Supplementary Material for Learned Event-based Visual Perception for Improved Space Object Detection

Code Available at:
https://unifieddatalibrary.com/sfm/?expandedFolder=/SupportingData/MISS/CognitiveEBS/v.1.0.0
Requires account creation at https://unifieddatalibrary.com/storefront

Appendix

0.1 Training

For overall hybrid model training, we optimize the loss function used for YOLOv3 training.

$$\begin{aligned} L_{YOLOv3} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} [(t_x - \hat{t}_x)^2 + \\ (t_y - \hat{t}_y)^2 + (t_w - \hat{t}_w)^2 + (t_h - \hat{t}_h)^2] \\ + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} [-log(\sigma(t_o)) + \sum_{k=1}^{C} BCE(\hat{y}_k, \sigma(s_k))] \\ + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{noobj} [-log(1 - \sigma(t_o))] \end{aligned} \tag{1}$$

As described in Equation (1), the loss function optimizes the sum of squared errors on the predicted bounding box dimensions, x/y coordinate and width/height transforms, as well as applying a binary cross-entropy loss to class probabilities. Loss is only calculated for objects appearing in the ith cell, referring to the 8x8 regions in the output prediction volume, and with the corresponding jth bounding box predictor, indicated by $1_{i,j}^{obj}$, where the jth bounding box predictor is the bounding box with the highest IoU calculated with respect to the ground-truth bounding box in cell i. Harshly penalizing the confidence scores of bounding boxes containing no ground-truth objects can lead to training instability, which motivates the inclusion of the parameters λ_{coord} and λ_{noobj}. For these parameters, we use the values of $\lambda_{coord} = 5$ and $\lambda_{noobj} = 0.5$ as presented in the original YOLO work. Due to the small size of targets in our context, we omit the multiple scales at which the original YOLOv3 architecture detects bounding boxes, instead opting to predict bounding boxes at only the smallest scale. In addition, we also perform unsupervised pretraining of each point-cloud model (PointNet, PointNet++, and PointConv), where a series of deconvolutional layers are used to reconstruct the initial point-cloud and trained using a mean squared error loss across the reconstructed point features. The hybrid framework training, and point-cloud model pretraining, was performed on SatSim with event-based simulator generated datasets with a range of sensor parameter and scene conditions. Two data sets were generated with the first having a narrower range of parameters closely matching the real conditions expected and the second having a larger range of parameters expected to be far more difficult for

detection. The full list of sensor parameters and scene conditions for the two purely simulated datasets can be found in Table 1.

Parameter	Dataset 1	Dataset 2
FoV	0.08	U(0.05,1.0)
Zero Point	21	21
Exposure Time	1.0	U(0.5,2.0)
Background MV	U(18,24)	U(12,20)
Objects per Image	U(3,9)	U(3,9)
Object MV	U(10,16)	U(10,16)
Object Velocity	U(-10,10)	U(-10,10)

Table 1: Relevant SatSim Parameters for Dataset Creation. U(x, y) indicates sampling from a uniform distribution with range [x,y] inclusive. Velocity values are sampled separately for x and y axes with sign indicating corresponding directional movement.

0.2 Contrast Threshold Optimization

Table 2 lists the range of simulation parameters used to optimize contrast threshold in simulation.

Parameter	Value Range	Step
FoV	[0.05, 1.55]	0.3
Exposure Time	[0.1, 1.1]	0.25
Background MV	[11, 25]	1
Contrast Threshold	[0.05, 5.0]	0.05

Table 2: Range of Parameter Sweep Values

Figures 1 and 2 detail the results of analysis referenced in Section 4.1 of the main text. Figure 1 displays the mean and standard deviation of EB-SNR (event-based signal to noise ratio) with respect to the corresponding parameter values. As mentioned in section 4.1, event-based samples were generated with each combination of the parameters shown in Table 2, with a polynomial regression fit to the contrast thresholds that maximized EB-SNR in each scenario. While background magnitude and temporal window length show interesting trends across the tested values, both parameters vary over a much smaller range of EB-SNR as compared to field-of-view (FoV). Furthermore, the variance seen in EB-SNR versus FoV is also significantly smaller, suggesting that it may be easier to mitigate the negative effects of increasing field-of-view by tuning the contrast threshold.

