PEPR: Privileged Event-based Predictive Regularization for Domain Generalization

Supplementary Material

Overview

In the supplementary material, we provide additional technical details, extended experiments, and qualitative results to further support the findings presented in the main paper. Specifically, we include:

- **Sec. A**: a detailed description of the training setup, architectural components, and implementation choices.
- **Sec. B**: additional qualitative results for detection are discussed here, highlighting the proposed method qualitites.
- **Sec. C**: in-domain evaluation on DSEC-DET, to demonstrate the effect of PEPR on data drawn from the same distribution as the training data.
- **Sec. D**: discussion regarding the qualitative results obtained under diverse adverse domains for the Cityscapes Adverse dataset.
- **Sec. E**: discussion of limitations and future works.

A. Implementation Details

In this section, we provide a detailed description of the experimental setup, architectural choices, and hyperparameter configurations adopted to evaluate PEPR on the tasks of Object Detection and Semantic Segmentation. Both for the detection and the segmentation task, the predictor g_ϕ is a transformer decoder with a depth of 4 and 8 attention heads. We attach the source code as supplementary material.

Object Detection. For the detection task, we employ the DETR architecture [6]. We utilize the implementation provided by the `transformers` Python library, initializing both the RGB and Event encoders with weights pretrained on the COCO dataset [31]. Event data are converted into Time Surface representations [65] and then processed by the model. The two encoders and the RGB decoder are then fine-tuned end-to-end during training. The loss of Eq. 2 is computed with the outputs of the encoders for the two modalities. At test time, only the RGB branch is used, discarding the event part. The model is optimized using AdamW [35] with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4}. Training is conducted for 20 epochs.

Semantic Segmentation. For the segmentation task, we adopt SegFormer [56] as our backbone, utilizing the implementation provided by MMSegmentation [13]. In this case, the predictive loss of Eq. 2 is computed using features from the last block of the encoder of dimension 64. To generate the privileged information for the Cityscapes dataset, we simulate event streams using the ESIM simulator [39] in conjunction with FILM frame interpolation [40], following the video-to-event pipeline established in [23]. These simulated events, as well as the real events for DSEC, are converted into Time Surface representations [65]. The model is trained using the AdamW optimizer with a learning rate of 6×10^{-5}, a weight decay of 0.01, and betas set to $(0.9, 0.999)$. We employ a `poly` learning rate scheduler with a power of 1.0, a warmup ratio of 1×10^{-6}, and 1500 warmup iterations. Following the original SegFormer protocol, we use separate learning rate multipliers for the decoder head. The model is trained for a maximum of 40 epochs with an early stopping patience of 7 epochs.

B. Qualitative Results Object Detection

We report additional qualitative results. In Fig. 4 we compare outputs of the standard RGB DETR model compared with the L2 baseline and PEPR. Detections are shown in green, while ground truth boxes are represented in blue. Similarly, in Fig. 5, we show outputs for Hard-DSEC-DET. For both datasets, it can be seen how PEPR strengthens the model into detecting difficult objects, in challenging conditions. The L2 baseline also often provides an improvement, yet still falls behind in the most difficult scenarios.

C. DSEC-DET Results

Alongside the hard test split, DSEC-DET also contains an in-domain split, where train and test scenarios are balanced domain-wise, resulting less challenging for RGB-based models. In fact, the difficulty in this split is not explicitly linked to light intensity or weather of the scene. Nonetheless, as a control experiment, we report the results obtained by PEPR also on this in-domain test split. Results are presented in Tab. 9. Interestingly, as observed for FRED

Model	Train Mod.	Test Mod.	$\text{mAP}_{50:95}$	mAP_{50}
EA-DETR [42]	RGB+E	RGB+E	14.7	27.2
DETR [6]	E	E	12.0	25.8
DAGr [20]	E	E	14.0	-
Faster-RCNN [41]	RGB	RGB	18.2	35.4
RetinaNet [32]	RGB	RGB	16.6	30.5
CenterNet [64]	RGB	RGB	10.4	35.1
YOLOv7-E6E [54]	RGB	RGB	18.2	31.5
YOLOv5-L [24]	RGB	RGB	20.9	33.2
DETR [6]	RGB	RGB	27.7	50.6
DETR_{L2} (*Ours*)	RGB+E	RGB	29.1 (*+1.4*)	50.3 (*-0.3*)
DETR_{PEPR} (*Ours*)	RGB+E	RGB	28.3 (*+0.6*)	**52.2** (*+1.6*)

Table 9. Object detection results on DSEC-DET in-domain split.

