2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Object-centric Cross-modal Feature Distillation for Event-based Object Detection

Lei Li[1,2], Alexander Linger[1], Mario Millhäusler[1], Vagia Tsiminaki[1], Yuanyou Li[1] and Dengxin Dai[1]

***Abstract*— Event cameras are gaining popularity due to their unique properties, such as their low latency and high dynamic range. One task where these benefits can be crucial is real-time object detection. However, RGB detectors still outperform event-based detectors due to the sparsity of the event data and missing visual details. In this paper, we propose a cross-modality feature distillation method that can focus on regions where the knowledge distillation works best to shrink the detection performance gap between these two modalities. We achieve this by using an object-centric slot attention mechanism that can iteratively decouple feature maps into object-centric features and corresponding pixel-features used for distillation. We evaluate our novel distillation approach on a synthetic and a real event dataset with aligned grayscale images as a teacher modality. We show that object-centric distillation allows to significantly improve the performance of the event-based student object detector, nearly halving the performance gap with respect to the teacher.**

I. Introduction

Event cameras asynchronously measure brightness changes at independent pixel [12], which stands in contrast to standard cameras that measure the brightness within a fixed time interval. These two measurement principles result in two sensors with different characteristics and advantages. Event cameras due to their asynchronous nature have sub-millisecond latency, no motion blur, and a high dynamic range (up to 140dB) [12]. In contrast, regular cameras can measure the absolute intensity information, resulting in richer more detailed images. Given how close the two modalities are in their representation, both are based on pixels, the question naturally arises, *how can we leverage one modality to improve the other?*

In this paper, we study the task of object detection with event cameras [36]. This is a task where the low latency of event cameras can help with fast moving objects, which helps to drastically improve reaction times in safety critical tasks such as autonomous driving. However, object detection with event cameras also has disadvantages, first, datasets for event camera based detection [36] are not as mature as those for frame based cameras [2], [17], [31], [42]. Second, the missing global illumination information results in inferior appearance features. Given these two disadvantages, frame based detectors generally perform better than event based detectors. To reduce this performance gap and maintain the property of event cameras, we investigate how cross-modal knowledge distillation during training can help to improve event-based detectors. We will show that this results in increased performance *without* the need for frame-based information at inference time.

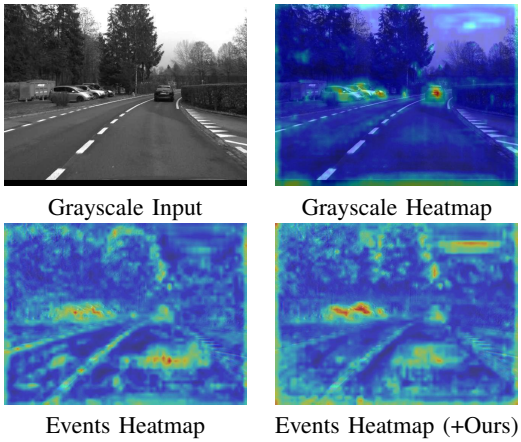

Fig. 1: Visualization of the spatial heat maps of the largest scale FPN feature in the YOLOX detector [13] using grayscale image and event frame as inputs. With our distillation model, foreground features are enhanced and background noise (lane markings) is weakened.

Knowledge distillation [21] is one of the most common tools to transfer knowledge between models. However, this is often done from a heavy to a lighter model [10], [19], [21], [25], [35], [47], [48]. We are interested in transferring knowledge between input modalities, which comes with its own set of challenges [7], [9], [22]. For cross modality knowledge distillation from event to frame cameras the main issue is that the event data is much sparser, thus it does not help to distill the information of the whole image but we propose to rather focus on the foreground objects that should be detected. The reason to focus on foreground objects can be directly understood when investigating the differences between the features generated from a detector using either events or image frames, shown as a heatmap in Fig. 1. It is clearly visible that event frames and grayscale images exhibit different pixel-level reactions. In event frames, objects appear as boundaries which is harder to distinguish from the background, while in RGB/grayscale images, they are displayed as shapes and textures. Thus, to distillate features between event frames and images, it is important to avoid to imitating the whole teacher's features directly at the pixel level. An additional problem is that due to the ego-motion, background areas can trigger a large number of events, additionally complicating the distillation of background regions.

