Supplementary Materials - A Closer Look at Temporal Ordering in the Segmentation of Instructional Videos

BMVC 2022 Submission # 669

1 Qualitative Results

We present the qualitative comparison with state-of-art method PDVC [3] in Figures.1 . The PDVC [3] model output a segment counter i.e. number of segments in an instructional video. The segment counter is utilized to select the top-*N* proposals from the proposal generation module. The qualitative results shows that the top-*N* proposals of PDVC [3] are overlapping and redundant. Ours model accurately predict the top-*N* proposals which align with ground truth segments and are sequentially ordered.

References

[1] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020.

[2] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021.

[4] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, 2018.

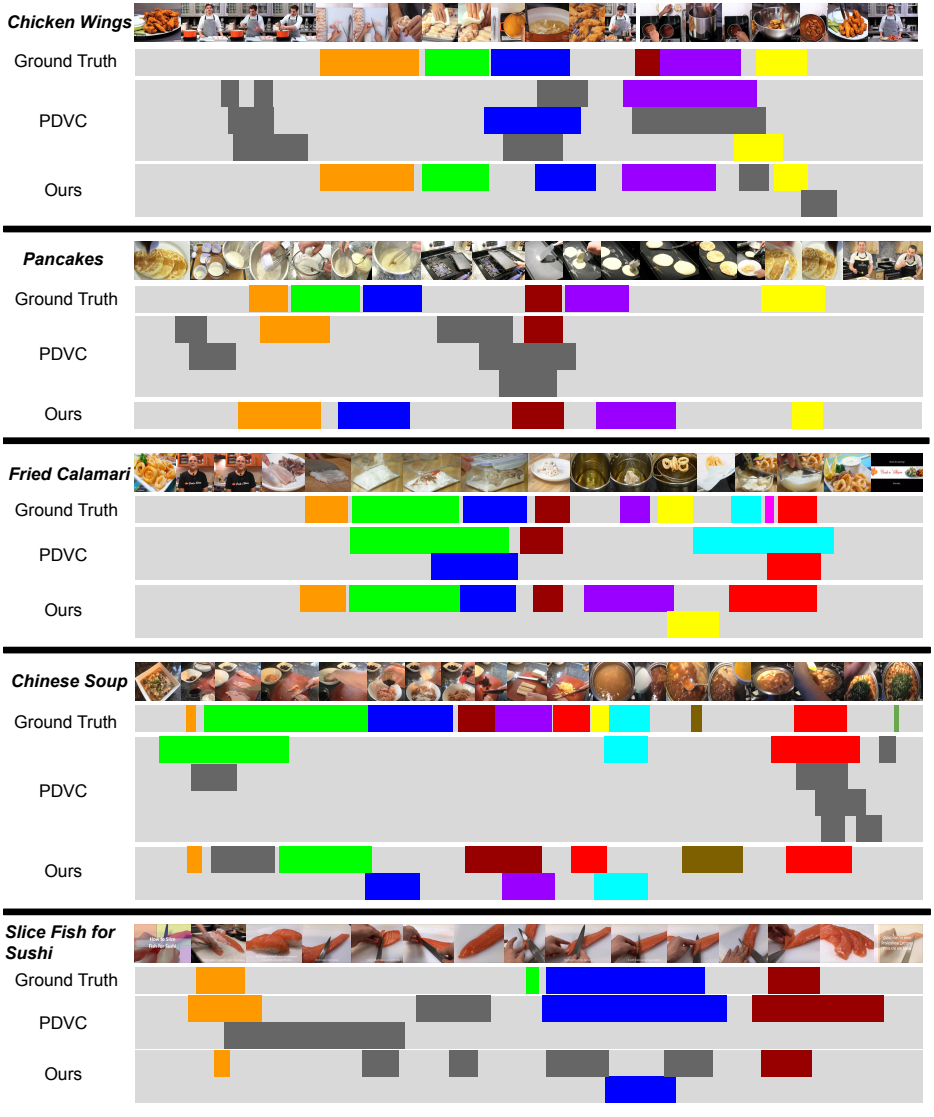

Figure 1: Qualitative Comparison on Samples from YouCook2 dataset [2]. Ours is the full model with proposed improvements, i.e. multi-modal features (S3D [1]) and differentiable matching (SoftSODA) algorithm considering the temporal order. PDVC [3] and Ours model predict the count of proposals (N). The top-N proposals are shown for samples with YouTube ID - *vLcBGs389k4*, *NjAtxfaLwCk*, *peld2w63tpM*, *vVZsj1t9R70* and *mi8NwUqf7nM*.

