

# EventFly: Event Camera Perception from Ground to the Sky

## Supplementary Material

### Table of Contents

<b>A The EXPo Benchmark</b>	<b>1</b>
A.1 Benchmark Overview . . . . .	1
A.2 Cross-Platform Configurations . . . . .	1
A.3 Benchmark Structure . . . . .	2
A.4 Semantic Definitions . . . . .	2
A.5 Platform-Specific Statistics . . . . .	3
A.6 License . . . . .	5
<b>B Event Activation Prior: Formulation</b>	<b>5</b>
B.1. Problem Formulation . . . . .	5
B.2 EAP: Motivation & Formulation . . . . .	5
B.3 Likelihood for Supervised Loss . . . . .	5
B.4 Formulating EAP . . . . .	6
B.5 Empirical Estimation of EAP . . . . .	6
B.6 Integrating EAP into the Training Objective . . . . .	6
<b>C Event Activation Prior: Observation</b>	<b>6</b>
C.1. Class Distribution Statistics . . . . .	6
C.2. Class Distribution Maps . . . . .	6
C.3. Event-Triggered Activation Maps . . . . .	8
<b>D Additional Experiment Results</b>	<b>10</b>
D.1. Class-Wise Adaptation Results . . . . .	10
D.2 Additional Qualitative Assessment . . . . .	10
D.3 Failure Cases . . . . .	10
D.4 Video Demos . . . . .	11
<b>E Broader Impact &amp; Limitations</b>	<b>11</b>
E.1. Broader Impact . . . . .	11
E.2. Societal Influence . . . . .	11
E.3. Potential Limitations . . . . .	11
<b>F. Public Resource Used</b>	<b>17</b>




### A. The EXPo Benchmark

In this section, we elaborate on the data structure, definitions, configurations, statistics, and visual examples of the proposed *EXPo* (Event-based *X*ross-*P*latform perception) benchmark.

#### A.1. Benchmark Overview

Our *EXPo* benchmark serves as the first comprehensive effort to tackle the challenging task of cross-platform adaptation for event camera perception. Building upon the newly launched M3ED dataset [1], our benchmark focuses on enabling robust, domain-adaptive perception across diverse robotic platforms. By incorporating a rich variety of event

Table A. The summary of platform-level statistics in *EXPo*.

Platform	 Vehicle	 Drone	 Quadruped
Frame (train)	30,321	13,458	17,302
Frame (val)	12,998	5,772	7,421
Res. ( $H, W$ )	$360 \times 640$	$360 \times 640$	$360 \times 640$
Res. ( $T$ )	20	20	20
Duration	5,000,000	5,000,000	5,000,000
Semantics	19 Classes / 11 Classes		
Environment	City, Urban, Suburban, Rural		

data and semantic labels, we aim to highlight key discrepancies among platforms and provide a robust testbed for evaluating cross-platform performance.

Tab. A provides the platform-level statistics of each platform. The overall benchmark consists of 89,228 frames collected from three distinct platforms – *vehicle*, *drone*, and *quadruped* – across 21 sequences: 6 from the vehicle, 7 from the drone, and 8 from the quadruped. The sequences capture a wide range of dynamic real-world scenarios and span diverse environments, including city, urban, suburban, and rural scenes. This diversity ensures that the benchmark covers both structured and unstructured environments, replicating real-world challenges faced by event cameras deployed across different robotic platforms.

#### A.2. Cross-Platform Configurations

The *EXPo* benchmark aims to highlight platform-specific discrepancies, such as motion dynamics, perspectives, and environmental interactions. Specifically, ground vehicles capture low-altitude perspectives with dense surface-level details, such as roads, curbs, and obstacles. Drones provide high-altitude views with sparse ground-level features, focusing on landscapes, buildings, and environmental structures. Quadrupeds, on the other hand, operate closer to human eye levels, capturing mixed indoor-outdoor dynamics and a wider range of semantic elements. These platform-specific variations make this benchmark a holistic resource for studying domain-specific adaptation and developing robust models capable of generalizing across diverse operational settings.

The event camera data in our benchmark is collected using the Prophesee Gen 4 (EVKv4) event camera [2], a state-of-the-art sensor known for its high temporal resolution and dynamic range. This sensor offers a spatial resolution of  $720 \times 1280$  pixels and a field of view of  $63^\circ \times 38^\circ$ . This consistent sensor setup is employed across all three platforms, ensuring that the observed domain gaps arise purely from platform-specific differences, such as variations in motion

patterns, viewpoint dynamics, and environmental interactions, rather than discrepancies in sensor specifications. By eliminating sensor-level variations, the benchmark ensures that the adaptation challenge remains focused on the core differences between the platforms. This configuration not only strengthens our validity for cross-platform adaptation but also facilitates meaningful comparisons of model performance across varied operational contexts.

### A.3. Benchmark Structure

The *EXPo* benchmark comprises 21 sequences distributed across three platforms: 6 sequences for the *vehicle*, 7 sequences for the *drone*, and 8 sequences for the *quadruped*. Tab. B provides a detailed breakdown of the dataset structure and sequence information for each platform.




Specifically, the benchmark includes 43,766 frames from the *vehicle* platform, 19,899 frames from the *drone* platform, and 25,563 frames from the *quadruped* platform, resulting in a total of 89,228 frames. The detailed information for each sequence across the three platforms is shown in Tab. B. This extensive collection makes *EXPo* the largest benchmark for event camera perception.

As shown in Tab. A, we split each platform into two subsets: training set and validation set. We sample for each sequence in each platform the last 40% of frames for validation, and use the remaining data for training. In total, there are 61,081 frames for training and 26,191 frames for validation. Since the original spatial resolution is high, we subsample it from  $720 \times 1280$  pixels to  $360 \times 640$  pixels, *i.e.*, resize both the height and width to half of the original values. Following the setting of DSEC-Semantic [5], the temporal resolution is set to 20 (bins). Additionally, the duration  $\Delta T$  is set to 5,000,000.

### A.4. Semantic Definitions

The *EXPo* benchmark consists of a total of 19 semantic classes, which ensure a holistic dense perception for the event camera scenes acquired by the three platforms. The specific definition of each class is listed as follows:

- ■ road (ID: 0): The drivable surface designed for vehicle travel, typically marked by lanes and boundaries.
- ■ sidewalk (ID: 1): Elevated pathways adjacent to roads, designated for pedestrian use.
- ■ building (ID: 2): Permanent structures designed for residential, commercial, or industrial purposes.
- ■ wall (ID: 3): Vertical structures that enclose or divide areas, often used for security or boundary delineation.
- ■ fence (ID: 4): Lightweight barriers, usually made of wood or metal, marking boundaries or containing areas.
- ■ pole (ID: 5): Vertical cylindrical objects, such as lamp posts or utility poles, used for lighting, signage, or power distribution.
- ■ traffic-light (ID: 6): Signal devices positioned































Table B. The dataset structure and sequence information among the  vehicle ( $\mathcal{P}^v$ ),  drone ( $\mathcal{P}^d$ ), and  quadruped ( $\mathcal{P}^q$ ) platforms, respectively, in the proposed *EXPo* benchmark.


Platform	Sequence Name	# Frames	Total
Vehicle	horse	714	43,766
	penno_small_loop	1,102	
	rittenhouse	9,752	
	ucity_small_loop	16,867	
	city_hall	7,453	
	penno_big_loop	7,878	
Drone	fast_flight_1	2,229	19,899
	fast_flight_2	4,077	
	penno_parking_1	2,810	
	penno_parking_2	2,713	
	penno_plaza	1,694	
	penno_cars	3,073	
	penno_trees	3,303	
Quadruped	penno_short_loop	2,942	25,563
	skatepark_1	2,305	
	skatepark_2	1,652	
	srt_green_loop	1,597	
	srt_under_bridge_1	5,083	
	srt_under_bridge_2	4,533	
	art_plaza_loop	3,615	
	rocky_steps	3,836	

at road intersections to manage traffic flow and ensure the safety of traffic participants.

- ■ traffic-sign (ID: 7): Informational or regulatory signs placed along roads to guide and control traffic behavior.
- ■ vegetation (ID: 8): Plant life, including trees, shrubs, and grass, typically forming natural surroundings in outdoor environments.
- ■ terrain (ID: 9): Unpaved ground surfaces such as dirt paths, grassy fields, or rocky areas.
- ■ sky (ID: 10): The open expanse above the ground, often capturing atmospheric and weather conditions.
- ■ person (ID: 11): Human individuals present in the scene, either stationary or in motion.
- ■ rider (ID: 12): Individuals on moving devices such as bicycles, motorcycles, or scooters, distinct from pedestrians.
- ■ car (ID: 13): Small to medium-sized motorized vehicles used for personal or commercial transport.
- ■ truck (ID: 14): Larger motorized vehicles designed for transporting goods or heavy materials.
- ■ bus (ID: 15): Large motorized vehicles used for mass public transportation of passengers.
- ■ train (ID: 16): Rail-based vehicles, including locomotives and wagons, used for transporting passengers or freight.
- ■ motorcycle (ID: 17): Two-wheeled motorized vehicles, often used for individual transport or recreation.

Table C. **The definitions of the semantic classes** in the *EXPo* benchmark. We provide two versions of label mappings, *i.e.*, the **19-class** setting and the **11-class** setting, to ensure a holistic dense perception of the scenes acquired by the event camera.