Thus, to get the best features to distill, we propose a novel distillation approach that uses slot attention [32] to

*Work done during internship at Huawei Zurich Research Center (ZRC).
[1] Huawei ZRC, Switzerland. firstname.lastname@huawei.com
[2] ETH Zurich, Switzerland. lllei.li0386@gmail.com

iteratively decouple feature maps into object-centric features and corresponding pixel features that allows to best distill the knowledge from the frame based teacher to the event based student. Our algorithm aligns the Feature Pyramid Network [30] (FPN) output features between the student and the teacher and is thus applicable to a large number of modern detection algorithms. We show that our algorithm helps both one-stage (YOLOX [13]) and two-stage detectors (Faster R-CNN [40]).

In summary, our contributions are threefold. First, we propose a novel slot attention based feature distillation algorithm that can iteratively decouple feature maps into object-centric features and corresponding pixel features for distillation. Second, we show that the attention matrices from the slot attention mechanism can be used as learning based mask for attention aided feature alignment and that considering the relation information between object-centric features can be exploited during distillation. Third, we show that the proposed distillation algorithm allows for improving the performance of event based detection algorithms, halving the performance gap to grayscale detectors.

II. Related Work

A. Object Detection

The addition of deep neural networks has significantly improved object detection over the last decade. Modern detectors are broadly classified into one-stage and two-stage detectors. One-stage detectors [13], [26], [37]–[39], [52] generate predictions directly on the feature map and generally come with the advantage of faster run times. In contrast, two-stage detectors [3], [40] typically employ a region proposal network to generate coarse object bounding boxes, which are then refined in a second stage. Recently, major advances in object detection have focused on anchor-free detectors [13], [26], [52], label assignment strategies [50] and end-to-end detectors [4].

In the sub-field of real-time detection, the engineering-heavy YOLO family [1], [13], [27], [39] integrates recent advancements in detection technologies, always seeking the best speed and accuracy trade-offs for real-time applications. Due to the low latency of event cameras, these types of detectors are of special interest, as they allow to generate very high rate predictions if combined with asynchronous event cameras. In this paper, we utilize YOLOX [13], which uses an anchor-free detection head and a sophisticated label assignment strategy.

By using a standard detector based on a Convolutional Neural Network (CNN), we follow the most common approach for event based object detection [6], [16], [24], [36]. Identical to ours, the idea is to first represent the event stream as a dense "frame", which then can be processed by a CNN. These methods are often coupled with recurrent network architectures [16], [28], [36] to deal with the sparse event data. An alternative approach to dense CNNs, is to process the events asynchronously using Graph Neural Networks (GNN) to achieve incremental detection [14], [29], [41], which is more aligned with the characteristics of event cameras. However, comes with limitations and normally requires a heavy downsampling of the event stream, thereby losing crucial information.

B. Knowledge Distillation for Object Detection

In recent years, knowledge distillation has been shown to be an effective approach for reducing the complexity of object detection models while maintaining high accuracy. It is usually classified into three types, response distillation [21], feature distillation [10], [19], [25], [47], [48], and label assignment distillation [35]. Response distillation [21], [51] uses the final response distribution of the teacher model as soft labels for the student. Label assignment distillation [35] leverages the teacher model to generate label assignments for the student. Most distillation methods perform feature distillation, which works directly on feature maps. Defeat [19] highlights the importance of region selection for distillation. ICD [25] proposes a learnable conditional decoding module that uses each object instance as a query to distill the feature map with dot-product attention FGD [47] focuses on balancing foreground and background objects in spatial and channel-wise feature spaces. MGD [48] utilizes a masked auto-encoder to align the feature space between the student and teacher.

Different from classical knowledge distillation for object detection, there exists a significant domain gap between grayscale images and event frames, making it challenging to directly distill information between these two modalities. Intuitively, high-level knowledge, such as object information, has the potential to alleviate the modality gap between grayscale images and event frames. To bridge the modality gap, recent studies have explored single-image-to-event translation techniques for feature representation [33], [45], segmentation [43], and pre-training [46]. These studies have shown that task performance on events can be significantly improved by leveraging information from grayscale images. However, the use of direct image-to-event translation for object detection tasks has not been extensively investigated. In contrast, cross-modality knowledge distillation has been utilized in other modalities to transfer knowledge, such as between RGB images and pseudo depth maps [9], and between RGB images and point clouds [7], [22]. Note that these methods have similar data requirements to ours as they need aligned cross-modal data during training but only use the student modality at inference.

