Journal of Computer and Knowledge Engineering, Vol. , No.. 2023.
DOI:

On the Challenges of Unsupervised Domain Adaptation for Object Detection in Event Streams

Ahmed Al-Fatlawi[1], Abedin Vahedian[2], and Ahad Harati[3]

Abstract—Event cameras asynchronously record local brightness changes, offering high temporal resolution, wide dynamic range, and robustness to motion blur, but their sparse event-based output differs fundamentally from intensity images used to train most object detectors. In addition, event-camera datasets with object-level annotations remain severely limited relative to the large labeled image datasets available for conventional vision. Together, this cross-modal mismatch and the limited availability of labeled target data make direct retraining less practical and motivate mechanisms for transferring knowledge from image-trained detectors to event data without target-domain detection labels. This study compares two practical transfer routes for a YOLOv5s detector: adapting the detector to event-derived input, the approach taken by ConfMix and SF-YOLO, or relinquishing the adaptation altogether by translating event streams into an intensity-like representation and applying the original pretrained detector. ConfMix retains labeled source supervision and adapts through confidence-guided source–target mixing, while SF-YOLO is a source-free teacher–student method driven by target-domain pseudo-labels. These adaptation strategies are evaluated against E2VID, a pretrained recurrent model that reconstructs events into intensity images. The Cityscapes dataset provides labeled conventional images for the source domain, whereas DSEC provides real automotive event-camera streams for the target domain. Overall, the results show that image-to-event transfer depends jointly on detector-facing representation, temporal support, and the adaptation mechanism. No single route is preferable under all conditions: reconstruction becomes the stronger option only when the temporal window is long enough to support it, and the ordering reverses in favour of detector adaptation at shorter windows and on an independent event dataset. Selecting a transfer route therefore requires matching it to the target data and to the temporal characteristics of the event representation, motivating future UDA strategies that address cross-modal discrepancy more explicitly at both input and feature levels.

***Index Terms*-- Event cameras; Object detection; Unsupervised domain adaptation; Source-free domain adaptation; Cross-modal transfer; Event-to-Video Reconstruction; DSEC; YOLOv5**

1- INTRODUCTION

Object detection is a fundamental component of visual perception systems, particularly in applications such as autonomous driving, mobile robotics, and intelligent transportation. Most modern detectors, including the YOLO family, are developed and pretrained using conventional intensity images, where visual information is represented as dense spatial arrays with familiar brightness, texture, and channel statistics. Event cameras operate according to a fundamentally different sensing principle. Rather than recording complete frames at fixed intervals, each pixel responds asynchronously to local brightness changes and generates events encoding spatial location, timestamp, and polarity. This sensing mechanism provides high temporal resolution, high dynamic range, and reduced sensitivity to motion blur, making event cameras attractive for visual perception under rapid motion and challenging illumination [1].

These advantages are accompanied by a substantial representation mismatch. An object detector trained on intensity images cannot operate effectively when its input is replaced by an event-derived representation. Before obtaining a standard image detector, asynchronous events must be accumulated or otherwise transformed into a dense tensor; the resulting intensity, temporal, polarity, and channel statistics can remain markedly different from those encountered during image-domain pretraining. Training an event-specific detector can address this mismatch when sufficiently annotated event data are available. However, event datasets with object-level annotations remain severely limited, and obtaining large-scale target-domain detection annotations reduces the practical benefit of reusing an existing image detector. The problem considered in this work is therefore how to transfer an image-trained object detector to event-camera data without using target-domain detection labels for adaptation.

Despite progress in event-based perception, domain-adaptive object detection, and event-to-intensity reconstruction, a practical question remains: when an image-trained detector is already available, how effectively can its knowledge be transferred across the image-to-event modality gap, and what can be learned from adapting the detector compared with reconstructing a more familiar input representation? This question cannot be answered reliably by directly comparing values reported in separate publications. Differences in source-model initialization, target data, class spaces, confidence thresholds, checkpoint-selection rules, and evaluation procedures can produce substantial variation independently of the underlying transfer strategy. A meaningful comparison therefore requires a common starting detector and a unified target-domain evaluation protocol.

To investigate these questions, we conduct a controlled Cityscapes-to-DSEC study centered on a common Cityscapes-trained YOLOv5s detector. Cityscapes provides labeled conventional urban imagery for the source domain, whereas DSEC provides real automotive event-camera observations for the target domain. Two practical transfer routes are compared. The first changes the detector through unsupervised domain adaptation. ConfMix [2] represents a source-accessible setting in which labeled source supervision is retained while confident

[1] Phd Candidate, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

[2] Corresponding author. Associate Professor, Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran. Email: vahedian@um.ac.ir

[3] Associate Professor, Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

target regions are introduced through source–target mixing. SF-YOLO [3] represents the stricter source-free setting, where adaptation proceeds from a pretrained detector using a teacher–student mechanism and pseudo-labels obtained from unlabeled target data. The second route changes the input rather than the detector. E2VID [4] is a pretrained recurrent event-to-intensity reconstruction model that transforms event streams into an intensity-like representation that can subsequently be processed by the unchanged image detector.

Detector adaptation attempts to make the learned model more suitable for event-derived target inputs, whereas reconstruction attempts to make the target observations more compatible with the image statistics already learned by the detector. The comparison made in this study is therefore functional rather than component- or prior-matched. Direct transfer, a continued source-only condition, and a supervised target model provide reference conditions.

Two additional factors are particularly important in this cross-modal setting. First, adaptation performance may vary substantially over the course of training. An adaptation method can reach an improved target-domain state and later lose part of that gain, making the ability to preserve adaptation performance a distinct issue from the ability to reach it. For this reason, we use predefined reporting epochs rather than target annotations to select an unsupervised checkpoint. Second, event-based transfer depends on the temporal interval over which asynchronous events are accumulated. Changing this interval alters the amount of motion evidence available in each detector-facing representation and, for recurrent reconstruction, also changes the temporal continuity of the event stream. Temporal support should therefore be treated as part of the image-to-event transfer problem rather than merely as an implementation parameter.

The experiments yield three principal findings. Under the canonical 50 ms protocol, reconstruction provides the strongest label-free result among the evaluated routes, raising mAP@0.5 from 0.0340 for direct transfer to 0.1787, ahead of 0.1030 for ConfMix and 0.0456 for SF-YOLO. This advantage is not general: it collapses at shorter temporal windows and reverses on the independent Gen1 dataset, where direct transfer exceeds reconstruction, and the same dataset-dependent relationship between direct event transfer and reconstruction is reproduced with an architecturally distinct second detector. Finally, both adaptation methods reach substantially stronger intermediate states than they retain at the predefined final epoch, indicating that the difficulty lies less in adapting an image-trained detector to event data than in preserving what adaptation achieves. Taken together, these findings indicate that the transfer route must be matched to the target data and to the temporal support available, rather than selected once and applied uniformly.

The main contributions of this work are threefold:

- We establish a controlled image-to-event transfer framework for comparing two practical strategies for reusing an image-trained detector: adapting the detector to event-derived observations and reconstructing the event stream into an intensity-like input. The comparison uses a common YOLOv5s source initialization, common target data, and a unified evaluation protocol, while direct transfer, source-only continuation, and supervised target training provide complementary reference conditions.
- We show that the severe image-to-event domain gap does not imply an absence of adaptability. Both evaluated UDA methods reach target-domain states above direct transfer during training, while their trajectories reveal that these states are not necessarily preserved throughout continued unsupervised optimization. This distinction separates the ability to adapt from the additional problem of maintaining a useful adapted state without target-label-based checkpoint selection.
- We investigate how detector-facing event representation and temporal support influence image-to-event transfer. Direct transfer and recurrent event-to-intensity reconstruction are evaluated across 5, 10, 20, and 50 ms, while ConfMix and SF-YOLO are re-adapted at each duration across three independent random seeds, so that every temporal condition is reported as a mean and sample standard deviation rather than as a single run. The experiments show that temporal support does not affect the evaluated transfer routes uniformly.

The remainder of this paper is organized as follows. Section II reviews event representations, domain-adaptive detection, and cross-modal reconstruction. Section III describes the experimental design and evaluation protocol, Section IV presents the results, Section V discusses their implications, Section VI states the limitations of the study, and Section VII concludes.

2- RELATED WORK

The problem addressed in this study lies at the intersection of two research directions: domain-adaptive object detection and the processing of event-camera data using models originally developed for conventional images. The relevant literature can be organized around several complementary routes for transferring knowledge from image-domain detectors to event data, including detector-facing event representations, self-training with pseudo-labels, adversarial feature alignment, cross-modal reconstruction and distillation, and more recent alignment with pretrained foundation models.

Event representations for image-trained detectors. Event cameras report asynchronous per-pixel brightness changes rather than synchronous intensity frames, providing high temporal resolution and wide dynamic range [1]. Because conventional object detectors expect dense tensor inputs, event streams must first be accumulated or transformed into a detector-facing representation. Common choices include event-count images, polarity-separated histograms, and voxel grids that preserve coarse temporal structure [5]. A substantial body of recent work instead develops detectors specifically for event data, including recurrent, transformer-based, spiking, and asynchronous architectures. That line is complementary to the present study because it generally assumes event-domain supervision during training. Here, we instead begin with an existing image-pretrained detector and examine how effectively its knowledge can be reused under an unlabeled image-to-event domain shift. The DSEC dataset [6] and its detection annotations [7] provide the automotive event benchmark used

throughout this study.

Unsupervised detector adaptation. Domain-adaptive object detection seeks to reduce the discrepancy between a labeled source domain and an unlabeled target domain without using target annotations. One family aligns features so that a domain discriminator cannot distinguish them, with multi-level variants operating at the pixel, instance, and category levels [10] and source-free variants aligning target subsets within a teacher-student framework [11]. Hybrid frameworks such as ALDI and ALDI++ [9] combine alignment with self-distillation, and their benchmarking shows that reported adaptation gains can shift substantially under stronger baselines and standardized implementations. Adversarial alignment is not included in our experimental comparison because representative methods are commonly formulated around detector-specific feature hierarchies and training pipelines that differ substantially from the common YOLOv5s setup adopted here [12], [13], [11]. Reimplementing these methods within the present detector would therefore introduce additional architecture- and implementation-dependent changes beyond the controlled comparison considered in this study.

The second family, self-training with pseudo-labels, is the setting examined here. A detector trained on the labeled source domain generates predictions on unlabeled target data, and sufficiently confident predictions are then used as supervision during continued adaptation. Mean-teacher formulations, in which a temporally averaged teacher supervises a student on unlabeled target data, were introduced for cross-domain detection by Cai et al. [8]. The two methods evaluated in this study represent distinct settings within this family with respect to source access. ConfMix [2] identifies target regions in which the current detector is most confident and mixes them into labeled source images, while labeled source supervision remains active throughout adaptation. SF-YOLO [3] instead operates source-free, without access to the original source images. It uses a teacher-student architecture driven by pseudo-labels on unlabeled target data, together with a target-specific style augmentation module and stabilization mechanisms intended to reduce the instability of source-free self-training.

This formulation makes SF-YOLO particularly relevant to image-to-event transfer, where the initial target-domain predictions may be weak and the teacher signal can evolve substantially during adaptation. Our experiments therefore examine not only whether ConfMix and SF-YOLO improve over direct transfer, but also whether the resulting adapted states persist throughout training under a severe image-to-event modality shift. The resulting conclusions are restricted to the evaluated methods and should not be generalized to all UDA families.

Event-to-intensity reconstruction. An alternative strategy is to reduce the modality mismatch before detection rather than adapting the detector itself. Event-to-intensity reconstruction provides such a cross-modal bridge by converting asynchronous events into an intensity-like representation that can be processed by a conventional image detector. E2VID [4], used in this study as the representative reconstruction route, is a pretrained recurrent model that reconstructs intensity video from event streams, after which the original image-trained detector is applied without target-domain detector optimization. Instead of modifying the detector to accommodate event-derived inputs, reconstruction modifies the input so that it more closely resembles the modality on which the detector was originally trained. The resulting advantage can therefore depend strongly on the temporal evidence available to the reconstruction model, especially when its recurrent state is expected to integrate information across successive event windows.

Several further lines of work bridge the image and event modalities without being evaluated here. Reconstruction has been used to generate surrogate intensity observations for source-free adaptation in object recognition [14]; event voxel grids have been converted into pseudo-image and pseudo-flow modalities processed by frozen RGB-trained encoders for motion segmentation [15]; and soft attention-based alignment and distillation have been applied to cross-modal semantic segmentation [16]. Distillation offers a further route, transferring knowledge from an RGB teacher to an event student on DSEC-DET [17], routing vision-language supervision into event streams for open-vocabulary detection [18], and interpolating RGB and event inputs during spiking-network training [19]. A more recent direction aligns event representations with the latent spaces of large pretrained image models, either by mapping events into a foundation-model latent space [20], by distilling transferable event representations [21], or by aligning an event encoder with a frozen vision model before autoregressive pretraining [22].

These approaches differ from the reconstruction experiment considered here because they train or adapt a target-domain model. In our reconstruction route, no model is optimized on the target-domain detection data. Instead, the externally pretrained E2VID network is placed in front of the unchanged source detector, allowing the effect of detector-facing input compatibility to be separated from target-domain detector optimization. The latent-alignment approaches differ in one further respect that is relevant to our results: the supervisory signal in these approaches does not originate primarily from confident detections produced by the source detector on unlabeled target samples. Instead, it is obtained from pretrained latent spaces or intermediate teacher representations. This distinction is relevant to the behavior observed in our experiments and suggests a promising direction for future image-to-event transfer methods that do not rely exclusively on self-generated detector predictions.

Research gap and scope of existing comparisons. The most closely related study to the central question addressed here is that of Hannan et al. [23], who investigate event-to-video conversion for overhead object detection. They report a substantial performance gap between detectors operating on dense event representations and those operating on the corresponding intensity frames. They attribute part of this difference to incompatibility between event-derived inputs and the data distributions used to pretrain conventional image detectors, and reduce the gap by converting the event stream into grayscale video. Their setting differs from ours in four respects: it is fully supervised, with detectors trained on annotated event data; their events are simulated from video rather than recorded; their scenes are captured from an overhead rather than an automotive perspective; and their comparison includes neither published domain-adaptation methods nor an analysis of how adaptation evolves during training.

