# Cross-Modal Filters for RGB-Neuromorphic Wormhole Learning

**Alessandro Zanardi**[1], Andreas Aumiller[1], Julian Zilly[1], Andrea Censi[1], Emilio Frazzoli[1]
[1]**Institute for Dynamic Systems and Control, ETH Zurich**

## 1 Introduction

The technique of **"wormhole learning"** [1] shows that even **temporary access to a different sensor** with complementary invariance characteristics **can be used to enlarge the operating domain of an existing object detector** without the use of additional training data.

## 2 Wormhole Learning

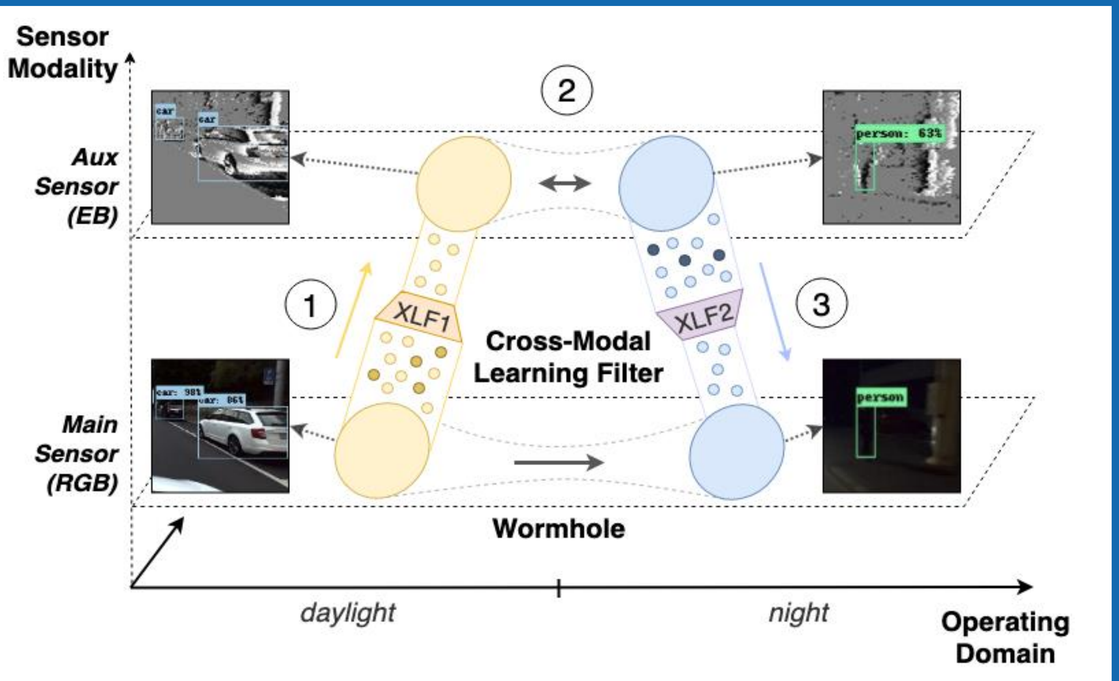


**Fig. 1.** The principle of wormhole learning is illustrated.
It begins with a detector that works only in a limited region of the operating environment, in this case, the bottom left circle (RGB detector at day time). The ultimate goal is to make it **more robust with respect to a certain nuisance**, hence, expanding its operating envelope. In this case, the nuisance is the ambient illumination. The goal is achieved in a **completely automated fashion with the temporary addition of an auxiliary sensor**. In the following the **three key points** are presented:

1. Starting from the bottom left, **transfer learning** is applied to learn a detector in the auxiliary domain. As training labels we adopt the inferences generated by the RGB detector. **The result is an even-based object detector**.
2. From there, **the inherent invariance of the new domain to the specific nuisance of illumination** allows us to "travel" across the operating domain. The event-based object detector is now able to **perform inference also at night**.
3. In the final step the student becomes the teacher, and symmetrically to step one, we retrain the RGB detector at daytime from the event-based inferred labels at night. We can now remove the auxiliary sensor and **we are left with an RGB detector that works both at day and night**.