19-Class		11-Class	
ID	Class Name	ID	Class Name
0	 road	5	 road
1	 sidewalk	6	 sidewalk
2	 building	1	 building
3	 wall	9	 wall
4	 fence	2	 fence
5	 pole	4	 pole
6	 traffic-light	10	 traffic-sign
7	 traffic-sign		
8	 vegetation	7	 vegetation
9	 terrain	0	 background
10	 sky		
11	 person	3	 person
12	 rider		
13	 car	8	 car
14	 truck		
15	 bus		
16	 train		
17	 motorcycle		
18	 bicycle		

-  bicycle (ID: 18): Non-motorized two-wheeled vehicles powered by pedaling, used for transport or leisure.

Our benchmark supports two versions of label mappings, *i.e.*, the **19-class setting** and the **11-class setting**, where the latter is consistent with the seminar event-based semantic segmentation work ESS [5]. Tab. C summarizes the relationship between these two label mappings. In our benchmark experiments, we adopt the 11-class setting for comparing different adaptation methods across platforms.

### A.5. Platform-Specific Statistics

Each of the three platforms in the *EXPo* benchmark represents a unique collection of event camera data. To better understand the domain gaps among these platforms, we calculate the following platform-specific statistics.

- **Platform-Specific Semantic Distributions:** The relative proportions of each semantic class across the three platforms are presented in Tab. D, with semantic occupations normalized to 1. Notable discrepancies are observed among the platforms.
  - For instance, the *drone* platform accounts for 45.75% of the *road* class, attributed to its high-altitude per-

spective that captures expansive ground surfaces. In contrast, the *vehicle* platform dominates classes such as *building*, *traffic-sign*, and all categories of *car*, reflecting its road-level viewpoint and focus on urban navigation. Similarly, all instances of *traffic-light* appear exclusively in the *vehicle* platform, as this class is inherently associated with vehicle-centric scenarios.

- On the other hand, the *quadruped* platform, with its low-height perspective, captures a higher proportion of *fence* (76.36%), *wall* (83.23%), and similar semantic categories. This aligns with its tendency to perceive surroundings closer to ground level, making it better suited for mixed indoor-outdoor environments.
  - As for the *drone* platform, a significant proportion of *terrain* (69.26%) is captured due to its elevated viewpoint, which provides a broader landscape perspective. This platform also includes a notable share of *car*-related classes, such as *truck* (19.20%), *bus* (7.89%), and *motorcycle* (45.45%), reflecting its ability to observe these objects from a unique vantage point that complements ground-level perspectives.
  - Each platform thus exhibits distinct semantic distributions, emphasizing the importance of tailored domain adaptation strategies for robust cross-platform event perception.
- **Absolute Semantic Distributions:** We calculate the absolute semantic occupations for each platform and present the statistics in Tab. E. As shown, the distributions for all three platforms exhibit a long-tailed nature, reflecting real-world event camera scenarios where certain static classes dominate while dynamic and small-object classes occur less frequently.
    - The majority classes for the *vehicle* platform are *building* (24.91%), *vegetation* (23.77%), and *road* (21.94%). These static classes dominate due to the platform’s road-level perspective, which frequently encounters large, continuous structures and roadside greenery. In contrast, small and dynamic classes, such as *rider* (0.02%) and *motorcycle* (0.01%), are underrepresented, underscoring the vehicle platform’s bias towards large, static objects in its operating environment.
    - The *drone* platform primarily captures *road* (34.51%), *terrain* (31.46%), and *vegetation* (14.52%). This is due to its high-altitude perspective, which provides expansive views of ground surfaces and surrounding landscapes. Dynamic classes, such as different categories of *car*, are underrepresented because they occupy less visual space from the drone’s viewpoint compared to static, large-area features.
    - We also observe that the *quadruped* platform exhibits notably higher proportions of *sky* (8.68%),

Table D. **The platform-specific semantic distributions** among the 🚗 vehicle ( $\mathcal{P}^v$ ), 🚁 drone ( $\mathcal{P}^d$ ), and 🐾 quadruped ( $\mathcal{P}^q$ ) platforms, respectively, in the proposed **EXPo** benchmark. We compare the relative proportions (normalized to 1) of each semantic class from three platforms. The distributions of *vehicle*, *drone*, and *quadruped* are denoted by the 🟢 green, 🔴 red, and 🔵 blue colors, respectively.

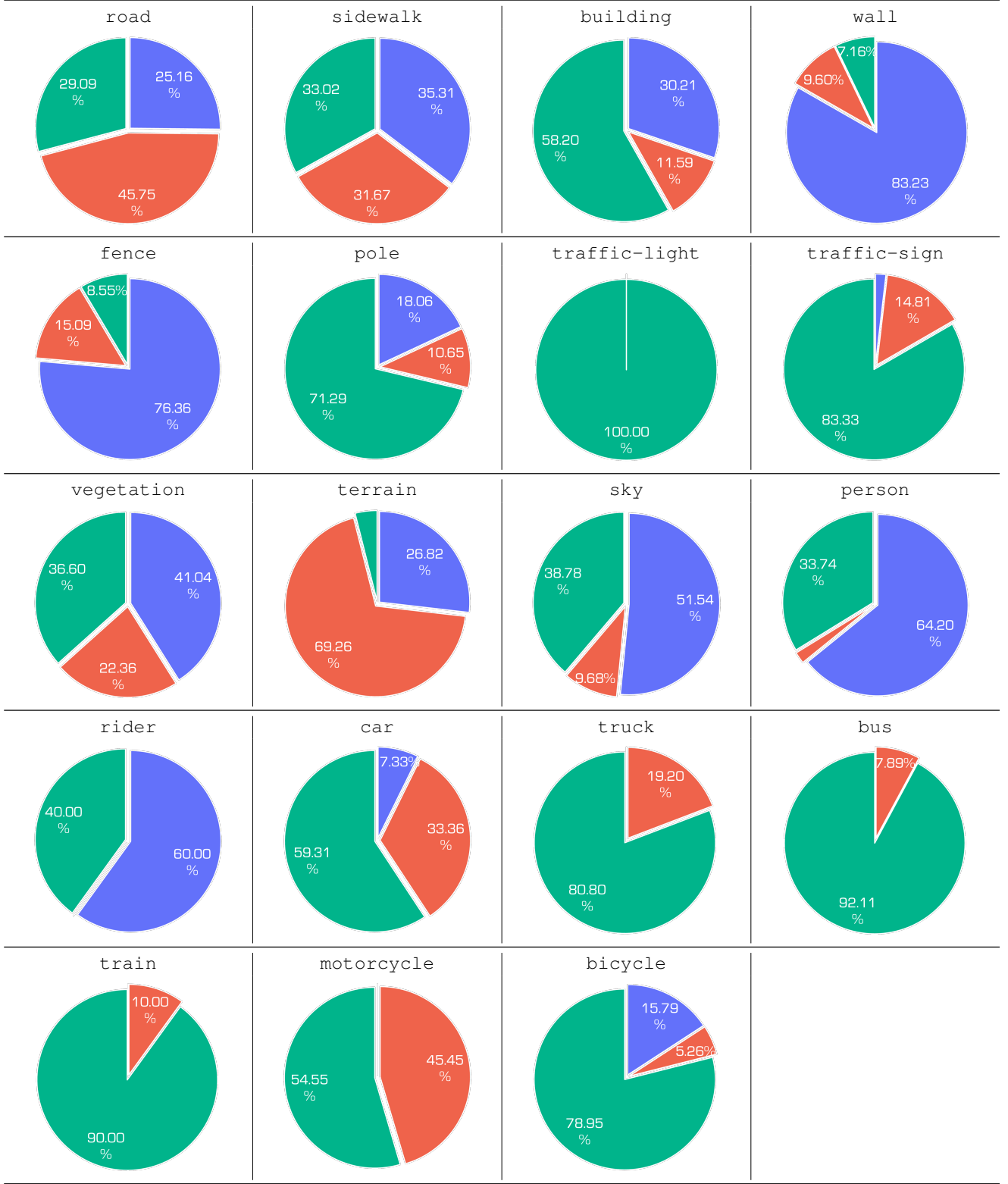


























Table E. The absolute platform-specific semantic distributions among the  vehicle ( $\mathcal{P}^v$ ),  drone ( $\mathcal{P}^d$ ), and  quadruped ( $\mathcal{P}^q$ ) platforms, respectively, in the proposed *EXPo* benchmark.

Class	Vehicle	Drone	Quadruped
 road	21.94%	34.51%	18.98%
 sidewalk	6.63%	6.36%	7.09%
 building	24.91%	4.96%	12.93%
 wall	0.47%	0.63%	5.46%
 fence	0.55%	0.97%	4.91%
 pole	2.21%	0.33%	0.56%
 traffic-light	0.22%	0.00%	0.00%
 traffic-sign	0.45%	0.08%	0.01%
 vegetation	23.77%	14.52%	26.65%
 terrain	1.78%	31.46%	12.18%
 sky	6.53%	1.63%	8.68%
 person	0.82%	0.05%	1.56%
 rider	0.02%	0.00%	0.03%
 car	7.36%	4.14%	0.91%
 truck	1.01%	0.24%	0.00%
 bus	1.05%	0.09%	0.00%
 train	0.09%	0.01%	0.00%
 motorcycle	0.01%	0.01%	0.00%
 bicycle	0.15%	0.01%	0.03%
<b>Total</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

wall (5.46%), and fence (4.91%) compared to the other two platforms. This is attributed to its low-altitude perspective, which captures more vertical structures and surrounding boundaries, as well as frequent mixed indoor-outdoor scenarios. Unlike the vehicle and drone platforms, quadruped data features a more balanced representation of close-range objects and environmental details.