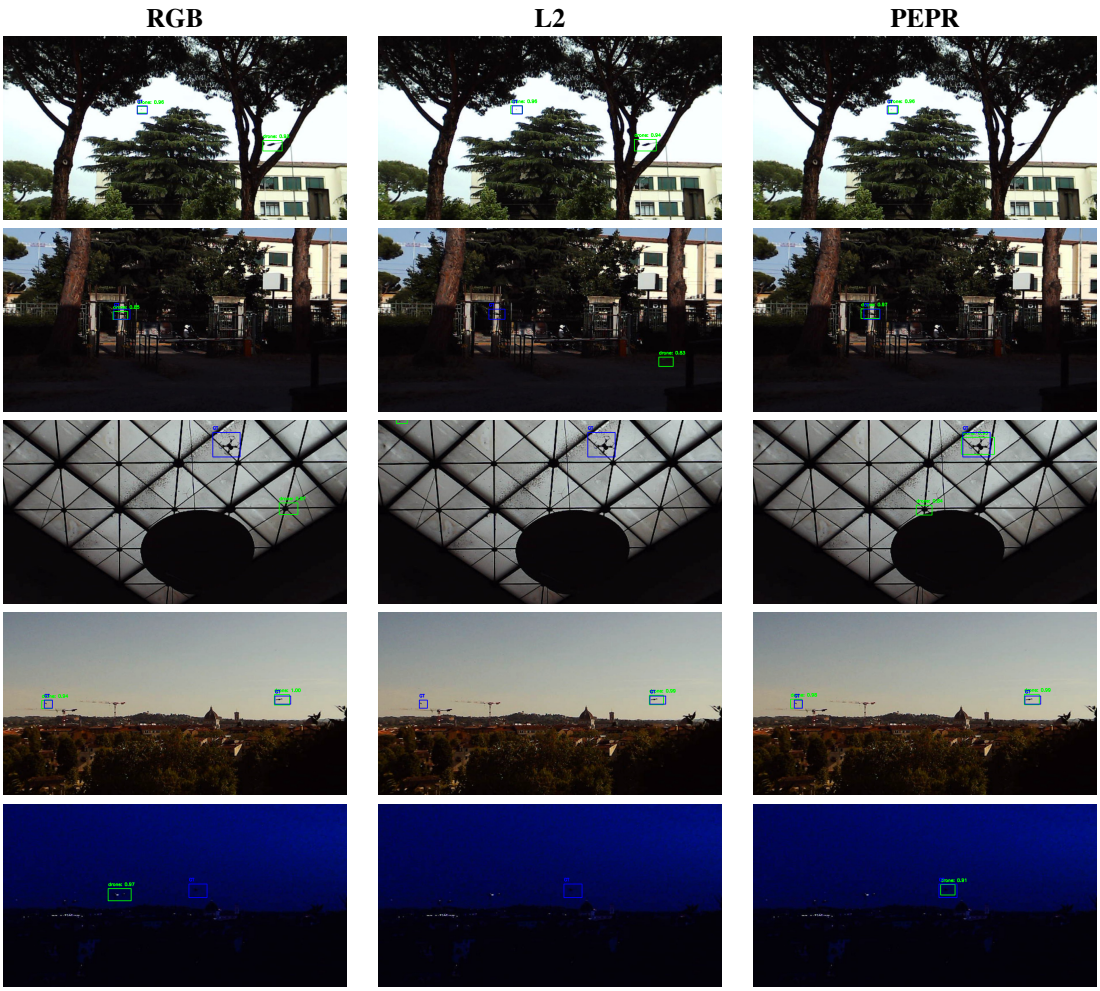

Figure 4. Qualitative results on FRED Challenging. Detections are shown in green, ground truth in blue.

Canonical in Tab. 4 of the main paper, adopting event data as privileged information also improves robustness when domain shifts are not present, with PEPR improving on the L2 baseline.

D. Qualitative Results Semantic Segmentation

In Fig. 6 we show segmentation results for Cityscapes Adverse. We highlighted details of interest with yellow circles, where the improvements obtained by PEPR are most noticeable. Interestingly, PEPR is capable of recovering errors in challenging conditions, such as sky at dawn, wet road and snowy vegetation.

