAD supervision in Table 4, STSA outperforms explicit reconstruction across all reported metrics. Additional component-level ablations in Appendix C.1 show that removing any component of STSA consistently degrades validation performance, supporting their complementary roles in the complete design.

The Role of Feature Distillation With the spike-to-VFM interface established by STSA, we next ask whether input adaptation alone is sufficient to transfer the pretrained visual priors. As shown in Table 4, adding PMAD improves all seven metrics, reducing AbsRel from 0.665 to 0.345 and SqRel from 24.791 to 6.983 with STSA. This confirms that input adaptation alone is insufficient and that explicit feature alignment is crucial for transferring pretrained visual priors. Under the same PMAD supervision, STSA also outperforms the reconstruction-based interface across all metrics. This suggests that directly learning a task-oriented spike representation provides a more effective interface for depth-prior transfer than first reconstructing an image-like representation.

We further analyze the learned representations in Fig. 4. Compared with the non-distilled student, PMAD produces patch-token structures closer to the RGB teacher in the shared PCA space and substantially reduces the teacher-student cosine discrepancy. The higher linear CKA further shows that the alignment improvement persists into later Transformer blocks without direct feature supervision. Further representation analyses are provided in the Appendix C.2.

Distillation Layers and Objectives Having established the necessity of feature distillation, we next examine where the teacher guidance should be applied. As shown by the validation-set results in Table 5, PatchEmbed supervision improves all four error metrics over the no-distillation baseline. The block-only variants exhibit a non-monotonic trend with depth, with no single block consistently

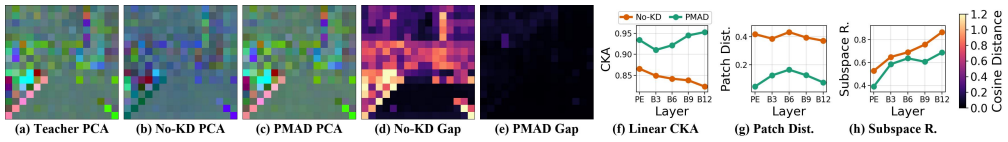

Figure 4: **Feature-space analysis of PMAD.** Shared-PCA patch-token projections (a)-(c), teacher-student cosine-distance maps (d)-(e), and layer-wise linear CKA, patch-wise cosine distance, and relative teacher-subspace residual (f)–(h). Darker maps and higher CKA indicate closer alignment.

Input interface	PMAD	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
STSA	-	0.665	24.791	0.637	0.327	0.706	0.777	0.840
Reconstruction	✓	0.491	17.053	0.493	0.198	0.707	0.839	0.914
STSA	✓	**0.345**	**6.983**	**0.445**	**0.165**	**0.735**	**0.848**	**0.919**

Table 4: **Effects of PMAD and spike-to-VFM interfaces on DENSE-Spike test set.** The distilled variants are compared under the same PMAD objective.

Location	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑	Location	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑
No distillation	0.304	6.251	0.376	0.126	**0.863**	PE + B3	**0.272**	4.444	**0.354**	**0.111**	0.845
PE only	0.278	4.528	0.359	0.115	0.843	PE + B6	0.275	4.215	0.364	0.117	0.839
B3 only	0.291	5.624	0.367	0.122	0.853	PE + B9	0.277	4.330	0.364	0.114	0.841
B6 only	0.299	5.902	0.371	0.125	0.851	PE + B12	0.283	**4.113**	0.369	0.120	0.835
B9 only	0.277	4.884	0.366	0.116	0.834	PE + B3 + B6	0.283	5.010	0.360	0.115	0.843
B12 only	0.286	5.431	0.362	0.118	0.848	All selected	0.307	5.763	0.378	0.127	0.838

Table 5: **Ablation of feature-distillation locations on DENSE-Spike validation set.** PE denotes the patch-embedding output and Bk denotes the k-th Transformer block. The best result for each metric is shown in bold.

dominating across all metrics. When combined with PatchEmbed supervision, PE+B3 achieves the lowest AbsRel, $\mathrm{RMSE}_{\log}$, and SILog, while PE+B12 gives the lowest SqRel. Adding supervision at more stages does not provide further gains and degrades the overall error profile. Together with the representation analysis in Fig. 4, these results suggest that effective cross-modal transfer favors selective shallow alignment rather than increasingly dense supervision across the encoder.

We also compare feature-matching objectives at the PatchEmbed output. On the validation split, normalized Smooth-L_1 and ℓ_2 regression provide the strongest error profiles. We empirically adopt normalized Smooth-L_1 for its competitive validation performance, with the held-out test results providing further support. Full objective comparisons are reported in Appendix C.4.

Data-Level Transfer and Mixed Supervision We finally examine whether DSEC-derived supervision provides a complementary data-level transfer signal. Validation results show that relative pseudo-depth supervision improves the error profile beyond using DSEC-derived pairs for feature supervision alone. Increasing the DSEC ratio further improves performance, with the 40%, $\lambda_{\mathrm{rel}} = 3$ setting providing a favorable trade-off between accuracy and auxiliary-data usage. These results support the complementarity of data-level supervision and PMAD. Detailed comparisons are provided in Appendix C.5.

6 Conclusion

In this work, we presented RODES, a cross-modal transfer framework that repurposes an RGB depth foundation model for monocular depth estimation from spike streams. RODES combines STSA for spike-to-VFM adaptation, PMAD for selective shallow feature alignment, and teacher-guided data expansion that complements metric spike-depth supervision with relative pseudo-depth on additional real-world scenes. Experiments on DENSE-Spike and DSEC-Spike demonstrate consistent improvements over spike-specific and foundation-model-based baselines, while results on unseen Outdoor-Spike indicate generalization to physical spike-camera recordings. Fine structures can still be smoothed, and quantitative evaluation on real spike-camera data remains limited by the lack of depth ground truth. Future work will focus on preserving fine-scale geometry and establishing larger real spike-depth benchmarks for more comprehensive evaluation.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this manuscript, including improving its clarity, concision, organization, and LaTeX presentation. They were also used to assist with literature retrieval and discovery. All suggested references and bibliographic information were independently verified against the original sources, and all AI-assisted text was reviewed, revised, and validated by the authors for technical accuracy and consistency with the reported methods and results. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main paper specifies the model architecture, training objectives, datasets, evaluation protocols, and core implementation settings. Appendices A and B provide the complete STSA and PMAD formulations, DSEC-to-spike construction procedure, dataset splits, preprocessing and evaluation protocols, baseline adaptations, and optimization hyperparameters.