A terminological clarification is also warranted. Mechler and Rojtberg [24] investigate the use of dense detection models on event data by replacing dense convolutions with sparse and asynchronous operations. In that work, "transferring" refers to porting the model architecture from dense to sparse computation rather than transferring knowledge from a labeled image domain to an unlabeled event domain. Their work is therefore relevant to the broader problem of applying image-derived detectors to event data, but it does not constitute domain adaptation under the setting examined in this study.

Taken together, prior work provides strong evidence for event-native detection, domain-adaptive detection, and event-to-intensity reconstruction as separate research directions. What remains less clearly characterized is how these routes behave when the same image-trained detector is transferred to event data under a common detector initialization, common target data, and a unified evaluation protocol, particularly with respect to the detector-facing representation, the persistence of adaptation, and the temporal support of the event input. This study is designed to examine these factors jointly rather than to propose a new event-specific detector architecture, and to use the resulting evidence to identify directions for more explicit cross-modal adaptation in future work.

3- EXPERIMENTAL DESIGN AND EVALUATION PROTOCOL

A. Data and Problem Setting

We study the transfer of an object detector trained on conventional intensity images to event-camera data without using target-domain detection annotations for unsupervised detector adaptation. The source domain is labeled, whereas the target data used by the adaptation methods are treated as unlabeled. Target annotations are withheld from all unsupervised optimization procedures and are used only for standardized evaluation and explicitly identified diagnostic analyses.

The labeled source domain is Cityscapes [25], a large-scale urban-driving dataset acquired with conventional frame cameras. All experiments that require an image-domain detector begin from the same YOLOv5s source model [26], initialized from the publicly released Cityscapes checkpoint distributed with ConfMix [2]. For methods that retain access to source data during adaptation, the source set contains 2968 labeled Cityscapes images. SF-YOLO, by contrast, operates in a source-free setting and receives only the pretrained source detector and unlabeled target-domain observations.

The target domain is derived from DSEC [6], a real automotive dataset recorded with event cameras. Object-detection annotations are taken from its DSEC-Detection (DSEC-Det) extension [7].[4] The annotations are defined in the distorted event-camera coordinate system; event inputs are therefore constructed in the same native geometry and are not geometrically rectified.

We use the six sequences of zurich_city_04 recording family. Sequences zurich_city_04_b, zurich_city_04_c, zurich_city_04_d, zurich_city_04_e, and zurich_city_04_f form the unlabeled target adaptation set and provide 3528 event windows. The complete zurich_city_04_a sequence is reserved for evaluation and contains 695 annotated samples with 6191 ground-truth bounding boxes.

The split is performed at the sequence level rather than by randomly partitioning temporally adjacent samples. This reduces leakage caused by the strong temporal correlation between neighboring observations from the same recording. The selected sequence family provides a controlled within-city setting, although it does not represent the full geographic and scene diversity of DSEC; this restriction is therefore treated as a limitation of the study.

The source checkpoint predicts eight categories in the following order: *person*, *car*, *train*, *rider*, *truck*, *motorcycle*, *bicycle*, and *bus*. The class order is read directly from the checkpoint rather than inferred from dataset conventions. For the controlled target-domain evaluation, only the *person* and *car* categories are retained. Predictions from the remaining source categories are not remapped to these classes. Consequently, an object outside the adopted two-class label space that is predicted as *person* or *car* cannot be matched to a corresponding ground-truth category and is counted as a false positive. This evaluation rule is applied identically to all methods and can affect the absolute precision measured on the target domain.

Three considerations motivate this two-class evaluation space. First, it is the only label space common to all evaluation sets used in this study: the Prophesee Gen1 dataset employed for cross-dataset generalization provides annotations for the car and pedestrian categories only, so a broader space would prevent the primary and generalization experiments from being reported on commensurable terms. Second, the two categories carry by far the strongest annotation support in DSEC-Detection. Across all available annotations, cars account for 71.16% and pedestrians for 19.32% of the labeled objects, so the two categories together constitute 90.48% of the annotated instances, and the same imbalance holds in the held-out sequence used here. Third, the DSEC-Detection annotations were generated by tracking on the rectified image frames followed by manual removal of false tracks [7]; this procedure suppresses spurious detections but does not systematically recover missed ones, so annotation completeness is likely to be least reliable for the sparsely represented categories. The restriction is therefore driven by evaluation comparability and annotation support rather than by semantic incompatibility between the source and target taxonomies, which correspond closely.

B. Event Data Construction and Representations

An event camera does not acquire complete images at a fixed frame rate. Instead, each pixel independently reports a brightness change once the log-intensity variation exceeds the sensor contrast threshold. An event in the raw DSEC stream can be written as

$$e_i = (x_i, y_i, t_i, p_i), \tag{1}$$

where (x_i, y_i) is the pixel location, t_i is the timestamp, and $p_i \in \{0,1\}$ is the polarity value stored in the DSEC data. In the

[4] The DSEC-Detection annotations and associated preparation tools are distributed through the official DSEC-Det repository: https://github.com/uzh-rpg/dsec-det.

conventional physical notation used in the event-camera literature, the two polarities are commonly represented as $\{-1,+1\}$; accordingly, when a signed representation is required, we convert the stored binary polarity as

$$s_i = 2p_i - 1 \in \{-1,+1\}. \quad (2)$$

This conversion is used only by representations that explicitly preserve polarity. The count-based representations described below instead ignore the polarity value and accumulate all events at each pixel.

For every annotated DSEC-Det frame used in the study, we collect the events within a canonical 50-ms temporal window aligned with the corresponding annotation timestamp. Unless a different duration is stated explicitly, every result reported in this paper uses this canonical window. Exactly the same construction is used for the five target-training sequences and for the held-out evaluation sequence, so the 3528 target-training samples described in Section III.A correspond to 3528 event windows, while the evaluation sequence provides 695 windows. Holding this temporal support fixed ensures that the representation study varies only the event encoding; the separate effect of the window duration itself is defined in Section III.B.3 and reported in Section IV.D.

All event representations are generated in the native 640×480 event-camera geometry. The events are not geometrically rectified before rendering, because the DSEC-Det bounding boxes used in this study are defined in the distorted event-camera view; rectifying the event coordinates without transforming the annotations would shift the event structures relative to their boxes [7].

A central distinction in the following definitions is between the internal event representation and the detector-facing input. Some encodings internally contain 10 or 20 temporal or polarity channels, but the pretrained YOLOv5s model is not modified to accept them: every direct-transfer representation is rendered as a three-channel image before inference. Figure 1 shows the four detector-facing representations for the same 50-ms event window, with the corresponding held-out annotations.

1) Primary Signed Voxel Representation: The primary detector-facing representation used for direct transfer, ConfMix, and SF-YOLO preserves event polarity while aggregating the complete temporal support of each event window. Internally, the interval is divided into ten temporal bins, and each signed event contributes to its neighboring bins through linear temporal interpolation, producing bin responses $V_k(x,y)$ for $k = 0,\dots,9$.

Rather than selecting individual temporal bins as detector channels, all ten signed bins are aggregated:

$$S(x,y) = \sum_{k=0}^{9} V_k(x,y). \quad (3)$$

This produces a single signed activity map that incorporates the complete event support of the window. Because the interpolation weights of each event sum to unity, the aggregation is equivalent to accumulating the net signed event count at every pixel; the bin construction therefore reflects the implementation pipeline rather than adding temporal information to the detector-facing input. To reduce sensitivity to isolated high-activity pixels, the nonzero absolute values of S are robustly normalized using their 99th percentile,

$$q_{0.99}^{S} = P_{99}(\{|S(x,y)| : S(x,y) \neq 0\}), \quad (4)$$

giving the normalized signed response

$$\bar{S}(x,y) = \mathrm{clip}\left(\frac{S(x,y)}{q_{0.99}^{S}}, -1, 1\right), \quad (5)$$

which is mapped to the 8-bit intensity range as

$$I_S(x,y) = 127\,\bar{S}(x,y) + 128. \quad (6)$$

Locations with zero event activity therefore map to approximately mid-gray, while positive and negative accumulated activity appear on opposite sides of this baseline. Because the pretrained YOLOv5s detector expects a three-channel input, the same signed map is replicated across all three channels,

$$I_{\mathrm{voxel}} = \mathrm{Repeat}_3(I_S) \in \mathbb{R}^{3\times H\times W}. \quad (7)$$

The detector therefore receives a three-channel rendering obtained from the complete signed event support of the temporal window.

2) Alternative Detector-Facing Representations: To determine whether direct image-to-event transfer depends on the rendering used to expose event information to the pretrained detector, we additionally evaluate three deterministic alternatives: an event-count image, an inverted event-count image, and a stacked polarity–recency histogram. These representations use the same temporal window, native geometry, detector checkpoint, and evaluation protocol as the primary signed voxel representation. They are used only in the representation study and do not replace the signed voxel representation used in the main adaptation experiments.

Event-count representation. The count representation deliberately removes polarity and explicit within-window timing. For each pixel, all events within the 50-ms interval are accumulated,

$$C(x,y) = \sum_{e_i \in \mathcal{E}} \mathbf{1}\left[(x_i,y_i) = (x,y)\right], \quad (8)$$

where $\mathcal{E}$ denotes the events belonging to the current window. The nonzero count values are normalized using their 99th percentile,

$$\bar{C}(x,y) = 255\,\mathrm{clip}\left(\frac{C(x,y)}{q_{0.99}^{C}}, 0, 1\right), \quad (9)$$

where $q_{0.99}^{C}$ denotes the corresponding percentile. As in the voxel representation, percentile clipping limits the influence of isolated high-activity pixels on the available intensity range. The resulting single-channel image is then replicated across the three detector input channels,

$$I_{\mathrm{count}} = \mathrm{Repeat}_3(\bar{C}) \in \mathbb{R}^{3\times H\times W}. \quad (10)$$

This encoding retains the spatial distribution of event activity but discards both event polarity and temporal ordering within the window.

Inverted-count representation. The inverted-count representation contains exactly the same event information as the ordinary count rendering. Only its brightness convention is reversed,

$$I_{\mathrm{inv}} = 255 - I_{\mathrm{count}}. \quad (11)$$

The inversion is applied identically to all three replicated channels. Consequently, event-active regions that are bright in the ordinary count representation become dark, while inactive regions become bright. This variant therefore isolates the effect of the image-intensity convention without adding temporal or polarity information.

Stacked event histogram. The stacked histogram preserves both polarity and coarse temporal structure internally. The 50-

ms interval is divided into ten temporal bins, and positive and negative events are accumulated separately, yielding an internal polarity–time tensor

$$\mathcal{H} \in \mathbb{R}^{20\times H\times W}, \tag{12}$$

corresponding to ten temporal bins for each polarity. This 20-channel tensor is not passed directly to YOLOv5. Instead, it is compressed into a three-channel color rendering.

For each polarity $p \in \{+,-\}$, the activity accumulated over the ten temporal bins is first computed as

$$\mathcal{A}_p(x,y) = \sum_{k=1}^{10} \mathcal{H}_{p,k}\,(x,y), \tag{13}$$

where $\mathcal{H}_{pk}(x,y)$ denotes the event count at pixel (x,y) for polarity p and temporal bin k. The two accumulated polarity maps are normalized independently using the 99th percentile of their nonzero values,

$$\bar{\mathcal{A}}_p(x,y) = 255\ \mathrm{clip}\left(\frac{\mathcal{A}_p(x,y)}{q_{0.99}^{\mathcal{A}_p}}, 0, 1\right), \tag{14}$$

where $q_{0.99}^{\mathcal{A}_p}$ denotes the corresponding percentile. As in the other renderings, percentile clipping prevents isolated high-activity pixels from dominating the available intensity range.

Temporal recency is encoded using the timestamp of the most recent event at each pixel. Let $t_{\mathrm{last}}(x,y)$ denote the latest event timestamp at pixel (x,y) within a window beginning at t_0 and ending at t_1. The normalized recency map is

$$R(x,y) = \begin{cases} \dfrac{t_{\mathrm{last}}(x,y) - t_0}{t_1 - t_0}, & \text{if at least one event occurs at } (x,y) \\ 0, & \text{otherwise,} \end{cases}, \tag{15}$$

with $R(x,y) \in [0,1]$. Its 8-bit form is

$$\bar{R}(x,y) = 255\,R(x,y). \tag{16}$$

Semantically, the final visualization assigns positive-polarity activity to red, negative-polarity activity to green, and recency to blue. Thus, in color-semantic notation,

$$I_{\mathrm{hist}}^{\mathrm{RGB}}(x,y) = [\bar{\mathcal{A}}_+(x,y), \bar{\mathcal{A}}_-(x,y), \bar{R}(x,y)]. \tag{17}$$

This notation describes the intended color semantics rather than the in-memory channel order. In the OpenCV implementation, images are stored in BGR order, so the corresponding array is written with the blue recency channel first and the red positive-polarity channel last before entering the common image-loading pipeline.

The detector-facing rendering therefore remains a three-channel image even though it is constructed from an internal 20-bin polarity-temporal representation: the bins are collapsed into two polarity-specific activity maps and a single recency channel, so the ten-bin temporal structure is not preserved explicitly.

This distinction is important for interpreting the later representation study. The stacked histogram is constructed from a comparatively rich polarity-temporal representation, but much of that structure is compressed before inference into an artificial three-channel color code whose semantics differ substantially from the RGB images used to pretrain the detector. The number of channels in the intermediate representation should therefore not be interpreted as the amount of temporal information available to the detector, and internal representation structure and compatibility with image-domain pretraining are treated as separate properties.

