Supplementary Material — From Chaos Comes Order: Ordering Event Representations for Object Recognition and Detection

Nikola Zubić* Daniel Gehrig* Mathias Gehrig Davide Scaramuzza
Robotics and Perception Group, University of Zurich, Switzerland

1. Appendix

Here we add additional qualitative results and proofs to support the work in the main manuscript. We will refer to sections, equations, figures, and tables in the main manuscript with the prefix "M-", while referring to those in the appendix with "A-". We start by providing additional details about ERGO-12 and GWD in Sec. A-1.1 and include two proofs regarding the robustness of Gromov-Wasserstein Discrepancy (GWD) in Sec. A-1.2. Afterward, we provide more results with fewer optimized channels in Sec. A-1.3. Finally, we show the qualitative results of our method on the Gen1 and 1 Mpx datasets in Sec. A-1.4.

1.1. Additional Details on ERGO-12 and GWD

ERGO-12 details: We provide more details of our optimized representation in Fig. A-2. As can be seen from the top sub-figure, we show the optimized channels in more detail than in Figure M-7. At each new step, there is a decrease in GWD, which demonstrates that additional channels reduce the distance. We calculated GWD on the Gen1 [1] validation dataset, which contained 100 samples, and plotted the results as dashed horizontal lines for chosen representations. The blue line shows the performance of the optimization process after each channel addition. We can observe that, for example, our optimized representation outperforms the Voxel Grid after seven channels and MDES after nine channels. Furthermore, we found that the optimization process initially selected the time function, which capitalizes on the high temporal resolution of event cameras to minimize GWD. Subsequently, counts and polarity were used.

In the bottom sub-figure of Fig. A-2, we visualize the channels of ERGO-12 (our optimized representation after 12 channels). For visualization, we min-max normalized the channels within the range of 0-255. Each channel emphasizes different parts of the image. For instance, the last channel highlights the left edges of the pedestrian, while the seventh channel emphasizes the right part. Our optimization process enables us to capture as much information as possible at different scales and resolutions (spatial and temporal), which is highly advantageous when training with common object detectors. The optimized representation achieves an mAP of over 50% on the Gen1 dataset, and it represents the first non-recurrent neural network architecture that scores over 40% mAP on the 1 Mpx [2] dataset.

Mathematical properties of the GWD: The GWD introduced in [3] and used in this work does not satisfy all axioms of a distance measure and is thus not a metric. It is a generalization of the GW *Distance* that is specifically designed for spaces where an L2 metric comparison is not suitable, as in this work where we compare raw events and representations. [3] showed that using KL-divergence (Eq. 9) with the kernel in Eq. 7 can effectively discard outliers, which we leverage in our work. Due to this more general formalism, the GW Discrepancy does not satisfy symmetry, or the triangle inequality (due to the KL-Divergence in Eq. 9), but ensures non-negativity, and is 0 only for equal sets. Absolute scalability is also not satisfied (see Eq. 7), but is not a common property of distance measures.

1.2. Invariances of the GWD for Events

In this section, we will go over some basic properties of the GWD for events. In particular, we will show that it is invariant to affine feature transformations, concatenation with a constant, and duplication of the features. For clarity, we repeat here the definition of the GWD for events, following Eq. M-5:

$$L(\mathcal{E}, \mathcal{F}) = \min_{T} \sum_{i,j,k,l} T_{ij} T_{kl} \mathcal{L}(C^e_{ik}, C^f_{jl}) \tag{1}$$

with similarity matrices for Eqs. M-7 and M-8.

$$C^e_{ik} = e^{\frac{-\|e_i - e_k\|^2}{2h^2\sigma_e^2}}, \quad C^f_{jl} = e^{\frac{-\|f_{\mathbf{x}_j} - f_{\mathbf{x}_l}\|^2}{2h^2\sigma_f^2}} \tag{2}$$

$$\sigma_e^2 = \underset{i<j}{\text{mean}} \|e_i - e_j\|^2, \quad \sigma_f^2 = \underset{i<j}{\text{mean}} \|f_{\mathbf{x}_i} - f_{\mathbf{x}_j}\|^2. \tag{3}$$