Equation 2 shows the final 2nd polynomial fit used to extrapolate optimal contrast thresholds from simulator parameters,

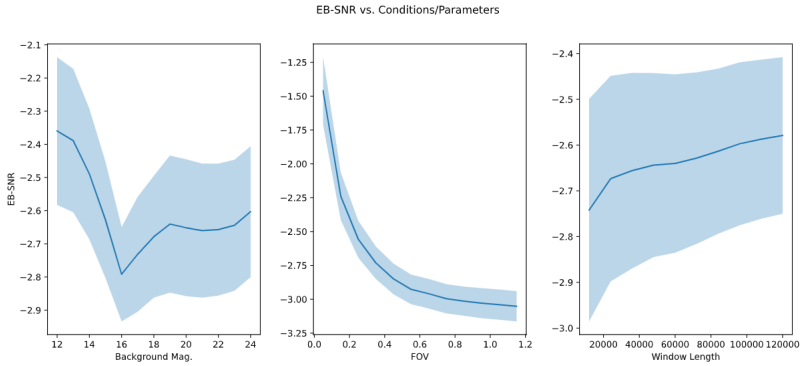
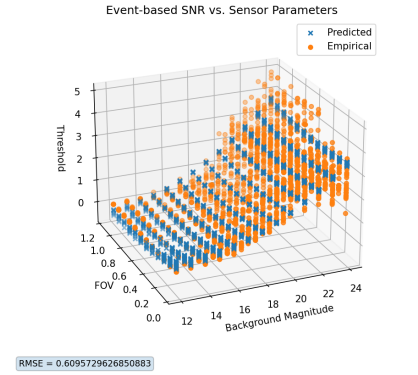


Figure 1: Mean observed EB-SNR values versus each parameter of interest in contrast threshold optimization.



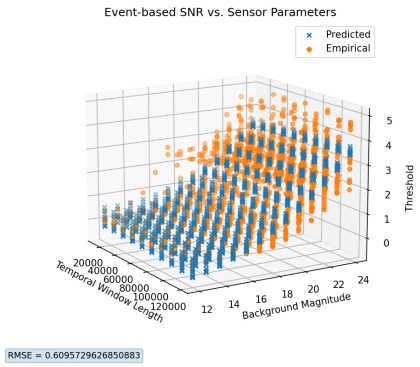


Figure 2: Regression fit for optimal contrast threshold compared to empirical values.

$$\begin{aligned}\Theta = 4.16 + -0.55 * x + -2.54 * y + -5.94e-06 * z + \\ 0.02 * x^2 + 0.07 * x * y + \\ 4.14e-07 * x * z + 0.77 * y^2 + \\ 5.76e-07 * y * z + -1.73e-11 * z^2\end{aligned} \tag{2}$$

where x is the background magnitude in MV, y is the field-of-view in degrees, and z is the overall exposure time in seconds. For our data generation process, we use a symmetric contrast threshold, i.e. the threshold for positive and negative event generation are equal and opposite, such that $\Theta^- = -\Theta^+$. The regression fit calculated in 2 resulted in an RMSE of 0.61 with respect to the empirical optimal contrast thresholds.

Figure 2 shows the optimal contrast threshold values both empirically determined and from the resulting polynomial regression. Although the resulting error (RMSE of 0.61) does diverge from the empirically determined optimal thresholds, the random variation in sample production may make a less certain optimal threshold more desirable when taken over the entirety of a generated data set.

0.3 Simulation-to-Real Data Generation

The provided real dataset used for sim-to-real gap analysis and model validation consisted of 640x480 sized samples taken all with 0.8°field-of-view and 20 second exposure times, but with unknown contrast thresholds, background magnitudes, and target magnitudes. As shown in Figure 3, the number of events generated by identified stars in samples was counted and fit with the corresponding star magnitudes obtained from the Hipparcos star catalog. While satellite magnitude was estimated by simply using the resulting polynomial fit shown in Figure 3, background magnitude was extrapolated by averaging the estimated magnitude obtained from events counted in several 10x10 pixel patches across the image space. Finally, the contrast threshold used to generate equivalent simulated samples was determined using the optimized threshold resulting from Equation 2. The significant variation in event counts and the unknown contrast threshold used for collection contributes heavily to the observed disparities in the event streams generated, as stated in the main text, and no doubt also confounds analysis of the simulation-to-real gap for our simulation pipeline.