Figure 1 also shows that the four encodings differ in the image statistics presented to the pretrained detector, while the identical box locations confirm that they differ through event rendering rather than spatial geometry. No rendering is assumed in advance to transfer better; the effect is measured in Section IV.

These four inputs are deterministic renderings of a fixed event window and do not attempt to infer the latent intensity image that generated the events. Event-to-intensity reconstruction is therefore a different transfer route, described in Section III.C.

Finally, no representation is spatially downsampled during event generation; all are produced at the native 640×480 event resolution. The common detector pipeline subsequently uses an input-size setting of 640, as detailed in Section III.D. By holding the temporal window, spatial geometry, source checkpoint, label set, and downstream detector fixed, the representation experiment isolates the effect of how the same event stream is rendered for an image-pretrained detector.

3) Temporal-Window Sensitivity: The canonical experiments use a 50-ms event window because the DSEC detection annotations are spaced at approximately 20 Hz, making successive 50-ms windows nearly contiguous in time. To examine whether the conclusions depend on this temporal support, we additionally construct detector inputs using $\Delta t \in \{5,10,20,50\}$ ms while keeping the annotation timestamps, spatial geometry, detector initialization, and evaluation protocol unchanged. This analysis is treated as a post-hoc temporal-sensitivity experiment rather than as hyperparameter selection.

All four transfer conditions are evaluated across the four window durations. Direct transfer and E2VID require no detector optimization and are evaluated directly. For ConfMix and SF-YOLO, adaptation is independently repeated using three random seeds at each temporal window, with the same predefined 50-epoch training schedule and standardized final evaluation protocol. Accordingly, the temporal adaptation analysis comprises three independent runs for each method at each of the four evaluated durations. The primary manuscript results remain those obtained under the canonical 50-ms protocol, while the shorter-window runs are used only to characterize temporal sensitivity and are not used to select an alternative operating configuration.

Additional generalization evaluation. To examine whether the observations obtained in the primary zurich_city_04 setting extend beyond the original evaluation family, we perform two additional zero-shot evaluations using the finalized models. The first uses the previously unseen DSEC sequence thun_02_a, which contains 3857 annotated samples and 41,803 person/car ground-truth boxes. The second uses the complete labeled Gen1 evaluation set considered in this study, containing 30,605 samples and 66,304 ground-truth boxes.

No training, domain adaptation, hyperparameter tuning, or checkpoint selection is performed on either additional evaluation set. Direct transfer and E2VID use the same fixed source detector as in the primary experiment, whereas ConfMix and SF-YOLO use their fixed seed-0 epoch-50 checkpoints. The canonical 50-ms temporal support and the same standardized detector confidence and NMS settings are retained. For Gen1, the pedestrian and car categories are

mapped to the person and car evaluation classes, respectively.

Detector-architecture control. To examine whether the detector-facing observations depend on the YOLOv5s architecture, we additionally evaluate a Cascade R-CNN detector [27] with a ResNet-50-FPN backbone, implemented in MMDetection [28]. Cascade R-CNN is a two-stage detector with a sequence of progressively stricter IoU-trained detection heads, and therefore differs substantially in both detection paradigm and training objective from the one-stage detector used throughout the primary experiments. The Cityscapes-trained checkpoint, denoted cascade_rcnn_r50_city, is obtained from the official implementation released with the work of Fu et al. [29], where it is identified as the Cityscapes-pretrained detector used to initialize their Foggy-Cityscapes and Rainy-Cityscapes experiments. The class ordering of the released checkpoint was verified empirically on Cityscapes source images, as the ordering recorded in the checkpoint metadata was found to be inconsistent with its learned outputs; all reported results use the verified ordering. The detector is kept frozen throughout: no target-domain training, adaptation, hyperparameter tuning, or checkpoint selection is performed, and it receives exactly the same finalized detector-facing inputs and the same standardized evaluation settings as the primary detector.

To assess the contribution of cross-sample recurrent-state propagation to the E2VID temporal sensitivity, we additionally perform a diagnostic control in which the recurrent state is reinitialized before every annotation-aligned sample. All event windows, reconstruction settings, detector parameters, and evaluation procedures remain unchanged. The original continuous-state configuration is retained as the principal E2VID protocol, while the reset-state condition is used only to examine recurrent-state confounding.

C. Compared Transfer Strategies

The experiments compare several ways of reusing the same image-trained detector under the image-to-event domain shift. They differ in where transfer is performed and in the information available during it. Direct transfer and continued source-only training provide reference conditions without target-domain adaptation; ConfMix and SF-YOLO update the detector using unlabeled target event data but differ in whether labeled source images remain available; E2VID leaves the detector unchanged and reconstructs the event stream into an intensity-like representation before inference.

Unless otherwise stated, the target-domain input used for direct transfer, ConfMix, and SF-YOLO is the signed voxel representation defined in Section III.B. The alternative deterministic representations are examined only in the representation study.

Table I summarizes the information and learned components used by each strategy. Its final column refers specifically to an additional learned model that remains in the target inference pipeline; training-only components are described with the corresponding methods.

TABLE I
SUMMARY OF THE TRANSFER STRATEGIES EVALUATED IN THE CONTROLLED IMAGE-TO-EVENT COMPARISON. THE SUPERVISED TARGET MODEL IS INCLUDED ONLY AS A REFERENCE CONDITION. TARGET EVENTS PROCESSED BY E2VID ARE USED FOR RECONSTRUCTION AND INFERENCE, NOT FOR OPTIMIZATION OF EITHER E2VID OR THE SOURCE DETECTOR.

Strategy	Source images during transfer	Unlabeled target used for learning	Target labels used for learning	Detector weights updated	Additional inference model
Direct transfer (signed voxel)	No	No	No	No	No
Continued source-only control	Yes	No	No	Yes	No
ConfMix	Yes	Yes	No	Yes	No
SF-YOLO	No	Yes	No	Yes	No
E2VID + source detector	No	No	No	No	Yes
Supervised target reference	No	No	Yes	Yes	No

1) Direct Transfer and Source-Only Control: The simplest transfer condition leaves the source detector unchanged. The Cityscapes-trained YOLOv5s checkpoint is applied directly to the detector-facing event representations defined in Section III.B. No target samples are used for optimization, no pseudo-labels are generated, and no target-specific detector parameters are introduced. We refer to this condition as *direct transfer*. It measures how much of the detector's image-domain knowledge remains immediately usable when only the input modality changes.

The principal direct-transfer reference uses the signed voxel representation. The alternative event-count, inverted event-count, and stacked-histogram renderings are evaluated separately to examine the sensitivity of direct transfer to the detector-facing event representation.

Direct transfer alone cannot distinguish the effect of target-domain adaptation from the effect of simply continuing detector optimization. We therefore include a *continued source-only control* in which the same source checkpoint is trained for an additional 50 epochs using only labeled Cityscapes data. No target event samples are observed during this training.

Direct transfer measures zero-adaptation compatibility between the pretrained image detector and event-derived inputs, whereas continued source-only training isolates the effect of additional optimization without target-domain information.

2) Source-Accessible Adaptation: ConfMix: ConfMix [2] represents the source-accessible UDA setting evaluated in this study. During adaptation, labeled Cityscapes source images remain available together with unlabeled target-domain samples represented using the signed voxel representation. The method introduces target-domain appearance into training through confidence-guided source–target region mixing while preserving ordinary supervised source detection.

For a labeled source image x_S and an unlabeled target image x_T, the detector first produces candidate target detections. The target image is evenly divided into four spatial regions. A predicted detection is assigned to a region when the center

coordinate of its bounding box lies inside that region. The confidence of each region is computed as the average confidence of the detections assigned to it, and the region with the highest average confidence is selected for mixing with the source image.

A central component of ConfMix is that pseudo-detection reliability is not estimated from the ordinary detector confidence alone. The method adopts a Gaussian representation of the four bounding-box coordinates. Let

$$\hat{\mathbf{b}}_{\Sigma} = \left[\Sigma_{b_x}, \Sigma_{b_y}, \Sigma_{b_h}, \Sigma_{b_w}\right] \tag{18}$$

denote the four predicted localization-variance terms. A sigmoid is applied to these variance predictions so that they lie in $[0,1]$. ConfMix defines the bounding-box confidence as

$$C_{\text{bbx}} = 1 - \text{mean}\left(\hat{\mathbf{b}}_{\Sigma}\right), \tag{19}$$

and combines it with the detector-dependent confidence C_{det} as

$$C_{\text{comb}} = C_{\text{det}} C_{\text{bbx}}. \tag{20}$$

Consequently, a prediction can receive a lower combined confidence even when its ordinary detector score is high if its predicted localization is uncertain.

ConfMix does not use this stricter confidence criterion at full strength from the beginning of adaptation. Instead, pseudo-label confidence transitions progressively from C_{det} toward C_{comb}:

$$C = (1-\delta) C_{\text{det}} + \delta C_{\text{comb}}, \tag{21}$$

where

$$\delta(r) = \frac{2}{1+\exp(-\alpha r)} - 1. \tag{22}$$

Here, $r \in [0,1]$ denotes normalized adaptation progress and α controls the rate of transition. Early in adaptation, pseudo-label filtering therefore depends predominantly on the ordinary detector confidence; as training progresses, the localization-uncertainty term receives progressively greater influence. The numerical settings used in our implementation are reported in Section III.D.

ConfMix combines the ordinary supervised source loss with a consistency loss on the mixed sample,

$$\mathcal{L}_{\text{ConfMix}} = \mathcal{L}_{\text{det}} + \gamma \mathcal{L}_{\text{cons}}, \tag{23}$$

where $\mathcal{L}_{\text{det}}$ preserves supervised source detection training and $\mathcal{L}_{\text{cons}}$ compares the predictions on the mixed image with a combined source–target detection set.

Specifically, let $\tilde{\mathbf{y}}_T^R$ denote the target pseudo-detections inside the selected target region and $\tilde{\mathbf{y}}_S^{R-}$ the source predictions outside that region. The consistency target is constructed as

$$\tilde{\mathbf{y}}_{S,T} = \tilde{\mathbf{y}}_T^R \cup \tilde{\mathbf{y}}_S^{R-}. \tag{24}$$

Predicted boxes that extend beyond the corresponding source or target mixing region are clipped at the region boundary before being used in this consistency target.

The weight γ is not a fixed loss coefficient. It reflects the estimated reliability of the detections used to supervise the consistency term. ConfMix defines

$$\gamma = \frac{\left|\{\tilde{y}_{S,T}^{i} \in \tilde{\mathbf{y}}_{S,T} : C^i \geq C_{\text{th}}^{\gamma}\}\right|}{\left|\tilde{\mathbf{y}}_{S,T}\right|}, \tag{25}$$

where C_{th}^{γ} is a stricter reliability threshold. The consistency loss therefore receives greater weight when a larger fraction of its supervisory detections satisfies this threshold. The pseudo-label filtering threshold and the reliability threshold used in our experiments are reported numerically in Section III.D.

ConfMix is therefore not driven exclusively by target pseudo-labels: labeled source supervision remains available throughout adaptation, while target-domain appearance is introduced through mixing.

ConfMix retains its uncertainty-aware detection head during adaptation. For standardized cross-method evaluation the localization-variance outputs are removed and the remaining box, objectness, and class outputs are evaluated using the standard YOLO confidence formulation, with the output dimensionalities stated in Section III.D. The adaptation procedure itself is unchanged.

3) Source-Free Adaptation: SF-YOLO: The second training-based route considers a stricter transfer condition. Once the source detector has been obtained, the original source images and their annotations are no longer accessible. Adaptation is therefore performed using only the pretrained source detector and unlabeled target samples. We instantiate this setting using Source-Free YOLO (SF-YOLO) [3].

SF-YOLO initializes teacher and student copies of the source-pretrained YOLO detector. For each target sample, the teacher produces candidate detections, and predictions below a confidence threshold are discarded. The surviving detections become pseudo-labels used to optimize the student on a modified view of the same target observation.

The modified student view is generated using a learned Target Augmentation Module (TAM). In the published SF-YOLO formulation, the input and style images are represented by VGG-16-encoded features within the TAM style-transfer formulation [3]. This feature-extraction path builds the training-time augmentation and is not part of the final adapted-teacher inference graph.

The style reference supplied to TAM can be either the average of the target-domain images when their backgrounds are sufficiently similar or a target image sampled from the target domain [3]. We use the sampled-image alternative; the sampling procedure and the corresponding image statistics are given in Section III.D.

During detector adaptation, the teacher and student receive two views of the same target sample. The teacher branch receives the target tensor before the student-specific TAM transformation, whereas the student receives the TAM-transformed counterpart. Common data-loader transformations that occur before this branch separation remain part of the training pipeline. Thus, "unmodified" in this context means that the teacher input is not subjected to the additional TAM transformation applied to the student.

Let θ and ϕ denote the teacher and student parameters, respectively. The student is updated by backpropagation from the pseudo-label detection loss. The teacher is updated as an exponential moving average of the student,

$$\theta \leftarrow \beta\theta + (1-\beta)\phi, \tag{26}$$

where β denotes the teacher EMA momentum. The teacher therefore evolves more smoothly than the directly optimized student and provides pseudo-labels for subsequent target samples.

A difficulty of mean-teacher self-training is that student errors can propagate into the teacher and subsequently into future pseudo-labels. SF-YOLO introduces a Student Stabilization Module (SSM) to provide teacher-to-student communication at a lower update frequency. At the end of each

epoch, the student parameters are moved toward the teacher,

$$\phi \leftarrow \eta\phi + (1-\eta)\theta, \quad (27)$$

where η denotes the SSM momentum. Thus, teacher EMA updates occur during minibatch training, whereas the SSM update is applied once per epoch [3].

Its detector-learning signal therefore comes entirely from teacher-generated target pseudo-labels, target-specific augmentation, and the teacher-student stabilization mechanism. The adapted teacher is used for target inference; the TAM, its feature-extraction path, and the student network are not additional components of the final deployed detector.

