E^2(GO)MOTION: Motion Augmented Event Stream for Egocentric Action Recognition Supplementary Material

Chiara Plizzari*[,1] Mirco Planamente*[,1,2] Gabriele Goletto[1] Marco Cannici[3] Emanuele Gusso[1]
Matteo Matteucci[3] Barbara Caputo[1,2]

[1] Politecnico di Torino [2] CINI Consortium [3] Politecnico di Milano
name.surname@polito.it name.surname@polimi.it

Abstract

This supplementary material is divided into seven sections. The first one provides an ablation on how performance varies based on values of α used to weigh the proposed distillation loss. The second one shows the results of the TSN architecture. Additional results obtained on the combination of all modalities, i.e., RGB, Event, and Flow are provided as third. It follows ablations on the number of channels for the event representation, alternative fusion strategies, and additional details about the event conversion. Finally, some qualitative results are presented, showing the Class Activation Maps resulting from the proposed approaches.

1. Ablation on α

We illustrate in Figure 1 how performance varies based on the weight α used to scale the distillation loss $\mathcal{L}_{dist}$ used in the proposed E^2(GO)MO. We set $\alpha = 50, 100, 150, 200$, and show the performance of $\mathcal{L}_{dist}$ when applied to both event and RGB modality. E^2(GO)MO outperforms the baseline for all values of α, demonstrating that $\mathcal{L}_{dist}$ is not sensitive to α variations. Moreover, it can be seen that E^2(GO)MO outperforms RGB+$\mathcal{L}_{dist}$ on unseen domains for all values of α.

2. Temporal Segment Network (TSN)

In Table 1 we show the performance of TSN [6], which we decided not to report in the main paper because it showed the lowest performance w.r.t. TSM [3] and I3D [1] networks. In line with the behavior of I3D and TSM, the event modality consistently outperforms the RGB modality on unseen domains while performing on-par with it on seen ones. The algorithmically generated (TV-L1) optical flow, as expected, provides the best results. Indeed, when combining the event modality with optical flow rather than RGB, superior results are produced, justifying the decision to use distillation between the two last. Most importantly, when combined to optical flow, the event modality outperforms the combination of RGB and optical flow on unseen test sets.

Model	Streams	Pretrain	Seen (%)	Unseen (%)
TSN	Event	ImageNet	59.82	35.24
TSN	RGB	ImageNet	60.88	31.55
TSN	Flow	ImageNet	**67.26**	**43.35**
TSN	Event+RGB	ImageNet	63.92	34.35
TSN	Event+Flow	ImageNet	65.57	**41.31**
TSN	RGB+Flow	ImageNet	**66.81**	38.81

Table 1. Accuracy results (%) of single- and multi-modal streams on TSN architecture. **Bold**: the best scores for single and multi-modal.

3. All Modalities: RGB, Flow, Event

We show in Table 2 the performance obtained by combining all the RGB, optical flow, and event modalities. Interestingly, combining the contribution of all modalities does not improve the best single-modal results on TSM and TSN. In fact, on TSM the combination of all modalities achieves 73.85% and 44.89% accuracy on seen and unseen domains respectively, while the optical flow alone achieves 73.23% and 53.98% (see main paper). This behavior has been already noticed in [7], where the authors attribute the problem to the fact that different modalities overfit and generalize at different rates, thus training them jointly with a unique optimization strategy leads to sub-optimal results.

*The authors equally contributed to this work.