*Equal contribution

Affine transformation: We expect that if we apply an affine transformation to the event representation, the score should not change since information in the representation remains distinctive. Moreover, we do not want the GWD to be sensitive to the scale of the feature. We see that replacing representation features with

$$f^*_{\mathbf{x}} = a f_{\mathbf{x}} + b \tag{4}$$

changes only the similarity matrix C^f_{jl} to

$$C^{f,*}_{jl} = \exp\left(\frac{-\|f^*_{\mathbf{x}_j} - f^*_{\mathbf{x}_l}\|^2}{2h^2\sigma^{2,*}_f}\right). \tag{5}$$

We see that the norms and data-dependent variances then transform as

$$\|f^*_{\mathbf{x}_j} - f^*_{\mathbf{x}_l}\|^2 = a^2\|f_{\mathbf{x}_j} - f_{\mathbf{x}_l}\|^2 \tag{6}$$

$$\sigma^{2,*}_f = a^2\sigma^2_f \tag{7}$$

We thus see that

$$C^{f,*}_{jl} = \exp\left(\frac{-\|f^*_{\mathbf{x}_j} - f^*_{\mathbf{x}_l}\|^2}{2h^2\sigma^{2,*}_f}\right) \tag{8}$$

$$= \exp\left(\frac{-a^2\|f_{\mathbf{x}_j} - f_{\mathbf{x}_l}\|^2}{2h^2a^2\sigma^2_f}\right) \tag{9}$$

$$= \exp\left(\frac{-\|f_{\mathbf{x}_j} - f_{\mathbf{x}_l}\|^2}{2h^2\sigma^2_f}\right) \tag{10}$$

$$= C^f_{jl} \tag{11}$$

which shows that the similarity matrix does not change. The minimizer of Eq. M-5 thus also does not change, which means the GWD is invariant to this affine transformation. This invariance is only possible through the use of a data-dependent variance, and thus highlights its advantage.

Invariances to Concatenation In the case of concatenation, we consider the following transformation:

$$f^*_{\mathbf{x}} = [f_{\mathbf{x}}\|c_{\mathbf{x}}] \tag{12}$$

where $[.\|.]$ denotes concatenation, and $c_{\mathbf{x}} \in \mathbb{R}^C$ denotes a pixel dependent additional feature. Again, we find that only the similarity matrix C^f_{jl} is affected, and in particular, only the norm and variance, which become:

$$\|f^*_{\mathbf{x}_j} - f^*_{\mathbf{x}_l}\|^2 = \|f_{\mathbf{x}_j} - f_{\mathbf{x}_l}\|^2 + \|c_{\mathbf{x}_j} - c_{\mathbf{x}_l}\|^2 \tag{13}$$

$$\sigma^{2,*}_f = \sigma^2_f + \underset{i<j}{\text{mean}}\|c_{\mathbf{x}_i} - c_{\mathbf{x}_j}\|^2 \tag{14}$$

We will consider two special cases: $c_{\mathbf{x}} = c$, a constant vector, and $c_{\mathbf{x}} = f_{\mathbf{x}}$ the same feature. In the first case, the additional terms above become 0, meaning that the norm does not change, and thus the metric stays the same. In the second case, the norm transforms as in the affine case, multiplying the squared norm and variance by 2. For the same reasons as before, the metric also stays the same. Generalizing this result to more general $c_{\mathbf{x}}$ remains future work.