These platform-specific statistics provide a comprehensive understanding of the challenges in cross-platform adaptation, emphasizing the need for robust event camera perception models capable of handling diverse semantic distributions and environmental contexts.

## A.6. License

The *EXPo* benchmark is released under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)<sup>1</sup> license.

## B. Event Activation Prior: Formulation

In event-based cross-platform adaptation, each platform introduces unique activation patterns due to variations in sensor perspectives, motion dynamics, and environmental conditions. The Event Activation Prior (EAP) captures these platform-specific activation patterns and encourages confident predictions by leveraging the classic entropy minimization framework. In this section, we elaborate on the formulation of our proposed EAP in more detail.

<sup>1</sup><https://creativecommons.org/licenses/by-sa/4.0/legalcode>.

## B.1. Problem Formulation

In our setting, we address cross-platform adaptation across three distinct event data domains: *vehicle*, *drone*, and *quadruped*, referred to as  $\mathcal{D} = \{\mathcal{P}^v, \mathcal{P}^d, \mathcal{P}^q\}$ , respectively. Each domain contains:

- **Event Voxel Grids:**  $\mathbf{V} \in \mathbb{R}^{T \times H \times W}$ , where  $T$  is the number of temporal bins and  $(H, W)$  are the spatial dimensions of the event sensor.
- **Semantic Labels** (source domain only):  $y \in \mathbb{R}^{H \times W}$ , where each pixel corresponds to one of  $C$  pre-defined semantic classes.

In our cross-platform adaptation problem, we assume access to fully labeled data from a source domain while only having access to unlabeled data from a target domain. The objective is to leverage both the labeled source data and the unlabeled target data to train an event camera perception model that can perform well on the target domain. This adaptation is challenging because each platform captures data from distinct perspectives, motion patterns, and environmental contexts.

## B.2. EAP: Motivation & Formulation

EAP is designed to guide cross-platform adaptation by leveraging platform-specific event activation patterns. Events are triggered by changes in brightness due to motion, making certain regions in the event data – characterized by frequent activations – highly informative. By minimizing entropy in these regions, we hope to encourage the model to make confident predictions that align with the target domain’s unique motion-triggered patterns, which in turn improve the perception performance.

## B.3. Likelihood for Supervised Loss

For labeled data from the source domain  $\mathcal{P}^{\text{src}} \in \mathcal{D}$ , we train our event camera perception model by maximizing the likelihood of the ground truth labels. This likelihood,  $P(y|\mathbf{V})$ , forms the supervised loss term:

$$L(\theta) = - \sum_{\mathbf{V} \in \mathcal{P}^{\text{src}}} \log P(y|\mathbf{V}; \theta), \quad (1)$$

where  $\theta$  represents the model parameters. This supervised loss anchors the model’s learning in well-labeled source data, providing a foundation for generalization.

Since we lack labeled data in the target domain, we define the EAP to help the model leverage unlabeled data by minimizing prediction uncertainty in *high-activation regions* of the target domain. These regions,  $\mathbf{S} \subset \{0, 1, \dots, H-1\} \times \{0, 1, \dots, W-1\}$ , are identified based on the characteristic event activations in each platform. To achieve this, EAP follows the principle of entropy minimization, where we aim to:

- Identify high-activation regions  $\mathbf{S}$  in the target domain.

- Minimize the conditional entropy  $H(y_S|\mathbf{V}_S, \mathbf{S})$  in these regions, promoting confident predictions that align with target-specific patterns.

#### B.4. Formulating EAP

To incorporate the EAP into the model, we enforce a prior on  $\theta$  that reduces entropy in high-activation regions  $\mathbf{S}$  of the target domain. Following the maximum entropy principle [3], we express this as a soft regularization:

$$\mathbb{E}_\theta [H(\mathbf{V}_S, y_S|\mathbf{S})] \leq c, \quad (2)$$

where  $c$  is a small constant enforcing high confidence in predictions. Using the principle of maximum entropy, we obtain:

$$P(\theta) \propto \exp(-\lambda H(\mathbf{V}_S, y_S|\mathbf{S})) , \quad (3)$$

$$\propto \exp(-\lambda H(y_S|\mathbf{V}_S, \mathbf{S})) , \quad (4)$$

where  $\lambda > 0$  is the Lagrange multiplier corresponding to constant  $c$ , which balances the effect of EAP on the model’s training objective.

#### B.5. Empirical Estimation of EAP

To implement the EAP, we estimate the conditional entropy  $H(y|\mathbf{V}, \mathbf{S})$  by focusing on high-activation regions  $\mathbf{S}$  in the target domain. This conditional entropy captures prediction uncertainty within the specific spatial region  $\mathbf{S}$ , allowing us to concentrate adaptation efforts on regions aligned with platform-specific activations. Using an empirical plug-in estimator, we approximate this entropy as:

$$H_{\text{emp}}(y|\mathbf{V}, \mathbf{S}) = \mathbb{E}_{\mathbf{V}, y, \mathbf{S}} \left[ \hat{P}(y|\mathbf{V}, \mathbf{S}) \log \hat{P}(y|\mathbf{V}, \mathbf{S}) \right], \quad (5)$$

where  $\hat{P}(y|\mathbf{V}, \mathbf{S})$  is the empirical prediction probability conditioned on the event voxel grid  $\mathbf{V}$  and restricted to region  $\mathbf{S}$ . By minimizing  $H_{\text{emp}}(y|\mathbf{V}, \mathbf{S})$ , we encourage confident predictions within these regions, aligning the model’s predictions with the target domain’s activation patterns.

#### B.6. Integrating EAP into the Training Objective

To incorporate EAP into the model’s training, we define the overall objective function as a maximum-a-posteriori (MAP) estimation:

$$C(\theta) = \mathcal{L}(\theta) - \lambda H_{\text{emp}}(y|\mathbf{V}, \mathbf{S}), \quad (6)$$

where  $\mathcal{L}(\theta)$  represents the supervised loss on source data.  $H_{\text{emp}}(y|\mathbf{V}, \mathbf{S})$  minimizes uncertainty in the target domain by leveraging EAP over high-activation regions.

By focusing on high-activation areas, the event camera perception model learns to adapt to the target domain’s unique event-triggered patterns, achieving robust adaptation across platforms. This approach captures and emphasizes platform-specific activation patterns, making EAP an effective regularization for confident adaptation in event-based cross-platform scenarios.

### C. Event Activation Prior: Observation

In this section, we provide concrete evidence supporting the proposed Event Activation Prior (EAP) by analyzing the platform-specific activation patterns in both static and dynamic regions. The evidence is presented through class distribution statistics and maps, which highlight the unique activation characteristics of each platform.

#### C.1. Class Distribution Statistics

As discussed in Sec. A.4 and Sec. A.5, the same semantic class exhibits notable discrepancies across the three platforms, influenced by their unique perspectives, motion dynamics, and environmental contexts. Such discrepancies emphasize the need for spatial priors, as formulated in EAP, to account for platform-specific variations.




For example, the class `road` dominates the *drone* platform (45.75%) due to its high-altitude perspective capturing extensive ground-level surfaces, while in *vehicle* (21.94%) and *quadruped* (15.42%) platforms, this class appears more localized. Dynamic classes such as `car` and `person` show higher prominence in the *vehicle* platform, consistent with its traffic-oriented scenarios, while being less frequent in *drone* and *quadruped* data due to limited proximity and perspectives for capturing such objects. Static classes like `vegetation` and `building` exhibit significant variation in coverage due to platform-specific viewpoints, with *drone* capturing broader fields of view compared to the ground-level perspectives of *vehicle* and *quadruped*.

These statistics reinforce the hypothesis that leveraging spatial priors informed by class-specific activation patterns can significantly enhance cross-platform adaptation.

#### C.2. Class Distribution Maps

Tab. F and Tab. G present the activation proportions for static and dynamic classes, respectively, across the *vehicle*, *drone*, and *quadruped* platforms. These heatmaps reveal distinct spatial coverage and density patterns for each platform, which serve as the foundation for the proposed EAP. These tables highlight the following key observations:

- In the *vehicle* platform, the `road` class is highly concentrated in the lower-central region, reflecting the ground-level perspective. In contrast, *drone* exhibits a broader, more evenly distributed pattern due to its high-altitude viewpoint capturing expansive ground surfaces. The *quadruped* platform shows a localized, narrower distribution, aligning with its lower vantage point.
- The *vehicle* platform exhibits dense, vertically structured priors for `building`, consistent with urban driving scenarios. Meanwhile, *drone* and *quadruped* display sparser coverage, with *drone* capturing larger landscape-level structures and *quadruped* focusing on closer, localized regions. A similar pattern applies to some `car` classes, such as `bus`, `train`, `motorcycle`, and `bicycle`.

Table F. **The class distribution maps** of static classes among the  vehicle ( $\mathcal{P}^v$ ),  drone ( $\mathcal{P}^d$ ), and  quadruped ( $\mathcal{P}^q$ ) platforms, respectively, in the proposed **EXPo** benchmark. The brighter the color, the higher the probability of occurrences. Best viewed in colors.