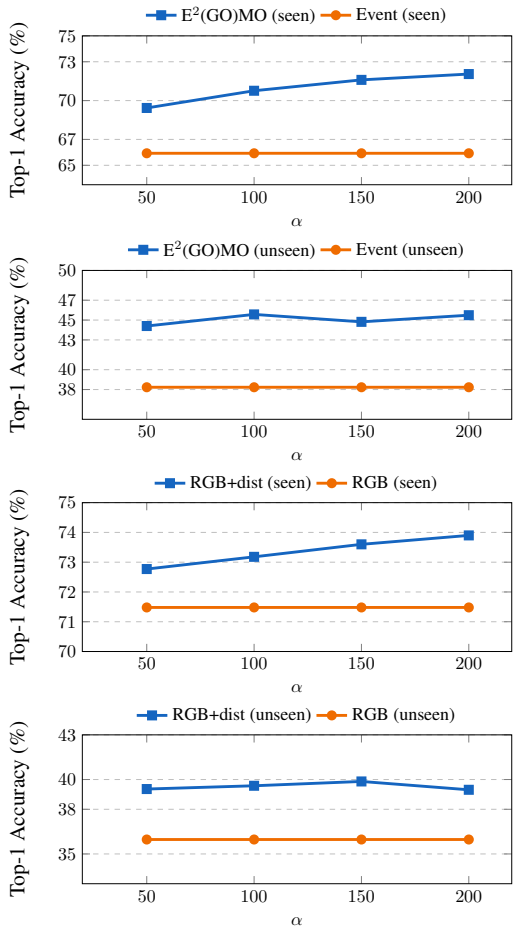

Figure 1. Difference in terms of performance (average Top-1 Accuracy (%)) based on the value of α used to weight $\mathcal{L}_{dist}$ on both seen and unseen test sets.

Performance on I3D, on the other hand, slightly increases, given the fact that 3D convolutions have a lower tendency to overfit on appearance, and hence do not prioritize RGB information as much as 2D-based networks do. This can also be seen in the multi-modal combinations reported in the original paper, where combining appearance and motion information (RGB+Flow) improves performance with respect to Flow alone in I3D but not in TSM.

4. Ablation on Number of Channels

In Table 3 we show the performance of the event modality depending on the number of channels used for the voxel representation [11]. It can be observed that extracting 3-channels Voxel Grid is the optimal choice and we used it in all the experiments in the main paper. In fact, it allows retaining the first ImageNet pre-trained convolution, which is otherwise trained from scratch when using a different number of channels. Indeed, the latter option is damaging on unseen domains. In fact, the first layers of the network are usually the ones that specialize the most on training data distribution [8], thus training them from scratch may lead the network to overfit on the training set, poorly generalizing on the unseen test. Instead, when exploiting pre-trained layers, the network can take advantage of robust low-level features.

Model	Streams	Pretrain	Seen (%)	Unseen (%)
I3D	RGB+Event+Flow	ImageNet	60.38	44.24
E²GO-3D	RGB+Event+Flow	ImageNet	**61.06**	**45.87**
TSM	RGB+Event+Flow	ImageNet	72.66	44.25
E²GO-2D	RGB+Event+Flow	ImageNet	**73.85**	**44.89**
TSN	RGB+Flow+Event	ImageNet	65.93	36.92

Table 2. Accuracy results (%) of the combination of all modalities. In **bold** the best results on both seen and unseen test sets.

Model	Voxel ch.	Testing	Seen acc	Unseen acc
I3D	9	Clip	49.84	34.52
		Video	**52.50**	**36.24**
	3	Clip	53.75	35.90
		Video	**55.54**	**37.52**
	1	Clip	49.34	34.93
		Video	**51.29**	**35.05**
TSN	9	Clip	57.28	31.74
		Video	**58.98**	**32.52**
	3	Clip	58.81	34.65
		Video	**59.82**	**35.24**
	1	Clip	52.59	30.94
		Video	**54.54**	**31.87**
TSM	9	Clip	65.02	37.65
		Video	**66.39**	**38.71**
	3	Clip	64.38	37.75
		Video	**65.93**	**38.23**
	1	Clip	60.76	34.66
		Video	**62.46**	**36.45**

Table 3. Accuracy results (%) on I3D, TSN and TSM architectures depending on the number of channels for the event representation.