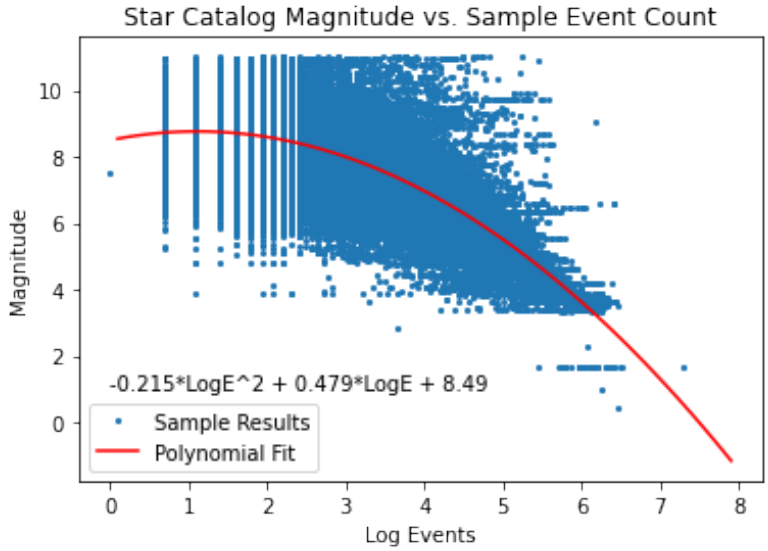


Figure 3: Magnitude extrapolation using log event counts from real data collections matched to known star catalog magnitudes

Figure 4 depicts the results of direct comparison between the real event streams collected and the equivalent simulations. The attributes that most greatly impact object detection results, namely number of events, ratio of positive to negative polarity events, and the resulting EB-SNR, are compared within each plot. Each plot also indicates the average value observed across real and simulated samples.

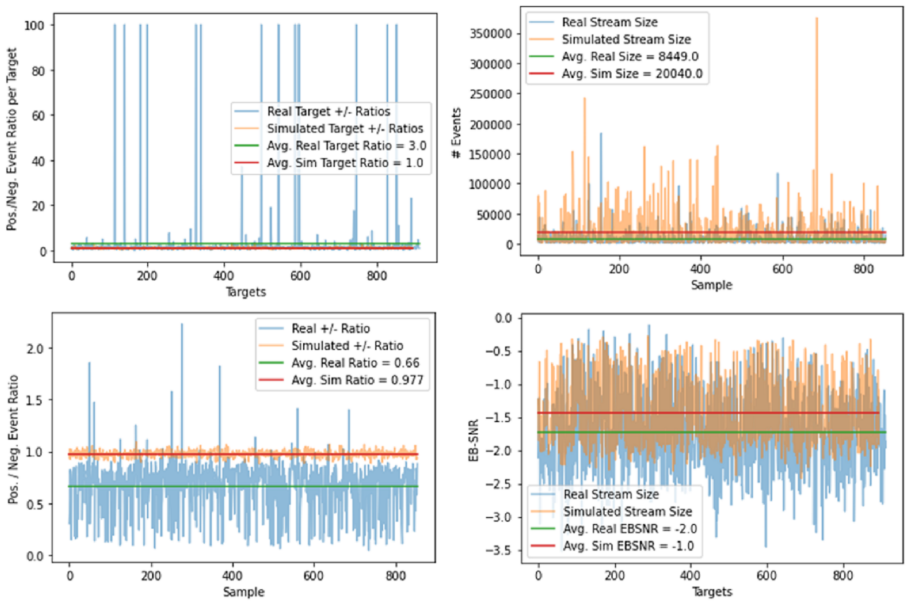


Figure 4: Comparison between real and equivalent simulated event streams. Upper-left: Ratio of positive to negative events per target; upper-right: total event stream size per sample; lower-left: overall ratio of positive to negative events; lower-right: EB-SNR per target

0.4 Additional Object Detection Results

Hybrid Architecture	TP	FP	FN	Precision	Recall	F_1^*	Conf.
DarkNet Only	66169	22989	22827	0.742154	0.743505	0.742829	0.93196
DarkNet-PointNet	70089	23235	18907	**0.751029**	0.787552	0.768857	0.975000
DarkNet-PointNet++	72425	24208	16571	0.749485	0.813801	**0.78032**	0.965000
DarkNet-PointConv	75580	32119	13416	0.701771	**0.849252**	0.768499	0.95999

Table 3: Performance comparison of hybrid frame and event stream architectures on Dataset 1.