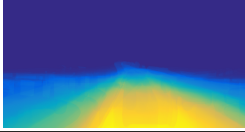
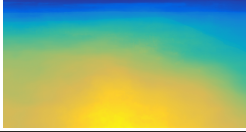
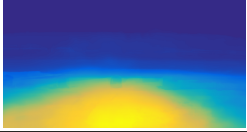
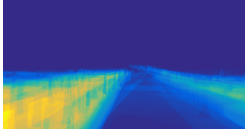
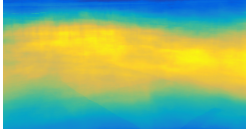
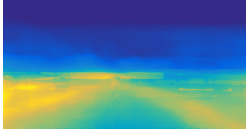
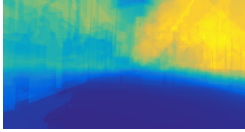
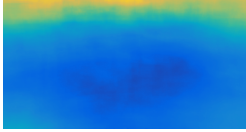
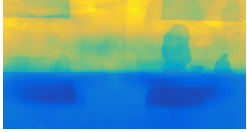
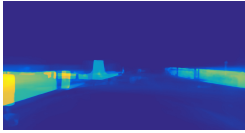

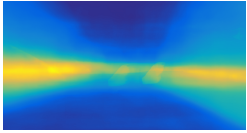
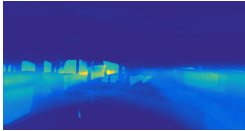
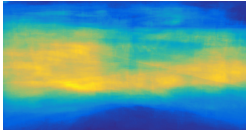
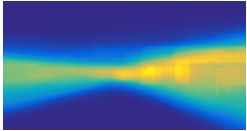
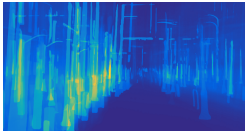
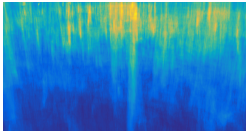
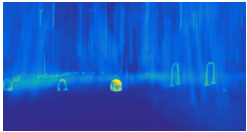
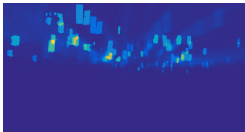
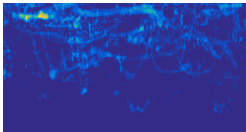


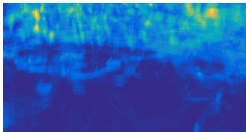
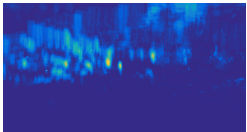
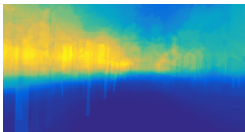

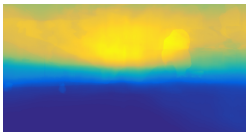


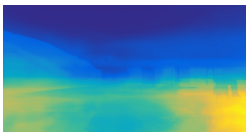









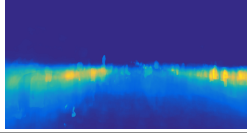
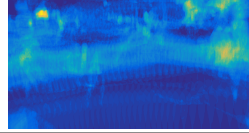
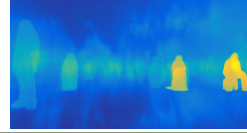
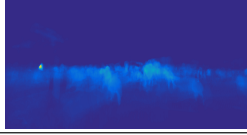
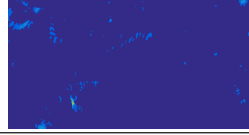
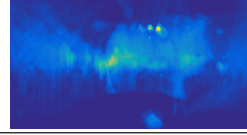
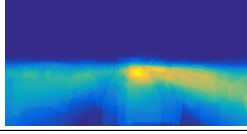
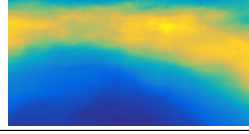
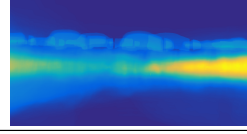
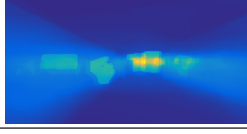
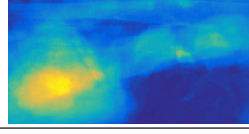
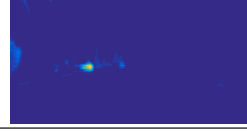
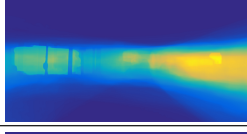
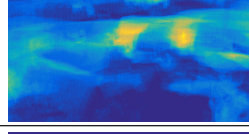
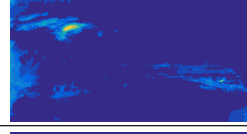
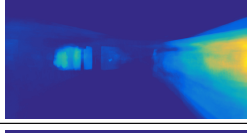
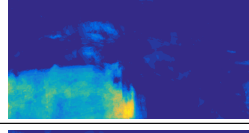
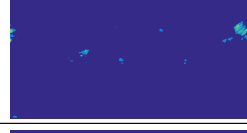
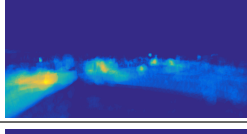
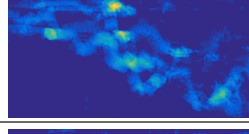
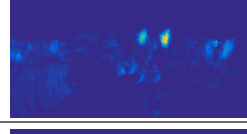
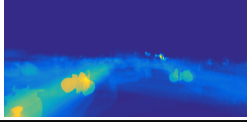
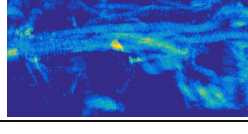
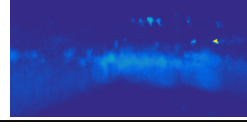
ID	Class	Type	 vehicle ( $\mathcal{P}^v$ )	 drone ( $\mathcal{P}^d$ )	 quadruped ( $\mathcal{P}^q$ )
0	road	static			
1	sidewalk	static			
2	building	static			
3	wall	static			
4	fence	static			
5	pole	static			
6	traffic-light	static			
7	traffic-sign	static			
8	vegetation	static			
9	terrain	static			
10	sky	static			

Table G. **The class distribution maps** of dynamic classes among the  vehicle ( $\mathcal{P}^v$ ),  drone ( $\mathcal{P}^d$ ), and  quadruped ( $\mathcal{P}^q$ ) platforms, respectively, in the proposed *EXPo* benchmark. The brighter the color, the higher the probability of occurrences. Best viewed in colors.

ID	Class	Type	 vehicle ( $\mathcal{P}^v$ )	 drone ( $\mathcal{P}^d$ )	 quadruped ( $\mathcal{P}^q$ )
11	person	dynamic			
12	rider	dynamic			
13	car	dynamic			
14	truck	dynamic			
15	bus	dynamic			
16	train	dynamic			
17	motorcycle	dynamic			
18	bicycle	dynamic			




- The pole and traffic-light classes are distinctly prominent in the *vehicle* platform due to urban driving environments. The *drone* platform shows certain occurrences, while the *quadruped* platform captures sporadic patterns that align with its lower viewpoint.
- For majority classes, such as vegetation, terrain, and sky, the spatial distribution for *vehicle* and *drone* is broader and denser, reflecting outdoor scenarios with natural elements. The *quadruped* platform captures localized vegetation mainly from the upper half of the field of view, often in close proximity to its route.




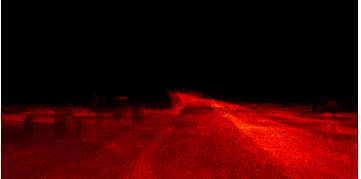
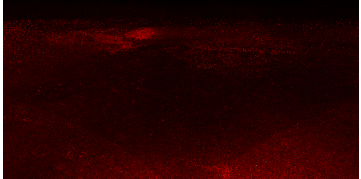
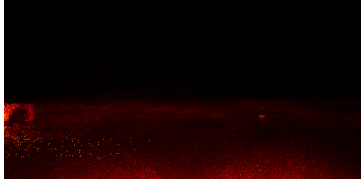
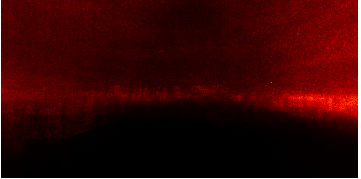
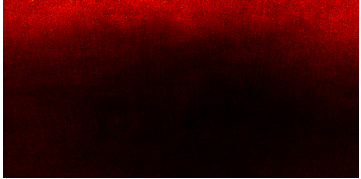
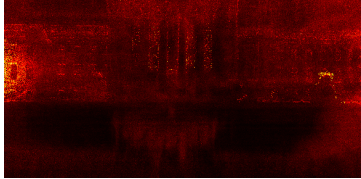
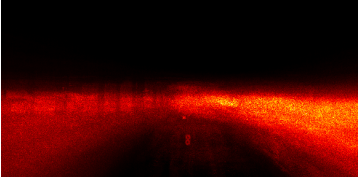
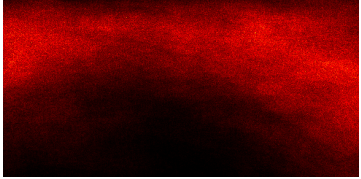