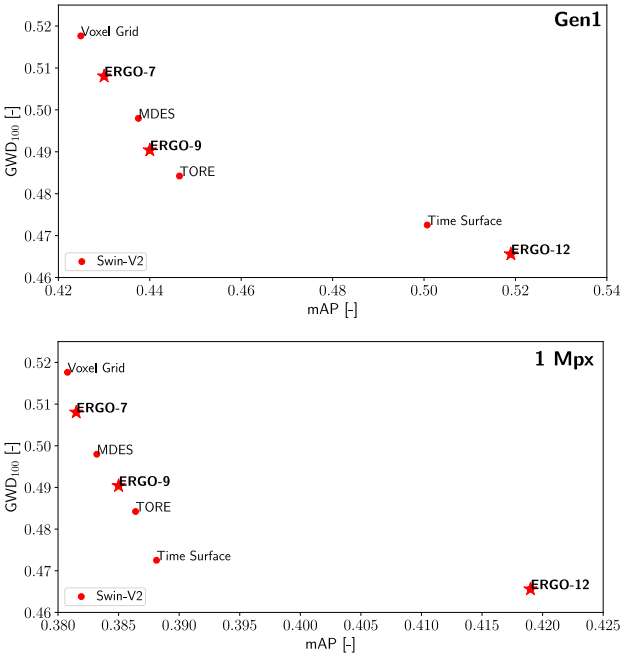

Figure 1: Correlation of the Gromov-Wasserstein Discrepancy with the mAP (higher is better) for object detection on Gen1 [1] (top) and 1 Mpx [2] (bottom) datasets. ERGO-12, ERGO-9, and ERGO-7 represent our optimized representations with twelve, nine, and seven channels. The mAP is reported on the validation set, while the Gromov-Wasserstein Discrepancy is reported on the Gen1 validation dataset with 100 chosen samples.

1.3. Fewer optimized channels

Figure 1 depicts a correlation between the GWD (given on the x-axis, computed on the Gen1 validation dataset with 100 samples) and the task performance (mAP on object detection task). Since the Swin V2 backbone outperforms all other backbones, it is the only backbone shown in the plot, and the 2D Histogram, which is the poorest-performing method, is omitted. The results demonstrate that our optimized representation with nine and seven channels performs better than MDES and Voxel Grid, respectively, which is consistent with the findings in Figure 2. Furthermore, we observe that the results on 1 Mpx correlate with GWD computed on the Gen1 validation dataset with 100 samples, which highlights the generalization capabilities of our approach.

1.4. Qualitative results

We present qualitative object detection results on the 1 Mpx and Gen1 datasets in Figs. 3 and 4, respectively. Our approach exhibits the ability to detect objects that are not present in the ground truth.