III. Method

A. Overview

Our cross-modality feature distillation pipeline consists of two main blocks, pixel-feature distillation, and object-feature distillation, which are illustrated in Fig. 2. Specifically, our approach distills knowledge from the Feature Pyramid Network [30] (FPN) output features, which enables us to apply the approach to both one-stage and two-stage detectors. Our method generates teacher features by passing grayscale images through a fully trained and frozen detection network. Meanwhile, our student model takes event voxel girds [12]

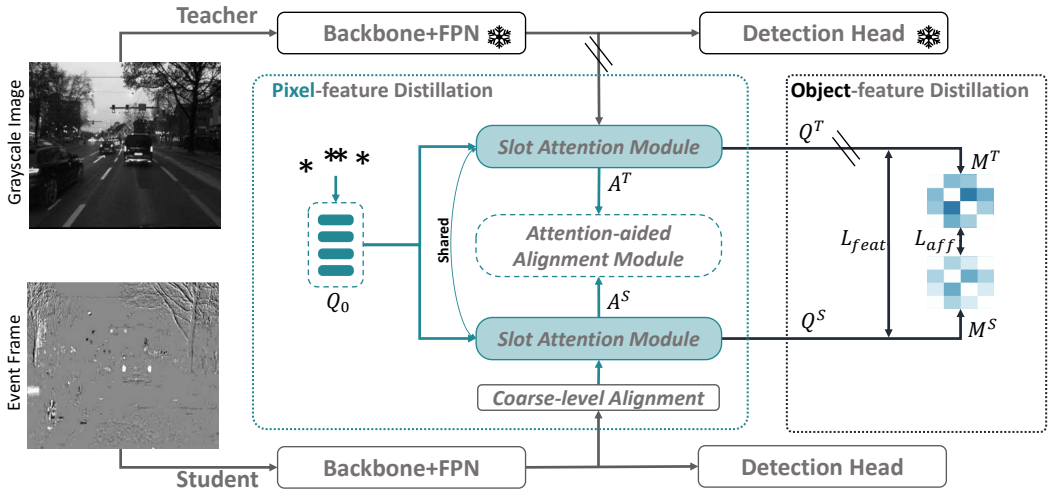

Fig. 2: **Framework overview**. Our distillation model comprises two components: pixel-feature distillation and object feature distillation. The auxiliary task module is not drawn in the pixel-feature distillation.

as input but otherwise uses an identical neural network. In the following sections, we will discuss each module in more detail and conclude with the combined loss function.

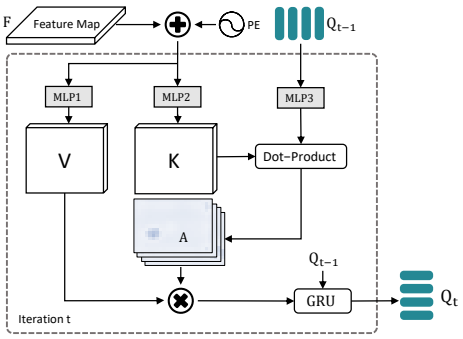

Fig. 3: An illustration of the slot attention module in the t-th iteration.

B. Input Representations

Given the asynchronous event stream $e_i = \{(x_i, y_i, p_i, t_i)\}_{k=1}^{N}$, which consists of the x and y pixel locations, $\{-1, 1\}$ polarity and time step respectively. Our method takes a time window of duration Δt and represents the event stream into a voxel grid representation following the approach in [54]. The resulting voxel grid has a size of $B \times H \times W$, where B is the number of temporal bins, and H and W are the height and width of the event frame, respectively. We also normalize the voxel grid values to the range $[0, 1]$ after clipping the pixel values to a fixed range to ensure a consistent input representation. The advantage of this event representation is that it can be efficiently implemented with a rolling buffer for real-time applications. As for the teacher network input, we use grayscale images to avoid introducing further differences between the two modalities since events are also grayscale.

C. Object-centric Feature Extraction

To learn well-suited object-centric features, we use slot attention [32]. The slot attention allows us to formulate an object-centric approach, with one slot feature per object. This idea is closely related to our task, where we try to find features of object instances (distinct entities) that align well across modalities. Besides, this module tolerates minor alignment discrepancies between modalities, enhancing method robustness.

