Enhancing Transformer Backbone for Egocentric Video Action Segmentation

Sakib Reza[1, 2], Balaji Sundareshan[1], Mohsen Moghaddam[1, 2], Octavia Camps[1]
[1]College of Engineering, Northeastern University, Boston, MA
[2]Khoury College of Computer Sciences, Northeastern University, Boston, MA
{reza.s, sundareshan.b, mohsen, o.camps}@northeastern.edu

Abstract

Egocentric temporal action segmentation in videos is a crucial task in computer vision with applications in various fields such as mixed reality, human behavior analysis, and robotics. Although recent research has utilized advanced visual-language frameworks, transformers remain the backbone of action segmentation models. Therefore, it is necessary to improve transformers to enhance the robustness of action segmentation models. In this work, we propose two novel ideas to enhance the state-of-the-art transformer for action segmentation. First, we introduce a dual dilated attention mechanism to adaptively capture hierarchical representations in both local-to-global and global-to-local contexts. Second, we incorporate cross-connections between the encoder and decoder blocks to prevent the loss of local context by the decoder. We also utilize state-of-the-art visual-language representation learning techniques to extract richer and more compact features for our transformer. Our proposed approach outperforms other state-of-the-art methods on the Georgia Tech Egocentric Activities (GTEA) and HOI4D Office Tools datasets, and we validate our introduced components with ablation studies. The source code and supplementary materials are publicly available on `https://www.sail-nu.com/dxformer`. *

1. Introduction

Automated detection and segmentation of human activities from an egocentric perspective have numerous applications in fields such as mixed reality, human behavior analysis, and robotics [9, 10]. However, egocentric action segmentation is particularly challenging due to several factors. First, the videos are untrimmed and can span several minutes, making it difficult to assign an action label to each frame accurately. Second, egocentric videos often have occlusions, where the camera wearer's body or objects in the foreground obstruct the view of the action. Random movement and camera motion can also lead to inconsistent viewpoints, making it challenging to track actions across frames. Third, active hand-object interactions, which are prevalent in egocentric videos, can result in actions slightly outside the frame, further complicating the segmentation task. To address these challenges, action segmentation methods for egocentric videos focus on modeling the temporal relations among frames using pre-extracted frame-wise feature sequences, while also considering the unique characteristics of egocentric videos.

Recent advances in action segmentation models have shown promising results through the use of cutting-edge visual language feature representations [11, 15]. However, there remains scope for improvement in the transformer backbone, which serves as the fundamental component of these models. Specifically, enhancing the transformer backbone can aid in precise action identification by effectively attending to salient regions within the video. A strengthened attention mechanism within the improved transformer plays a pivotal role in capturing temporal dependencies, ultimately leading to improved model accuracy.

In this paper, we introduce DXFormer, a new transformer-based architecture that improves the state-of-the-art transformer backbone for action segmentation by introducing a dual dilated attention mechanism and cross-connections between encoder and decoder blocks. In addition, we build on advanced visual-language representation learning approaches to extract richer and more compact characteristics for the transformer. We evaluate DXFormer on two challenging egocentric video datasets and show that it outperforms other state-of-the-art action segmentation methods both quantitatively and qualitatively. Furthermore, we conduct an ablation study to validate the effectiveness of newly added components in improving the results with respect to various metrics.

2. Method

The proposed DXFormer is an improved backbone model for action segmentation, based on the ASFormer

*This work is supported by the NSF Grant No. 2128743. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