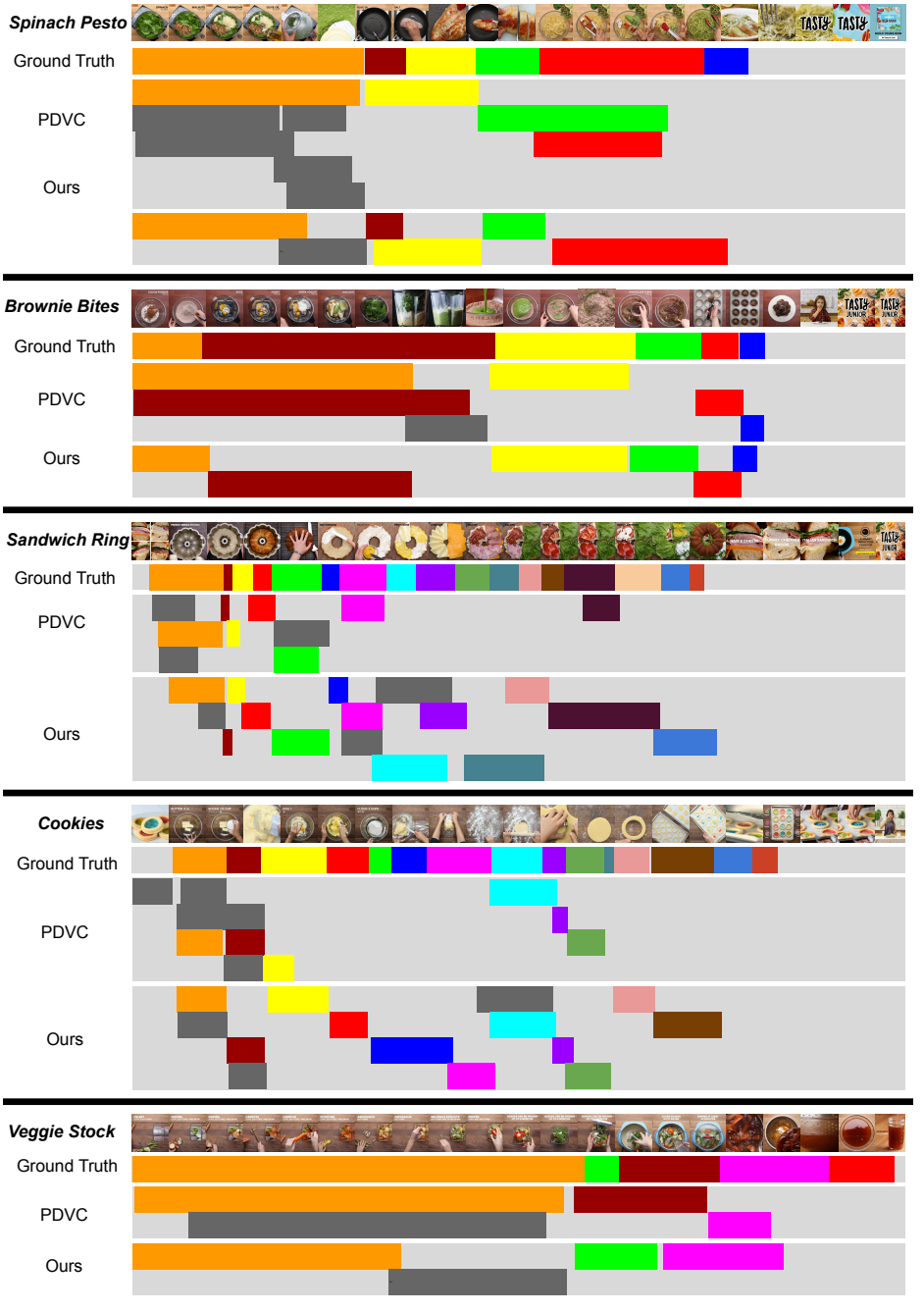

Figure 2: Qualitative Comparison on Samples from Tasty dataset [7]