Learning to Exploit Multiple Vision Modalities by Using Grafted Networks
Supplementary Material

Anonymous ECCV submission

Paper ID 2499

1 Media Files in the Supplementary Material

1. `NGA_video.mp4`: This video demonstrates the use of the NGA on the thermal and event camera datasets.
2. `NGA_code`: This folder contains a partial code release for illustrating the NGA training procedure. The complete code base will be released upon paper acceptance.

2 RGB Frame Correction for the Thermal Dataset

The RGB frames have a resolution of 1800×1600. A RGB frame I_{RGB} is first re-scaled into 720×640. Then the corrected RGB frame is corrected for the spatial displacement mentioned in the text, *e.g.*, $I_{\text{corr}} = I_{\text{RGB}}\,[65:705, 58:570]$. The final corrected frame has a resolution of 640×512 which is the same size as the thermal frame resolution.

3 Data to Produce Figures 8 and 9 in the Paper

The data used for generating the plots in Figs. 8 and 9 are described in Table 1 and 2.

Table 1. AP_{50} scores from the different grafted network variants. The results for each configuration variant is obtained from 5 runs (`mean` ± `std` reported). Best results in bold.

Thermal			**DVS-3**			**DVS-10**		
	S4	S5		S4	S5		S4	S5
S1	30.54±1.20	29.06±0.52	S1	58.00±0.47	56.85±0.72	S1	58.40±0.59	57.43±0.26
S2	39.81±0.88	37.70±0.59	S2	68.35±0.70	66.34±0.67	S2	68.42±0.66	67.16±0.42
S3	**45.27±1.14**	43.81±1.34	S3	**70.14±0.36**	69.40±0.67	S3	**70.35±0.51**	69.76±0.59

Table 2. Thermal and DVS camera object detector average detection precision (AP_{50}) trained with different loss configurations. The results for each loss configuration are from 5 runs (`mean` ± `std` reported). The improvements (%) are relative to the models trained with FRL alone .

Loss Config	**Dataset**			
	Thermal		DVS-10	
	AP_{50}	Improvement	AP_{50}	Improvement
FRL	40.27±1.86		62.43±1.35	
FEL	43.98±0.74	+9.21%	67.10±1.18	+7.48%
FRL+FEL	42.52±3.59	+5.59%	69.65±0.89	+11.56%
FRL+FSL	44.48±1.40	+10.45%	70.49±0.36	+12.91%
FEL+FSL	44.85±1.21	+11.37%	65.86±1.43	+5.49%
FRL+FEL+FSL	45.27±1.14	+12.42%	70.35±0.51	+12.69%

4 Technical Details for N-MNIST Experiments

The N-MNIST dataset consists of 70,000 event camera recordings corresponding to the 70,000 handwritten digits in the MNIST dataset. Each event volume is prepared by setting $D = 3$ and using all events in the recording. The event volumes are processed by the LeNet-5 architecture described in Table 3. The original LeNet-5 received grayscale input image from MNIST and the grafted network received the event volume as input. The front end consists of the first three sets of convolution and pooling layers. The middle net is the fourth convolution layer. Because the spatial size of the features of the front end is small (5×5), the hyperparameters γ_h and γ_r for the FSL loss terms are set to a low value, in this case, 100.

Table 3. LeNet-5 partitioning into front end, middle net and remaining layers.

Layer	# input channels	# filters	Filter size	Output feature size
	Front end			
`Conv`	1/3	6	5×5	28×28
`ReLU`				
`MaxPool`			2×2	14×14
`Conv`	6	16	5×5	10×10
`ReLU`				
`MaxPool`			2×2	5×5
`Conv`	6	16	5×5	10×10
`ReLU`				
`MaxPool`			2×2	5×5
	Middle net			
`Conv`	16	120	5×5	1×1
`ReLU`				
	Remaining layers			
`Flatten`				
`Linear`	120	84		84
`ReLU`				
`Linear`	84	10		10
`LogSoftmax`				

5 Additional Examples of Grafted Network Front end Feature Decoding

Fig. 1 shows eight additional examples from both datasets of RGB images decoded from GN features. The left two columns show examples from the thermal dataset and the right two columns show examples from the event camera dataset. The implementation is in `NGA_code/decode_thermal.py`.

In the thermal examples (1-4), the decoded GN images have colors representing typical statistics from the RGB image training set, *e.g.*, blue sky, yellow lane, gray color of the cars, even though the thermal input has only one channel. In example 2, the cars are well represented in the decoded intensity frame while the original intensity frame was over-exposed by the front lights of the cars.

Example 5 shows that a decoded intensity frame has a car on the left of the scene while the original intensity frame was so underexposed that it is barely visible. Example 6 shows a case where the car in front of the camera was absent from the decoded intensity frame because the car was not moving and so no events were triggered.

Examples 3-4, 7-8 show scenes that were well-captured by both modalities (thermal/event camera v.s. standard camera). We can see that the decoded GN features represent the scene very well.

Fig. 1. Eight additional examples of decoded features for the thermal dataset (left) and the event camera dataset (right).