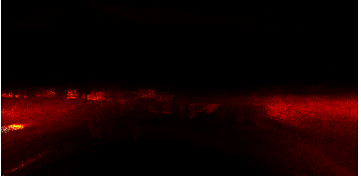
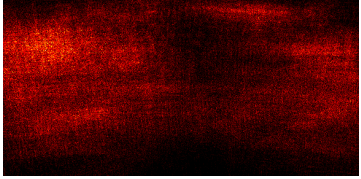
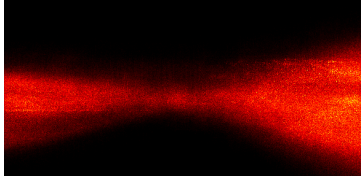
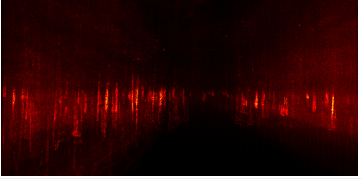
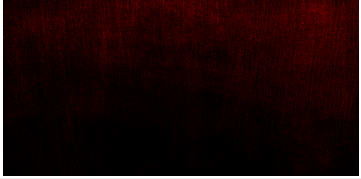
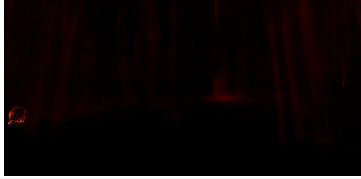
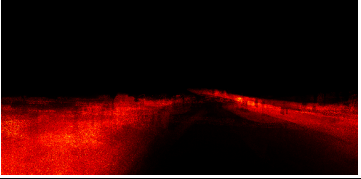
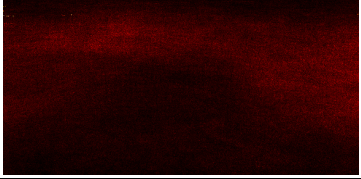
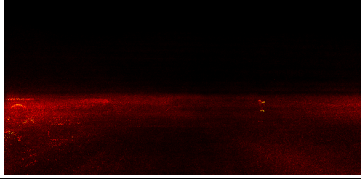
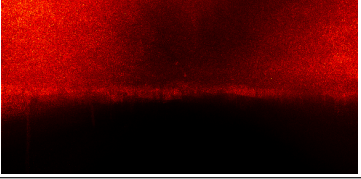
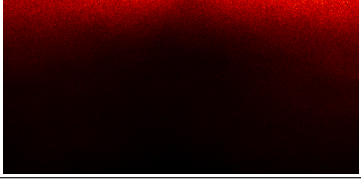
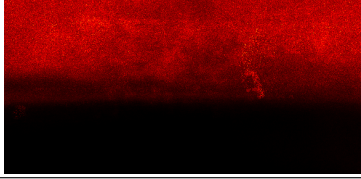
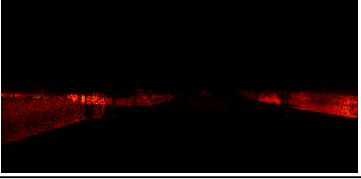
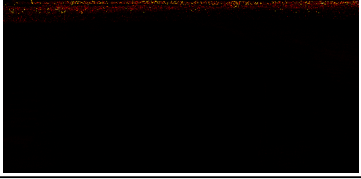
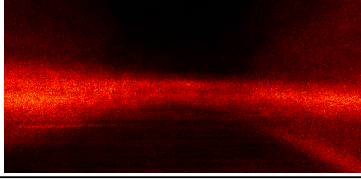
These heatmaps demonstrate the inherent semantic and spatial discrepancies across platforms, highlighting the necessity of incorporating spatial priors into the cross-platform adaptation process. By leveraging these platform-specific semantic distributions, the EAP enables more confident and domain-aligned predictions, ensuring effective adaptation across diverse operational contexts.

### C.3. Event-Triggered Activation Maps

Our EAP-driven event data mixing technique builds on the assumption that event-triggered activations are closely



Table H. **The event-triggered activation maps** among the  vehicle ( $\mathcal{P}^v$ ),  drone ( $\mathcal{P}^d$ ), and  quadruped ( $\mathcal{P}^q$ ) platforms, respectively, in the proposed *EXPo* benchmark. The brighter the color, the higher the probability of occurrences. Best viewed in colors.

Class	 vehicle ( $\mathcal{P}^v$ )	 drone ( $\mathcal{P}^d$ )	 quadruped ( $\mathcal{P}^q$ )
road			
building			
car			
fence			
pole			
sidewalk			
vegetation			
wall			



linked to semantic distributions, as these activations reflect dynamic and structural changes captured by event cameras. To validate this assumption, we calculate probability maps of event-triggered activations for all semantic classes and present the results in Tab. H.

These maps reveal a striking correlation between event-triggered activations and semantic class distributions. Specifically, the event-triggered activations in static classes such as *road*, *building*, and *vegetation* demonstrate strong spatial consistency across platforms. For example, in the *vehicle* platform, *road* activations are concentrated in the lower-central region, reflecting the expected viewpoint of ground-level sensors. Similarly, *building* activations align vertically, consistent with urban environments. This correlation underscores the utility of EAP in capturing spatially consistent priors for static classes.

For dynamic classes such as *car*, activations are more sporadic but still exhibit platform-specific patterns. The *vehicle* platform shows dense activations in traffic-heavy areas, while the *drone* platform captures broader distributions due to its high-altitude perspective. The *quadruped* platform highlights localized activations near dynamic objects encountered in its immediate surroundings.

These observations reinforce the premise of EAP: that leveraging platform-specific activation patterns can guide adaptation by aligning predictions with the unique event-triggered dynamics of each platform. By incorporating these patterns into the adaptation process, EAP enhances confidence in predictions, particularly for challenging classes or underrepresented regions.

## D. Additional Experiment Results

In this section, we provide additional results from our comparative and ablation experiments to further demonstrate the effectiveness and superiority of the proposed *EventFly* framework.

### D.1. Class-Wise Adaptation Results

In the main body of this paper, due to space limits, we provide only the class-wise cross-platform adaptation results for the *vehicle* ( $\mathcal{P}^v$ ) to *drone* ( $\mathcal{P}^d$ ) and the *vehicle* ( $\mathcal{P}^v$ ) to *quadruped* ( $\mathcal{P}^q$ ) settings.

In this supplementary file, we further provide the cross-platform adaptation results from the following settings:

- Tab. I: Adaptation from the *drone* ( $\mathcal{P}^d$ ) platform to the *vehicle* ( $\mathcal{P}^v$ ) platform.
- Tab. J: Adaptation from the *drone* ( $\mathcal{P}^d$ ) platform to the *quadruped* ( $\mathcal{P}^q$ ) platform.
- Tab. K: Adaptation from the *quadruped* ( $\mathcal{P}^q$ ) platform to the *vehicle* ( $\mathcal{P}^v$ ) platform.
- Tab. L: Adaptation from the *quadruped* ( $\mathcal{P}^q$ ) platform to the *drone* ( $\mathcal{P}^d$ ) platform.

Across all adaptation settings, our framework consistently achieves the highest accuracy (Acc), mean accuracy (mAcc), and mean Intersection over Union (mIoU), demonstrating its robustness in adapting event-based perception across platforms. Notably, *EventFly* outperforms prior methods such as MIC [4] and PLSR [8] by significant margins, particularly in complex settings such as the adaptation from *drone* ( $\mathcal{P}^d$ ) to *quadruped* ( $\mathcal{P}^q$ ), and from *quadruped* ( $\mathcal{P}^q$ ) to *drone* ( $\mathcal{P}^d$ ).

Our approach demonstrates superior performance in static classes, such as *road* and *vegetation*, which are critical for general scene understanding. This aligns with the strengths of EAP, which captures spatially consistent patterns. Dynamic classes often pose greater challenges due to motion and variability across domains. However, we observe that our approach achieves competitive results, surpassing existing methods in most cases. For example, in the *quadruped* ( $\mathcal{P}^q$ ) to *vehicle* ( $\mathcal{P}^v$ ) scenario, our approach provides notable improvements in *car* and *person* classes, highlighting its ability to transfer motion-sensitive information effectively.

Additionally, the adaptation results emphasize the domain discrepancies between platforms. For instance, in the *drone* ( $\mathcal{P}^d$ ) to *vehicle* ( $\mathcal{P}^v$ ) setting, static classes such as *road* and *building* are better aligned, while smaller, dynamic classes like *pole* and *traffic-light* show more variation. This reflects the inherent viewpoint differences between high-altitude drone perspectives and ground-level vehicle data.

Similarly, in the *quadruped* ( $\mathcal{P}^q$ ) to *drone* ( $\mathcal{P}^d$ ) scenario, our framework’s performance in *vegetation* and *terrain* highlights its ability to adapt between the low-altitude, close-proximity view of quadrupeds and the expansive aerial coverage of drones.

The additional results reinforce the effectiveness of the *EventFly* framework across diverse cross-platform settings. By addressing both static and dynamic class distributions and leveraging platform-specific activation patterns, our framework demonstrates superior generalization and robust adaptation capabilities. These insights further validate the suitability of our approach for real-world, multi-platform event camera perception applications.



### D.2. Additional Qualitative Assessment

In addition to the visual comparisons provided in the main body of this paper, we include more qualitative examples in this supplementary file. Please kindly refer to Fig. A, Fig. B, Fig. C, and Fig. D for the cross-platform adaptation results of the state-of-the-art adaptation methods.

### D.3. Failure Cases

Although the proposed approach demonstrates promising cross-platform adaptation performance, there are certain







failure cases that highlight the limitations and challenges of the approach.

Classes that are inherently dynamic and less frequently represented in the datasets, pose significant challenges. Classes such as `traffic-sign`, which occupy small regions in the voxel grid, exhibit higher misclassification rates. This is particularly evident in the adaptation from  drone ( $\mathcal{P}^d$ ) to  vehicle ( $\mathcal{P}^v$ ), where high-altitude drone perspectives fail to capture the fine details necessary for distinguishing these classes in ground-level data. Additionally, in scenarios involving dense vegetation or crowded urban areas, occlusions lead to reduced prediction confidence.

#### D.4. Video Demos

To provide a more comprehensive illustration of the proposed *EventFly* framework and the *EXPo* benchmark, we have attached three video demos with this supplementary material. Please kindly find the `demo1.mp4`, `demo2.mp4`, and `demo3.mp4` files in the attachment.