ETHICS STATEMENT

This work uses existing research datasets and does not involve new data collection from human participants. The DSEC-derived spike data are constructed and used in accordance with the terms of the original dataset. The proposed method is intended for research use. Any deployment in safety-critical applications would require further evaluation under application-specific conditions.

REFERENCES

Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19669–19678, 2025.

Shiyan Chen, Jiyuan Zhang, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. Enhancing motion deblurring in high-speed scenes with spike streams. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70279–70292, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dead3d8ff3f9198e38a36a950ebbcafd-Abstract-Conference.html.

Zhiwen Chen, Junhui Hou, Zhiyu Zhu, Jinjian Wu, and Guangming Shi. Scaling dense event-stream pretraining from visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8011–8022, 2026.

Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *2017 Data Compression Conference*, pp. 437–437, 2017. doi: 10.1109/DCC.2017.69.

Siwei Dong, Lin Zhu, Daoyuan Xu, Yonghong Tian, and Tiejun Huang. An efficient coding method for spike camera using inter-spike intervals. In *2019 Data Compression Conference*, pp. 568–568, 2019. doi: 10.1109/DCC.2019.00080.

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*, volume 27, 2014.

Zhuoheng Gao, Yihao Li, Jiyao Zhang, Rui Zhao, Tong Wu, Hao Tang, Zhaofei Yu, Hao Dong, Guozhang Chen, and Tiejun Huang. Spikestereonet: A brain-inspired framework for stereo depth estimation from spike streams. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lPMPFeioCZ.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. doi: 10.1109/LRA.2021.3068942.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2827–2836, 2016.

Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *International Conference on 3D Vision*, pp. 534–542, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

Liwen Hu, Rui Zhao, Ziluo Ding, Lei Ma, Boxin Shi, Ruiqin Xiong, and Tiejun Huang. Optical flow estimation for spiking camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17844–17853, 2022.

Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, Jianing Li, Shanshan Jia, Yihua Fu, Boxin Shi, Si Wu, and Yonghong Tian. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 25:110–119, 2023. doi: 10.1016/j.eng.2022.01.012.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. doi: 10.1109/JSSC.2007.914337.

Yongrui Ma, Shi Guo, Yutian Chen, Tianfan Xue, and Jinwei Gu. TimeLens-XL: Real-time event-based video frame interpolation with large motion. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 178–194. Springer, 2024. doi: 10.1007/978-3-031-72907-2_11.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6550.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. ESS: Learning event-based semantic segmentation from still images. In *Computer Vision – ECCV 2022*, volume 13694 of *Lecture Notes in Computer Science*, pp. 341–357. Springer, 2022. doi: 10.1007/978-3-031-19830-4_20.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 608–619, 2021. doi: 10.1109/CVPR46437.2021.00067.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024b. doi: 10.52202/079017-0688.

Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *European Conference on Computer Vision*, pp. 34–52, 2022. doi: 10.1007/978-3-031-20071-7_3.

Jiyuan Zhang, Kang Chen, Shiyan Chen, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. SpikeGS: 3d gaussian splatting from spike streams with high-speed camera motion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9194–9203, 2024. doi: 10.1145/3664647.3680996.

Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11996–12005, 2021. doi: 10.1109/CVPR46437.2021.01182.

Rui Zhao, Ruiqin Xiong, Jing Zhao, Zhaofei Yu, Xiaopeng Fan, and Tiejun Huang. Learning optical flow from continuous spike streams. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7905–7920, 2022. doi: 10.52202/068431-0574.

Rui Zhao, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Shuyuan Zhu, Lei Ma, and Tiejun Huang. Spike camera image reconstruction using deep spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5207–5212, 2023.

Rui Zhao, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Zhaofei Yu, and Tiejun Huang. Boosting spike camera image reconstruction from a perspective of dealing with spike fluctuations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24955–24965. IEEE, 2024.

Rui Zhao, Ruiqin Xiong, Dongkai Wang, Shiyu Xuan, Jian Zhang, Xiaopeng Fan, and Tiejun Huang. Spike camera optical flow estimation based on continuous spike streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4756 – 4770, 2026.

Yajing Zheng, Jiyuan Zhang, Rui Zhao, Jianhao Ding, Shiyan Chen, Weijian Wu, Ruiqin Xiong, Zhaofei Yu, and Tiejun Huang. Spikecv: open a continuous computer vision era. *Science China Information Sciences*, 69(3):132106, 2026.

Jinjing Zhu, Tianbo Pan, Zidong Cao, Yexin Liu, James T. Kwok, and Hui Xiong. Depth any event stream: Enhancing event-based monocular depth estimation via dense-to-sparse distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5146–5155, 2025.

Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1438–1446, 2020. doi: 10.1109/CVPR42600.2020.00151.

A Additional Method Details

This section consolidates the complete operator definitions and data-construction details underlying the model-side and data-side prior-transfer pathways of RODES. It expands the STSA tensor flow, the PMAD feature-alignment objective, the DSEC-to-spike generation pipeline, and the source-specific supervision used for joint training.

A.1 STSA Architecture and Tensor Flow

Let $\mathbf{S} \in \{0,1\}^{T\times H\times W}$ denote a spike window. The Spatio-Temporal Spike Adapter (STSA) partitions $\mathbf{S}$ into K contiguous, non-overlapping temporal chunks $\{\mathbf{S}_k\}_{k=1}^{K}$. A lightweight local encoder with shared parameters processes each chunk, after which a temporal chunk mixer exchanges information between neighboring chunks through depth-wise temporal mixing and point-wise channel transformation. In parallel, a global branch processes the complete spike window to capture context over the full observation interval. Denoting the local encoder, chunk mixer, and global encoder by ϕ_{loc}, ϕ_{mix}, and ϕ_{glob}, respectively, the tensor flow is

$$\mathbf{U}_k = \phi_{\mathrm{loc}}(\mathbf{S}_k), \qquad \{\widetilde{\mathbf{U}}_k\}_{k=1}^{K} = \phi_{\mathrm{mix}}\big(\{\mathbf{U}_k\}_{k=1}^{K}\big), \qquad \mathbf{F}_{\mathrm{global}} = \phi_{\mathrm{glob}}(\mathbf{S}). \tag{8}$$

The shared local pathway preserves short-range temporal variations, whereas the global pathway summarizes the complete observation window; the two pathways therefore provide complementary temporal information (Zhang et al., 2022).