E. Limitations and Future Work

While PEPR shows strong empirical performance, our study also has limitations that suggest natural directions for future research. Our experiments mainly target appearance shifts such as day-to-night and related changes in illumination and weather. We also observe gains in several adverse conditions, but there are regimes where the event signal can become noisy or less informative (for example, under rain or flickering neon lights), which may reduce the benefit of event-based supervision. In such cases, the robustness of PEPR is inherently bounded by the quality of the event stream.

Moreover, we currently treat events as the only privileged modality. This is a natural choice in scenarios where events remain relatively stable under domain shift, but it may be suboptimal when the event sensor itself is strongly affected by domain shift. In those settings, other sensing modalities (such as thermal imaging) or combinations of multiple privileged signals could provide a more reliable training signal than events alone. In addition, our implementation relies on a relatively simple frame-based representation of events. While this keeps the framework easy to integrate into existing architectures, it may not fully exploit the fine-grained spatio-temporal structure of the event

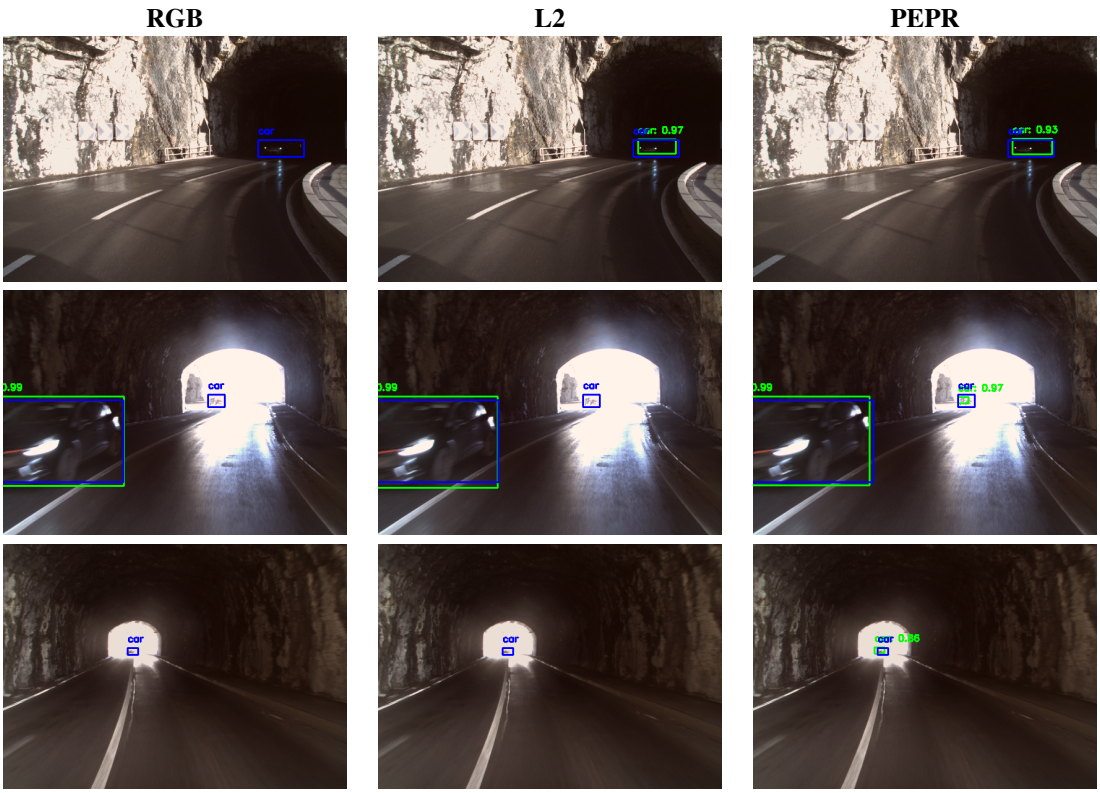

Figure 5. Qualitative results on Hard-DSEC-DET. Detections are shown in green, ground truth in blue.

stream. Alternative encodings, such as voxel grids or point-cloud–like representations, could provide richer supervision signals and further strengthen the predictive regularization effect.