Specifically, these three video demos contain the following visual content:

- **Demo #1:** The first demo consists of 813 frames from the `penco_parking_2` sequence, illustrating the cross-platform adaptation from the  vehicle ( $\mathcal{P}^v$ ) platform to the  drone ( $\mathcal{P}^d$ ) platform.
- **Demo #2:** The second demo consists of 1013 frames from the `art_plaza_loop` sequence, illustrating the cross-platform adaptation from the  vehicle ( $\mathcal{P}^v$ ) platform to the  quadruped ( $\mathcal{P}^q$ ) platform.
- **Demo #3:** The third demo consists of 1,000 frames from the `city_hall` sequence, illustrating the cross-platform adaptation from the  drone ( $\mathcal{P}^d$ ) platform to the  vehicle ( $\mathcal{P}^v$ ) platform.

### E. Broader Impact & Limitations

In this section, we elaborate on the broader impact, societal influence, and potential limitations of the proposed *EventFly* framework and the *EXPo* benchmark.

#### E.1. Broader Impact

Our approach and benchmark have the potential to redefine event camera perception across diverse operational platforms, including vehicles, drones, and quadrupeds. By enabling robust cross-platform adaptation, our framework could accelerate advancements in autonomous navigation, disaster response, and robotics, particularly in dynamic and unstructured environments. These contributions could enhance safety, efficiency, and adaptability in real-world applications, such as autonomous driving in dense urban areas, aerial surveillance in remote regions, and robotic assistance in disaster zones.

Moreover, the emphasis on domain-invariant learning for event-based perception addresses a critical gap in current technologies, facilitating the fairer deployment of AI systems across varied socioeconomic and geographical contexts. By creating a benchmark with diverse samples and settings, we aim to foster transparency and reproducibility in the evaluation of event-based systems, contributing to the broader research community’s understanding of event-camera capabilities and limitations.

#### E.2. Societal Influence

The societal influence of our approach and benchmark spans multiple domains:

- **Improved Safety:** Enhanced perception capabilities in dynamic environments can improve safety in autonomous systems, reducing the risk of accidents in transportation and industrial applications.
- **Environmental Monitoring:** The adaptability of our framework to drones and quadrupeds facilitates ecological and environmental monitoring, promoting sustainability and conservation efforts.
- **Accessibility:** The cross-platform design lowers barriers for deploying event camera solutions in resource-constrained settings, democratizing access to advanced vision technologies.



Despite its benefits, it is essential to consider potential ethical implications, including misuse in surveillance and privacy-intrusive applications. Researchers and practitioners should adhere to ethical guidelines to mitigate risks associated with deploying these technologies.

#### E.3. Potential Limitations

While our approach and benchmark demonstrate substantial advancements, there are inherent limitations. For example, the reliance on domain-specific activation patterns might struggle in highly heterogeneous environments with atypical dynamics, such as extreme weather or chaotic lighting conditions. Besides, the reliance on pseudo-labels in unsupervised settings may propagate errors, especially when source-to-target domain gaps are substantial.

Additionally, although our benchmark is comprehensive, it might not encompass all possible scenarios, such as multi-agent coordination or environments with severe occlusions, necessitating further expansions. The current version of the benchmark also does not include settings of multi-source or multi-target adaptation.

In future work, we aim to address these challenges by optimizing the framework for real-time applications, expanding the benchmark to include more diverse scenarios, and investigating advanced self-supervised learning techniques to minimize reliance on pseudo-labels. By acknowledging these limitations, we hope to inspire continued innovation and improvement in event-based perception systems.

Table I. **Benchmark results of platform adaptation** from  drone ( $\mathcal{P}^d$ ) to  vehicle ( $\mathcal{P}^v$ ). Target is trained with ground truth from the target domain. All scores are given in percentage (%). The **second best** and **best** scores under each metric are highlighted in colors.





Method	Acc	mAcc	mIoU	fIoU	ground	build	fence	person	pole	road	walk	veg	car	wall	sign
Source-Only 	57.91	29.97	20.79	11.72	52.64	30.04	0.82	0.35	11.27	46.96	7.48	46.51	31.99	0.00	0.68
AdaptSegNet [7]	68.29	39.99	29.55	13.72	41.79	56.46	0.53	2.80	20.75	68.86	34.47	58.42	40.70	0.00	0.23
DACS [6]	71.78	48.58	36.10	14.34	47.65	60.00	0.00	32.97	23.57	69.89	37.76	63.69	43.45	6.53	11.60
MIC [4]	72.46	49.54	36.88	14.42	48.15	60.68	0.00	30.87	24.95	70.33	39.47	65.17	44.51	6.36	15.21
PLSR [8]	72.46	49.84	37.18	14.45	44.94	62.15	2.55	35.60	23.98	72.59	41.99	61.18	47.92	3.87	12.24
<b>EventFly (Ours)</b>	75.50	52.90	39.92	15.08	53.93	65.14	6.43	31.61	23.93	72.18	46.22	68.68	47.90	4.12	19.01
Target 	86.12	66.02	55.93	16.18	87.07	75.41	22.70	52.59	39.41	79.49	58.82	77.75	69.63	14.79	37.61

Table J. **Benchmark results of platform adaptation** from  drone ( $\mathcal{P}^d$ ) to  quadruped ( $\mathcal{P}^q$ ). Target is trained with ground truth from the target domain. All scores are given in percentage (%). The **second best** and **best** scores under each metric are highlighted in colors.





Method	Acc	mAcc	mIoU	fIoU	ground	build	fence	person	pole	road	walk	veg	car	wall	sign
Source-Only 	66.83	34.05	23.06	17.24	59.62	42.17	2.76	0.24	8.20	48.56	8.55	66.11	17.12	0.00	0.27
AdaptSegNet [7]	67.57	49.51	33.99	14.64	42.75	51.73	33.04	33.32	14.33	54.05	19.71	73.43	20.56	30.91	0.00
DACS [6]	67.73	51.73	36.11	14.49	42.10	55.10	36.25	34.55	15.00	50.45	21.54	75.77	26.54	39.87	0.01
MIC [4]	67.29	50.91	36.27	14.53	44.15	51.15	34.40	37.99	14.43	45.74	23.09	75.38	30.36	41.41	0.89
PLSR [8]	67.83	50.57	36.21	14.67	42.62	53.73	30.80	28.39	15.70	50.15	20.94	75.82	36.48	43.70	0.00
<b>EventFly (Ours)</b>	69.68	51.03	37.37	15.30	44.92	53.12	34.16	39.34	16.95	53.85	17.59	75.10	33.03	41.98	0.97
Target 	80.02	60.55	49.84	19.58	74.80	56.23	46.08	55.28	21.79	59.90	30.31	77.24	58.38	62.47	5.81

Table K. **Benchmark results of platform adaptation** from  quadruped ( $\mathcal{P}^q$ ) to  vehicle ( $\mathcal{P}^v$ ). Target is trained with ground truth from the target domain. All scores are given in percentage (%). The **second best** and **best** scores under each metric are highlighted in colors.






Method	Acc	mAcc	mIoU	fIoU	ground	build	fence	person	pole	road	walk	veg	car	wall	sign
Source-Only 	57.49	33.39	21.30	12.00	56.09	30.76	1.16	13.67	8.84	37.18	13.08	56.39	15.79	1.08	0.30
AdaptSegNet [7]	66.74	41.78	30.65	13.54	43.30	57.25	2.11	23.74	14.85	66.78	34.40	59.86	33.61	1.12	0.11
DACS [6]	71.20	48.04	34.78	14.30	45.05	63.55	3.44	28.44	23.52	67.79	39.25	63.47	43.46	4.56	0.00
MIC [4]	72.46	47.54	35.22	14.59	47.87	64.23	4.17	30.35	21.61	70.35	40.20	63.88	42.85	1.91	0.00
PLSR [8]	72.93	49.82	36.38	14.48	48.51	64.69	3.92	30.15	23.91	71.16	43.34	65.40	46.13	2.97	0.00
<b>EventFly (Ours)</b>	73.93	49.56	37.70	14.93	50.94	66.17	4.90	35.48	26.13	66.73	32.53	69.77	46.93	2.49	12.68
Target 	86.12	66.02	55.93	16.18	87.07	75.41	22.70	52.59	39.41	79.49	58.82	77.75	69.63	14.79	37.61

Table L. **Benchmark results of platform adaptation** from  quadruped ( $\mathcal{P}^q$ ) to  drone ( $\mathcal{P}^d$ ). Target is trained with ground truth from the target domain. All scores are given in percentage (%). The **second best** and **best** scores under each metric are highlighted in colors.