STSA aggregates the mixed chunk features using sample-dependent relevance scores:

$$\alpha_k = \frac{\exp\Big(q\Big(\mathrm{GAP}(\widetilde{\mathbf{U}}_k)\Big)\Big)}{\sum_{j=1}^{K}\exp\Big(q\Big(\mathrm{GAP}(\widetilde{\mathbf{U}}_j)\Big)\Big)}, \qquad \mathbf{F}_{\mathrm{local}} = \sum_{k=1}^{K}\alpha_k\widetilde{\mathbf{U}}_k, \tag{9}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling over the spatial dimensions and $q(\cdot)$ is a learned scoring function. Each α_k is a scalar shared across spatial locations, allowing the adapter to change the contribution of a temporal segment without destroying its spatial organization.

The aggregated local feature and the global feature are concatenated and processed by a fusion module. A final projection produces

$$\mathbf{X}^s = \phi_{\mathrm{proj}}(\phi_{\mathrm{fuse}}(\mathrm{Concat}[\mathbf{F}_{\mathrm{local}}, \mathbf{F}_{\mathrm{global}}])) \in \mathbb{R}^{3\times H\times W}. \tag{10}$$

The projected representation is passed through the original DAv2 patch-embedding layer, leaving the downstream encoder–decoder topology unchanged. The three-channel projection therefore preserves the pretrained image interface rather than merely reducing dimensionality. It is optimized jointly by the depth and feature-distillation objectives so that its patch embeddings become compatible with the pretrained representation space.

A.2 Full PMAD Formulation

For each synchronized RGB–spike pair $(\mathbf{I}, \mathbf{S})$, the same spatial transform is applied to both modalities to preserve pixel–token correspondence. The RGB image is processed by a frozen teacher encoder E_ξ, while STSA first maps the spike sequence to $\mathbf{X}^s$, which is then processed by the trainable student encoder E_θ. The teacher and student use the same ViT-S architecture and are initialized from the same pretrained checkpoint. Their token grids and channel dimensions therefore agree directly, so PMAD does not require a learned feature projector.

Let

$$\mathcal{A} = \{\mathrm{patch}, \mathrm{blk3}\}$$

denote the selected alignment locations, corresponding to the patch-embedding output and Transformer block 3. For $\ell \in \mathcal{A}$, let $\mathbf{F}_\ell^s, \mathbf{F}_\ell^t \in \mathbb{R}^{B\times N_\ell\times C_\ell}$ be the student and teacher patch-token features. PMAD first normalizes every token along its channel dimension:

$$\overline{\mathbf{F}}^s_{\ell,bn} = \frac{\mathbf{F}^s_{\ell,bn}}{\max\Big(\|\mathbf{F}^s_{\ell,bn}\|_2, \epsilon_f\Big)}, \qquad \overline{\mathbf{F}}^t_{\ell,bn} = \frac{\mathbf{F}^t_{\ell,bn}}{\max\Big(\|\mathbf{F}^t_{\ell,bn}\|_2, \epsilon_f\Big)}. \tag{11}$$

The normalized Smooth-L_1 loss at location ℓ is

$$\mathcal{L}_{\mathrm{NSL1}}^{(\ell)} = \frac{1}{BN_\ell C_\ell} \sum_{b=1}^{B} \sum_{n=1}^{N_\ell} \sum_{c=1}^{C_\ell} \rho_1\Big(\overline{F}^{s}_{\ell,bnc} - \overline{F}^{t}_{\ell,bnc}\Big), \tag{12}$$

where the Smooth-L_1 penalty uses threshold $\beta = 1$:

$$\rho_1(e) = \begin{cases} \frac{1}{2}e^2, & |e| < 1, \\ |e| - \frac{1}{2}, & |e| \geq 1. \end{cases} \tag{13}$$

The complete PMAD objective is

$$\mathcal{L}_{\mathrm{PMAD}} = \lambda_{\mathrm{patch}} \mathcal{L}_{\mathrm{NSL1}}^{(\mathrm{patch})} + \lambda_{\mathrm{early}} \mathcal{L}_{\mathrm{NSL1}}^{(\mathrm{blk3})}. \tag{14}$$

Token-wise normalization reduces sensitivity to modality-dependent feature magnitudes and focuses the loss on the channel pattern of each patch token. Alignment is restricted to shallow stages because the modality discrepancy is introduced at the input interface and during early feature formation. Anchoring these stages allows subsequent blocks to reuse pretrained geometric priors, while leaving deeper decoder-reading representations free to adapt to spike-specific depth prediction. Extending the same constraint to deeper layers would require those representations to imitate the RGB teacher while simultaneously serving the spike-depth objective.

A.3 DSEC-to-Spike Data Construction

Source-sequence preparation. The expanded data are constructed from real-world DSEC driving sequences, which provide additional appearance, motion, and scene-geometry diversity (Gehrig et al., 2021). For each temporal interval, TimeLens-XL densifies the associated image sequence (Ma et al., 2024). The densified sequence is linearly resampled at $3840\,\mathrm{Hz}$, converted to grayscale, and resized to the spike-camera resolution.

Spike simulation. We apply the integrate-and-fire mechanism in Equation 1 to the resampled intensity sequence. Rather than resetting the simulator independently for every interval, the accumulator state is propagated across adjacent intervals to preserve temporal continuity. The resulting samples contain 128-frame simulated spike windows paired with their temporally aligned reference RGB observations. Dataset composition, sequence splits, and sampling ratios are reported in Appendix B.