As illustrated in Fig. 3, the input to the slot attention module is the FPN feature F, which is augmented with a positional embedding and then flattened into vectors. The resulting feature vector is denoted as $\overline{F}$, with shape of $HW \times D$. The second set of inputs consists of the initial query vectors, Q_0, derived from the features at L object center locations to enhance the efficiency of iterative updates. Each query vector is D-dimensional, resulting in $Q_0 \in \mathbb{R}^{L \times D}$.

At the t-th step, the query $q_{t-1} \in \mathbb{R}^{L \times D}$, key $K \in \mathbb{R}^{HW \times D}$, and value $V \in \mathbb{R}^{HW \times D}$ are generated via three distinct linear transformations,

$$q_{t-1} = Q_{t-1} \cdot W^Q, \quad K = \overline{F} \cdot W^K, \quad V = \overline{F} \cdot W^V, \quad (1)$$

where W^Q, W^K, $W^V \in \mathbb{R}^{D \times D}$. Given the key, query, and value, we compute the output of the module using an attention update. Slot attention coefficients are normalized across the slot dimension [32]. This fosters competition among object slots for explaining different input segments, and also allows the computation of the attention matrix,

$$\text{attn}_{i,j} = \frac{e^{Mi,j}}{\sum_l e^{M_{i,l}}}, \quad \text{with} \quad M = \frac{K \cdot q_{t-1}^T}{\sqrt{D}}, \quad (2)$$

where attn $\in \mathbb{R}^{HW \times L}$. The input values and corresponding query vectors are aggregated based on their co-attention weights, calculated as a weighted mean,

$$U = A^T \cdot V, \quad \text{with} \quad A_{i,j} = \frac{\text{attn}i, j}{\sum_l \text{attn}l, j}, \quad (3)$$

where $U \in \mathbb{R}^{L \times D}$. In the final step T, the query slots Q_{t-1} are updated with a Gated Recurrent Unit [8] (GRU) to generate the input slot query for the next step $Q_t = \text{GRU}(Q_{t-1}, U)$.

D. Pixel Feature Distillation

Coarse-level Feature Alignment. To alleviate the modality gap between grayscale images and event frames, we first introduce a shallow network $\mathcal{G}_c$ that maps student features $F^{\mathcal{S}}$ to an intermediate latent space $\widetilde{F}^{\mathcal{S}}$. Mapping the student feature to the latent space forces the student model to adapt to the coarse feature distribution of the teacher. All distillations are performed between the newly mapped student feature $\widetilde{F}^{\mathcal{S}}$ and the teacher feature $F^{\mathcal{T}}$.

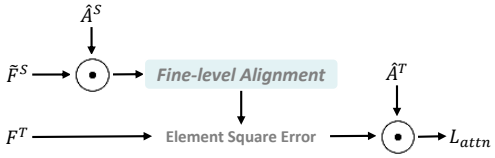

Fig. 4: An illustration of the attention-aided fine-level feature alignment module.

Attention-aided Fine-level Feature Alignment. During the slot attention module, the attention maps of the teacher and the student $\text{attn}^{\mathcal{T}}$ and $\text{attn}^{\mathcal{S}}$, learned to focus on the important region of the corresponding objects by interacting with the slot features. Thus, they contain crucial spatial information about the importance of the feature map for all objects. Technically, our attention-aided fine feature alignment module takes the latest attention maps and reshapes them into attention matrices for each object, $\mathcal{A}_i^{\mathcal{T}}$ and $\mathcal{A}_i^{\mathcal{S}}$, where the subscript i refers to the ith object. These attention matrices are then used to mask the FPN feature maps $\widetilde{F}^{\mathcal{S}}$ and $F^{\mathcal{T}}$. As illustrated in Fig. 4, we first mask the student's feature map using the student attention map,

$$\tilde{F}_i^{\mathcal{S}} = \hat{\mathcal{A}}_i^{\mathcal{S}} \odot \widetilde{F}^{\mathcal{S}} . \tag{4}$$

Where $\odot$ is the Hadamard product, and the output is a masked feature map per object. "hat" denotes the gradient detach. Given the masked student feature, we pass them through a fine-level feature alignment module $\mathcal{G}_f$, which is similar to [48]. Given this refined masked student feature, we compute the feature alignment loss with the teacher feature map using element squared error ℓ_{ese}. Thus, we can compute the attention-aided fine feature alignment loss as,

$$L_{\text{attn}} = \sum_i^L \left(\ell_{ese} \left(\mathcal{G}_f(\tilde{F}_i^{\mathcal{S}}), F^{\mathcal{T}} \right) \odot \hat{A}_i^{\mathcal{T}} \right) / \sum_i^L \hat{A}_i^{\mathcal{T}} , \tag{5}$$

where the element wise square error is masked with the teacher's attention map.