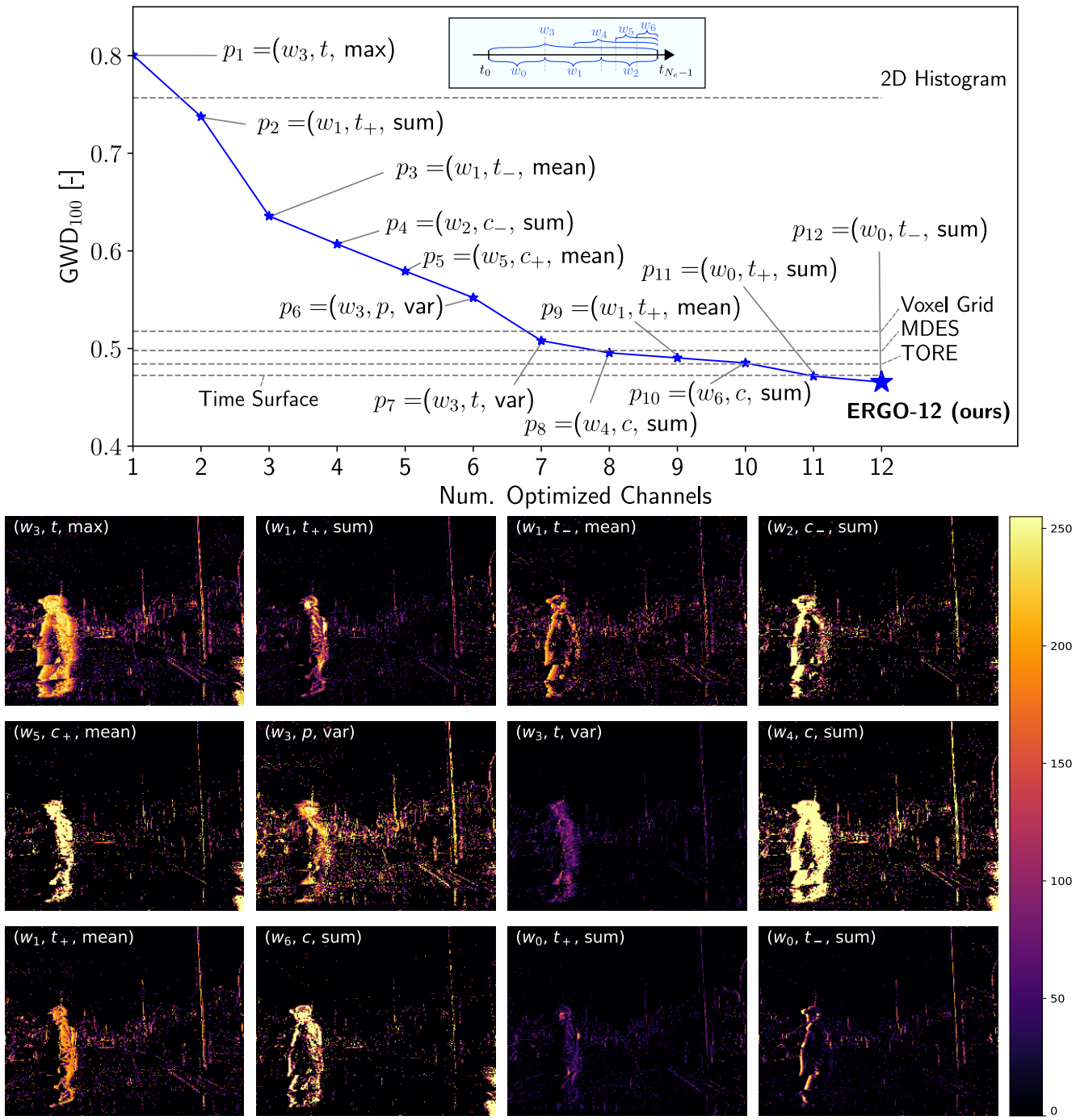

Figure 2: Visualization of the channels of ERGO-12, min-max normalized in the range 0-255. The channels are ordered in row-major order, and the hyperparameters selected are shown in the top left of each subfigure.