Teacher targets. A frozen DAv2 ViT-L model processes each reference RGB image offline and produces a relative pseudo-depth target (Yang et al., 2024b). The paired RGB observation is also processed by the separate frozen ViT-S feature teacher that supplies online PMAD targets. The two teachers therefore have distinct roles: ViT-L provides offline relative-depth supervision for data expansion, whereas ViT-S provides representation-level supervision at the selected student layers. Neither teacher nor the paired RGB observation is required at inference.

A.4 Full Supervision Objectives

The original and expanded data provide supervision in different scale spaces. DENSE-Spike supplies metric ground-truth depth, whereas DSEC-Spike supplies relative pseudo-depth without a valid metric scale. We retain the pseudo-depth teacher in its original relative-depth regime rather than adapting it to the limited metric training set, and assign a source-appropriate objective to each subset.

Metric ground-truth supervision. For a DENSE-Spike sample, let Ω_{D} be its valid pixel domain and define

$$e_p = \log \widehat{D}_p - \log D_p^{\star},$$

where $\widehat{D}_p$ and $D_p^{\star}$ denote predicted and ground-truth metric depth. We use the scale-invariant logarithmic loss (Eigen et al., 2014):

$$\mathcal{L}_{\mathrm{SiLog}} = \sqrt{\frac{1}{|\Omega_{\mathrm{D}}|} \sum_{p \in \Omega_{\mathrm{D}}} e_p^2 - \lambda_{\mathrm{var}} \left(\frac{1}{|\Omega_{\mathrm{D}}|} \sum_{p \in \Omega_{\mathrm{D}}} e_p \right)^2}. \tag{15}$$

Relative pseudo-depth supervision. Because the pseudo-depth target is defined only up to scale and shift, the student's inverse-depth prediction and the teacher target are normalized independently for every sample. Following Ranftl et al. (2022), define

$$m_\Omega(\mathbf{Z}) = \underset{u\in\Omega}{\text{median}}\, Z(u), \qquad a_\Omega(\mathbf{Z}) = \frac{1}{|\Omega|}\sum_{u\in\Omega} |Z(u) - m_\Omega(\mathbf{Z})|\,, \qquad \mathcal{R}_\Omega(\mathbf{Z}) = \frac{\mathbf{Z} - m_\Omega(\mathbf{Z})}{a_\Omega(\mathbf{Z})}. \tag{16}$$

For a relative-supervision sample j, the normalized prediction and target are

$$\overline{\mathbf{Q}}_j = \mathcal{R}_{\Omega_j}\left(\frac{1}{\widehat{\mathbf{D}}_j}\right), \qquad \overline{\mathbf{R}}_j = \mathcal{R}_{\Omega_j}\left(\widetilde{\mathbf{R}}_j\right), \tag{17}$$

where $\widetilde{\mathbf{R}}_j$ is the frozen teacher's relative pseudo-depth. For the relative subset $\mathcal{B}_{\mathrm{R}}$ of a mixed batch, we use

$$\mathcal{L}_{\mathrm{SSI}} = \frac{1}{|\mathcal{B}_{\mathrm{R}}|}\sum_{j\in\mathcal{B}_{\mathrm{R}}} \underset{u\in\Omega_j}{\text{mean}}\, \rho_1\big(\overline{Q}_j(u) - \overline{R}_j(u)\big)\,. \tag{18}$$

Independent normalization removes the incompatible scale and offset while preserving the relative geometric structure supplied by the teacher.

Joint objective and inference. The model-side and data-side transfer pathways are optimized jointly:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SiLog}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{SSI}} + \mathcal{L}_{\mathrm{PMAD}}. \tag{19}$$

The final configuration uses $\lambda_{\mathrm{var}} = 0.5$, $\lambda_{\mathrm{rel}} = 3$, $\lambda_{\mathrm{patch}} = 2000$, and $\lambda_{\mathrm{early}} = 400$. $\mathcal{L}_{\mathrm{SiLog}}$ is evaluated only on valid metric-depth pixels from DENSE-Spike, whereas $\mathcal{L}_{\mathrm{SSI}}$ is evaluated only on valid DSEC-Spike relative pseudo-labels. Because both sources contain aligned RGB–spike pairs during training, $\mathcal{L}_{\mathrm{PMAD}}$ is evaluated over the complete mixed batch. At inference, the two teachers, all RGB observations, and the offline pseudo-labels are removed; prediction uses only STSA and the trained spike encoder–decoder.

B Experimental Protocol and Reproducibility

This section expands the experimental protocol summarized in Section 5.1. It specifies the dataset splits, evaluation procedures, baseline adaptations, and training configuration used throughout the experiments.

B.1 Datasets and Splits

DENSE-Spike. The benchmark contains eight sequences, split at the sequence level into five training, two validation, and one test sequence, corresponding to 4,995/1,998/999 samples. Each sample comprises 128 spike frames at 346×260, synchronized RGB observations, and metric-depth ground truth. All modalities are center-cropped to 224×224. Valid ground-truth depths lie in $(10^{-6}, 1000]$.

DSEC-Spike. The auxiliary training split is constructed from 41 DSEC source sequences, with three 50-sample clips drawn from each sequence, giving 6,150 samples in total. Each sample contains a 128-frame simulated spike window, its aligned RGB observation, and a relative pseudo-depth label. The test split contains 1,800 samples from 12 additional held-out source sequences and is sequence-disjoint from the auxiliary training split. Because its references are generated relative-depth targets rather than sensor metric depth, evaluation on this split measures agreement with the teacher-derived geometry, not metric-depth accuracy.

Outdoor-Spike. Outdoor-Spike contains physical spike-camera recordings at 400×250 and 40,000 Hz. It is excluded from training and model selection and, because synchronized depth ground truth is unavailable, is used only for zero-shot qualitative evaluation.

Configuration	$\mathrm{AbsRel}\downarrow$	$\mathrm{SqRel}\downarrow$	$\mathrm{RMSE}_{\log}\downarrow$	$\mathrm{SILog}\downarrow$	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
TFI + Frozen DAv2-S	0.417	4.129	1.584	1.780	0.483	0.644	0.714
TFI + Fine-tuned DAv2-S	0.669	27.339	0.565	0.277	0.725	0.806	0.875
TFI + PMAD	0.491	17.053	0.493	0.198	0.707	0.839	0.914
RGB + Fine-tuned DAv2-S	0.349	5.931	0.487	0.194	0.732	0.817	0.883

Table 6: **Additional DAv2-S reference configurations on DENSE-Spike.** The frozen TFI reference is evaluated after per-image scale-and-shift alignment in inverse-depth space, whereas the trained configurations are evaluated directly in metric depth.