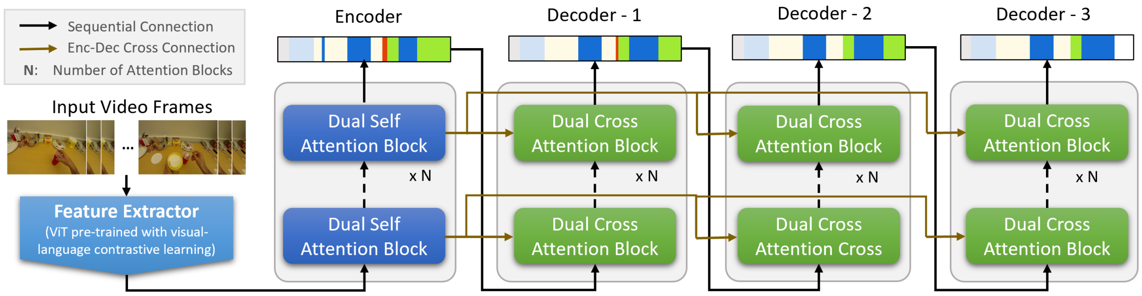

Figure 1. The proposed DXFormer model for temporal action segmentation uses a multi-decoder approach to capture temporal dependencies more effectively.

[17] baseline model. It introduces two innovative techniques, dual dilated attention (DA) and encoder-decoder cross-connection (CC), to elevate its performance. The overall architecture of DXFormer, depicted in Figure 1, is presented to provide a high-level overview. The proposed model comprises a single encoder and three decoders. Initially, video frames are passed through a feature extractor model, which is a vision transformer pre-trained using visual language contrastive learning methods (e.g., BridgePrompt [11] and CLIP [14]). The encoder leverages the extracted frame-wise features to generate initial action predictions for each frame. Subsequently, the three decoders refine these initial outputs, yielding improved frame action labels as the final output. This multi-decoder approach effectively enhances the accuracy of the action segmentation by capturing temporal dependencies in the egocentric video data.

2.1. Dual Dilated Attention Mechanism

The baseline model, ASFormer [17], maintains a hierarchical pattern representation to focus first on local information and then gradually expand the attention span to capture the global feature. In this approach, while the attention span is very large for the higher attention blocks, lower blocks may suffer from a small attention span. Due to the smaller attention windows, the initial attention blocks capture the local context from neighboring frames quite well, but miss the global contexts. On the other hand, the higher-attention blocks capture the global context quite well because of the larger attention span, but they miss out on the local contexts. To address the limitations mentioned above, we propose a dual-dilated attention (DA) approach, to help the model to capture both global and local contexts adaptively in both lower and higher attention blocks. Figure 2 shows the architecture of the proposed DA module.

The new DA module consists of two different attention branches, one with an increasing window size and the other with a decreasing window size. Each attention branch starts with a dilated convolution layer followed by an attention layer. Following [17], we keep the dilation size of the convolution layers and the window size of its attention layer the same. For the first attention branch, the window size is doubled at each block (e.g., 2^i, where i = 1, 2,...). On the other hand, for the second branch, the window size is halved in each block (e.g., 2^{N-i}, where N = 9 and i = 1, 2,...). Hence, in the initial attention blocks, the first branch has a small attention span and captures local context from neighboring frames. In contrast, the second branch starts with a large attention span and captures global context exploring both near and distant frames. These two branches are then merged through concatenation, followed by a convolution operation. This adaptive combination enables the model to dynamically learn the emphasis placed on each branch, facilitating the effective capture of both local and global contexts.

The proposed DA uses different attention mechanisms for the encoder and the decoders. The encoder uses the self-attention mechanism and the output of the previous layer as the query Q, the key K, and the value V. On the other hand, the decoders use a cross-attention mechanism where the query Q and the key K are generated by concatenating the outputs from its previous layer and the corresponding encoder attention block, and the value V is taken from the previous layer as self-attention.

2.2. Encoder-Decoder Cross Connections

Connecting the encoder and decoder sequentially in the transformer backbone can lead to a semantic gap between their features. For example, in the ASFormer [17], the first decoder takes the output from the last block of the encoder as input, while the subsequent decoders take the output from their previous decoder's last block as input. Although the output of any encoder or decoder's last block already contains global context of the input, sharing only this feature with the decoder blocks can result in a loss of local information. This can hinder the learning process of both local and global patterns using DA and limit the effectiveness of the overall architecture.