E. Object Feature Distillation

Direct Object-centric Feature Distillation. The slot attention module generates a slot feature for each ground truth object for the teacher $Q_{T,i}^{\mathcal{T}}$ and the student $Q_{T,i}^{\mathcal{S}}$, where the subscript i refers to the ith object. To explicitly encourage the student model to mimic the teacher at a high level, we formulate an ℓ_1 loss between the student and teacher slot feature. Thus, the feature distillation loss can be written as

$$L_{\text{feat}} = \frac{1}{L} \sum_i^L \ell_1 \left(Q_{T,i}^{\mathcal{S}}, \hat{Q}_{T,i}^{\mathcal{T}} \right) , \tag{6}$$

where L is the number of ground truth objects and slot features, and the "hat" $\hat{Q}$ indicates that the gradient is stopped, ensuring the student model has cleaner gradients.

Object-centric Relation Distillation. Given that student features can never be perfectly aligned with teacher features, exploiting the inter-object relations can further facilitate the distillation process. Inspired by [23], we construct an affinity matrix, denoted as $\mathbf{M}$, to represent the relationships between object-centric features. This affinity matrix quantifies the similarity between each pair of object-centric features. Specifically, the entry of the affinity matrix is calculated as follows

$$\mathbf{M}(m,n) = \frac{{Q_m}^\top \cdot Q_n}{\|Q_m\|_2 \cdot \|Q_n\|_2} , \tag{7}$$

where m, n denote the m-th and n-th object-centric features, respectively. To encourage the student model to learn the teacher's inter-object relationships, we apply a Mean Square Error (MSE) loss between the student's affinity matrix, $\mathbf{M}^{\mathcal{S}}$, and the teacher's affinity matrix, $\mathbf{M}^{\mathcal{T}}$,

$$L_{\text{aff}} = \ell_{mse} \left(\mathbf{M}^{\mathcal{S}}, \hat{\mathbf{M}}^{\mathcal{T}} \right) . \tag{8}$$

Auxiliary Task. The slot attention module allows us to generate both informative slot features and attention maps that are the core building blocks of our distillation approach. However, the distillation module risks converging to trivial solutions due to the under-constrained teacher's slot feature. To avoid this issue, we introduce two auxiliary tasks to keep the teacher's slot feature informative [25]. We use a classification task that is supervised using Binary Cross Entropy (BCE) and a bounding box size regression task that is supervised by an L1 loss, resulting in the following cost function,

$$L_{\text{aux}} = \frac{1}{L} \sum_i^L \ell_{bce}(C_i^{\mathcal{T}}, C_i^{gt}) + \frac{1}{L} \sum_i^L \ell_1(B_i^{\mathcal{T}}, B_i^{gt}) , \tag{9}$$

where C^{gt} and B^{gt} denotes ground-truth classification and bounding box center location. $C_i^{\mathcal{T}}$ and $B_i^{\mathcal{T}}$ are the outputs of the auxiliary MLP, $\mathcal{G}_{\text{aux}}(Q^{\mathcal{T}})$.

F. Overall Loss

The total loss of the student model can be written as

$$L_{\text{total}} = L_{\text{det}} + \lambda_1 L_{\text{feat}} + \lambda_2 L_{\text{attn}} + \lambda_3 L_{\text{aff}} + \lambda_4 L_{\text{aux}} , \tag{10}$$

where L_{det} is the original detection loss of the student and $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are hyperparameters to balance the losses.

IV. Experiments

A. Dataset Setting

We evaluate our method on two datasets. The first is a synthetic dataset generated with the CARLA Simulator [11], recorded across seven CARLA towns under good weather conditions, with one town designated for the validation sequence. This CARLA dataset comprises 29,600 labeled frames, and contains three classes: vehicle, pedestrian, and two-wheeler. The second dataset is the real-world event

dataset DSEC [15]. We generate bounding boxes by annotating the RGB images using Cascade-Faster R-CNN [3], trained on BDD100k [49]. The RGB detection results are then projected to the event camera using a mean depth assumption, which works well due to the small baseline between the cameras. Similarly, to align images between the two modalities, we warp the RGB images to the event camera given a mean depth. Note that we label the DSEC training split, which consists of 53 driving sequences, and use a subset of 5 sequences as a validation set. The resulting DSEC dataset has 52,645 labeled frames and contains seven classes: pedestrian, rider, car, truck, bus, bicycle, and motorcycle.