Implementation-specific quantities—including the teacher pseudo-label threshold, EMA and SSM momenta, TAM training duration, detector training duration, random seeds, style-image sampling procedure, and image-processing settings—are reported in Section III.D. In particular, the TAM training budget used in our experiments is stated there explicitly rather than being treated as part of the defining SF-YOLO method.

4) Reconstruction-Based Transfer: E2VID: The preceding strategies attempt to make the detector itself more suitable for the target domain. The reconstruction route takes the opposite approach: the detector remains unchanged, while the event stream is transformed into an intensity representation that can be processed by the existing image detector. We implement this route using the lightweight pretrained E2VID reconstruction model [4].

E2VID is a recurrent event-to-video reconstruction model that converts asynchronous event measurements into frame-like intensity images. Its recurrent state permits preceding event observations to influence subsequent reconstructions. The reconstruction model was pretrained independently of the DSEC detection task using simulated event data and therefore does not require DSEC-Det object annotations.

For the present experiment, we use the pretrained lightweight model distributed with the official E2VID implementation. The same 50-ms event windows defined in Section III.B are processed in chronological order. The windows remain aligned with the same DSEC-Det annotation timestamps used by the direct-event evaluation conditions. Consequently, the reconstruction and direct-event routes are evaluated on the same 695 held-out target samples.

For each event window, E2VID produces a single-channel grayscale intensity reconstruction. The resulting grayscale image is replicated across the three detector input channels,

$$I^{(3)}_{\mathrm{E2VID}} = \mathrm{Repeat}_3(I_{\mathrm{E2VID}}), \quad (28)$$

and is then passed through the standard image-detector preprocessing path before inference with the unchanged Cityscapes source detector. Neither the E2VID reconstruction network nor YOLOv5 is optimized on the DSEC-Det target data in this route. No pseudo-labels are generated and no target detection annotations are supplied to either model.

Because E2VID is recurrent, reconstruction depends not only on the events contained in the current temporal interval but also on the state accumulated from preceding observations. The temporal-window sensitivity analysis of Section III.B.3 therefore varies the event support supplied to this recurrent reconstruction process while keeping the downstream detector unchanged. The exact output normalization/tone-mapping procedure and the recurrent-state initialization policy used in our evaluation are treated as part of the reproducible implementation protocol and are specified in Section III.D.

This route changes the detector-facing input rather than the detector parameters: unlike the deterministic renderings of Section III.B, E2VID infers a latent intensity image, so the detector no longer observes event timing or polarity directly. The routes also differ in their auxiliary learned components and in when those components are used. E2VID places an externally pretrained reconstruction network in the target inference chain, whereas the SF-YOLO augmentation module is used only during training and the ConfMix uncertainty head is removed before evaluation. The comparison is therefore functional rather than component- or prior-matched, and a strong E2VID result indicates compatibility with the image-domain detector rather than intrinsic superiority of reconstruction.

5) Supervised Target Reference: Finally, we train the same detector family using target-domain detection annotations to provide a supervised reference. Unlike the preceding label-free transfer routes, this condition has direct access to DSEC detection labels during optimization and therefore does not satisfy the unsupervised transfer setting.

It indicates the target-domain performance attainable when the annotation constraint is removed, and provides context for the remaining gap between unsupervised transfer and direct target supervision. Its result at the predefined final epoch is used as the reference; any trajectory analysis is reported explicitly as diagnostic.

D. Training and Unified Evaluation Protocol

All models are assessed on the same held-out target sequence using a common label space and a unified YOLOv5 evaluation pipeline, so that differences between the compared strategies are not confounded by method-specific evaluation procedures. Method-specific training mechanisms are preserved.

Canonical experimental setting. The canonical target-domain protocol uses event windows of $\Delta t = 50$ ms. The unlabeled target adaptation set contains 3528 windows from the five training sequences described in Section III.A, while the held-out evaluation set contains 695 windows from the complete zurich_city_04_a sequence. Detector-facing inputs are generated in the native distorted 640×480 event-camera geometry and are evaluated with an image-size setting of 640.

The detector-training conditions in the main controlled comparison are run for 50 epochs. These conditions include continued source-only training, ConfMix, SF-YOLO, and the supervised target reference. Direct transfer and E2VID do not optimize the detector and therefore have no corresponding detector-training schedule.

For the two UDA methods, the canonical 50-ms experiments are repeated with random seeds 0, 1, and 2. Data, source initialization, training duration, method-specific hyperparameters, and evaluation procedure are held fixed across the three runs, so that the random seed is the only intended experimental variable. Aggregate results are reported as the mean and sample standard deviation across these runs.

Predefined reporting epochs are used to analyze adaptation trajectories. The 50-epoch checkpoint is treated as the principal final reporting point, while earlier fixed epochs are used only

for diagnostic analysis of training behavior. Target-domain annotations are not used to select the checkpoint reported as the unsupervised result.

The separate temporal-window sensitivity experiments defined in Section III.B.3 vary the window duration while leaving this canonical protocol otherwise unchanged, and the repeated temporal runs are treated as sensitivity experiments rather than as a new primary training protocol.

Common detector optimization. Both ConfMix and SF-YOLO use stochastic gradient descent with Nesterov momentum. The initial learning rate is 0.01, the momentum coefficient is 0.937, and the weight decay is 0.0005. Training uses a batch size of 16 with gradient accumulation over four steps. A linear learning-rate schedule is applied with a three-epoch warm-up.

These settings are retained across the canonical and temporal-window experiments unless stated otherwise; the temporal experiments modify the event support supplied to the target domain, not the adaptation algorithms.

ConfMix configuration. ConfMix retains access to the 2968 labeled Cityscapes source images together with the unlabeled DSEC target-adaptation samples. Adaptation preserves the uncertainty-aware ConfMix formulation described in Section III.C.2, including the localization-uncertainty outputs used during training.

Candidate target detections are filtered using a confidence threshold of $C_{\text{th}} = 0.25$, while the stricter reliability threshold used in consistency weighting is $C_{\text{th}}^{\gamma} = 0.5$. Non-maximum suppression within the adaptation procedure uses an IoU threshold of 0.5. The progressive-confidence coefficient in Equation (22) is set to $\alpha = 5$.

The canonical ConfMix experiment is repeated using random seeds 0, 1, and 2. No target-domain annotations are used for adaptation or for selecting the reported checkpoint.

SF-YOLO configuration. SF-YOLO is adapted without access to Cityscapes images or annotations. Its detector-learning inputs consist only of the source-pretrained detector and the unlabeled DSEC target samples. Teacher detections are retained as pseudo-labels using a confidence threshold of 0.40.

The teacher exponential-moving-average momentum is set to 0.999, and the Student Stabilization Module momentum is set to 0.5. The Target Augmentation Module is trained for 25,000 iterations before detector adaptation, which is then performed for 50 epochs.

For the target-specific style reference used by the augmentation module, one target image is sampled randomly from the current training batch and used as the common style reference for that batch; no independent target-image pool is queried. This avoids the nearly spatially uniform global target mean produced by averaging all target-adaptation samples, which has a mean pixel intensity of 127.9 and a spatial standard deviation of only 1.09.

E2VID inference protocol. E2VID performs no target-domain detector training. The pretrained lightweight E2VID model processes the same 50-ms target windows used by the other evaluation conditions. The windows are processed sequentially in chronological order. The recurrent state is initialized once at the beginning of the held-out sequence and is then propagated continuously across consecutive windows, without reinitialization between samples. The reconstructed single-channel intensity image is replicated across three channels and passed to the unchanged Cityscapes-trained YOLOv5s detector. Standard reconstruction and output-processing settings are retained throughout the experiment, and neither E2VID nor the downstream detector is optimized using DSEC-Det detection annotations.

Unified detector evaluation. All reported cross-method results are recomputed using the same corrected YOLOv5 evaluation procedure. The held-out zurich_city_04_a sequence contains 695 evaluation samples and 6191 ground-truth bounding boxes.

The source detector retains its eight-class output space, while quantitative target-domain reporting is restricted to the *person* and *car* classes defined in Section III.A. Predictions from the remaining source categories are not remapped into either evaluated class. The principal reported metrics are mAP@0.5 and mAP@0.5:0.95, supplemented by class-wise AP and diagnostic precision and recall where relevant.

Every standardized evaluation uses a common detection-confidence floor of 0.001, a non-maximum-suppression IoU threshold of 0.60, and an image-size setting of 640. This threshold is used only to retain predictions for metric computation and is distinct from the confidence thresholds used internally for adaptation. In particular, the SF-YOLO threshold of 0.40 determines which teacher detections are passed directly to the student as pseudo-label supervision, whereas the ConfMix threshold of 0.25 filters candidate target detections before confidence-based region selection while supervised source training remains active. The two values therefore serve different algorithmic roles and should not be interpreted as directly comparable pseudo-label operating points.

For ConfMix, the evaluation-time conversion described in Section III.C.2 is applied before metric computation. The four localization-variance channels are removed from the 17-output-per-anchor prediction head, producing the standard 13-output form consisting of four box coordinates, one objectness score, and eight class scores. Detection confidence is then computed using the ordinary YOLO objectness–class formulation, with no uncertainty-based confidence modulation. The same 17-to-13 conversion is applied to every ConfMix checkpoint included in the standardized comparison.

Evaluator audit. The evaluation pipeline was audited before the final cross-method comparison. A legacy bookkeeping behavior could discard the target annotations of an image when no detections survived for that image. Under severe domain shift this could exclude 562 of the 6191 ground-truth boxes and inflate the scores of weak detectors. The corrected evaluator retains target annotations independently of whether detections are present, so every standardized evaluation contains all 6191 boxes, and no reported result is taken from the affected legacy evaluation. Changes required for execution in the current software environment were limited to compatibility handling and did not alter the trained checkpoints, adaptation objectives, or dataset split.

Checkpoint reporting policy. The primary comparison uses the checkpoint obtained after the predefined 50-epoch training schedule and therefore requires no target labels for checkpoint selection. Earlier predefined reporting epochs are used only to characterize adaptation trajectories; they are reported as diagnostic evidence of adaptation potential and are not treated

as a valid unsupervised model-selection procedure.

All reported checkpoint results are recomputed using the unified standardized evaluation protocol. Logged intermediate metrics are used only to characterize training dynamics and are not substituted for standardized checkpoint evaluations. Epoch identities are reported only when they can be verified unambiguously from the training records.

Historical exploratory runs lacking complete configuration and execution provenance are excluded from the finalized comparison; all reported multi-seed results are derived from the fully audited three-run experiment set.

Table II summarizes the principal settings.

TABLE II

PRINCIPAL TRAINING AND EVALUATION SETTINGS USED IN THE CONTROLLED COMPARISON. THE CONFMIX AND SF-YOLO CONFIDENCE THRESHOLDS HAVE DIFFERENT ALGORITHMIC ROLES: THE FORMER FILTERS TARGET DETECTIONS BEFORE CONFIDENCE-BASED REGION SELECTION WHILE SOURCE SUPERVISION REMAINS AVAILABLE, WHEREAS THE LATTER DETERMINES WHICH TEACHER PREDICTIONS ARE SUPPLIED DIRECTLY TO THE STUDENT AS PSEUDO-LABEL SUPERVISION.

Method	Target-training labels	Detector epochs	Method-specific confidence threshold	Runs used for reported result
Direct transfer	No	—	—	1
Continued source-only control	No	50	—	1
ConfMix	No	50	0.25	3
SF-YOLO	No	50	0.40	3
E2VID + source detector	No	—	—	1
Supervised target reference	Yes	50	—	1

Unified evaluation uses the complete 695-sample held-out sequence with 6191 ground-truth boxes, the *person*/*car* evaluation space, and a common detection-confidence floor of 0.001.

Experiments were executed on two CUDA-enabled GPU environments, an NVIDIA Tesla T4 and an NVIDIA GeForce RTX 3090 Ti workstation with 24 GB of GPU memory, an Intel Core i9-14900K CPU, and 64 GB of system memory. No hardware-normalized runtime comparison across the evaluated transfer strategies is claimed.

4- EXPERIMENTAL RESULTS

This section examines what remains transferable from the image-trained detector after the transition to event data. We first isolate the effect of the detector-facing representation under direct transfer, then compare the adaptation and reconstruction routes, analyze the persistence of the adapted states, and finally examine how the conclusions change with temporal support.

A. Detector-Facing Representation and Direct Transfer

Before comparing adaptation and reconstruction, we first examine how much of the source detector's knowledge remains immediately usable when the event stream is presented through different deterministic renderings. All representations are constructed from the same 50-ms event windows with the detector unchanged, so the comparison isolates the effect of the detector-facing representation rather than differences in training or scene content.

Table III reports the direct-transfer results. The signed voxel representation defined in Section III.B, which is used as the primary event input in the subsequent adaptation experiments, provides the strongest overall compatibility with the image-trained detector. It reaches an mAP@0.5 of 0.0340 and an mAP@0.5:0.95 of 0.0152. The event-count representation reaches 0.0135 and 0.0058, respectively, while inversion of the count image does not improve average precision despite producing a higher recall. The stacked histogram yields the lowest direct-transfer performance among the evaluated renderings.

TABLE III

DIRECT-TRANSFER PERFORMANCE OF THE FIXED CITYSCAPES-TRAINED YOLOV5S DETECTOR UNDER DIFFERENT DETERMINISTIC EVENT REPRESENTATIONS. NO DETECTOR ADAPTATION OR TARGET-DOMAIN OPTIMIZATION IS PERFORMED.