B.2 Evaluation Protocols

On DENSE-Spike, all methods use the same valid-depth mask and metric implementation. Predictions are clipped to $(10^{-6}, 1000]$ and evaluated directly in metric depth, without per-image scale or shift alignment. We report AbsRel, SqRel, $\mathrm{RMSE}_{\log}$, SILog, and threshold accuracies δ_k for $k \in \{1, 2, 3\}$; lower errors and higher accuracies are better.

On DSEC-Spike, each prediction and reference is independently median-centered and normalized by its mean absolute deviation before computing scale-and-shift-invariant Smooth-L1 and Spearman's ρ. Ordinal accuracy is estimated from 10,000 valid pixel pairs. Lower SSI Smooth-L1 and higher rank correlation and ordinal accuracy are better.

B.3 Baseline Adaptation

All baselines are retrained with 128-frame spike inputs, 224×224 crops, and the same valid-depth mask. The superscript * denotes our spike-adapted DAE and DAE-R implementations, in which the original Tencode representation is replaced by TFI encoding. RODES uses STSA and PMAD with DENSE-Spike training only, whereas RODES+ additionally uses DSEC-derived mixed supervision.

B.4 Additional VFM Reference Configurations

To further characterize the contribution of the pretrained depth model, we evaluate several additional DAv2-S reference configurations on DENSE-Spike. The TFI-based variants use reconstructed spike observations as input and consider frozen inference, task-specific fine-tuning, and feature distillation. The RGB variant directly uses the synchronized RGB observation as an image-domain reference.

For TFI + Frozen DAv2-S, we retain the original relative-depth model without spike-domain training. Since its predictions are not metrically calibrated, the predicted and ground-truth inverse depths are scale-and-shift aligned per image before conversion back to metric depth for evaluation. The remaining trained configurations are evaluated directly in metric depth without such alignment.

As shown in Table 6, directly applying the frozen VFM to reconstructed spike observations remains limited even with scale-and-shift alignment, while fine-tuning alone does not provide a reliable transfer solution. Feature distillation substantially improves the reconstruction-based interface, and the RGB-input reference further indicates the remaining gap between reconstructed spike observations and native image inputs.

B.5 Training Configuration

The RODES student uses a Depth Anything V2 ViT-S/14 encoder and a DPT decoder. A frozen ViT-S/14 encoder supplies online feature targets for PMAD, while a frozen ViT-L model supplies offline relative pseudo-depth for DSEC-derived samples. Paired modalities undergo identical spatial transforms to preserve pixel–token correspondence. The PMAD terms at the patch-embedding output and third Transformer block are weighted by 2,000 and 400, respectively, and the DSEC relative-depth objective uses $\lambda_{\mathrm{rel}} = 3$.

We train for 20 epochs with AdamW using $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, and initial learning rates of 5×10^{-7} for the pretrained encoder and 5×10^{-5} for all remaining parameters. Training uses batch size 4 and iteration-wise polynomial learning-rate decay with power 0.4 and no

Method	Params	Latency ↓	FPS ↑
Spike-T	20.546M	24.803 ms	40.32
DAE-R*	26.113M	**9.798 ms**	**102.07**
RODES	25.076M	12.093 ms	82.69

Table 7: **Computational efficiency on 128-frame spike inputs.** Latency denotes the inference time per input window, and FPS is the corresponding throughput.

Variant	$\mathrm{AbsRel}\downarrow$	$\mathrm{SqRel}\downarrow$	$\mathrm{RMSE}_{\log}\downarrow$	$\mathrm{SILog}\downarrow$	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
Full STSA	**0.304**	**6.251**	**0.376**	**0.126**	**0.863**	**0.911**	**0.940**
w/o global branch	0.334	7.831	0.397	0.143	0.852	0.906	0.937
w/o local mixer	0.314	6.704	0.396	0.138	0.847	0.901	0.935
Uniform aggregation	0.323	7.109	0.386	0.134	0.857	0.909	0.938

Table 8: **Component analysis of STSA on DENSE-Spike validation set.** "w/o" denotes removal of the corresponding component. The best result for each metric is shown in bold.

warm-up. Each training pass traverses all 4,995 DENSE-Spike samples and draws 3,330 reshuffled DSEC-Spike samples, corresponding to an auxiliary-data ratio of approximately 40%. Training is performed on one NVIDIA GeForce RTX 5080 GPU with 16 GB memory. RGB observations, frozen teachers, and pseudo-depth labels are used only during training; all models receive spike streams alone at inference.

B.6 Computational Efficiency and Temporal Footprint

Table 7 compares the inference efficiency of representative spike-based and VFM-based depth models. RODES contains 25.1M parameters and processes a 128-frame spike window in 12.1 ms, corresponding to 82.7 FPS. This is approximately $2.05\times$ faster than Spike-T, while remaining close to DAE-R* in both model size and inference latency. At the 40 kHz sampling rate of Outdoor-Spike, 128 consecutive spike frames span approximately 3.2 ms of physical time. This acquisition interval is distinct from network inference latency and characterizes the temporal extent of the spike observation used for each depth prediction.

C Extended Ablation Studies

This section complements the ablations in Section 5.4 with component-level analysis of STSA, detailed representation analysis of PMAD, complete PMAD configuration sweeps, and a detailed comparison of DSEC-derived mixed supervision.

C.1 STSA Component Analysis

Table 8 isolates the global-context branch, local mixer, and adaptive aggregation in STSA on the DENSE-Spike validation split. Removing any of the three components consistently degrades all reported metrics, with the complete STSA achieving the best performance throughout. In particular, removing the global branch produces the largest increases in AbsRel and SqRel, highlighting the importance of global spatio-temporal context. Removing the local mixer also consistently worsens performance, confirming the contribution of cross-chunk temporal interaction. Finally, replacing adaptive aggregation with uniform averaging degrades all metrics, supporting the use of input-dependent temporal weighting. Together, these results validate the complementary roles of the three components in the complete STSA.

