# Semantics Guided Contrastive Learning of Transformers for Zero-shot Temporal Activity Detection
# Supplementary Material

Sayak Nag*, Orpaz Goldstein†, Amit K. Roy-Chowdhury*
*University of California, Riverside, USA, †Amazon, USA
{snag@ece, amitrc@ece.}ucr.edu, orpgol@cs.ucla.edu

## 1. Dataset Class Splits

### 1.1. Thumos'14

The class splits for Thumos'14 [3] are kept the same as that of [9], whereby 12 out of the 20 classes are considered as seen and the remaining 8 are considered unseen. The names of the seen and unseen classes are showcased in Table I.

### 1.2. Charades

The class splits for Charades [3] are kept the same as that of [9], whereby 120 out of the 157 classes are considered as seen and the remaining 37 are considered unseen. For brevity we only show the names of the 37 unseen classes in Table II.

## 2. Additional Implementation Details

### 2.1. Transformer parameters

The internal parameters of the transformer are listed in Table III. For both datasets a dropout of 0.1 is used during model training.

### 2.2. Network Initialization

For the experiments with I3D features we initialized the network with Xavier Normal initialization [2] where else for the experiments with C3D features we use Xavier Uniform initialization [2].

Table I: Thumos'14 seen and unseen class splits.

| **Seen Classes** | **Unseen Classes** |
|---|---|
| Basketball Dunk | Baseball Pitch |
| Billiards | Cricket Bowling |
| Clean and Jerk | Diving |
| Cliff Diving | Hammer Throw |
| Cricket Shot | Long Jump |
| Frisbee Catch | Shotput |
| Golf Swing | Soccer Penalty |
| High Jump | Tennis Swing |
| Javelin Throw | |
| Pole Vault | |
| Volleyball Spiking | |
| Throw Discus | |

Table II: Charades unseen classes.

| Unseen Classes | | |
|---|---|---|
| Throwing clothes | Eating a sandwich | Tidying on the floor |
| Opening a door | Taking shoes | Holding medicine |
| Sitting at a table | Holding a pillow | Taking a vacuum |
| Talking on a phone | Tidying a shelf | Lying on a bed |
| Holding a bag | looking at a picture | Watching television |
| Taking a book | Closing a window | Fixing a doorknob |
| Reading at a book | Taking a broom | Opening a refrigerator |
| Holding a towel/s | Holding a mirror | Someone is eating |
| Taking from a box | Turning off a light | Someone is dressing |
| Closing a box | Washing a cup | |
| Taking a laptop | Opening a closet | |
| Tidying up a blanket | Taking paper | |
| Sitting in a chair | Wash a dish | |
| Putting food somewhere | Sitting on sofa | |

Table III: Parameters of transformer. The encoder layers are that of a simple MLP.

| | |
|---|---|
| Number of attention heads | 8 |
| Numer of nodes in the feed-forward network | 1024 |
| Hidden dimension, $v$ | 512 |
| Numer of Encoder Layers, $L_E$ | 3 |
| Numer of Decoder Layers, $L_D$ | 6 |

Table IV: Performance of TranZAD with different number of decoder layers. For all cases, I3D features are used. For Thumos'14 the performance is shown in terms of mAP@tIoU=0.5, and for Charades, the mAP metric of [4] is used.

(a) Results of TranZAD-G

| | Number of Decoder layers | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 4 | 6 | 8 |
| Thumos'14 | 9.85 | 11.91 | 13.42 | **14.17** | 13.95 |
| Charades | 8.43 | 10.36 | 12.88 | **13.56** | 13.27 |

(b) Results of TranZAD-W

| | Number of Decoder layers | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 4 | 6 | 8 |
| Thumos'14 | 9.79 | 12.07 | 13.38 | **13.84** | 13.61 |
| Charades | 8.36 | 10.14 | 12.77 | **13.21** | 13.03 |

### 2.3. Feature Extraction

For each temporal window, we collect features at 5 fps i.e. chunks of 5 non-overlapping frames. Thus for a window of 500 frames, the extracted features have a temporal length $l_T$ of 100. However, the 3D backbones require a minimum number of frames to be supplied (frame stride) for feature generation, 8 for I3D [1] and 16 for C3D [7]. Therefore, we follow the feature extraction strategy of Tan et. al. [6]. For the training videos, following [9], we first remove any segment containing unseen class activities, then for the remaining video, we extract features with the minimum frame stride (non-overlapping) corresponding to each backbone. Following that, we map the extracted features to each of the 5 frames in a given temporal window as per the strategy of [6]. In this way we stack the features of each window $\hat{\mathbf{x}}$ to get $\mathbf{f}(\hat{\mathbf{x}}) \in \mathbb{R}^{l_T \times f_d}$, where $f_d = 2048$ for I3D and $4096$ for C3D.