Method	Acc	mAcc	mIoU	fIoU	ground	build	fence	person	pole	road	walk	veg	car	wall	sign
Source-Only 	52.62	29.38	16.85	15.45	50.85	15.47	1.65	2.24	15.48	36.88	9.98	35.84	15.20	1.50	0.23
AdaptSegNet [7]	57.07	33.15	20.96	16.49	31.15	24.78	2.71	0.08	19.90	58.22	4.43	53.49	20.42	15.41	0.00
DACS [6]	60.74	38.60	24.50	17.92	32.17	26.42	3.56	2.01	23.57	60.32	11.57	56.01	29.39	24.50	0.00
MIC [4]	64.49	40.02	26.11	18.65	40.50	29.26	0.70	3.02	20.52	62.66	21.37	57.58	36.20	15.36	0.00
PLSR [8]	63.57	42.62	27.34	18.08	40.71	26.42	0.42	3.39	24.07	62.16	18.07	57.80	29.16	38.50	0.00
<b>EventFly (Ours)</b>	65.78	41.91	28.79	19.01	40.74	30.90	1.50	2.63	24.76	64.11	18.22	61.85	33.44	38.23	0.29
Target 	79.57	52.25	42.90	23.30	74.48	39.40	7.10	0.33	31.67	71.96	31.64	67.87	57.51	66.14	23.79

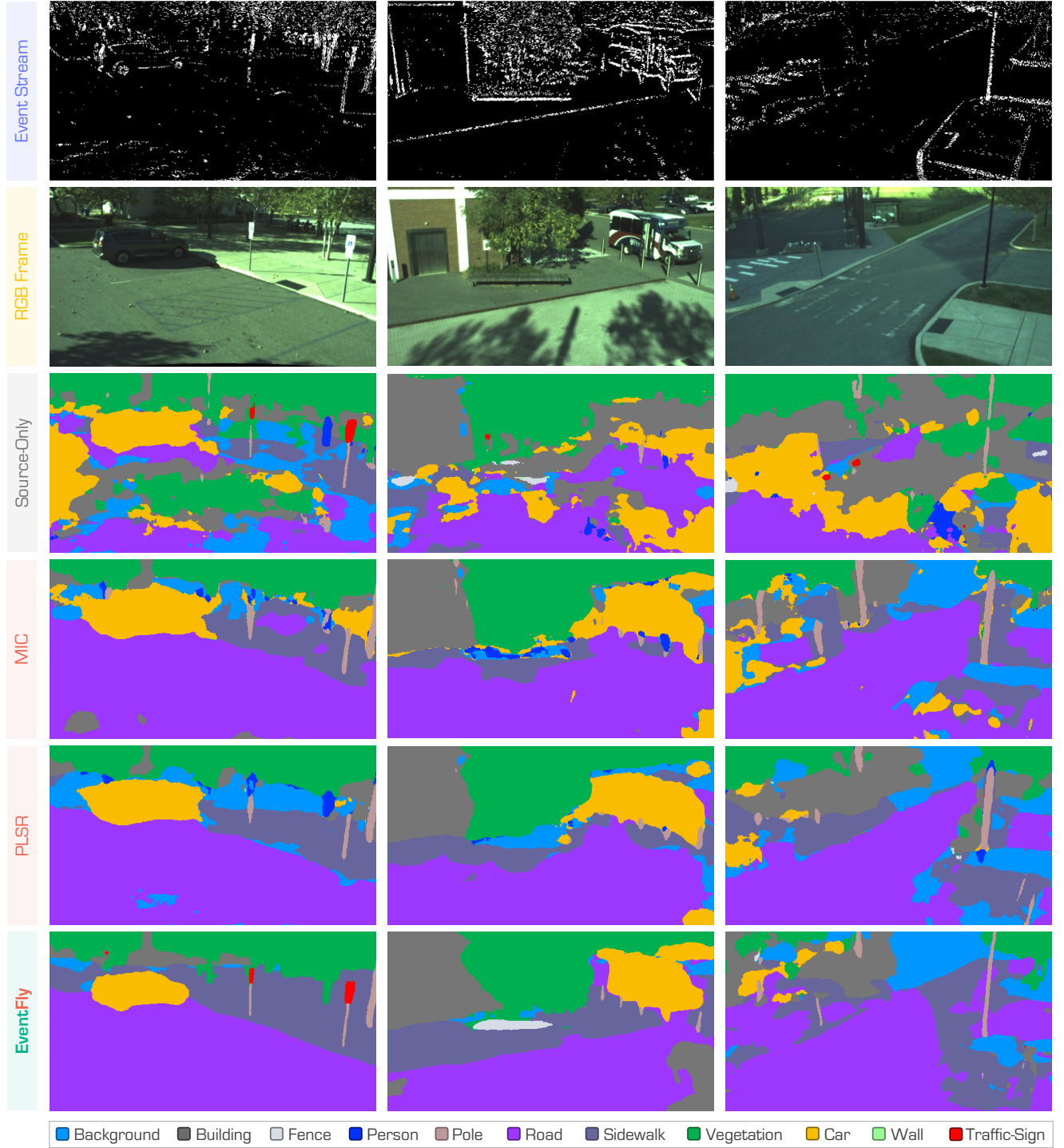


Figure A. **Additional qualitative assessments** of cross-platform adaptation from the  $\mathcal{P}^v$  vehicle ( $\mathcal{P}^v$ ) platform to the  $\mathcal{P}^d$  drone ( $\mathcal{P}^d$ ) platform. We use grayscale event images for better visibility. The RGB frames are for reference purposes only. Best viewed in colors.



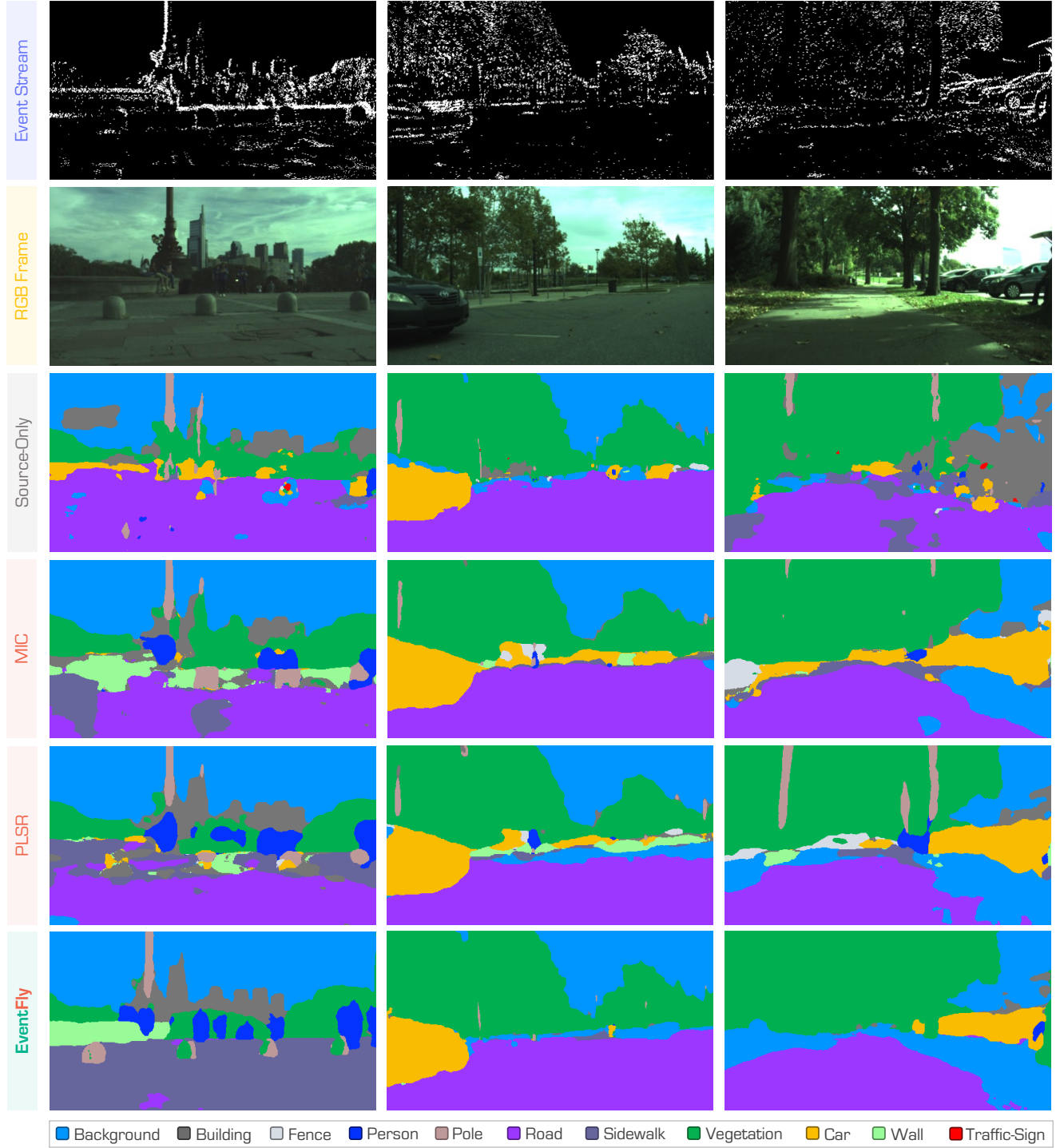




Figure B. **Additional qualitative assessments** of cross-platform adaptation from the  vehicle ( $\mathcal{P}^v$ ) platform to the  quadruped ( $\mathcal{P}^q$ ) platform. We use grayscale event images for better visibility. The RGB frames are for reference purposes only. Best viewed in colors.