To address the issue of losing local context, we propose cross-connections between the encoder and decoder blocks. These cross-connections connect each decoder block with its corresponding encoder block (as depicted in Figure 1)

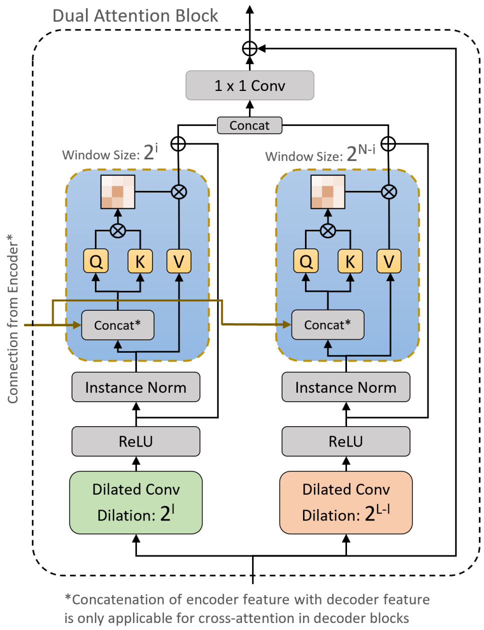

Figure 2. Dual Dilated Attention (DA) block.

and pass the local information from the lower encoder block to its corresponding decoder blocks, preventing the model from losing important local information at the decoders. Each cross-connection takes the output of an encoder block and passes to its corresponding decoder block. In the decoder blocks, the features are concatenated with the input of their cross-attention layer, and the result is used as the query Q and the value V, as shown in Figure 2.

3. Experiments

3.1. Datasets and Experiment Setup

We utilize two challenging egocentric video datasets to assess the performance of the proposed approach. The first dataset, **Georgia Tech Egocentric Activities (GTEA)** [5], comprises 28 instructional videos recorded from an egocentric perspective. It includes 11 distinct action classes representing daily kitchen activities. The second dataset, **HOI4D Office Tools**, is a subset of the larger HOI4D dataset [13] and consists of 553 egocentric videos depicting hand interactions with office tools such as scissors, pliers, and staplers. It covers 12 action classes related to office tool usage. To evaluate the performance of our approach, we conducted a four-fold cross-subject validation for both datasets. This evaluation technique ensures robustness and mitigates the influence of subject-specific biases on the results.

For our experiments, we closely follow the settings employed by the ASFormer model [17], our baseline. However, for the large HOI4D dataset, we make specific adjustments for efficient training. We utilized a batch size of 8, a learning rate of 0.001, and incorporated 7 attention blocks in each encoder and decoder.

Table 1. Performance comparison of different configurations of our proposed method vs. the baseline (ASFormer) on the GTEA dataset (using only BrPrompt feature representation).

Backbone Model	**F1@{10, 25, 50}**			**Edit**	**Acc**
Baseline	91.4	89.9	81.4	88.2	80.7
Baseline + DA	91.7	**90.5**	**82.6**	88.8	**81.7**
Baseline + CC	**91.8**	90.3	80.6	88.3	80.9
Baseline + DA + CC	91.0	89.8	81.9	**89.0**	81.3

* DA = Dual Dilated Attention, CC = Cross-Connection

Table 2. Performance comparison of different feature representations with DXFormer on the GTEA dataset.

Feature	**F1@{10, 25, 50}**			**Edit**	**Acc**
I3D	89.1	88.0	78.1	84.9	80.3
ViT	85.3	83.2	72.6	83.4	74.1
BrPrompt	**91.4**	**89.8**	**81.9**	**89.0**	**81.3**

3.2. Effect of Dual-Attention and Cross-Connection

To verify the influence of the new components in our DXFormer model, we perform an ablation study on the GTEA dataset. We specifically test various combinations of these elements and assessed how they affect the performance indicators. The addition of dual dilated attention (DA) increases the F1@50 score by 1.2% and accuracy by 1%, according to the findings of the ablation study. The best F1@10 score is achieved by including cross-connections (CC) between the encoder and decoder blocks. Finally, adding both DA and CC allows us to achieve the highest edit score. The findings in Table 1 show how well the newly incorporated components work and how they help the action segmentation models function better.