Note that for other real-world datasets, a similar approach can be used to align the RGB and event data, as long as the event and RGB cameras are mounted close to each other. This dataset requirement is fulfilled for a large number of existing event datasets [5], [15], [34], [44], [53], which contain both RGB and event data.

B. Implementation Details

We use a normalized voxel gird as our input representation, in the case of CARLA we use a time window of $\Delta t = 50$ms with $B = 5$ temporal bins. For DSEC we use a larger temporal window of $\Delta t = 100$ms with $B = 10$ temporal bins, thus in both cases each temporal bin contains 10ms of event data. In our distillation module, $\mathcal{G}_c$ consists of two 3×3 convolutional layers with ReLU activation, $\mathcal{G}_f$ comprises three 3×3 convolutional layers with ReLU activation, and $\mathcal{G}_{\text{aux}}$ includes four-layered MLPs. The positional embedding in the slot attention module is learned through two-layer MLPs with FPN features F as inputs. We set $T = 3$ iterations for the slot attention module.

For our baseline detector, we use YOLOX [13], a fast one-stage anchor-free detector, since it well represents modern real-time detectors [1], [13], [27], [39]. Furthermore, to show that our distillation approach generalizes to other detector architectures, we use Faster R-CNN [18] which is an anchor-based two-stage detector. For YOLOX, we use the "small" version, which uses 9M parameters and achieves run times of below 10ms in PyTorch eager mode and less than 2ms using TensorRT on an Nvidia 3090. This allows for real-time inference using a rolling buffer in our event representation with time bins of 10ms. For Faster R-CNN we use a ResNet-50 [20] as backbone.

The set of λ in Eq. 10 is set to $\{0.1, 1.0, 0.002, 1.0\}$ for YOLOX and $\{0.01, 0.1, 0.0002, 0.1\}$ for Faster R-CNN. All detectors are trained using SGD for 60 epochs, with a momentum of 0.937 and a weight decay of 0.0005, and initialized with COCO pre-trained model weights. A cosine learning rate schedule with a 5-epoch warm-up period is utilized, and a learning rate of 0.000075 per sample in the batch is applied. The batch size is set to 32 for YOLOX and 16 for Faster R-CNN. Our data augmentation strategy, which includes random horizontal flips, and contrast and brightness changes, is consistently applied to both images and event frames. All experiments are performed on a single Nvidia RTX 3090 GPU.

TABLE I: Comparison with SOTA methods on the CARLA val set. † denotes without response distillation.

Method	mAP	AP_{50}	AP_{75}
YOLOX-S (Teacher-GS)	59.6	81.1	70.8
YOLOX-S (Student-Event)	53.8	72.6	58.4
+ ICD [25]	53.1 (-0.7)	73.7	60.5
+ FGD [47]	53.8 (+0.0)	75.9	61.2
+ MGD [48]	54.7 (+0.9)	73.2	60.8
+ MonoDistill† 1001[9]	54.1 (+0.3)	72.5	61.7
+ Ours	56.4 (+2.6)	76.4	63.2
Faster R-CNN (Teacher-GS)	49.6	71.6	57.5
Faster R-CNN (Student-Event)	41.8	63.4	45.7
+ ICD [25]	41.6 (-0.2)	62.0	46.5
+ FGD [47]	42.7 (+0.9)	64.0	46.4
+ MGD [48]	40.1 (-1.7)	60.7	44.3
+ MonoDistill† 1001[9]	38.6 (-3.2)	58.8	43.2
+ Ours	44.0 (+2.2)	63.4	49.5

C. Main Results

We compare our novel cross-modality distillation method with four state-of-the-art distillation methods for object detection, including three single modality distillation methods, ICD [25], FGD [47], MGD [48], and one cross-modality distillation method, MonoDistill [9]. For all methods, we perform feature distillation on the multi-scale FPN features. The results of our comparison experiment on the CARLA dataset are presented in Table I. The majority of the state-of-the-art methods result in a decrease in the performance of the student network. However, our distillation method significantly outperforms the other four methods. The student detector with our distillation module demonstrates significant mAP improvement. Our method results in 2.6 mAP and 2.2 mAP improvements on YOLOX and Faster R-CNN, respectively. When comparing this to the grayscale teacher model, our distillation method is able to nearly half the performance gap to the teacher model.

