

OWL (Observe, Watch, Listen): Audiovisual Temporal Context for Localizing Actions in Egocentric Videos (*Supplementary Material*)

Supplementary Material

We complement our work with the following: (i) The details on the proposal generation (Sec. 1), (ii) fusion experiments (Sec. 2), and (iii) qualitative examples (**please check the attached .ppt slides**).

1. Action proposals

This section analyzes the action proposals for EPIC-KITCHENS produced by the proposal generator, as explained in Sec. 3.1 and Fig. 2 (*cf.* the main manuscript). We measure the quality of the proposals with average recall (AR). [2] It is worth noting that proposals are class-agnostic and require further classification. AR measures the localization quality of the action proposals. We consider the limited number of predicted proposals when computing AR and compute it for several tIOU thresholds. In the following sections, we investigate which feature encoders to use and how to treat the input sequence.

Features	Modality	AR (%)
TBN	RGB, flow, audio	64.61
SlowFast	visual	64.09
SlowFast	audio	56.38
SlowFast	visual, audio	65.66

Table 1. **Average Recall (AR) on EPIC-KITCHENS** for the proposals using TBN and SlowFast features in uni-modal and multi-modal scenarios.

1.1. Feature encoders

Our focus is to investigate audiovisual inputs; thus, we consider the encoders that process auditory and visual signals. We consider TBN [4] and SlowFast [3, 5] networks as our feature encoders. TBN operates on RGB, Flow, and spectrogram. Visual and Auditory SlowFast take video frames and spectrogram, respectively, as inputs. In Tab. 1 we compare the performance of the proposal generator on EPIC-KITCHENS with TBN and SlowFast fea-

tures. To demonstrate the effect of audiovisual features, we also provide the results of a uni-modal proposal generator with visual-only or audio-only inputs. To create audiovisual SlowFast features, we concatenate the visual and auditory features of the corresponding SlowFast backbones. We notice that audiovisual SlowFast features outperform TBN (65.66% vs. 64.61%). Furthermore, we can observe that multi-modal SlowFast features outperform uni-modal (65.66% for audiovisual vs. 64.09% for visual and 56.38% for audio).

Window Size	AR (%)
200	65.52
300	65.66
400	63.75

Table 2. **Average Recall (AR) on EPIC-KITCHENS for the proposals using different sliding window sizes.**

1.2. Window size

While processing the input sequence with a sliding window, we aim for the most effective window size. As observed in [8], over 98% of annotated action instances in EPIC-KITCHENS [1] are shorter than 20 seconds. We extracted features at 5 fps; thus, to capture 98% of actions, we should aim for a minimal stride $s = 20 \times 5 = 100$. In our experiments, we always make the window size w double the stride s . In Tab. 2 we investigate the best window size, starting with $w = 200$ and $s = 100$. We keep increasing w and s until the performance degrades. That ensures that at least one sliding window will cover any action that does not exceed $\frac{w}{2}$. We reach the highest performance with $w = 300$ (and $s = 150$). This is because increasing the window size to 300 incorporates some relevant context to the model. However, further increasing the window size to 400 degrades the performance, suggesting that faraway context becomes irrelevant (similar to OWL’s temporal context).

2. Fusing audio and visual modalities

In this section, we explain our preliminary experiments on multi-modal fusion strategies. First, we elaborate on our terminology of the proposal generator and classifier.

Proposal generator \mathcal{G} . Given the visual features \mathbf{x}^v and the audio features \mathbf{x}^a of the video sequence, the proposal generator \mathcal{G} predicts a set of candidate segments with temporal boundaries, namely, proposals $\Phi = \{\phi_m = (t_{s,m}, t_{e,m}, s_m)\}_{m=1}^M$, where ϕ_m represents an action proposal, M is the number of proposals, and $t_{s,m}$, $t_{e,m}$ and s_m are its start time, end time and confidence score, respectively. Note that proposals do not have class labels.

Proposal classifier \mathcal{C} . Given the set of proposals Φ , the snippet-level visual features \mathbf{x}^v , and audio features \mathbf{x}^a , we first extract visual features \mathbf{x}_m^v and audio features \mathbf{x}_m^a for the m^{th} proposal by max-pooling the snippets within its start/end boundaries¹. Then, the proposal classifier \mathcal{C} predicts from \mathbf{x}_m^v and \mathbf{x}_m^a verb and noun class labels c^{verb} and c^{noun} , as well as their respective scores s^{verb} and s^{noun} . Based on the predicted verbs and nouns, we generate action predictions $\Psi = \{\psi_n = (t_{s,n}, t_{e,n}, c_n, s_n)\}_{n=1}^N$.