## 3. Additional Ablations

### 3.1. Number of Decoder Layers

The performance of TranZAD w.r.t. varying numbers of transformer decoder layers, $L_D$ is shown in Table IV. The performance increases as the number of decoder layers are increased but stagnates after 6 layers, and therefore, we restrict to $L_D = 6$.

Table V: Performance of TranZAD with MLP and transformer encoder. For all cases, I3D features are used. For Thumos'14 the performance is shown in terms of mAP@tIoU=0.5, and for Charades, the mAP metric of [4] is used.

(a) Results of TranZAD-G

| | MLP Encoder | Transformer Encoder |
|---|---|---|
| Thumos'14 | **14.17** | 10.94 |
| Charades | **13.56** | 9.61 |

(b) Results of TranZAD-W

| | MLP Encoder | Transformer Encoder |
|---|---|---|
| Thumos'14 | **13.84** | 10.66 |
| Charades | **13.21** | 9.53 |

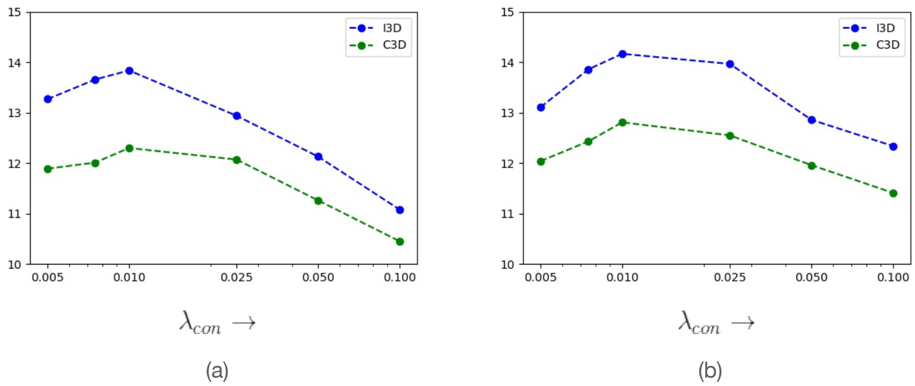


Figure I: mAP at tIoU $= 0.5$ for different values of $\lambda_{con}$ on Thumos'14. Figure (a) shows the results for TranZAD-W and figure (b) shows it for TranZAD-G. The blue line reflects sensitivity of TranZAD with I3D features and the green line is for TranZAD with C3D features.

### 3.2. MLP Encoder vs Transformer Encoder

Our choice of using an MLP encoder is inspired by recent studies [6], which show that the inherent slowness of video features makes the traditional transformer encoder prone to over-smoothing them, thereby reducing their discriminability. This phenomenon is exacerbated in our zero-shot setting, where the video features of the test classes are completely unseen during testing. We empirically evaluate this by replacing the MLP encoder with a traditional transformer encoder having the same number of layers, $L_E = 3$, and the results are shown in Table V.

### 3.3. Sensitivity analysis of $\lambda_{con}$

We perform a sensitivity analysis on $\lambda_{con}$ by varying it within $[0.005, 0.0075, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1]$ and recording the mAP@0.5 value as shown in Fig. I. The performance of TranZAD slowly improves when $\lambda_{con}$ is increased from $0$ and stagnates after $0.05$. Notably, setting $\lambda_{con}$ to $0.01$ yields the best results in our experiments. It can be observed that using $\mathcal{L}_{con}$ in conjunction with GLoVE embeddings gives the best results, which is also reflected in the overall performance of TranZAD on both datasets.