3.3. Effect of Feature Representations

We experiment with various feature representations of video frames as input to our transformer-based DXFormer model. The results on the GTEA dataset show that DXFormer performs best with BridgePrompt feature representation, trained using visual-language contrastive learning with custom prompts [11]. We also evaluate the performance of two commonly used pre-trained models, I3D [2] and ViT [15], which are pre-trained on the Kinetics400 dataset [8]. The results demonstrate that BridgePrompt achieves the highest performance, followed by I3D, while ViT exhibits comparatively lower performance. Detailed comparison of these results is provided in Table 2.

3.4. Comparison with the State-of-the-Art

Table 3 presents a comprehensive comparison between the proposed DXFormer model and other state-of-the-art approaches for temporal action segmentation on the GTEA dataset. Notably, our DXFormer model outperforms all other models, achieving state-of-the-art performance with an F1@50 score of 82.6% and a frame-wise accuracy of

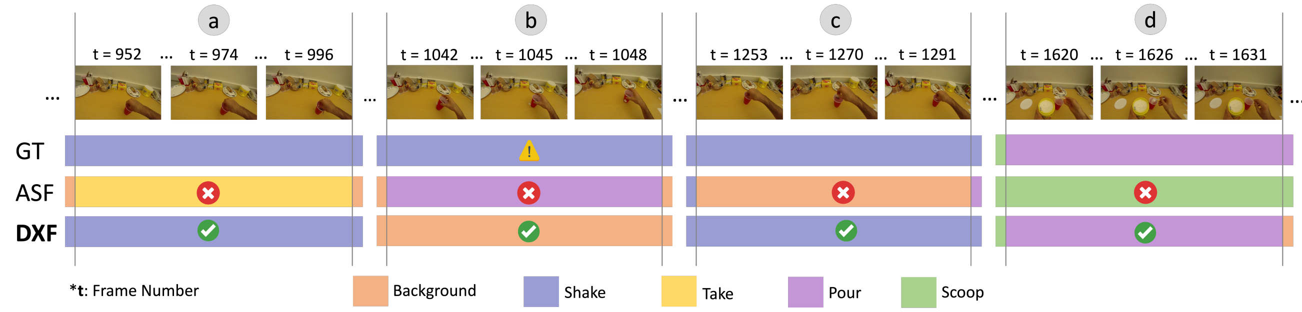

Figure 3. Qualitative Evaluation: DXFormer outperforms the baseline (ASFormer) approach in challenging scenarios, as demonstrated on a representative video ('S1_Tea_C1') from the GTEA dataset.

Table 3. Action segmentation results on GTEA dataset.

Method	F1@{10, 25, 50}			Edit	Acc
MS-TCN++ [12]	88.8	85.7	76.0	83.5	80.1
BCN [16]	88.5	87.1	77.3	84.4	79.8
G2L [6]	89.9	87.3	75.8	84.6	78.5
ASRF [7]	89.4	87.8	79.8	83.7	77.3
SSTDA [3]	90.0	89.1	78.0	86.2	79.8
SSTDA+HASR [1]	90.9	88.6	76.4	87.5	78.7
I3D+ASFormer [17] *	89.9	88.3	78.2	85.4	79.3
BrPrompt+ASFormer [11] *	91.4	89.9	81.4	88.2	80.7
BrPrompt+**DXFormer** (DA)	**91.7**	**90.5**	**82.6**	88.8	**81.7**
BrPrompt+**DXFormer**	91.4	89.8	81.9	**89.0**	81.3

* Reproduced results on our hardware configuration
DA = Dual Dilated Attention Only

Table 4. Action segmentation results on HOI4D (Office Tools) dataset.