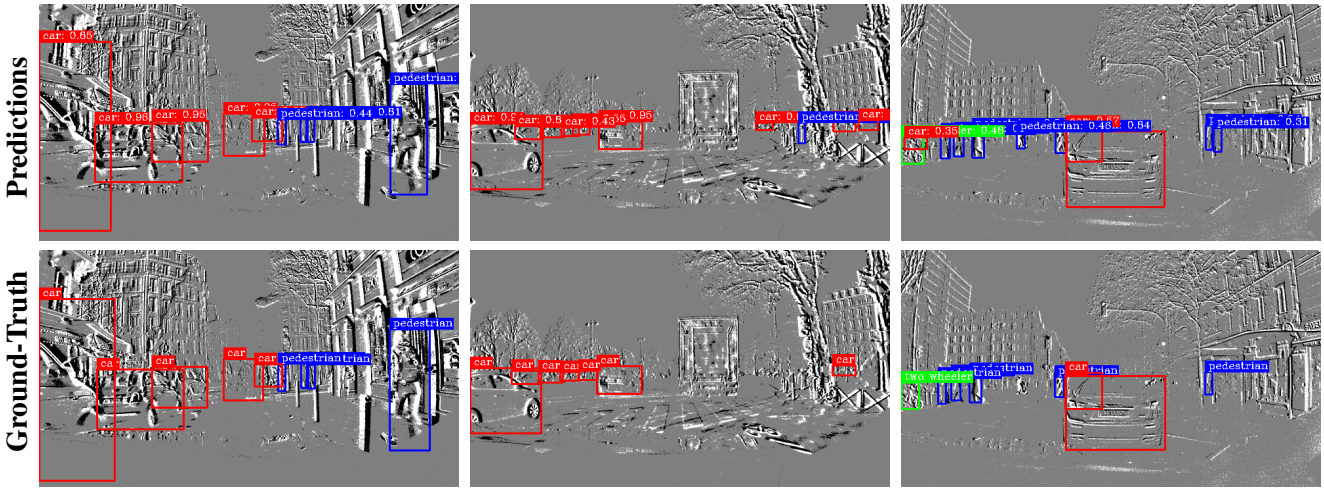

Figure 3: Qualitative results of our method with ERGO-12 input on the 1 Mpx [2] dataset. (top row) predictions, and (bottom row) ground truth. Note that sometimes our method detects objects that do not appear in the ground truth.

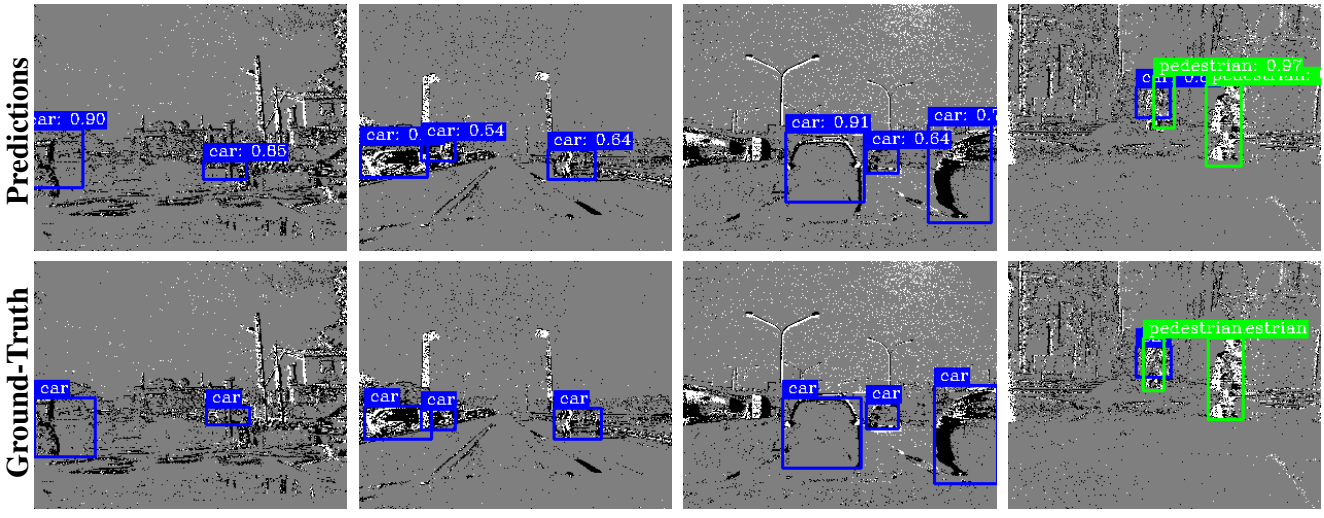

Figure 4: Qualitative results of our method with ERGO-12 input on the Gen1 [1] dataset. (top row) predictions, and (bottom row) ground truth. Note that sometimes our method detects objects that do not appear in the ground truth.

References

[1] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv e-prints*, abs/2001.08499, 2020. 1, 2, 4

[2] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020. 1, 2, 4

[3] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *ICML 2016*, June 2016. 1