## 4. Analysis of Inference time

Although the authors of ZS-RC3D [9] did not release their code it is still possible to gauge its inference time by analyzing the same for the underlying RC3D [8] model on top of which it is built. On average, RC3D [9] takes about $3.06$s to perform inference on a $3.5$ min video. In comparison, TranZAD takes $0.24$s to perform inference on the same $3.5$ min video, where the inference is conducted on a single NVIDIA GeForce RTX 3090 and excludes the feature extraction time for both models. Due to the direct detection procedure, TranZAD is nearly $13\times$ faster since it is free of post-processing, such as non-maximum suppression, which is exclusively required for two-stage detectors like ZS-RC3D.

## 5. Additional Results on Charades

Temporal localization performance on charades [5] is commonly obtained in terms of Sigurdsson *et al.*'s [4] standard and post-processed mAP (mean average precision). The results shown in Table 3 of the main paper are in terms of the post-processed mAP of [4]. The authors of ZS-RC3D [9] also showed results in terms of the standard mAP of [4], along with

Table VI: Charades per unseen class standard AP(%), following Sigurdsson *et al.* [4]. The overall standard mAP(%), following [4], is shown at the very bottom of the table.

| | ZS_RC3D | TranZAD-W | TranZAD-G |
|---|---|---|---|
| Throwing clothes | 10.80 | 11.07 | **11.87** |
| Opening a door | 11.53 | **11.82** | 10.66 |
| Sitting at a table | **16.44** | 12.77 | 14.83 |
| Talking on a phone | 5.28 | 5.79 | **6.62** |
| Holding a bag | 7.86 | 11.62 | **12.74** |
| Taking a book | 3.93 | 5.10 | **5.91** |
| Reading at a book | 11.66 | 13.38 | **15.40** |
| Holding a towel/s | 12.87 | 10.91 | **13.65** |
| Taking from a box | 3.58 | 3.44 | **3.72** |
| Closing a box | **4.08** | 3.96 | 3.03 |
| Taking a laptop | 3.45 | 10.94 | **12.11** |
| Tidying up a blanket | **5.93** | 3.77 | 4.18 |
| Sitting in a chair | **18.09** | 16.81 | 17.46 |
| Putting food somewhere | **10.94** | 8.94 | 8.53 |
| Eating a sandwich | **7.96** | 7.57 | 6.81 |
| Taking shoes | **10.88** | 8.83 | 9.29 |
| Holding a pillow | 7.91 | **11.26** | 9.44 |
| Tidying a shelf | 4.84 | **5.43** | 4.81 |
| looking at a picture | **5.64** | 4.33 | 3.64 |
| Closing a window | 3.67 | 4.01 | **4.86** |
| Taking a broom | **10.35** | 9.89 | 9.07 |
| Holding a mirror | 2.69 | **3.04** | 2.73 |
| Turning off a light | **4.97** | 4.56 | 4.88 |
| Washing a cup | 4.05 | 4.68 | **5.16** |
| Opening a closet | 7.54 | **11.63** | 9.34 |
| Taking paper | 4.11 | 5.71 | **6.19** |
| Wash a dish | **9.59** | 3.68 | 4.36 |
| Sitting on sofa | **14.41** | 12.35 | 13.41 |
| Tidying on the floor | 8.14 | 8.48 | **10.70** |
| Holding medicine | 5.04 | 6.31 | **6.83** |
| Taking a vacuum | **5.63** | 4.75 | 3.97 |
| Lying on a bed | 10.10 | 11.55 | **12.79** |
| Watching television | **11.12** | 9.04 | 9.84 |
| Fixing a doorknob | 2.87 | **3.66** | 2.29 |
| Opening a refrigerator | 4.50 | **6.29** | 5.11 |
| Someone is eating | 5.32 | 5.18 | **6.18** |
| Someone is dressing | **14.90** | 9.23 | 9.18 |
| **Standard mAP** | 7.91 | 7.88 | **8.15** |

per-unseen class standard average precision (AP). We also compute the performance of TranZAD in terms of Sigurdsson *et al.*'s [4] standard mAP, and it is shown in Table VI along with the per-unseen class standard AP. For brevity, we show the best results obtained using the I3D features. The overall results show that TranZAD achieves comparable performance to ZS-RC3D for most of the unseen classes. In many cases, it also outperforms ZS-RC3D.

## 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.

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[3] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[4] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017.

[5] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[6] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13526–13535, 2021.

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[8] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.

[9] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander Hauptmann. Zstad: Zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2020.