Representation	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Signed voxel	**0.2359**	0.0299	**0.0340**	**0.0152**
Stacked histogram	0.0039	0.0135	0.0022	0.0009
Event count	0.0636	0.0198	0.0135	0.0058
Inverted event count	0.0109	**0.0561**	0.0092	0.0040

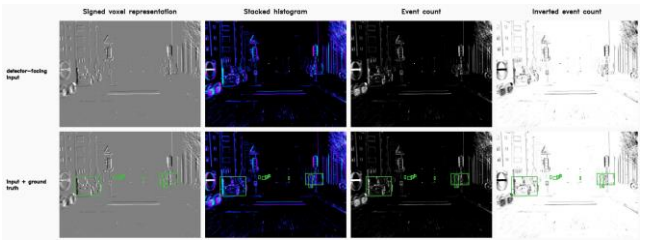


Fig. 1. Detector-facing renderings of the same 50-ms event window from the held-out DSEC target sequence. Columns show, from left to right, the signed voxel representation, the stacked histogram, the event count, and the inverted event count. The upper row shows the detector-facing input alone; the lower row overlays the corresponding ground-truth *person* and *car* annotations, which are used only for evaluation. All panels share the same events and the same distorted event-camera geometry; only the representation presented to the image-trained detector differs.

The precision–recall behavior provides additional insight into these differences. The signed voxel representation attains the highest precision (0.2359), whereas the inverted event count attains the highest recall (0.0561) but with a precision of only 0.0109. Thus, increasing the number of recovered target instances does not necessarily improve detector compatibility: in this case, the additional responses are accompanied by many detections that do not translate into higher average precision. The event count occupies an intermediate position, while the stacked histogram remains weak across both precision and recall.

Figure 1 provides a visual counterpart to this quantitative comparison. Because the four renderings are generated from the same event window and retain identical ground-truth geometry, the corresponding performance differences cannot be attributed to changes in scene content or bounding-box geometry, but are associated with the representation presented to the image-trained detector.

At the same time, the strongest deterministic representation still reaches only 0.0340 mAP@0.5. The representation study therefore establishes two points for the subsequent experiments: detector-facing rendering matters, but a more compatible rendering alone does not remove the image-to-event gap. The following comparison consequently asks whether that remaining discrepancy is better addressed by adapting the detector to unlabeled event data or by reconstructing an intensity-like input while keeping the detector unchanged.

B. Canonical Transfer Comparison at 50 ms

Table IV presents the central quantitative comparison. All values correspond to the predefined final reporting point: for the training-based conditions this is the checkpoint obtained after the 50-epoch schedule, while direct transfer and E2VID perform no detector optimization and are evaluated directly. No target annotations are used to select any reported checkpoint.

TABLE IV

LABEL-FREE TRANSFER PERFORMANCE ON THE COMPLETE ZURICH_CITY_04_A TARGET SEQUENCE UNDER THE CANONICAL 50-MS PROTOCOL. ALL VALUES USE THE UNIFIED EVALUATION PROTOCOL WITH 6191 GROUND-TRUTH BOXES AND A CONFIDENCE FLOOR OF 0.001, AND CORRESPOND TO THE PREDEFINED FINAL REPORTING POINT. CONFMIX AND SF-YOLO ARE REPORTED AS MEAN ± SAMPLE STANDARD DEVIATION OVER THREE INDEPENDENT RUNS. THE SUPERVISED TARGET REFERENCE IS NOT A LABEL-FREE CONDITION AND IS THEREFORE DISCUSSED IN THE TEXT RATHER THAN LISTED HERE.

Method	mAP@0.5	mAP@0.5:0.95
Direct transfer	0.0340	0.0152
Continued source-only	0.0169	0.0078
ConfMix	0.1030 ± 0.0303	0.0396 ± 0.0098
SF-YOLO	0.0456 ± 0.0173	0.0183 ± 0.0063
E2VID + unchanged source detector	**0.1787**	**0.0918**

Direct transfer retains only limited detection capability under the image-to-event shift, reaching 0.0340 mAP@0.5. Continued optimization on the source domain alone does not improve target compatibility: after the same 50-epoch training horizon, the source-only control reaches 0.0169 mAP@0.5. This control is important because it shows that additional optimization of the source detector, without target-domain adaptation, is not sufficient to account for the gains observed with the adaptation methods.

Both adaptation strategies exceed the direct-transfer reference at the predefined final reporting point, although to different degrees and with different levels of evidential support. ConfMix reaches 0.1030 ± 0.0303 mAP@0.5 and 0.0396 ± 0.0098 mAP@0.5:0.95, demonstrating that useful target-domain compatibility can be recovered while labeled source supervision remains available. SF-YOLO reaches 0.0456 ± 0.0173 and 0.0183 ± 0.0063, respectively. Its margin over direct transfer is smaller than its own across-seed standard deviation, so this final advantage should be read as a descriptive tendency rather than an established improvement. The smaller final gain indicates that the source-free teacher–student setting is more difficult under the present cross-modal shift, but the predefined final checkpoint alone does not establish whether the method was unable to adapt or whether stronger adapted states occurred earlier in training. This distinction is examined explicitly in Section IV.C.

The reconstruction route exhibits a different behavior. E2VID followed by the unchanged source detector reaches 0.1787 mAP@0.5 and 0.0918 mAP@0.5:0.95, providing the strongest label-free result. This result shows that presenting the source detector with an intensity-like reconstruction can substantially improve its compatibility with target event data. It should not, however, be interpreted as a prior-matched demonstration that reconstruction is universally superior to detector adaptation: E2VID contributes an independently pretrained reconstruction model, whereas the adaptation methods operate under different training and information constraints. Accordingly, the comparison should be interpreted as a comparison of practical transfer routes under a common downstream evaluation protocol, rather than as a capacity- or information-matched algorithmic comparison.

For additional context, a detector trained with target-domain annotations reaches 0.3093 mAP@0.5 at the predefined final reporting point. The gap between this supervised reference and all label-free conditions confirms that none of the evaluated transfer routes fully removes the cross-modal discrepancy. At the same time, the difference between direct transfer and the adapted or reconstructed conditions shows that a nontrivial portion of the source detector's knowledge can be recovered without using target detection labels.

The canonical comparison therefore gives only the first part of the answer. The remaining question, particularly for ConfMix and SF-YOLO, is whether the final epoch accurately represents what the adaptation process was able to achieve.

C. Adaptation Dynamics and Persistence

The predefined final checkpoint provides a reproducible endpoint for unsupervised adaptation, but it does not reveal whether the adaptation process remains weak throughout training or instead passes through more useful target-domain states before the final epoch. We therefore examine ConfMix and SF-YOLO at the same fixed reporting points, {10,20,30,40,50} epochs. These intermediate evaluations are diagnostic: they are inspected after training under a common schedule and are not used to select a target-label-informed checkpoint for the primary comparison.

Table V reports the three-seed mean and sample standard deviation at these fixed epochs. Both methods exhibit non-monotonic target-domain trajectories.

TABLE V

TARGET-DOMAIN PERFORMANCE AT FIXED REPORTING EPOCHS UNDER THE CANONICAL 50-MS PROTOCOL. VALUES ARE MEAN ± SAMPLE STANDARD DEVIATION OVER THREE RANDOM SEEDS. EPOCHS 10–40 ARE DIAGNOSTIC REPORTING POINTS; EPOCH 50 IS THE PREDEFINED FINAL REPORTING POINT. NO INTERMEDIATE CHECKPOINT IS SELECTED AS THE PRIMARY RESULT USING TARGET-DOMAIN ANNOTATIONS.

Method	Epoch	mAP@0.5	mAP@0.5:0.95
ConfMix	10	0.1351 ± 0.0219	0.0561 ± 0.0102
	20	0.1657 ± 0.0091	0.0696 ± 0.0034
	30	0.1228 ± 0.0417	0.0506 ± 0.0204
	40	0.1467 ± 0.0137	0.0604 ± 0.0077
	50	0.1030 ± 0.0303	0.0396 ± 0.0098
SF-YOLO	10	0.0704 ± 0.0199	0.0303 ± 0.0092
	20	0.0749 ± 0.0206	0.0307 ± 0.0094
	30	0.0612 ± 0.0196	0.0249 ± 0.0084
	40	0.0506 ± 0.0184	0.0205 ± 0.0075
	50	0.0456 ± 0.0173	0.0183 ± 0.0063

ConfMix shows the clearest separation between the state reached during adaptation and the predefined final state. Its three-seed mean increases from 0.1351 mAP@0.5 at epoch 10 to 0.1657 at epoch 20, which is the highest mean among the inspected fixed diagnostic points. The corresponding mAP@0.5:0.95 reaches 0.0696. Performance subsequently varies rather than improving monotonically, and the final epoch reaches only 0.1030 mAP@0.5 and 0.0396 mAP@0.5:0.95. Thus, the final ConfMix result understates the strongest target-domain state observed at the predefined intermediate reporting points.

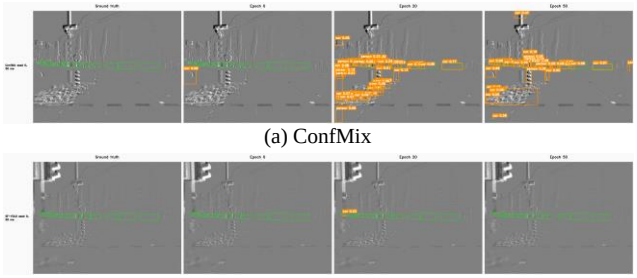
This distinction is particularly relevant to the interpretation of the canonical comparison. At epoch 20, ConfMix reaches

0.1657 mAP@0.5, approaching the 0.1787 obtained by E2VID under the same 50-ms target evaluation. This observation does not alter the primary comparison: epoch 20 is a diagnostic reporting point rather than a deployable checkpoint selected without target labels, and E2VID remains stronger in mAP@0.5:0.95. Instead, the comparison shows that source-accessible adaptation is capable of reaching a substantially more useful cross-modal state than its final checkpoint alone would suggest.

SF-YOLO exhibits the same general persistence problem at a lower performance level. Its three-seed mean reaches 0.0749 mAP@0.5 and 0.0307 mAP@0.5:0.95 at epoch 20, compared with 0.0456 and 0.0183 at epoch 50. Importantly, the epoch-20 mAP@0.5 exceeds the corresponding epoch-50 value in each of the three random seeds. The repeated direction of this change indicates that the difference between the intermediate and final states is not attributable to a single anomalous run. Nevertheless, these fixed-epoch observations should not be interpreted as evidence that epoch 20 is an optimal stopping point, since the target labels required to recognize that state are unavailable in the unsupervised setting.

The continued source-only and supervised controls provide useful context for this behavior. Continued source-only training does not produce a corresponding target-domain improvement, ending at 0.0169 mAP@0.5 after 50 epochs. Conversely, the target-supervised reference reaches 0.3093 mAP@0.5 at the predefined final epoch, while its retrospective target-label-selected diagnostic maximum is only slightly higher at 0.3115. The small separation between these two supervised states suggests that the pronounced late-stage decrease observed for the unsupervised adaptation methods is not simply an inevitable consequence of using a 50-epoch training horizon. Because the supervised and unsupervised objectives differ, however, this comparison is contextual rather than a causal test of the degradation mechanism.

Figure 2 provides a qualitative view of the same phenomenon. The ConfMix example shows substantially more target-domain responses at the intermediate state, although these also include false positives and imperfect localization, while the SF-YOLO example shows a target detection that appears at the intermediate reporting point and is absent again at the final state. These examples are not used for checkpoint selection; rather, they illustrate that the numerical trajectory corresponds to a visible change in detector behavior over the course of adaptation.



(a) ConfMix

(b) SF-YOLO

Fig. 2. Qualitative evolution of (a) ConfMix and (b) SF-YOLO under the canonical 50-ms protocol, both shown for seed 0. In each panel the leftmost column shows the ground-truth annotations, and the remaining columns follow the same held-out DSEC sample from the common source initialization (epoch 0), through the fixed epoch-20 diagnostic reporting point, to the predefined final state at epoch 50. Green boxes denote ground truth and orange boxes denote predictions. Predictions are displayed at a confidence threshold of 0.05 for visual clarity, whereas the quantitative evaluation uses the common confidence floor of 0.001. Epoch 20 is shown only to illustrate adaptation dynamics and is not a target-label-selected deployment checkpoint.

Taken together, the fixed-epoch and qualitative results change the interpretation of the adaptation experiments. The principal difficulty is not that the evaluated UDA methods are incapable of improving an image-trained detector on event data. Both methods reach target-domain states that are more useful than their source initialization, and ConfMix in particular approaches the reconstruction result at an intermediate reporting point. The unresolved problem is that these adapted states are not reliably preserved through the predefined training schedule, while the unsupervised setting provides no target labels with which to recognize when such a state has been reached.

1) Adaptation-Signal Diagnostics: The preceding trajectories motivate a second question: what supervisory signal is available to the adaptation procedures while these target-domain states are being formed? We examine this question diagnostically rather than causally. The analysis can characterize the pseudo-label signal supplied to the models, but it cannot by itself establish why the later performance decrease occurs.

For SF-YOLO, the teacher directly determines which target detections become pseudo-label supervision for the student. This diagnostic audit is computed on the 3528-window unlabeled target-adaptation set, whereas the standardized detector metrics reported in the main comparison are computed on the separate 695-sample held-out evaluation sequence. Under the adopted confidence threshold of 0.40, the implementation audit retains 1208 teacher detections, of which 1130 belong to the two evaluated classes (439 person and 691 car). Within the diagnostic teacher view, 899 of these person/car detections match target objects and 231 are false positives, corresponding to a precision of approximately 0.796 among the retained person/car predictions. The mean and median IoU of the matched detections are approximately 0.768 and 0.792, respectively.

This result is important because it rules out an overly simple explanation of the SF-YOLO trajectory. The source teacher is not devoid of usable signal after the image-to-event shift; a substantial fraction of its accepted predictions are geometrically consistent with target objects. At the same time, the signal remains selective: only predictions exceeding the adopted confidence threshold contribute directly to student supervision,

and the teacher's accepted predictions do not cover the full target object set. Both the precision and the recall measured in this teacher-view diagnostic are therefore not directly comparable with the standardized detector values reported elsewhere in the paper: the diagnostic is computed only over predictions retained above the pseudo-label threshold of 0.40 on the adaptation set, whereas the standardized evaluation operates at the common confidence floor of 0.001 over the complete held-out target object set.