5. Other Fusion Strategies

In the main paper, we report results by aggregating multiple modality streams using a late fusion approach as in [4], consisting in summing the prediction logits from both modalities. In Table 4 we validate the choice of a late fusion approach over other existing ones, i.e., TBN [2] and

	Late fusion		TBN [2]		TRN [9]	
	Seen	**Unseen**	**Seen**	**Unseen**	**Seen**	**Unseen**
Event (TSM)	65.93	38.23	64.96.	37.70	64.81	37.53
E^{2}GO-2D	65.40	40.33	64.24	38.51	64.03	38.54
Event (I3D)	55.54	37.52	55.05	35.69	55.88	36.37
E^{2}GO-3D	57.87	38.76	58.77	39.17	58.98	39.87
E^{2}GO-MO	**70.76**	**45.57**	**70.70**	**44.36**	**70.48**	**44.96**

Table 4. Ablation on different fusion strategies alternative to traditional late fusion. **Bold**: highest result for each setting.

TRN [9]. The first one combines the feature embeddings of each modality before temporal aggregation through a mid-level fusion, while the second one models multi-scale relations. As it can be seen from results, the standard late fusion is the one achieving better results on almost all configurations, and thus we used it for all experiments in the main paper.

6. Additional Details on Event Conversion

A crucial parameter in the conversion pipeline is the conversion factor indicating the number of consecutive upsampled RGB frames to be used to create a single voxel representation. We observed the average number of events generated for each sample depending on the conversion factor (Figure 2). We considered that (i) if the number of frames used is too low, there are not enough events to generate a proper voxel representation; (ii) increasing the conversion factor would also decrease substantially the dimensionality of the dataset and (iii) in an hypothetical online setting, an high conversion factor is equivalent to increase the time before the generation of a voxel. Taking into account all the above aspects, we chose a trade-off conversion factor of *6 RGB frames per voxel*. Indeed, considering that EPIC-Kitchens's videos have all been sampled at 60 FPS, using a conversion factor of 6 also implies we are considering a temporal interval of 100 ms for the creation of a single voxel, which is consistent with the N-Cars dataset [5].

7. Qualitative Results

To conclude this supplementary material, we present in Figure 3 and Figure 4 some additional qualitative results based on the Class Activation Maps [10] obtained with standard TSM architecture for RGB, Event modality and E^2(GO)MO variation on both seen and unseen test sets. We show that the event modality, especially in the E^2(GO)MO variation, focuses on parts of the scene that are highly correlated with the motion of the action, allowing the network to be more robust when tested in unseen scenarios (Figure 4). The RGB modality, on the other hand, appears to be more focused on kitchen elements, such as the sink, rather than the user's hands. This demonstrates that RGB has difficulties in focusing on elements of the motion that are discriminative, limiting its capacity to generalize to unknown settings. Indeed, as indicated in the main paper, emphasizing on motion helps the event modality to generalize better on unknown test sets since it does not overfit on the environment, which differs the most from one kitchen to another.

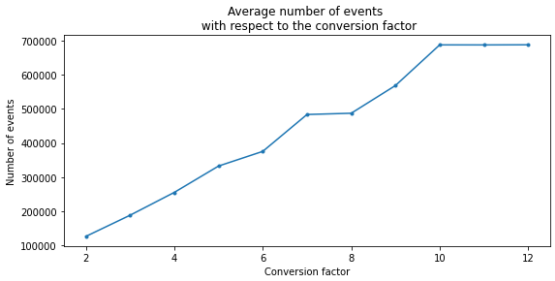

Figure 2. Conversion factor *vs* average number of generated events on a single sample.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2, 3