C.2 Representation-Level Effects of PMAD

Figure 5 expands the compact analysis in Fig. 4. The shared-PCA projections in panels (a)–(c) compare patch-token representations of the frozen RGB teacher, the non-distilled student, and the PMAD student using a common basis. PMAD more closely recovers the teacher's spatial token organization, while the teacher-gap maps in panels (d)–(e) show lower cosine discrepancy over most spatial positions. The residual maps in panels (f)–(g) further show that PMAD moves the spike representation closer to the principal subspace of the teacher features.

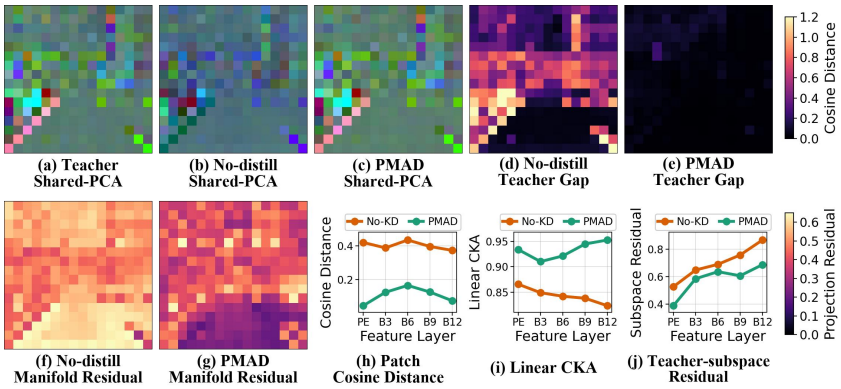

Figure 5: **Detailed feature-space analysis of PMAD.** Shared-PCA patch-token projections (a)–(c), spatial teacher–student cosine-distance maps (d)–(e), teacher-subspace residual maps (f)–(g), and layer-wise cosine distance, linear CKA, and relative teacher-subspace residual (h)–(j). Lower cosine distance and residual and higher CKA indicate closer alignment.

Objective	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑
ℓ_2 regression	0.284	**4.522**	**0.359**	**0.113**	0.840
Cosine distance	0.298	5.465	0.381	0.129	0.841
Normalized Smooth-L_1	**0.278**	4.528	**0.359**	0.115	0.843
Correlation-matrix matching	0.311	6.392	0.374	0.130	**0.854**
Masked feature generation	0.307	6.300	0.372	0.128	0.852

Table 9: **Comparison of feature-distillation objectives on DENSE-Spike validation set.** All objectives are evaluated at the PatchEmbed output. The best result for each metric is shown in bold.

The layer-wise measurements in panels (h)–(j) provide consistent aggregate evidence. At the PatchEmbed output, PMAD reduces the mean patch-wise cosine distance from 0.419 to 0.044, increases linear CKA from 0.866 to 0.934, and reduces the relative teacher-subspace residual from 0.527 to 0.389. The alignment advantage remains visible in later Transformer blocks that receive no direct feature supervision, supporting the conclusion that early cross-modal guidance propagates through the pretrained encoder.

C.3 Feature-Distillation Locations

Table 5 in the main paper reports the complete validation-set sweep over feature-distillation locations. The clearest trend is that introducing feature distillation provides a much larger gain in the continuous depth-error metrics than the differences among individual alignment locations. All single-location variants improve AbsRel, SqRel, $\mathrm{RMSE}_{\log}$, and SILog over no distillation, although the validation-set δ_1 does not improve over the non-distilled baseline. Their relative performance otherwise varies non-monotonically with depth. The shallow PatchEmbed location remains consistently competitive, and adding a second alignment location yields only modest, metric-dependent improvements. Together with the representation analysis in Fig. 5, these results suggest that, in our setting, effective cross-modal transfer favors selective shallow alignment over increasingly dense supervision across the encoder.

Extending supervision to more stages also substantially enlarges the configuration space, since each additional location introduces further choices of layer combinations and loss weights. We therefore evaluate representative broader settings rather than exhaustively searching all multi-layer combinations. Neither PE+B3+B6 nor supervision over all selected decoder-read stages improves consistently over the two-location setting, and both degrade several primary error metrics. These results suggest that broader layer-wise constraints are unnecessary in our setting and support restricting PMAD to a small number of shallow representations.

C.4 Feature-Matching Objectives

Table 9 compares five feature-matching objectives at the PatchEmbed output on the DENSE-Spike validation split. To reduce the influence of differences in numerical scale, we first inspect the magni-

Objective	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑
ℓ_2 regression	0.378	8.376	**0.454**	**0.170**	0.727
Normalized Smooth-L_1	**0.352**	**6.466**	0.456	0.171	**0.729**

Table 10: **Comparison of the two strongest objectives on DENSE-Spike test set.** Test results are reported to verify that the selected choice generalizes to the held-out DENSE-Spike test sequence.

Configuration	DSEC ratio	λ_{rel}	AbsRel↓	SqRel↓	$\mathrm{RMSE}_{\log}$↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
No DSEC	0%	–	0.272	4.444	0.354	0.111	0.845	0.912	0.945
+ DSEC feature supervision	10%	0	0.275	4.627	0.358	0.114	0.842	0.912	0.944
+ DSEC relative supervision	10%	1	0.271	4.546	0.354	0.111	0.845	0.913	0.945
+ DSEC relative supervision	10%	3	0.269	4.494	0.354	0.110	0.838	0.914	0.947
+ DSEC relative supervision	40%	1	0.272	4.778	0.355	0.112	0.844	0.917	0.948
+ DSEC relative supervision	40%	3	**0.260**	**3.980**	0.352	0.109	0.847	0.917	0.948
+ DSEC relative supervision	60%	3	**0.260**	4.520	**0.344**	**0.105**	0.846	**0.918**	**0.950**

Table 11: **Effect of DSEC-derived supervision on top of complete PMAD on DENSE-Spike validation set.** Setting $\lambda_{\mathrm{rel}} = 0$ disables only the relative-depth loss on DSEC-derived samples. The best result for each metric is shown in bold.

tude of each objective and adjust its loss coefficient accordingly, such that the weighted distillation terms operate at approximately comparable scales during training. At this fixed alignment location, the two direct regression objectives, ℓ_2 regression and normalized Smooth-L_1, provide the most favorable error profiles among the evaluated alternatives. Normalized Smooth-L_1 achieves the lowest AbsRel and remains comparable to ℓ_2 regression on SqRel and $\mathrm{RMSE}_{\log}$. We therefore adopt normalized Smooth-L_1, combining competitive validation performance with token-wise normalization that reduces sensitivity to modality-dependent feature magnitudes.