The audit does not establish that pseudo-label quantity, pseudo-label errors, or teacher feedback cause the subsequent decrease in SF-YOLO performance. Changing the acceptance threshold would alter both the number and reliability of pseudo-labels, and no controlled threshold ablation is used here to separate these effects. The supported conclusion is narrower: under a severe image-to-event shift, SF-YOLO begins from a teacher signal that contains useful target information but provides incomplete and self-generated supervision, and the resulting adapted state is not consistently preserved through later training.

A directly comparable internal pseudo-label audit is not reported for ConfMix because the available diagnostic record cannot be tied with sufficient provenance to all three finalized runs. We therefore do not attribute the ConfMix trajectory to pseudo-label quantity or correctness. Structurally, however, ConfMix differs from SF-YOLO in an important respect: labeled source supervision remains active throughout adaptation, whereas SF-YOLO must adapt without access to the original source images. The stronger intermediate and final ConfMix results are consistent with this difference in available supervision, but the present experiments do not isolate it as the causal source of the performance gap.

The adaptation-signal analysis therefore reinforces the main trajectory finding without explaining it away: useful cross-modal adaptation can occur, but obtaining such a state and preserving it are distinct problems. This motivates the next analysis, where we test whether the evolution of the adapted state also depends on the amount of temporal evidence used to construct each target observation.

D. Temporal-Window Sensitivity

We perform a post-hoc sensitivity analysis using $\Delta t \in \{5,10,20,50\}$ ms while keeping the annotation timestamps, target geometry, source detector, and standardized evaluation protocol unchanged. The shorter windows are used only to examine whether the conclusions depend on temporal support.

Figure 3 first illustrates what this change means at the input level. As the window expands, the signed voxel representation contains progressively richer event structure, while the effect on E2VID is more pronounced: the shorter-window reconstructions contain substantially less stable scene appearance than the coherent intensity-like structure obtained at 50 ms. The recurrent E2VID state is carried chronologically across the complete evaluation sequence for every condition; it is not reset independently for the displayed windows.

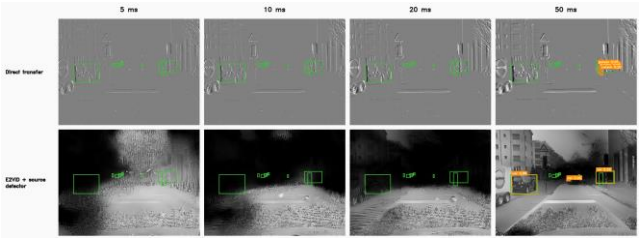


Fig. 3. Qualitative effect of temporal-window duration on the detector-facing input and the resulting predictions for the same held-out DSEC timestamp. Columns correspond to 5, 10, 20, and 50 ms. The upper row shows direct transfer using the signed voxel representation, whereas the lower row shows E2VID reconstruction followed by the unchanged source detector. Green boxes denote ground truth and orange boxes denote predictions displayed at a confidence threshold of 0.05; quantitative evaluation uses the common confidence floor of 0.001. The annotation timestamp is identical in every column, so only the temporal interval $(t - \Delta t, t]$ changes. For E2VID, recurrent state is initialized once at the beginning of the sequence and carried chronologically through all 695 evaluation samples.

The quantitative sweep in Fig. 4 shows that the four transfer strategies do not respond to this change in the same way. Direct transfer improves progressively with increasing temporal support, rising from 0.0084 mAP@0.5 at 5 ms to 0.0159, 0.0264, and 0.0340 at 10, 20, and 50 ms, respectively. The corresponding mAP@0.5:0.95 values increase from 0.0044 to 0.0078, 0.0124, and 0.0152. Thus, even without any adaptation, presenting the image-trained detector with a representation constructed from a longer temporal interval is associated with improved direct-transfer performance.

E2VID exhibits a much stronger temporal dependence. Its mAP@0.5 values are 0.0013, 0.0078, and 0.0235 at 5, 10, and 20 ms, respectively, before rising sharply to 0.1787 at 50 ms. The same pattern appears for mAP@0.5:0.95, which changes from 0.0003, 0.0042, and 0.0117 to 0.0918. Consequently, reconstruction is weaker than direct transfer at each of the three shorter windows in this experiment, but becomes substantially stronger under the canonical 50-ms condition. The reconstruction advantage observed in the canonical comparison is therefore not invariant to the amount of temporal evidence supplied to the reconstruction process.

A reset-state diagnostic provides a direct check on the recurrent-state confounding in this comparison. When the E2VID recurrent state is reinitialized independently before every annotation-aligned sample, mAP@0.5 reaches 0.0248, 0.0579, 0.1126, and 0.1911 at 5, 10, 20, and 50 ms, respectively; the corresponding mAP@0.5:0.95 values are 0.0120, 0.0290, 0.0574, and 0.0940. Thus, the strong dependence on temporal support persists after cross-sample recurrent memory is removed. Interestingly, resetting the state substantially improves the shorter-window results, whereas the difference between the two state conditions becomes small at 50 ms. This indicates that E2VID's temporal behavior reflects an interaction between the event support presented in each sample and cross-sample recurrent-state propagation rather than either factor alone.

The adaptation methods show different temporal responses. Every ConfMix and SF-YOLO temporal condition is repeated across three independent random seeds, so that no observation in this analysis depends on a particular random initialization. The points plotted in Fig. 4(a,b) are therefore three-seed means with error bars indicating the sample standard deviation, and the shaded bands in Fig. 4(c,d) show the corresponding

dispersion of the SF-YOLO adaptation trajectories. Table VI lists the exact mean and sample standard deviation for every method and window at the predefined final reporting epoch.

For ConfMix, mean mAP@0.5 increases monotonically across the four evaluated windows, from 0.0104 ± 0.0052 at 5 ms to 0.0359 ± 0.0198 at 10 ms, 0.0511 ± 0.0081 at 20 ms, and 0.1030 ± 0.0303 at 50 ms. The corresponding mAP@0.5:0.95 values are 0.0040 ± 0.0024, 0.0149 ± 0.0093, 0.0206 ± 0.0049, and 0.0396 ± 0.0098, respectively. The three-seed analysis therefore strengthens the tendency observed in the original matched sweep: stronger ConfMix performance is associated with increasing temporal support, and the non-monotonic step seen in the single-seed sweep does not persist under replication.

SF-YOLO exhibits a different, non-monotonic temporal response. Its mean mAP@0.5 changes from 0.0118 ± 0.0079 at 5 ms to 0.0612 ± 0.0118 at 10 ms and reaches its highest mean among the evaluated windows at 0.0766 ± 0.0222 at 20 ms, before decreasing to 0.0456 ± 0.0173 at 50 ms. The corresponding mAP@0.5:0.95 values are 0.0056 ± 0.0047, 0.0251 ± 0.0070, 0.0338 ± 0.0137, and 0.0183 ± 0.0063, respectively. Importantly, the 20-ms final mAP@0.5 remains higher than the corresponding 50-ms value in each of the three paired seeds. The expanded experiment therefore confirms that the previously observed intermediate-window behavior is not specific to seed 0, while also revealing appreciable run-to-run variability.

These repeated runs are used descriptively to characterize variability and directional consistency. With three seeds per condition, we do not interpret the differences as formal evidence of statistical significance, nor do we identify any temporal window as generally optimal.

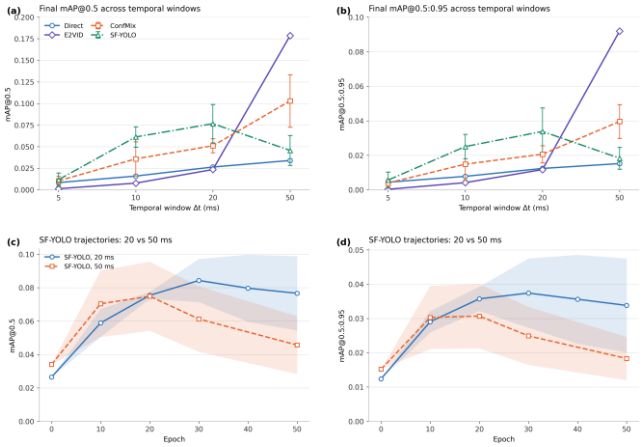


Fig. 4. Temporal-window sensitivity of the evaluated image-to-event transfer strategies. (a) mAP@0.5 and (b) mAP@0.5:0.95 at the predefined final reporting point for temporal windows of 5, 10, 20, and 50 ms. Direct transfer and E2VID require no detector optimization and are therefore shown without error bars, whereas the ConfMix and SF-YOLO points are means over three independent random seeds at every window, with error bars indicating the sample standard deviation. (c) and (d) show the corresponding SF-YOLO adaptation trajectories at 20 and 50 ms, plotted as the three-seed mean with shaded sample standard deviation. The exact final-epoch values are listed in Table VI. The 50-ms setting remains the canonical protocol; the temporal sweep is a post-hoc sensitivity analysis and is not used to select an alternative primary configuration.

TABLE VI

THREE-SEED TEMPORAL-SENSITIVITY RESULTS FOR THE TWO ADAPTATION METHODS AT THE PREDEFINED FINAL REPORTING EPOCH. VALUES ARE MEAN ± SAMPLE STANDARD DEVIATION OVER THREE INDEPENDENT RUNS.

Method	Δt	mAP@0.5	mAP@0.5:0.95
ConfMix	5 ms	0.0104 ± 0.0052	0.0040 ± 0.0024
	10 ms	0.0359 ± 0.0198	0.0149 ± 0.0093
	20 ms	0.0511 ± 0.0081	0.0206 ± 0.0049
	50 ms	0.1030 ± 0.0303	0.0396 ± 0.0098
SF-YOLO	5 ms	0.0118 ± 0.0079	0.0056 ± 0.0047
	10 ms	0.0612 ± 0.0118	0.0251 ± 0.0070
	20 ms	0.0766 ± 0.0222	0.0338 ± 0.0137
	50 ms	0.0456 ± 0.0173	0.0183 ± 0.0063

The per-seed fixed-epoch records further show that the 20-ms result cannot be explained simply by slower convergence, while Fig. 4(c,d) summarizes the corresponding mean trajectories. Among the fixed diagnostic points, the three 50-ms runs reach their highest mAP@0.5 at epoch 20 and subsequently decrease toward the final reporting point. At 20 ms, the corresponding highest inspected states occur at epochs 40, 20, and 30 for seeds 0, 1, and 2, respectively. The shorter-window trajectories therefore do not share a common delayed-convergence pattern. Instead, changing temporal support is associated with changes in both the final state and the evolution of the teacher–student adaptation process.

An important qualification is that Δt does not modify only the amount of event activity contained in an individual representation. The DSEC detection timestamps are separated by approximately 50 ms, so the canonical windows provide nearly continuous temporal coverage of the underlying event stream, whereas shorter windows leave progressively larger intervals between successive observations unrepresented. Because the recurrent-state contribution is examined separately by the reset-state control, the remaining ambiguity concerns local event accumulation and stream coverage, which the present experiment cannot separate; no causal claim is made about which of these factors produces the observed differences.

Taken together, the temporal analysis changes the interpretation of the canonical 50-ms comparison. E2VID is the strongest label-free strategy under the canonical protocol, but its advantage depends strongly on the temporal support available to the reconstruction process. ConfMix broadly benefits from longer support in the matched sweep, whereas SF-YOLO exhibits a replicated final advantage at an intermediate temporal window. Temporal-window duration should therefore not be treated as a neutral preprocessing choice in image-to-event transfer; its effect is strongly method dependent across the evaluated transfer mechanisms. This result cautions against ranking adaptation and reconstruction strategies from a single temporal configuration and motivates future transfer methods that explicitly account for the temporal structure of event data.

E. Class-Wise and Qualitative Analysis

The aggregate results do not indicate whether the recovered performance is distributed uniformly across object categories or scene conditions. We therefore examine class-wise average precision, representative qualitative detections, and the relationship between local event activity and object matching.

Table VII reports AP@0.5 separately for person and car. Direct transfer produces similarly low performance for the two classes, reaching 0.0315 AP@0.5 for person and 0.0366 for car. The two adaptation methods behave differently. At the predefined final epoch, ConfMix reaches only 0.0254 ± 0.0152 for person but 0.1806 ± 0.0460 for car. SF-YOLO shows the same qualitative asymmetry, reaching $0.0067 \pm$

0.0042 for person and 0.0846 ± 0.0308 for car. Consequently, the aggregate final improvements of the evaluated adaptation methods are driven predominantly by improved car detection rather than by uniform improvement across both target classes.

TABLE VII
CLASS-WISE AP@0.5 UNDER THE CANONICAL 50-MS PROTOCOL. CONFMIX AND SF-YOLO ARE REPORTED AS MEAN ± SAMPLE STANDARD DEVIATION OVER THREE RANDOM SEEDS AT THE PREDEFINED FINAL EPOCH.

Method	Person AP@0.5	Car AP@0.5
Direct transfer	0.0315	0.0366
ConfMix	0.0254 ± 0.0152	0.1806 ± 0.0460
SF-YOLO	0.0067 ± 0.0042	0.0846 ± 0.0308
E2VID + unchanged source detector	**0.0584**	**0.2990**

E2VID improves both categories relative to direct transfer, reaching 0.0584 AP@0.5 for person and 0.2990 for car, but the improvement remains much larger for car. Person detection therefore remains difficult across all evaluated label-free transfer routes. The present experiments do not isolate the cause of this class-dependent behavior. Differences in class frequency, object scale, motion, visibility, and event structure may all contribute, and none is varied independently here. We therefore treat the class-wise results as descriptive evidence rather than attributing them to class imbalance or any single scene factor.

Figure 5 provides a visual counterpart to these class-wise and aggregate results. Each row shows the same held-out DSEC sample across ground truth, direct transfer, the predefined final ConfMix and SF-YOLO states from seed 0, E2VID followed by the unchanged source detector, and the supervised target reference. The comparison deliberately includes both successful and unsuccessful cases rather than only favorable examples.