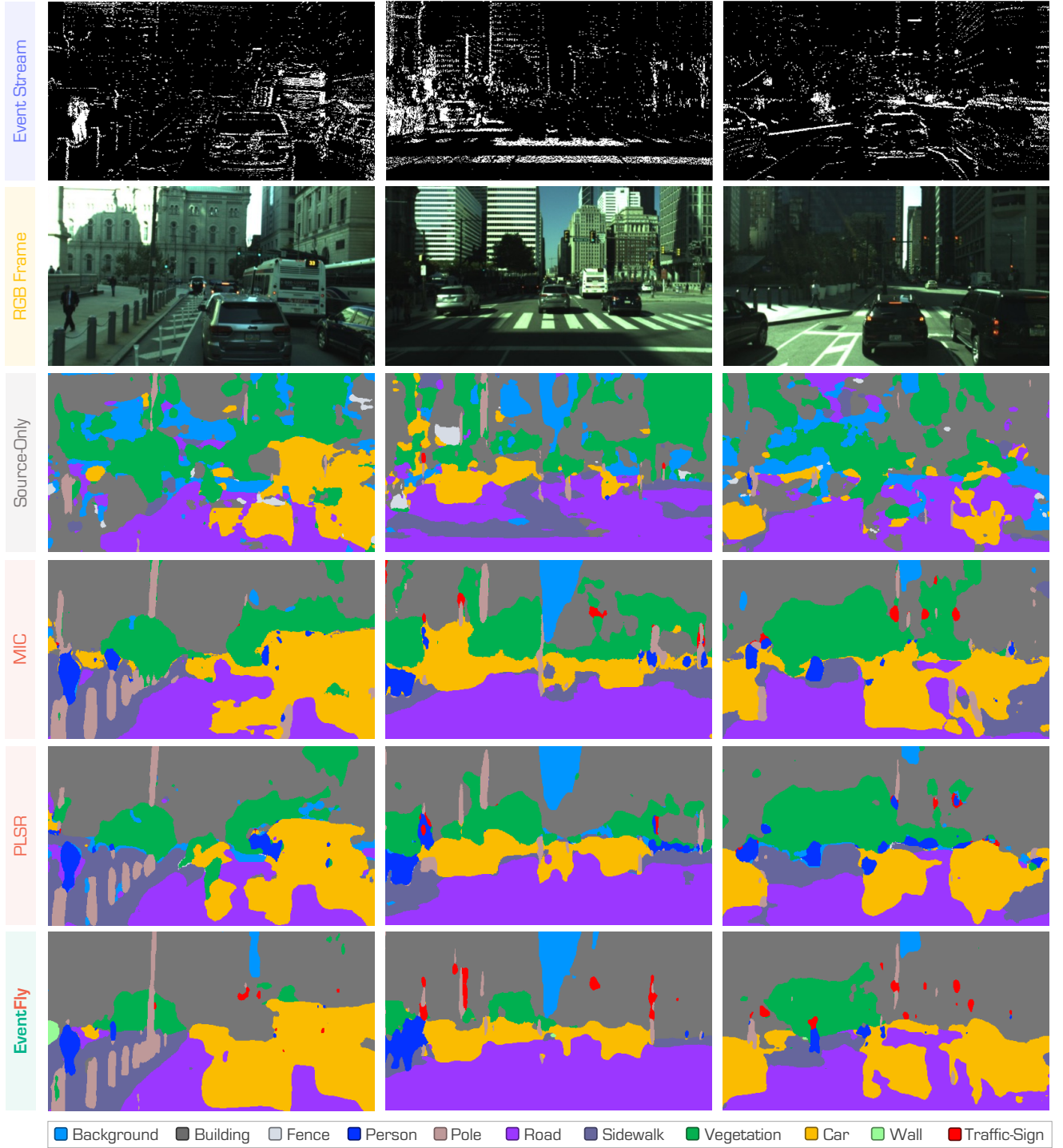


Figure C. **Additional qualitative assessments** of cross-platform adaptation from the drone ( $\mathcal{P}^d$ ) platform to the vehicle ( $\mathcal{P}^v$ ) platform. We use grayscale event images for better visibility. The RGB frames are for reference purposes only. Best viewed in colors.

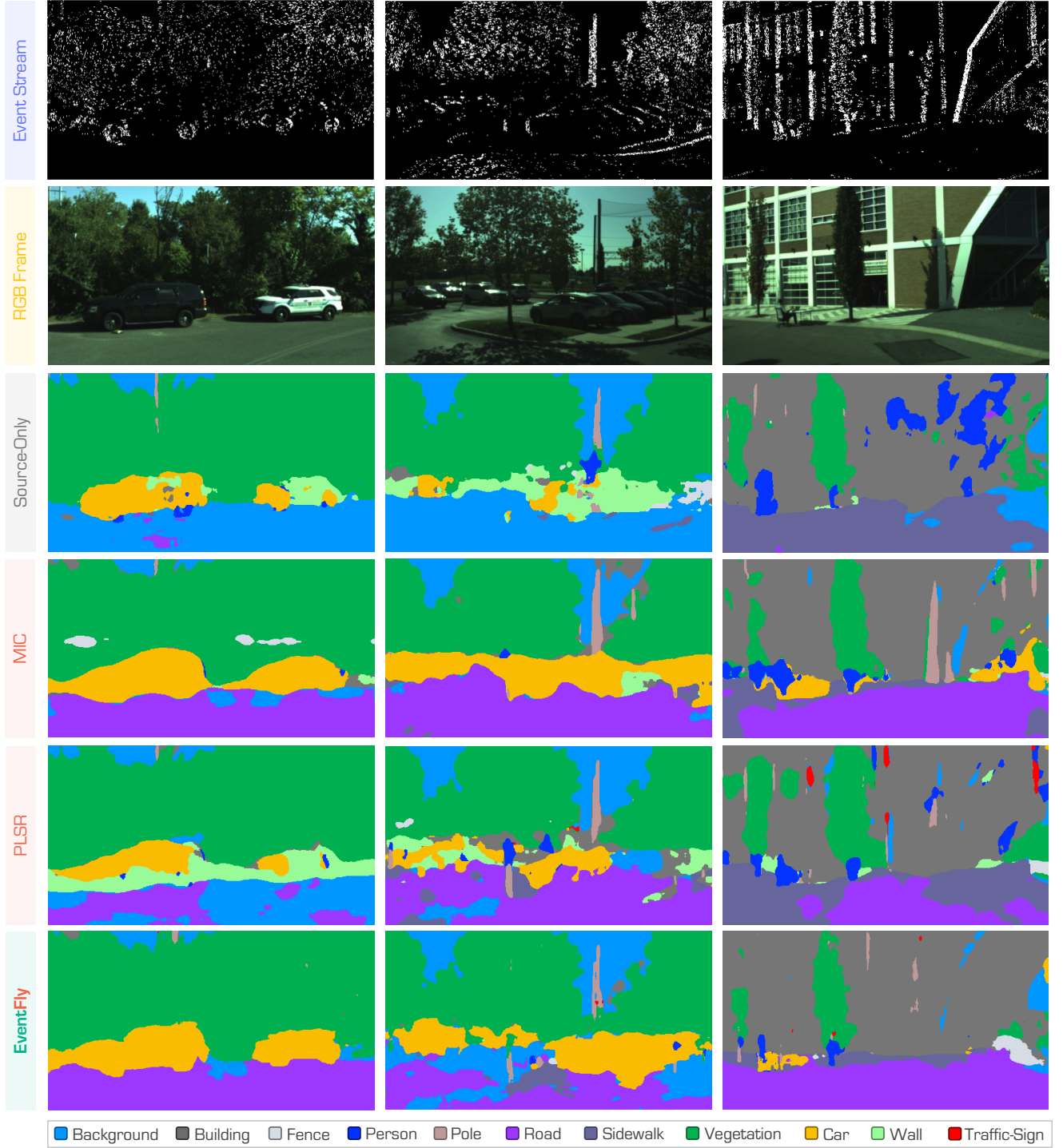


Figure D. **Additional qualitative assessments** of cross-platform adaptation from the  $\mathcal{P}^q$  quadruped ( $\mathcal{P}^q$ ) platform to the  $\mathcal{P}^v$  vehicle ( $\mathcal{P}^v$ ) platform. We use grayscale event images for better visibility. The RGB frames are for reference purposes only. Best viewed in colors.

## F. Public Resource Used

In this section, we acknowledge the use of the following public resources, during the course of this work:

- M3ED<sup>2</sup> ..... CC BY-SA 4.0
- ESS<sup>3</sup> ..... GNU General Public License v3.0
- E2VID<sup>4</sup> ..... GNU General Public License v3.0
- AdaptSegNet<sup>5</sup> ..... Unknown
- CBST<sup>6</sup> ..... CC BY-SA 4.0
- IntraDA<sup>7</sup> ..... MIT License
- DACS<sup>8</sup> ..... MIT License
- MIC<sup>9</sup> ..... Unknown
- Pytorch<sup>10</sup> ..... Pytorch License
- Pytorch3D<sup>11</sup> ..... BSD-Style License
- Open3D<sup>12</sup> ..... MIT license

## References

- [1] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4016–4023, 2023. 1
- [2] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Poo-ria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, et al. 5.10 a 1280× 720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86  $\mu\text{m}$  pixels, 1.066 gepps readout, programmable event-rate controller and compressive data-formatting pipeline. In *IEEE International Solid-State Circuits Conference*, pages 112–114, 2020. 1
- [3] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2004. 6
- [4] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 10, 12
- [5] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357, 2022. 2, 3
- [6] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 12
- [7] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 12
- [8] Xingchen Zhao, Niluthpol Chowdhury Mithun, Abhinav Ravanshi, Han-Pang Chiu, and Supun Samarasekera. Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2399–2409, 2024. 10, 12

---

<sup>2</sup><https://m3ed.io>.

<sup>3</sup><https://github.com/uzh-rpg/ess>.

<sup>4</sup>[https://github.com/uzh-rpg/rpg\\_e2vid](https://github.com/uzh-rpg/rpg_e2vid).

<sup>5</sup><https://github.com/wasidennis/AdaptSegNet>.

<sup>6</sup><https://github.com/yzou2/CBST>.

<sup>7</sup><https://github.com/feipanir/IntraDA>.

<sup>8</sup><https://github.com/vikolss/DACS>.

<sup>9</sup><https://github.com/lhoyer/MIC>.

<sup>10</sup><https://github.com/pytorch/pytorch>.

<sup>11</sup><https://github.com/facebookresearch/pytorch3d>.

<sup>12</sup><https://github.com/is1-org/Open3D>.