TABLE II: Comparison with SOTA methods on the DSEC val dataset. All student models are updated by Exponential Moving Average.

Method	mAP	AP_{50}	AP_{75}
YOLOX-S (Teacher-GS)	47.5	65.7	53.3
YOLOX-S (Student-Event)	27.4	42.9	29.0
+ ICD [25]	28.1 (+0.7)	44.6	29.9
+ FGD [47]	29.8 (+2.4)	46.5	31.9
+ MGD [48]	28.8 (+1.4)	45.9	30.8
+ MonoDistill† 1001[9]	29.4 (+2.0)	46.3	30.7
+ Ours	30.8 (+3.4)	49.3	32.4

In Table II we compare our distillation method on the DSEC dataset. In the case of DSEC, the grayscale teacher model is clearly better than the event based detector. We attribute this to the auto-labeling, which potentially misses objects that are hard to detect in the RGB images, giving frame based detectors an advantage. When investigating state-of-art-methods, we can see that distillation is definitely beneficial for event based detectors, with all methods improving the detection performance. However, our method results clearly the largest improvement with an increase of 3.4 mAP.

D. Ablation Studies and Analysis

To evaluate the individual components of our proposed method, we conduct a series of ablation experiments using the YOLOX detector on the CARLA dataset.

TABLE III: Ablation on feature distillation components.

Coarse	Fine+Direct	Relation	Auxiliary	mAP	AP_{50}	AP_{75}
✗	✗	✗	✗	53.8	72.6	58.4
✓	✓	✗	✓	55.3	77.1	63.0
✗	✓	✗	✓	52.0	73.1	61.9
✗	✓	✗	✗	37.9	64.7	39.7
✓	✓	✓	✓	56.4	76.4	63.2

Module Component. To gain a better understanding of our model component, we conduct ablation studies on the various components. Starting with a baseline mAP of 53.8 without distillation, we add a base module comprising of coarse-level feature alignment, attention-aided fine-level feature alignment, direct object-centric feature distillation, and auxiliary task, resulting in a 1.5 mAP improvement. Removing the coarse-level feature alignment module from the basic module causes a significant 3.3 mAP drop in performance, demonstrating the crucial role of this module in aligning event and grayscale features. Furthermore, removing the auxiliary task causes the model's performance to drop drastically to 37.9 mAP, indicating that without the auxiliary task, our object-centric feature extraction module learns a shortcut resulting in a trivial problem. Finally, adding object-centric relation distillation to our basic module improves performance by 1.1 mAP, showing that incorporating relational information between instance features can help bridge the modality gap and enhance performance.

TABLE IV: Ablation on attention types.

Attention Type	mAP	AP_{50}	AP_{75}
-	53.8	72.6	58.4
Full Region	54.8 (+1.0)	76.5	62.0
FG Region	55.0 (+1.2)	77.0	60.6
FGD Full	53.8 (+0.0)	75.9	61.2
FGD FG	54.6 (+0.8)	72.8	60.8
Ours	55.3 (+1.5)	77.1	63.0

Attention Type. To investigate the different attention types, we test several alternatives to our method that use the attention matrix of the slot attention module, see Table IV. "Full Region" is the simplest approach, which distills the entire feature map only using a coarse alignment network. "FG Region" improves the previous approach by only distilling the Foreground (FG) regions using the ground truth bounding boxes as masks. We also employ an advanced distillation method based on FGD [47]. Similar to the two previous methods, we use original FGD over the full feature maps "FGF Full" and over the foreground regions "FGD FG". Ours is the one without object-centric relation distillation.

We can see that focusing on the foreground regions generally performs better independent of the method, validating our argument that foreground features align better between the two modalities. We argue that this is the case because foreground objects contain more salient and discriminating features compared to background regions, and hence, foreground regions should be given more attention during feature distillation. Conversely, background regions may introduce noise and affect the distillation performance adversely.

Object-centric Relation Distillation. To better understand our object-centric relation distillation, we visualize the affinity matrix between slot features. In Fig. 5, we see that the affinity matrix generated by the teacher model exhibits high discrimination, while the affinity matrix from the student model appears to be sparse and noisy. With. relation distillation module, the student model produces a more structured affinity matrix that closely corresponds to the teacher. This indicates that our affinity distillation module enables a better transfer of structural knowledge from teacher to student.