2.1. Where and how to fuse the modalities in the classifier?

We categorize the modality fusion into the following: early, late, and intermediate fusion, as shown in Fig. 1. *Early fusion* happens at the input feature level (Fig. 1 a). Given the proposal’s visual features \mathbf{x}_m^v and audio features \mathbf{x}_m^a , we first fuse them and obtain one single feature vector $\mathbf{x}_m = \mathcal{F}_{\text{early}}(\mathbf{x}_m^v, \mathbf{x}_m^a)$. We feed \mathbf{x}_m to the following layers of operations (e.g., MLP) and classify them into different noun and verb classes. *How to choose the fusing function $\mathcal{F}_{\text{early}}$?* In our analysis, we simply fuse the modalities by concatenating the visual and audio features along the channel dimension. This doesn’t require extra computations and counts on the following network layers to learn from the fused features.

Intermediate fusion happens at the intermediate feature level (Fig. 1 b). We process the audio and video features independently for certain layers and generate intermediate features (\mathbf{z}_m^v and \mathbf{z}_m^a). We fuse them to one feature via $\mathbf{z}_m = \mathcal{F}_{\text{inter}}(\mathbf{z}_m^v, \mathbf{z}_m^a)$. The fused features \mathbf{z}_m as well as the visual and audio intermediate features \mathbf{z}_m^v and \mathbf{z}_m^a are processed independently in the following layers and correspondingly predict three groups of classification scores. We use them all for training and only use the scores from the fused features for inference. Similarly to early fusion, we use concatenation for $\mathcal{F}_{\text{inter}}$ in our experiments (Tab. 3). Our proposed model OWL uses intermediate fusion; however, instead of concatenation, it adaptively fuses audio features to visual by correlating to the context (more in Sec. ??).

¹We round the start/end values to the nearest snippets indices.

Late fusion happens at the output score level (Fig. 1 c). The visual and audio features of all proposals are independently processed until they produce classification scores $\mathbf{s}_m^v = \{s_m^{\text{verb},v} \in \mathbb{R}^V, s_m^{\text{noun},v} \in \mathbb{R}^U\}$, and $\mathbf{s}_m^a = \{s_m^{\text{verb},a} \in \mathbb{R}^V, s_m^{\text{noun},a} \in \mathbb{R}^U\}$ where V and U are the numbers of verb and noun classes, respectively. We fuse the scores from both modalities via $\mathbf{s}_m = \mathcal{F}_{\text{late}}(\mathbf{s}_m^v, \mathbf{s}_m^a)$, and apply *softmax* to \mathbf{s}_m to generate the final prediction for nouns and verbs. For late fusion, there is no straightforward way to concatenate. Naively averaging or multiplying the corresponding scores of the two modalities is not effective due to the imbalance between the modalities. While audio can be a complementary source of information, it doesn’t contribute equally to the visual modality to solve the task. We observe that either modality ‘specializes’ in different classes, and it’s beneficial to combine the scores with different weights per class. For example, the action of ‘taking something’ is usually not evident from the sound, but ‘turning on’ a kitchen device is.

For effective late fusion, motivated by [6, 7], we design a gating module to weigh the per-class scores before fusing them. The gating module Θ is composed of a fully-connected layer followed by a sigmoid activation function. It learns from the concatenated intermediate features of the two modalities $\mathbf{z}_m = [\mathbf{z}_m^v; \mathbf{z}_m^a]$ to predict weights for the verb and noun classes for both modalities: $\mathbf{w}_m^v = \Theta^v(\mathbf{z})$, $\mathbf{w}_m^a = \Theta^a(\mathbf{z})$. The weights are applied to the classification scores \mathbf{s}_m^v and \mathbf{s}_m^a of two modalities for linear combination, and generate the final scores via $\mathbf{s}_m = \mathbf{s}_m^v \odot \mathbf{w}_m^v + \mathbf{s}_m^a \odot \mathbf{w}_m^a$. We call the gating strategy *cross-gating*. Alternatively, we also experiment with a *self-gating* strategy, where the weights for each modality are learned only from its own features: $\mathbf{w}_m^v = \Theta^v(\mathbf{z}^v)$, $\mathbf{w}_m^a = \Theta^a(\mathbf{z}^a)$.