In order to cope with sensors that are radically different, such as RGB cameras and event-based neuromorphic sensors, we need a more careful selection of which samples to transfer. Thus **we design "cross-modal learning filters" which represents a first step in the relatively unexplored territory of multi-modal observability.**

## 3 Cross-Modal Filters

**Wormhole Learning Algorithm recap:**

1) Obtain samples $\{z\}$ in a domain where the initial detector $\mathrm{p}(\mathrm{Y}|Z_a)$ is accurate
2) Use $z_a^k$ to generate the labels $y^k$ from the initial detector
3) Use the pair $\langle z_b^k, y^k \rangle$ to learn $\mathrm{p}(\mathrm{Y}|Z_b)$
4) Once $\mathrm{p}(\mathrm{Y}|Z_b)$ is learned we proceed in reverse in the new domain

Legend:
Data $Z$
Task $Y$
Sensor $a, b$


**Multi-modal observability!**
$\mathrm{p}(\mathrm{Y}|Z_b) \neq \mathrm{p}(\mathrm{Y}|Z_a)$

We generalize the wormhole learning algorithm by introducing **cross-modal learning filters (XLFs)**, which are functions of the type:
$xlf_{a\to b}$: $Z_a \times Z_b \times Y \to Bool$

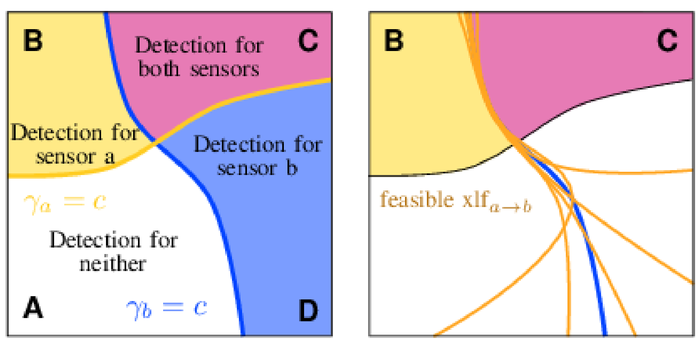





**Fig. 2** Given a detection in one domain, **cross-modal filters tell us whether or not that label should be used as training sample for the other domain**. Note that this is **not the same as a detector since it is conditioned on having had a detection** on one sensor. Graphically, among the four mutually exclusive cases we are reconstructing the decision boundary only between region B and C.

## 4 RGB- Neuromorphic Results

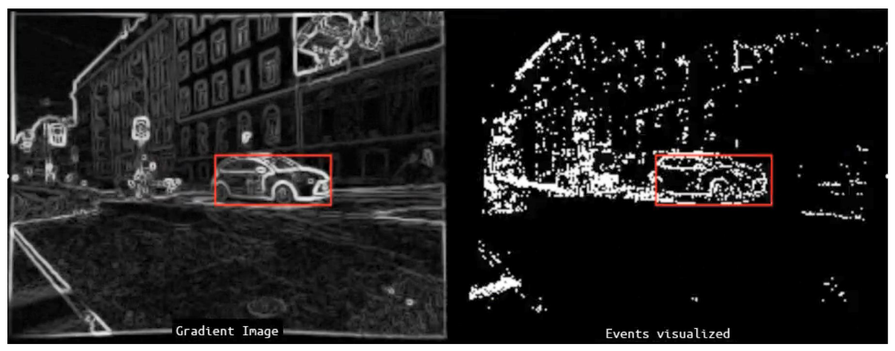


**Fig. 3.** To compute a cross-modal filter between RGB and event-based camera it is sufficient to check whether or not the generation model for the two sensors is respected. Events are expected to appear at edge location and vice-versa. Thus, we introduce an **edge overlap score ($S_{eo}$) to quantify how much the two models are fulfilled**. This can be **consider a proxy evaluating how much the objects are observable** in both domains. In the figure above the gradient of the RGB image is pointwise compared with the events' activity.
Hence, we have $\boldsymbol{xlf_{a\to b} = S_{eo} > \underline{S_{eo}}}$ where $\boldsymbol{S_{eo} = \frac{\sum_{BOX}\sqrt{|I_{EB}|}|\nabla I_{RGB}|}{\sum_{BOX}\sqrt{|}|\nabla I_{RGB}|}}$