To further verify that this choice generalizes beyond the validation sequences, Table 10 further compares the two strongest validation candidates on the held-out DENSE-Spike test split. Normalized Smooth-L_1 retains lower AbsRel and SqRel, while ℓ_2 regression remains slightly better on $\mathrm{RMSE}_{\log}$ and SILog.

C.5 DSEC-Derived Mixed Supervision

Table 11 evaluates the DSEC sampling ratio and relative-depth weight on the DENSE-Spike validation split while retaining PMAD. At a fixed 10% DSEC ratio, setting $\lambda_{\mathrm{rel}} = 0$ uses the additional RGB-spike pairs only for feature supervision. Introducing relative pseudo-depth supervision yields modest, metric-dependent changes at this ratio, with improvements in AbsRel and SILog but no consistent gain across all metrics.

At $\lambda_{\mathrm{rel}} = 3$, increasing the DSEC ratio from 10% to 40% reduces AbsRel from 0.269 to 0.260 and SqRel from 4.494 to 3.980. Increasing the ratio further to 60% retains the same AbsRel while improving $\mathrm{RMSE}_{\log}$, SILog, δ_2, and δ_3, but increases SqRel to 4.520. We therefore use the 40%, $\lambda_{\mathrm{rel}} = 3$ configuration as a balanced setting that achieves the lowest SqRel, ties for the lowest AbsRel, and requires less auxiliary data than the 60% alternative. Overall, these results show that relative pseudo-depth supervision contributes beyond additional feature-supervision pairs and that its effect depends on both the supervision weight and the auxiliary-data ratio.

C.6 Fine-Tuning Strategy

We additionally examine how the pretrained DAv2 student should be adapted to the spike domain. To isolate the effect of optimization strategy, all variants use STSA without feature distillation or DSEC-derived supervision and are evaluated on the DENSE-Spike validation split.

As shown in Table 12, training only the adapter is insufficient to adapt the frozen pretrained model to spike inputs. Updating the encoder substantially improves performance, whereas the LoRA variant remains less effective than direct encoder fine-tuning. Within our STSA-based setting, full fine-tuning performs best among the evaluated strategies, suggesting that restricting adaptation to the adapter or encoder alone is suboptimal for our setting.

Strategy	AbsRel↓	SqRel↓	$\text{RMSE}_{\log}$↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Adapter only	29.487	2762.545	2.867	2.155	0.015	0.029	0.041
Adapter + encoder	0.376	8.084	0.440	0.162	0.824	0.890	0.922
Adapter + encoder (LoRA)	0.408	9.271	0.469	0.180	0.777	0.872	0.912
Full fine-tuning	**0.304**	**6.251**	**0.376**	**0.126**	**0.863**	**0.911**	**0.940**

Table 12: **Comparison of fine-tuning strategies on DENSE-Spike validation set.** All variants use STSA without feature distillation or DSEC-derived supervision. Best in bold.

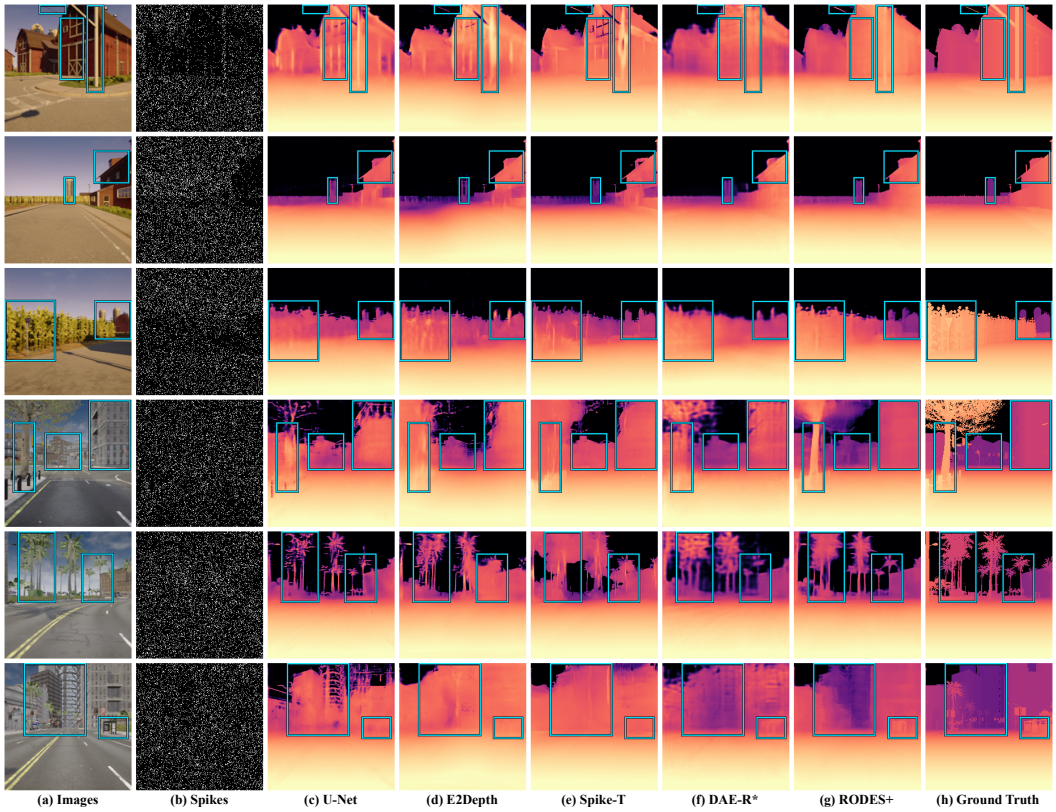

Figure 6: **Additional qualitative comparison on the DENSE-Spike test set.** From left to right: RGB scene reference, spike input, U-Net, E2Depth, Spike-T, DAE-R*, RODES+, and metric ground-truth depth.