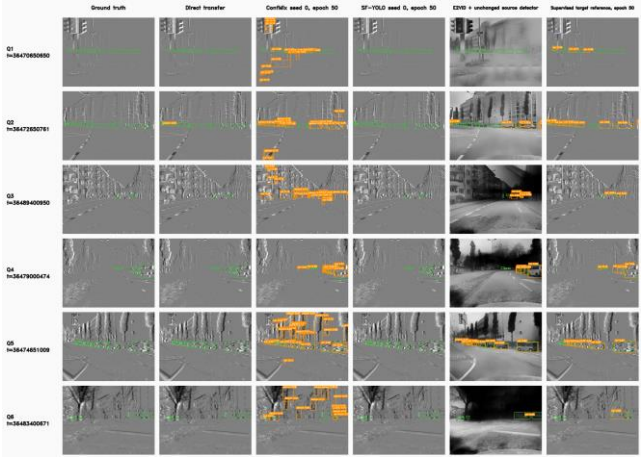


Fig. 5. Qualitative comparison on selected held-out DSEC samples under the canonical 50-ms protocol. Columns show, from left to right, ground truth, direct transfer, ConfMix seed 0 at the predefined final epoch, SF-YOLO seed 0 at the predefined final epoch, E2VID followed by the unchanged source detector, and the supervised target reference at epoch 50. Green boxes denote ground-truth annotations and orange boxes denote predictions. Predictions are displayed at a confidence threshold of 0.05 for visual clarity, whereas standardized quantitative evaluation uses the common confidence floor of 0.001. The displayed samples are used only for qualitative analysis and do not determine checkpoint selection.

Several aspects of the quantitative comparison are visible in these examples. Direct transfer leaves many annotated objects unmatched, reflecting the large detector-facing modality gap identified earlier. ConfMix produces substantially more target-domain responses in several scenes, but the additional responses also include false positives and imperfect localization. The predefined final SF-YOLO state remains comparatively sparse in the selected examples. E2VID recovers substantially more object detections while keeping the original source detector unchanged, although missed objects, false detections, and localization errors remain visible. The supervised reference provides context for the remaining gap without being treated as a label-free transfer method.

The examples also show why aggregate mAP alone is insufficient to describe the transfer problem. Closely spaced objects are not always separated correctly, some predicted boxes cover only part of the corresponding object extent, and small or distant targets can remain undetected even under the stronger transfer conditions. These observations are qualitative and are not used to establish object scale or distance as independent causes of failure. They instead motivate a more controlled diagnostic question: whether the amount of event activity locally available around a target is associated with its probability of being matched.

Event activity and object matching. For this diagnostic analysis, event activity is measured inside every ground-truth bounding box over the same 50-ms interval used to construct the detector input. To reduce the direct dependence of raw event count on object size, we use event density

$$\rho_{\text{event}} = \frac{N_{\text{event}}}{A_{\text{box}}}, \tag{29}$$

where N_{event} is the number of events falling inside the ground-truth box and A_{box} is the box area in pixels. All 6191 ground-truth objects from the 695 held-out samples are retained. For this diagnostic, an object is considered matched when a prediction of the correct evaluated class overlaps it at IoU ≥ 0.5 under the standardized prediction protocol. This matched-object analysis is distinct from the single operating-point precision and recall values reported by the YOLO evaluator.

Table VIII summarizes the relationship for direct transfer and E2VID. Under direct transfer, the 445 matched objects have a median event density of 2.2206 events/pixel, compared with 1.2932 for the 5746 unmatched objects. The same descriptive separation appears within each class: matched person instances have a median density of 1.9398 compared with 1.0429 for unmatched instances, while the corresponding car medians are 2.3826 and 1.3179.

TABLE VIII
EVENT-DENSITY DIAGNOSTIC ON THE HELD-OUT DSEC SEQUENCE UNDER THE CANONICAL 50-MS PROTOCOL. DENSITY IS MEASURED IN EVENTS PER PIXEL WITHIN EACH GROUND-TRUTH BOUNDING BOX. "MATCHED" DENOTES A CORRECT-CLASS PREDICTION WITH IOU ≥ 0.5.

Method	Class	Matched objects	Unmatched objects	Median pevent (matched)	Median pevent (unmatched)
Direct	All	445	5746	2.2206	1.2932
	Person	122	643	1.9398	1.0429
	Car	323	5103	2.3826	1.3179
E2VID	All	2837	3354	2.3838	0.7567
	Person	149	616	2.1250	0.9439
	Car	2688	2738	2.4116	0.7209

The separation is stronger for E2VID. Its 2837 matched objects have a median event density of 2.3838 events/pixel, whereas the 3354 unmatched objects have a median of 0.7567. For person, the corresponding medians are 2.1250 and 0.9439,

and for car they are 2.4116 and 0.7209. Thus, for both the direct and reconstruction routes, objects that are successfully matched tend to contain more local event activity than those that are missed.

Event density is not, however, a sufficient explanation of detection success. When the full target set is divided into event-density quartiles, matching does not increase monotonically through the highest-density group. For direct transfer, the matched-object fraction increases from approximately 0.005 in the lowest quartile to 0.046 and 0.171 in the second and third quartiles, but decreases to approximately 0.067 in the highest quartile. E2VID shows the same non-monotonic detail at the upper end: the corresponding fractions are approximately 0.116, 0.397, 0.675, and 0.646. High event activity is therefore associated with improved detectability over much of the observed range, but activity alone does not determine whether an object is detected.

This qualification is important because event density can co-vary with object size, relative motion, edge content, occlusion, and scene structure. The present analysis does not independently control these factors and therefore does not establish low event density as a causal failure mechanism. Instead, it provides complementary evidence that the information available within an event window is related to transfer difficulty while leaving substantial variation unexplained.

Taken together, the class-wise, qualitative, and event-activity analyses show that the image-to-event gap is not expressed uniformly across objects or transfer strategies. The adaptation methods recover substantial car performance but provide much smaller gains for person, the reconstruction route improves both classes while remaining imperfect, and local event activity is associated with successful matching without fully explaining it. Combined with the preceding representation, adaptation-dynamics, and temporal-window experiments, these results indicate that successful image-to-event transfer depends jointly on how event information is presented, how adaptation evolves and is preserved, and how much temporally supported target evidence is available to the transfer mechanism.

Cross-family and cross-dataset generalization. Table IX reports the additional zero-shot generalization results. On the unseen DSEC thun_02_a sequence, direct transfer reaches 0.0929 mAP@0.5, while the fixed ConfMix and SF-YOLO models reach 0.1653 and 0.1065, respectively. E2VID remains the strongest evaluated route on this sequence at 0.2648 mAP@0.5. The corresponding mAP@0.5:0.95 values are 0.0396, 0.0712, 0.0451, and 0.1278. Thus, the qualitative ordering observed on the primary DSEC evaluation family is preserved on this previously unseen DSEC recording family.

The independent Gen1 evaluation produces a different ordering. Direct transfer reaches 0.0114 mAP@0.5, while ConfMix and SF-YOLO reach 0.0197 and 0.0398, respectively. E2VID reaches 0.0049. The corresponding mAP@0.5:0.95 values are 0.0054, 0.0093, 0.0173, and 0.0021. Importantly, both adaptation methods remain above direct transfer, whereas the reconstruction route no longer retains the advantage observed on DSEC.

TABLE IX

ADDITIONAL ZERO-SHOT GENERALIZATION EVALUATION USING FIXED FINALIZED MODELS. EACH ENTRY REPORTS MAP@0.5 / MAP@0.5:0.95. NO TRAINING, ADAPTATION, HYPERPARAMETER TUNING, OR CHECKPOINT SELECTION IS PERFORMED ON THUN_02_A OR GEN1. CONFMIX AND SF-YOLO USE THE FIXED SEED-0 EPOCH-50 CHECKPOINTS.

Evaluation set	Direct	Conf Mix	SF-YOLO	E2VID
DSEC thun_02_a	0.0929 / 0.0396	0.1653 / 0.0712	0.1065 / 0.0451	0.2648 / 0.1278
Gen1	0.0114 / 0.0054	0.0197 / 0.0093	0.0398 / 0.0173	0.0049 / 0.0021

These results provide two complementary forms of generalization evidence. The adaptation gains are not restricted to the original zurich_city_04 evaluation family and remain observable on the independent Gen1 dataset. At the same time, the relative ranking of practical transfer routes is not invariant across datasets, particularly for E2VID. The generalization results therefore support the transferability of the adaptation effect while reinforcing the need to avoid universal rankings from a single target dataset.

Detector-architecture sensitivity. All results reported so far use a single detector family. To examine whether the detector-facing observations depend on that choice, we additionally evaluate the frozen Cityscapes-trained Cascade R-CNN defined in Section III.B on exactly the same finalized target inputs, under the same standardized evaluation protocol. Table X reports direct signed-event transfer and E2VID reconstruction for both detectors on all three evaluation sets.

Three observations follow. First, the direction of the reconstruction effect is preserved under both architectures: E2VID improves over direct signed-event transfer on both DSEC evaluation families, and on Gen1 it does not, with direct transfer remaining the stronger of the two routes for each detector. Second, the ordering of the deterministic event renderings on zurich_city_04_a is identical for the two detectors, with the signed voxel representation strongest, followed by the event count, the inverted event count, and the stacked histogram (0.0147, 0.0033, 0.0025, and 0.0005 mAP@0.5 for Cascade R-CNN). The representation conclusions of Section IV.A therefore hold for both detectors.

Third, the weaker event-domain result of Cascade R-CNN does not follow from a weak source model. Evaluated on the labeled Cityscapes source images under the same protocol and the same label space, YOLOv5s reaches 0.6186 mAP@0.5 and 0.3717 mAP@0.5:0.95, and Cascade R-CNN reaches 0.7705 and 0.5608. These are source training images for both detectors, so the values are optimistic and establish only that both provide functional source-domain baselines. Cascade R-CNN nevertheless retains a smaller fraction of its source performance on the primary DSEC evaluation, showing that stronger source-domain performance does not necessarily translate into stronger image-to-event transfer.

This control concerns the frozen detector-facing conditions only. ConfMix and SF-YOLO were not reimplemented with Cascade R-CNN, so no claim of architecture invariance is made for the adaptation methods themselves, and the absolute performance of the two detectors is not capacity matched.

TABLE X

DETECTOR-ARCHITECTURE SENSITIVITY OF DIRECT SIGNED-EVENT TRANSFER AND E2VID RECONSTRUCTION. EACH ENTRY REPORTS MAP@0.5 / MAP@0.5:0.95. THE CASCADE R-CNN DETECTOR IS EVALUATED FROZEN, WITH NO TARGET-DOMAIN TRAINING, ADAPTATION, HYPERPARAMETER

TUNING, OR CHECKPOINT SELECTION, AND RECEIVES THE SAME FINALIZED DETECTOR-FACING INPUTS AS THE PRIMARY DETECTOR.

Dataset	Detector	Signed	E2VID
zurich_city_04_a	YOLOv5s	0.0340 / 0.0152	0.1787 / 0.0918
	Cascade R-CNN	0.0147 / 0.0058	0.1285 / 0.0635
thun_02_a	YOLOv5s	0.0929 / 0.0396	0.2648 / 0.1278
	Cascade R-CNN	0.0787 / 0.0318	0.2306 / 0.1088
Gen1	YOLOv5s	0.0114 / 0.0054	0.0049 / 0.0021
	Cascade R-CNN	0.0148 / 0.0070	0.0031 / 0.0015

5- ANALYSIS AND DISCUSSION

The experiments provide a relatively subtle answer to the central question of this study, rather than supporting a single ranking of transfer methods. Reusing an image-trained detector on event data can be approached either by adapting the detector to an event-derived representation or by transforming the event stream into a representation that is more compatible with the detector's original image-domain training. Neither route dominates in general: which is preferable depends on the target dataset, on the temporal support available, and, for reconstruction, on the detector it feeds.

Detector-facing compatibility is a first-order factor. The direct-transfer experiments clarify the first component of this problem: detector-facing representation matters even before learning is introduced. Among the deterministic representations considered here, the signed voxel representation provides the strongest direct-transfer performance, yet its absolute accuracy remains low. This combination is informative. It shows that an event representation can preserve more transferable structure than other renderings without being sufficiently compatible with the statistics expected by an image-pretrained detector. E2VID addresses the same mismatch from the input side, and at the canonical window it allows the unchanged source detector to recover substantially more target-domain detections than direct event-to-detector transfer. This should not be read as evidence that intensity reconstruction is intrinsically superior to detector adaptation: E2VID introduces an independently pretrained recurrent reconstruction model that remains part of the inference pipeline, whereas ConfMix and SF-YOLO modify the detector through unlabeled target-domain adaptation. The comparison is therefore functional rather than prior- or component-matched.

The additional evaluations show where this benefit stops being general. Improved input compatibility persists on the unseen DSEC thun_02_a family, but reverses on the independent Gen1 dataset, where direct transfer exceeds reconstruction while both adaptation methods remain above it. The repeated reversal on Gen1 under two substantially different detector architectures supports a dataset-dependent interpretation and shows that the effect is not specific to YOLOv5s. Improving the detector-facing input can therefore be highly effective, but its benefit is conditional on the target domain rather than intrinsic to the reconstruction route, and the evaluated methods are better understood as alternative transfer routes than as a universally ordered ranking.

Reaching a useful adapted state and preserving it are different problems. Reporting only the final epoch would substantially understate what the evaluated adaptation methods are able to reach during training. Both ConfMix and SF-YOLO pass through more useful target-domain states before their predefined final checkpoints. For ConfMix the strongest inspected state approaches the canonical E2VID result under mAP@0.5, although E2VID remains stronger under the more stringent mAP@0.5:0.95 criterion. Because these intermediate states are identified using retrospective target-domain evaluation, they are diagnostic rather than deployable unsupervised checkpoints. This separates two questions that final-only reporting conflates. The first is whether detector adaptation can produce a useful target-domain state. The present experiments show that it can. The second is whether that state can be preserved, or recognized without target labels before later training degrades it. The evaluated adaptation methods address the first question more successfully than the second. The central limitation observed here is therefore not insufficient adaptability, but adaptation persistence and unsupervised model selection.