**Table 1**
Detection performance in mAP@0.5IoU

| Testset | RGB (D) | Event-based | RGB (D+N) | Relative Gain |
|---|---|---|---|---|
| **Day (D)** | 59.1 | 26.2 | 58.1 | **-2%** |
| **Night (N)** | 32.2 | 16.2 | 41.5 | **+29%** |

**Table 2**
Detection performance of **event-based detector as a function of the XLF** threshold [mAP@0.5IoU]

| Testset | None | Low threshold | High threshold |
|---|---|---|---|
| **Day** | 20.5 | **26.2(+27.5%)** | 22.5(+9.4%) |
| **Night** | 8.23 | 16.2(+97.1%) | **18.6(+126%)** |

- Same pattern observed in [1] with RGB-IR, **we compromise a bit of performance at day** (-2%) **to significantly improve at night** (+29%).
- We observe a **positive gain in spite of the middle modality being not particularly apt to the task**.
- Tab. 2 shows the **effectiveness of XLF** for the first transfer learning step. Instead, we did not experience particular need for any filter when going from events to RGB.

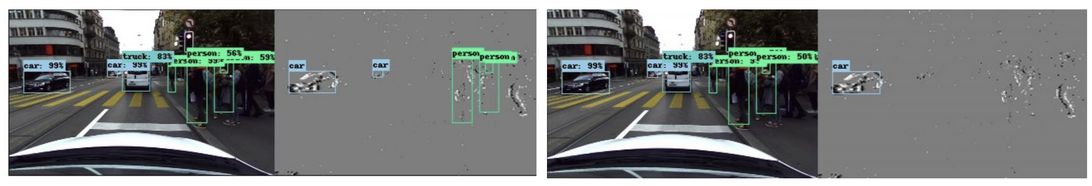

**Fig. 4.** On the left the **effect of XLF** with low threshold, on the right with high threshold. Increasing the threshold progressively removes objects that appear less observable.

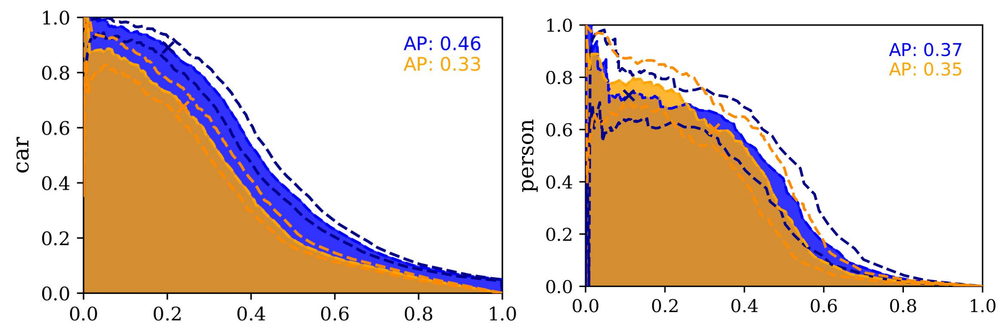


**Fig. 5. Recall-precision** curves for category *car* and *person*. In **blue** the detector after WHL, in **orange** the detector before WHL. For *cars* we can appreciate a **significant improvement in performance at night** that exceeds the 95% confidence bound (dashed-line). On the other hand, *persons* do not exhibit an equally remarkable improvement, we could speculate that the representation of a car changes more when switching from day to night than a person which does not exhibit new features such as headlights and shimmering lights.

## 5 Conclusion

These results on wormhole learning show that **there are many creative ways to combine the data from heterogeneous sensors**, and an additional sensor can be useful, even if you only have it during training, and even if it is not particularly good at the task at hand, or, equivalently, even if we do not know how to use it well for the task at hand.
The results suggest that we are still in the **early days of multi-modal perception** and many questions are still to be answered.

## 6 References

1. **A. Zanardi, J. Zilly, A. Aumiller, A. Censi, E. Frazzoli, "Wormhole learning," IEEE International Conference on Robotics and Automation (ICRA), 2019.**
2. **A. Zanardi, A. Aumiller, J. Zilly, A. Censi, E. Frazzoli, "Cross-Modal Learning Filters for RGB-Neuromorphic Wormhole learning," Robotics: Science and Systems. (RSS), 2019.**