Tables 3 and 4 show the full evaluation results for each architecture combination on the purely simulated datasets, while Tables 5 and 6 show recall over each of the varied parameters of interest overlaid on the distribution of the parameter in the corresponding data set. Note that FoV and temporal window length were both set to a static value in Dataset 1, hence the single recall and frequency points present in those plots. Referring to the parameters listed in Table 1, evaluations on dataset 1 show far superior detection compared to dataset 2 due to the much narrower ranges of sensor parameters/conditions, as expected. Conversely, the recall results on dataset 2 (displayed in Table 6), though far worse, demonstrate the effects of varied conditions on detection

Hybrid Architecture	TP	FP	FN	Precision	Recall	F_1^*	Conf.
DarkNet Only	1531	2015	1985	0.431754	0.435438	0.433588	0.92469
DarkNet-PointNet	1720	2652	1796	0.393413	**0.489192**	0.436105	0.88998
DarkNet-PointNet++	1581	2122	1935	0.426951	0.426951	0.438011	0.91500
DarkNet-PointConv	1616	2106	1900	**0.434175**	0.459613	**0.446532**	0.91998

Table 4: Performance comparison of hybrid frame and event stream architectures on Dataset 2

performance. As mentioned in the main text, high background magnitude (i.e. smaller MV values), low EB-SNR, and low target magnitude all contribute to reduced recall; however, increasing FoV has the largest adverse effect on recall due to it significantly increasing the number of noise events.

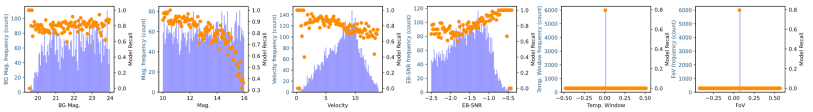
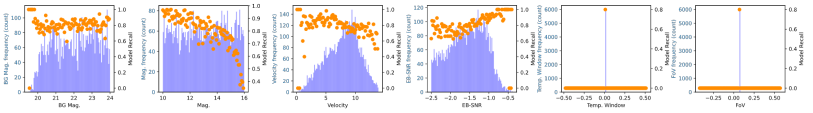
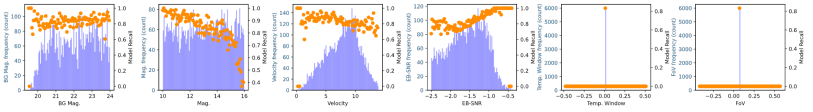
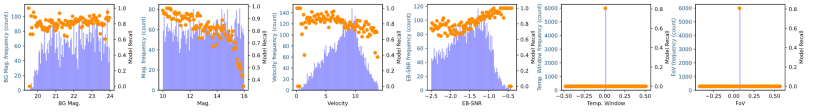
Backbone Arch.	Background Mag.	Target Mag.	Velocity	EB-SNR	Temporal Window	FoV
DarkNet Only						
With PointNet						
With PointNet++						
With PointConv						

Table 5: Recall at max F1 score versus sensor parameters and scene conditions for Dataset 1 (only one value of temporal window length and FoV was used for this dataset)

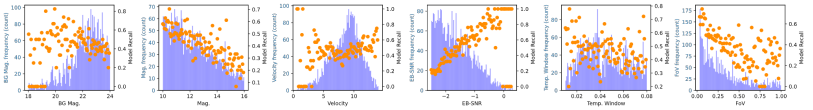
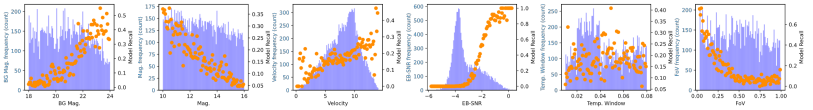
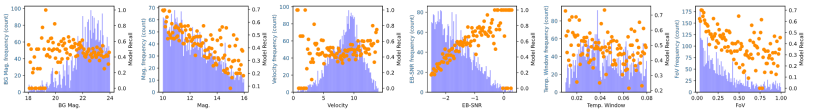
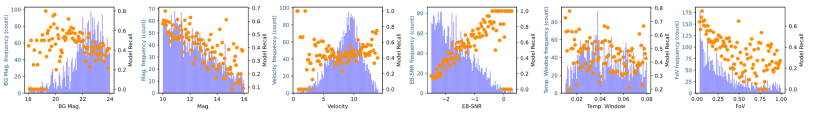
Backbone Arch.	Background Mag.	Target Mag.	Velocity	EB-SNR	Temporal Window	FoV
DarkNet Only						
With PointNet						
With PointNet++						
With PointConv						

Table 6: Recall at max F1 score versus sensor parameters and scene conditions for Dataset 2