D Additional Qualitative Results

This section provides additional qualitative comparisons on synthetic, real-scene-derived, and physical-camera spike data.

D.1 Additional Results on DENSE-Spike

Figure 6 extends the main-paper comparison with additional held-out test scenes. Across these examples, RODES+ preserves coherent scene geometry while recovering clearer object boundaries and thin structures than the compared baselines.

D.2 Additional Results on DSEC-Spike

Figure 7 provides additional comparisons on the sequence-disjoint DSEC-Spike test split. RODES preserves coherent relative scene structure without observing DSEC-derived samples during training, while RODES+ produces sharper road, vehicle, and foreground–background structure after incorporating DSEC-derived supervision. Because the references are generated by the frozen RGB depth teacher, these results measure qualitative agreement with teacher-derived geometry rather than accuracy against metric sensor depth.

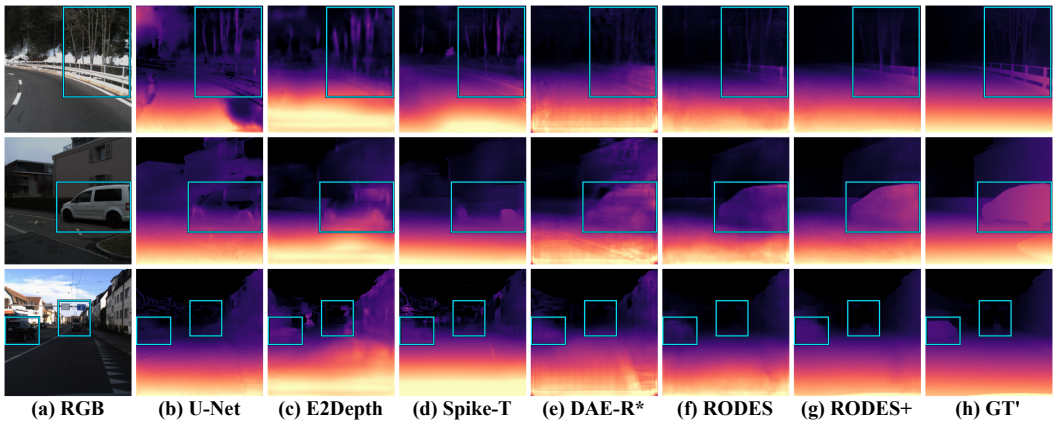

Figure 7: **Additional qualitative comparison on the DSEC-Spike test split.** From left to right: RGB scene reference, U-Net, E2Depth, Spike-T, DAE-R*, RODES, RODES+, and the DAv2 pseudo-depth reference (GT′).

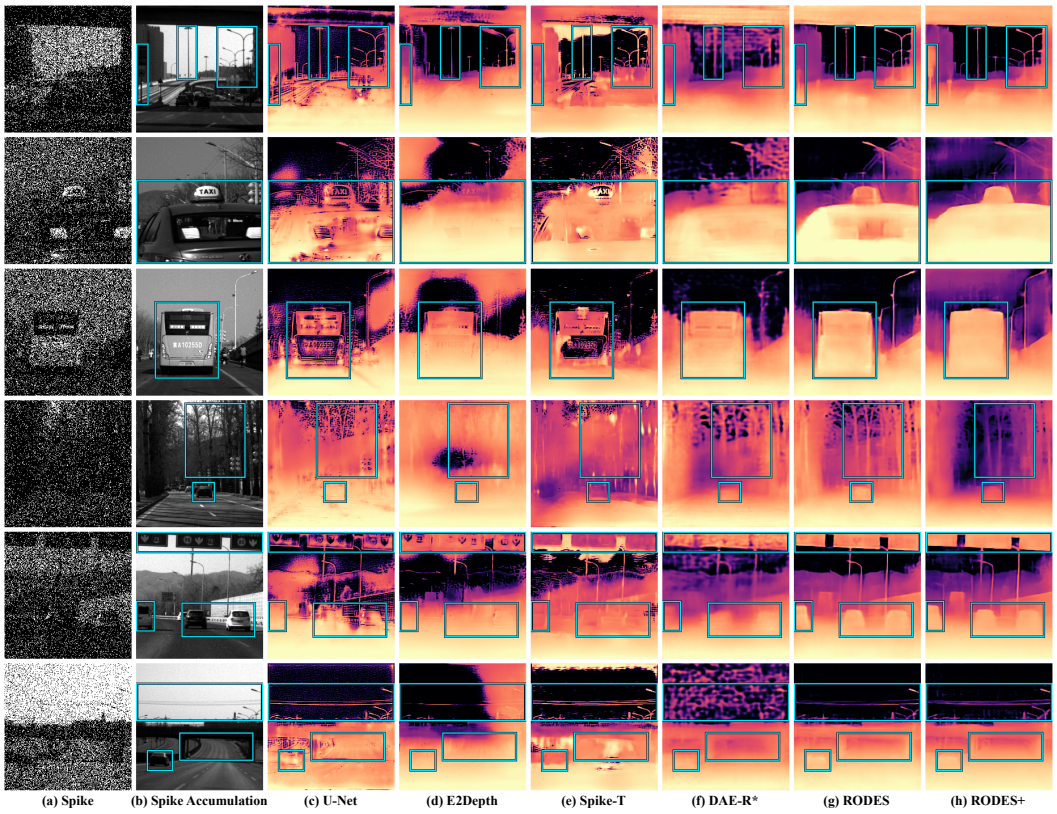

Figure 8: **Additional qualitative results on unseen Outdoor-Spike recordings.** From left to right: raw spike input, spike accumulation, U-Net, E2Depth, Spike-T, DAE-R*, RODES, and RODES+.

D.3 Additional Results on Outdoor-Spike

Figure 8 evaluates additional unseen physical-camera recordings spanning diverse scene layouts and spike densities. RODES and RODES+ recover stable road surfaces, foreground–background transitions, vehicles, and surrounding structures despite the domain shift from simulated training streams to physical spike-camera measurements. Ground-truth depth is unavailable for these recordings, so the comparison is qualitative and all methods use spike-only input at inference.