Method*	F1@{10, 25, 50}			Edit	Acc
MS-TCN [4]	88.1	84.2	71.2	91.4	74.7
MS-TCN++ [12]	88.6	84.9	72.6	92.0	75.1
ASFormer [17]	89.4	85.6	74.3	93.0	76.1
DXFormer (CC)	**89.8**	**86.0**	**74.7**	**93.4**	76.4
DXFormer	**89.8**	85.9	73.8	93.0	**76.5**

* CLIP is used here for frame-wise feature extraction
CC = Cross-Connection Only

81.7%. The incorporation of dual dilated attention and cross-connections in our architecture proves highly effective in improving the performance of the baseline. Moreover, on the HOI4D Office Tools dataset, where the action labels are more fine-grained, the cross-connection mechanism shows a particularly positive impact. In Table 4, DXFormer consistently outperforms other state-of-the-art models on this dataset as well. These results further validate the efficacy of the proposed approach, showcasing its potential to advance action segmentation and facilitate the development of robust video understanding models.

3.5. Qualitative Evaluation

To further examine the performance of DXFormer, we conducted a visual analysis of the model outputs for several videos. Although DXFormer performs similarly to the baseline in most general cases, our findings indicate that it outperforms the baseline in challenging cases. For instance, in Segment (a) of Figure 3, DXFormer accurately predicts the shaking action by capturing subtle local movements and global patterns in consecutive frames, while the baseline misclassifies it as a 'take' action. In Segment (b), we discover an error in the ground truth labeling of an action as 'shake', which should have been labeled as a 'background'. Despite the incorrect ground truth label, our model correctly predicts the label, demonstrating reliability in subtle action scenarios. Segment (d) presents a case where the pouring action is followed by the scooping action in the video. The baseline ASFormer model misclassifies the pouring action as 'scoop' due to the influence of the previous action, and irrelevant global patterns. However, DXFormer effectively captures more critical local contexts and correctly predicts the pouring action, demonstrating its ability to accurately segment actions in the presence of complex temporal dependencies. Overall, our model demonstrates greater flexibility and robustness by adaptively incorporating global and local context through the proposed dual dilated attention and cross-connection components, leading to better performance in challenging cases compared to baseline.

4. Conclusion

In this paper, we proposed two novel ideas, dual dilated attention and encoder-decoder cross-connection, to enhance the performance of the transformer backbones for the egocentric action segmentation task. We also conduct a comprehensive comparison of different feature representations for action segmentation models. Our proposed model demonstrates performance comparable to that of the baseline in general scenarios, while outperforming it in more challenging cases due to its ability to adaptively focus on temporally local and global contexts. Moving forward, we will continue to explore innovative ideas to further enhance the transformer backbone and extend our evaluations to a wider range of datasets, aiming to push the boundaries of egocentric action segmentation research.

References

[1] Hyemin Ahn and Dongheui Lee. Refining action segmentation with hierarchical video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16302–16310, 2021. 4

[2] 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. 3

[3] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 4

[4] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 4

[5] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 3

[6] Shang-Hua Gao, Qi Han, Zhong-Yu Li, Pai Peng, Liang Wang, and Ming-Ming Cheng. Global2local: Efficient structure search for video action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16814, 2021. 4

[7] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2322–2331, 2021. 4

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[9] Taein Kwon, Bugra Tekin, Siyu Tang, and Marc Pollefeys. Context-aware sequence alignment using 4d skeletal augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8172–8182, 2022. 1

[10] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 1

[11] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19880–19889, 2022. 1, 2, 3, 4

[12] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 4

[13] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 3

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[15] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 3

[16] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 34–51. Springer, 2020. 4

[17] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *The British Machine Vision Conference (BMVC)*, 2021. 2, 3, 4