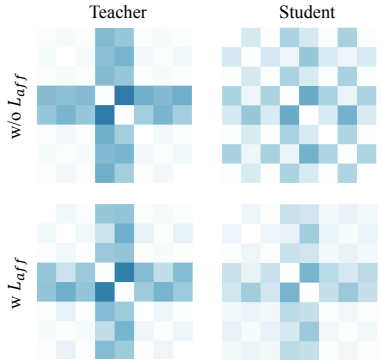

Fig. 5: Visualization of the affinity matrix without and with object-centric relation loss. Affinity values are cosine similarity scores between object-centric features.

Teacher Modality. To validate our decision to use grayscale images as the teacher modality, we also perform an ablation using RGB images. Using RGB images improves the teacher to 62.7 mAP on CARLA compared to 59.6. However, the distillation does not work well and the performance is even false below not using distillation to 51.1 mAP.

TABLE V: Model compression task on the DSEC dataset.

Method	YOLOX	mAP	AP_{50}	AP_{75}
Teacher(GS)	S	47.5	65.7	53.3
Student(GS)	Tiny	43.5	61.0	50.1
+ ICD [25]	Tiny	45.7 (+2.2)	65.7	51.3
+ FGD [47]	Tiny	45.8 (+2.3)	64.7	53.3
+ MGD [48]	Tiny	45.4 (+1.9)	64.5	50.8
+ MonoDistill† 1001[9]	Tiny	45.9 (+2.4)	65.1	51.6
+ Ours	Tiny	46.0 (+2.5)	64.4	51.8

E. Model Compression

Although our method is not designed for model compression, it is still interesting to investigate the detection performance within a single modality. Surprisingly, as demonstrated in Table V, our method performs competitively, with the best mAP improvement, showing that the method has also the potential for model compression.

V. Conclusion

In this paper, we introduced a novel knowledge distillation approach that leverages slot attention to reduce the detection performance gap between events and images. Our method distills object-centric slot features with corresponding attention maps from grayscale images to events, rather than directly mimicking image features. We conducted extensive experiments on both synthetic and real event datasets, demonstrating that our approach significantly enhances event-based detection performance. Moreover, we show that our proposed method boosts the single-modality model compression task.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2, 5

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2, 5

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[5] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4015–4022, June 2023. 5

[6] Nicholas FY Chen. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 644–653, 2018. 2

[7] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022. 1, 2

[8] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014. 3

[9] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*, 2022. 1, 2, 5, 6

[10] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 1, 2

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 4

[12] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1, 2

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 2, 5

[14] Daniel Gehrig and Davide Scaramuzza. Pushing the limits of asynchronous graph-based object detection with event cameras. *arXiv preprint arXiv:2211.12324*, 2022. 2

[15] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 5

[16] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. *arXiv preprint arXiv:2212.05598*, 2022. 2

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[18] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 5

[19] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021. 1, 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2

[22] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 87–104. Springer, 2022. 1, 2

[23] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022. 4

[24] Massimiliano Iacono, Stefan Weber, Arren Glover, and Chiara Bartolozzi. Towards event-driven object detection with off-the-shelf deep learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018. 2

[25] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34:16468–16480, 2021. 1, 2, 4, 5, 6

[26] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[27] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 2, 5

[28] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. 2

[29] Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Graph-based asynchronous event processing for rapid object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 934–943, October 2021. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[32] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 1, 3

[33] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 2

[34] Anton Mitrokhin, Chengxi Ye, Cornelia Fermüller, Yiannis Aloimonos, and Tobi Delbruck. Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6105–6112. IEEE, 2019. 5

[35] Chuong H Nguyen, Thuy C Nguyen, Tuan N Tang, and Nam LH Phan. Improving object detection by label assignment distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1005–1014, 2022. 1, 2

[36] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020. 1, 2

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[38] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[39] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 5

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[41] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12371–12381, 2022. 2

[42] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1

[43] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 341–357. Springer, 2022. 2

[44] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 5

[45] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2

[46] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. *arXiv preprint arXiv:2301.01928*, 2023. 2

[47] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 1, 2, 5, 6

[48] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022. 1, 2, 4, 5, 6

[49] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 5

[50] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 2

[51] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9407–9416, 2022. 2

[52] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[53] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 5

[54] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3