The teacher-signal audit indicates why this is difficult rather than resolving it. The source-initialized teacher does supply usable target predictions after the image-to-event shift, so adaptation does not begin from an uninformative signal; but that supervision is incomplete and self-generated, which is consistent with the pseudo-label signal shaping how the adapted detector evolves. A controlled intervention on the supervisory signal would be required to establish pseudo-label degradation as the cause of the later decrease. The comparison between the two methods is constrained in a similar way: ConfMix retains labeled source supervision while introducing unlabeled target appearance during adaptation, whereas SF-YOLO operates in the stricter source-free teacher-student setting so the stronger ConfMix result cannot be attributed solely to continued source access. What can be concluded is that both forms of adaptation recover useful cross-modal information and both exhibit limitations in maintaining that recovery across the full training schedule.

Temporal support is part of the transfer mechanism, not neutral preprocessing. The temporal experiment further limits any universal interpretation of the reconstruction advantage. Reconstruction performs poorly at the three shorter windows and becomes substantially stronger only at the canonical window, whereas the two adaptation methods show distinct temporal responses: across the three-seed adaptation study, ConfMix attains its highest mean final performance at 50 ms, whereas SF-YOLO attains its highest mean final performance at 20 ms and exceeds its paired 50-ms result in each of the three seeds. These repeated directional differences are evidence of method-dependent temporal sensitivity, but they do not establish an optimal operating point or identify its cause,

because shortening the window simultaneously changes the amount of local event evidence and the temporal coverage of the underlying stream. The reset-state control narrows the explanation without closing it: removing cross-sample recurrent memory does not remove the dependence on temporal support, so the canonical advantage cannot be attributed solely to recurrent information carried across samples. The temporal behavior of E2VID therefore appears to depend jointly on local event support, stream coverage, and recurrent-state propagation.

Transfer difficulty is also uneven within the target task. The final improvements of ConfMix and SF-YOLO are substantially larger for car than for person, and E2VID likewise achieves considerably higher AP for car. Direct transfer, in contrast, remains weak for both classes. The experiments do not isolate whether this asymmetry arises from class frequency, object scale, relative motion, appearance, event structure, or a combination of these factors so it should be read as evidence of class-dependent transfer difficulty rather than of class imbalance alone. At the object level, successfully matched objects tend to exhibit higher local event density than unmatched objects, but matching is not monotonic in density, so local event activity is associated with detectability without being sufficient to determine it.

These three dimensions interact, which is why a single final checkpoint or a single temporal configuration is insufficient for characterizing cross-modal adaptability. They also indicate a direction for future unsupervised domain adaptation. Rather than addressing only the detector input or only the adaptation of high-level predictions, future methods could explicitly reduce the image-to-event discrepancy at multiple levels. Low-level alignment can target the appearance and representation mismatch presented to the detector, while feature-level alignment can encourage modality-invariant task representations. Such a design would ideally be complemented by a target-label-free mechanism for assessing or stabilizing adaptation progress so that useful intermediate states can be preserved without retrospective access to target annotations. The present results do not establish the effectiveness of such a framework, but they motivate multi-level cross-modal UDA as a particularly relevant direction for further investigation.

6- LIMITATIONS

The conclusions of this study should be interpreted within several experimental and methodological limitations.

First, the revised study extends the primary zurich_city_04 evaluation through zero-shot testing on the unseen DSEC thun_02_a family and on the independent Gen1 dataset, and it evaluates the detector-facing transfer and reconstruction conditions with an architecturally distinct Cascade R-CNN detector. The generalization analysis nevertheless uses a single source domain and finalized checkpoints rather than repeated adaptation across multiple source-target pairs. The architecture control moreover covers only the frozen detector-facing conditions: ConfMix and SF-YOLO remain implemented with YOLOv5s, so their adaptation behavior is not established to be architecture invariant, and the two detectors are not capacity matched. The resulting evidence is therefore cross-recording-family, cross-dataset, and partially cross-architecture rather than comprehensive generalization across event-camera sensors, cities, datasets, detector architectures, or adaptation settings.

Second, quantitative evaluation is restricted to the person and car categories. As described in Section III.A, these categories form the only label space shared by all evaluation sets used here, they carry over 90% of the DSEC-Detection annotations, and they are the categories for which the semi-automatic annotation procedure is most reliable. This restriction nevertheless limits direct comparison with studies that report results over the complete eight-class DSEC-Detection label space, and no conclusion of this study should be extended to the less represented DSEC categories, for which transfer behavior may differ from that observed for person and car.

Third, the detector-adaptation comparison covers only two representative families : a source-accessible, confidence-guided mixing strategy and a source-free teacher-student setting. These expose different adaptation constraints but do not exhaust unsupervised or source-free domain-adaptive object detection, and the absence of adversarial, multi-level, distillation-based, and event-specific alternatives from this protocol should not be read as evidence regarding their relative effectiveness.

Fourth, the comparison between detector adaptation and E2VID reconstruction is functional rather than component- or prior-matched: the reconstruction route adds an independently pretrained model that remains in the inference pipeline and is not shared by the adaptation methods. Its stronger 50-ms result should consequently be interpreted as the strongest practical result among the evaluated transfer routes rather than as evidence that reconstruction is intrinsically superior to unsupervised domain adaptation.

Fifth, the stochastic adaptation experiments are evaluated using three independent random seeds: this applies to both ConfMix and SF-YOLO under the canonical 50-ms protocol and at every temporal window in the sensitivity analysis. Direct transfer and E2VID use fixed pretrained models and therefore do not involve adaptation-training variability, whereas the continued source-only and supervised target controls are single-run training references. The three-seed adaptation results characterize run-to-run variability through mean and sample standard deviation, but n = 3 remains insufficient for high-powered inferential statistical claims. We therefore interpret these statistics descriptively rather than as formal significance tests.

Sixth, changing Δt modifies more than the number of events contained within an individual detector input: it also changes how much of the underlying event stream is represented between successive annotated observations and, for E2VID, the temporal evidence supplied to the recurrent reconstruction process. The reset-state control removes recurrent information carried across annotation samples, but it does not separate the amount of local event evidence from the coverage of the underlying stream, so the experiment does not provide a fully isolated causal estimate of temporal-window duration.

Seventh, the fixed intermediate checkpoints used to analyze adaptation dynamics are retrospective diagnostics rather than deployable unsupervised model-selection decisions. Target annotations are used only after training to measure the states reached at predefined epochs; they are not used to update the detector or to select the primary reported checkpoint. The

finding that stronger intermediate target-domain states occur therefore does not provide an operational rule for recognizing them without target labels, and developing such a target-label-free criterion remains an open practical problem.

Eighth, the adaptation-signal analysis is diagnostic rather than causal. The SF-YOLO teacher audit shows that useful target predictions survive the initial image-to-event shift while the accepted pseudo-label set remains incomplete, but it does not establish pseudo-label quality, thresholding, or teacher-student feedback as the cause of the later performance decrease; a controlled intervention would be required to distinguish among these mechanisms.

Finally, the object-level analyses are observational rather than causal. The qualitative results and class-wise analysis indicate difficulty with small, distant, closely spaced, and weakly represented objects, and successfully matched objects exhibit higher median event density than unmatched objects, although matching does not increase monotonically across event-density quartiles. Object scale, distance, motion, occlusion, local contrast, event activity, and class identity covary within natural driving scenes and are not independently manipulated here, so these observations describe observed failure modes and detection difficulty rather than controlled estimates of their individual effects.

These limitations restrict the scope of the numerical conclusions, but they also clarify the questions that remain open : broader cross-dataset validation, more extensive stochastic replication, controlled temporal studies, and explicit investigation of multi-level cross-modal adaptation.

7- CONCLUSION

This study examined how an image-trained object detector can be reused on event-camera data when target-domain detection labels are unavailable for adaptation. Rather than treating image-to-event transfer as a single representation problem, we compared two practical routes: adapting the detector to event-derived inputs, and reconstructing the event stream into an intensity-like representation before applying the unchanged detector, both under a common Cityscapes-to-DSEC protocol and a unified target-domain evaluation pipeline.

Direct transfer of an image-trained detector to event representations remains difficult, and no deterministic rendering closes the modality gap on its own. Both evaluated unsupervised adaptation methods recover useful target-domain states without target labels, and reconstruction provides the strongest label-free result under the canonical DSEC protocol. That ordering is not general. It does not survive a reduction in temporal support, and it reverses on the independent Gen1 dataset, where direct transfer exceeds reconstruction while both adaptation methods remain above it. The additional Cascade R-CNN control further shows that the dataset-dependent relationship between direct event transfer and reconstruction is not specific to the YOLOv5s detector used in the primary experiments.

Two properties of the problem emerged more clearly than any ranking of methods. First, the ability to reach a useful adapted state and the ability to preserve it are distinct: both adaptation methods pass through stronger target-domain states than they retain at their predefined final checkpoints, and recognizing such a state without target annotations remains an open problem rather than a matter of longer training. Second, the temporal window over which events are accumulated is part of the transfer mechanism rather than neutral preprocessing, and it affects the evaluated routes in substantially different ways.

Transfer difficulty is also uneven across object categories, so aggregate metrics alone do not characterize it. The findings therefore motivate future cross-modal UDA approaches that address domain discrepancy explicitly at multiple levels. In particular, combining low-level alignment of image-event appearance statistics with feature-level alignment of task-relevant representations, together with a target-label-free mechanism for monitoring or stabilizing adaptation progress, appears to be a promising direction for improving the robustness of image-to-event object detection.

REFERENCES

[1] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.

[2] G. Mattolin, L. Zanella, E. Ricci, and Y. Wang, "ConfMix: Unsupervised domain adaptation for object detection via confidence-based mixing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 423–433, 2023.

[3] S. Varailhon, M. Aminbeidokhti, M. Pedersoli, and E. Granger, "Source-free domain adaptation for YOLO object detection," in *Computer Vision – ECCV 2024 Workshops*, vol. 15640 of *Lecture Notes in Computer Science*, pp. 218–235, Springer, 2025.

[4] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1964–1980, 2021.

[5] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5633–5643, 2019.

[6] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "DSEC: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.

[7] D. Gehrig and D. Scaramuzza, "Low-latency automotive vision with event cameras," *Nature*, vol. 629, no. 8014, pp. 1034–1040, 2024.

[8] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11457–11466, 2019.

[9] J. Kay, T. Haucke, S. Stathatos, S. Deng, E. Young, P. Perona, S. Beery, and G. Van Horn, "Align and distill: Unifying and improving domain adaptive object detection," *Transactions on Machine Learning Research*, 2025. Featured Certification.

[10] L. Zhang, W. Zhou, H. Fan, T. Luo, and H. Ling, "Robust domain adaptive object detection with unified multi-granularity alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9161–9178, 2024.

[11] Q. Chu, S. Li, G. Chen, K. Li, and X. Li, "Adversarial alignment for source free object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 452–460, 2023.

[12] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3339–3348, 2018.

[13] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6956–6965, 2019.

[14] X. Zheng and L. Wang, "EventDance: Unsupervised source-free cross-modal adaptation for event-based object recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17448–17458, 2024.

[15] M. Jeryo and A. Harati, "Learning domain-invariant representations for event-based motion segmentation: An unsupervised domain adaptation approach," *Journal of Imaging*, vol. 11, no. 11, p. 377, 2025.

[16] C. Xie, W. Gao, and R. Guo, "Cross-modal learning for event-based semantic segmentation via attention soft alignment," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2359–2366, 2024.

[17] D. Rossi, P. Vasseur, F. Morbidi, C. Demonceaux, and F. Rameau, "Event-aware distilled DETR for object detection in an automotive context," in *2025 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1849–1855, 2025.

[18] J. Zhang, Z. Li, J. Lin, and G. Lu, "Adaptive event stream slicing for open-vocabulary event-based object detection via vision-language knowledge distillation." arXiv preprint arXiv:2510.00681, 2025.

[19] Y. Xie, S. Ye, C. Wang, J. Xu, L. Shen, Y. Qian, and J. Qian, "Time-step mixup for efficient spiking knowledge transfer from appearance to event domain." arXiv preprint arXiv:2509.12959, 2025.

[20] V. Polizzi, D. B. Lindell, and J. Kelly, "REALM: An RGB and event aligned latent manifold for cross-modal perception." arXiv preprint arXiv:2605.00271, 2026.

[21] H. Liu, G. Yu, H. Cao, S. Qu, F. Lu, Y. Zhong, Z. Lu, L. Leng, and G. Chen, "I2EKD: Efficient and versatile image-to-event knowledge distillation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 1, pp. 292–303, 2026.

[22] J. Cao, J. Xing, N. Messikommer, and D. Scaramuzza, "Generative event pretraining with foundation model alignment." arXiv preprint arXiv:2603.23032, 2026.

[23] D. Hannan, R. Arnab, G. Parpart, G. T. Kenyon, E. Kim, and Y. Watkins, "Event-to-video conversion for overhead object detection," in *2024 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pp. 89–92, 2024.

[24] V. Mechler and P. Rojtberg, "Transferring dense object detection models to event-based data," in *Advanced Intelligent Virtual Reality Technologies*, vol. 330 of *Smart Innovation, Systems and Technologies*, pp. 35–47, Springer Singapore, 2023.

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[26] G. Jocher, A. Chaurasia, A. Stoken, *et al.*, "ultralytics/yolov5: v7.0 – YOLOv5 SOTA realtime instance segmentation," 2022. Version v7.0.

[27] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6154–6162, 2018.

[28] K. Chen, J. Wang, J. Pang, et al., "MMDetection: Open MMLab detection toolbox and benchmark." arXiv preprint arXiv:1906.07155, 2019.

[29] C. Fu, J. Xiao, W. Yuan, R. Liu, and X. Fan, "Learning cruxes to push for object detection in low-quality images," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 12, pp. 12233–12243, 2024.