[3] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 1

[4] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020. 2

[5] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018. 3

[6] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 1

[7] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 1

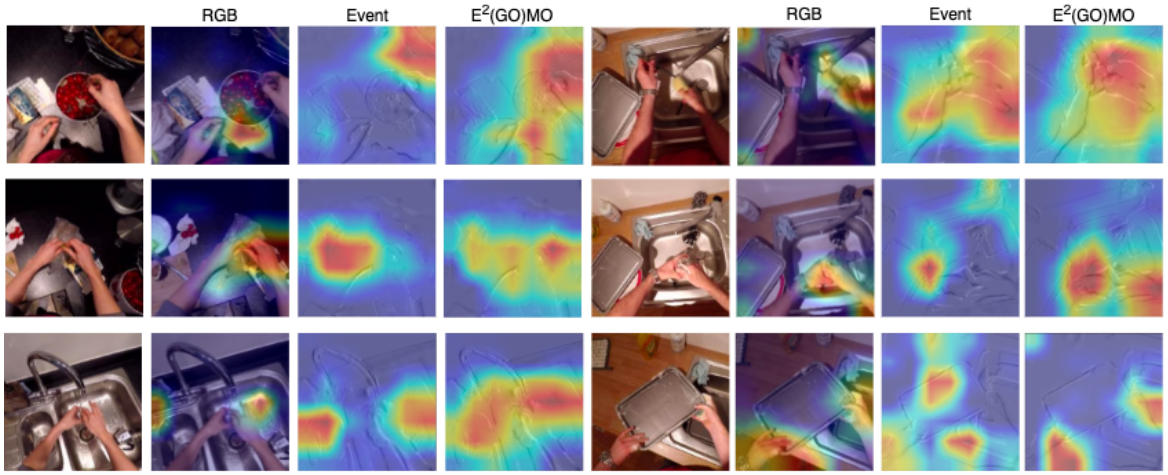

Figure 3. **Class Activation Maps** of RGB (TSM), Event (TSM), and E^2(GO)MO obtained by training on one kitchen, and testing on the same (*seen*). Red regions correspond to part of the image which activated most, while the blue ones correspond to those which are less activated.

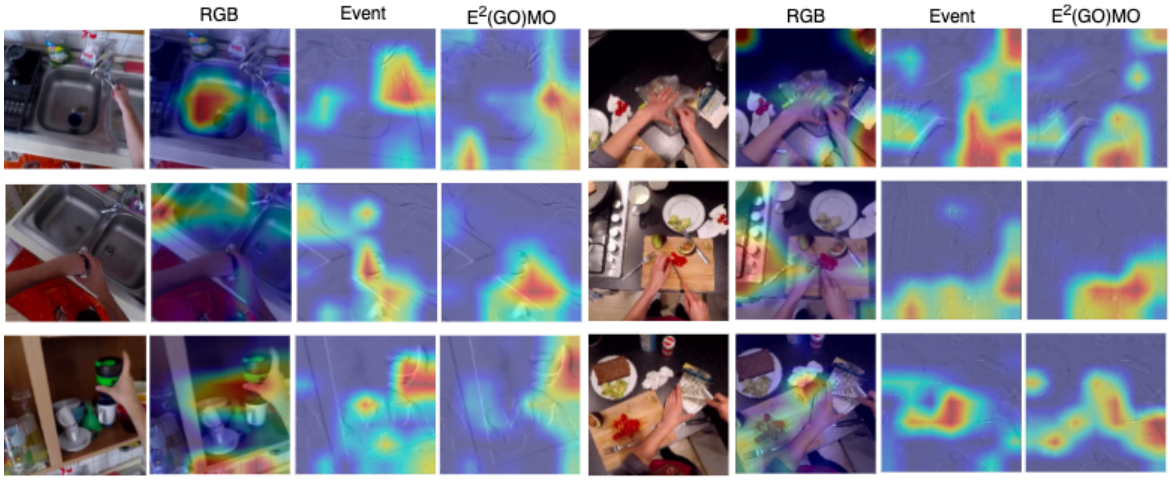

Figure 4. **Class Activation Maps** of RGB (TSM), Event (TSM), and E^2(GO)MO obtained by training on one kitchen, and testing on a different one (*unseen*). Red regions correspond to part of the image which activated most, while the blue ones correspond to those which are less activated.

[8] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014. 2

[9] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 3

[10] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[11] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 2