Table 3. **Fusion methods performance on EPIC-KITCHENS, measured by the average mAP (%)**. SG and CG correspond to the self-gating and cross-gating scenarios described in Sec. 2.1, respectively. We also show the modality streams being supervised in the second column.

Method	Supervision	Noun	Verb	Action
Early F	AV	12.63	11.47	8.35
Intermediate F	AV	12.55	11.66	8.24
Intermediate F	V, A, AV	13.66	12.90	8.75
Late F SG	V, A	11.51	10.84	7.99
Late F CG	V, A	12.66	12.89	8.82

2.2. Results

We compare several fusion strategies in Tab. 3. All experiments were run on audiovisual proposals (\mathcal{G} -AV). Early fusion results in a significant improvement over the visual-only model (VM). The intermediate fusion with only audio-visual supervision (8.24% action mAP) does not perform better than early fusion. However, we can achieve better

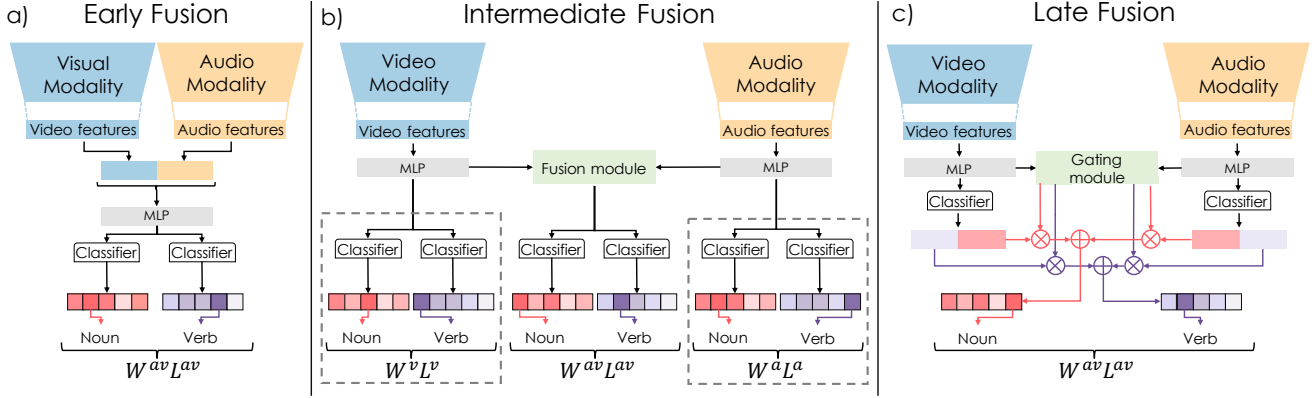


Figure 1. **Fusion methods for the audio and video streams.** Early fusion (a) does features aggregation. Intermediate fusion (b) combines intermediate representations of each modality. The model can be trained jointly by optimizing for three losses, or we can simplify it by setting $W_v = W_a = 0$ (the affected branches are highlighted with dashed lines). Late fusion (c) combines scores of two modalities. The gating module produces per-class weight for the scores generated by each modality. The weighted scores are aggregated by summation. Here we illustrate *cross-gating*, in which the gating module takes representations of both modalities as the input.

results by jointly training with the supervision from the visual and audio streams (8.75%). Doing late fusion with self-gating weights does not perform well (only 7.99%), but late fusion with cross-gating (Late F CG) achieves 8.82%. This finding is expected as cross-gating has richer representations of both modalities for weighting the class scores.

References

- [1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2021. 1
- [2] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Nieves, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1
- [3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1
- [4] 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. 1
- [5] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021. 1
- [6] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 279. BMVA Press, 2019. 2
- [7] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2
- [8] Zhiwu Qing, Ziyuan Huang, Xiang Wang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Changxin Gao, Marcelo H Ang Jr, and Nong Sang. A stronger baseline for ego-centric action detection. *arXiv preprint arXiv:2106.06942*, 2021. 1