These considerations point to several promising research directions. One avenue is to extend our predictive framework beyond domain generalization to settings such as (semi-)supervised domain adaptation and unsupervised domain adaptation, where unlabeled or sparsely labeled target data are available during training. Another is to explore multi-modal privileged supervision, leveraging multiple predictive targets (e.g., events and thermal) that can compensate for each other in challenging conditions where a single modality is unreliable. Finally, investigating more expressive event representations within our predictive framework is an interesting direction for better exploiting the temporal dynamics of event cameras in real-world deployments.

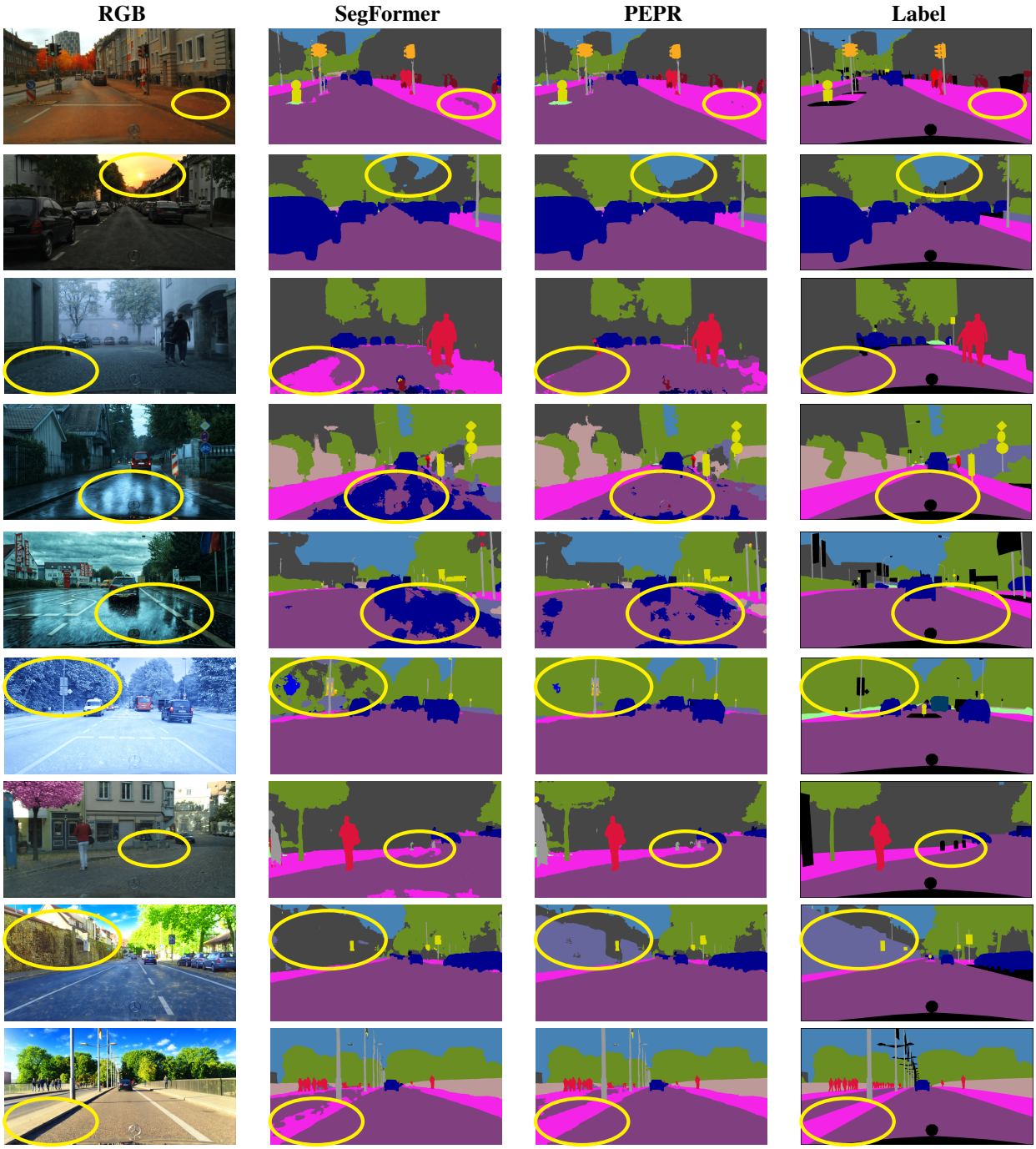

Figure 6. Qualitative results on Cityscapes Adverse. Details of interest are highlighted with yellow circles.