TRANSTL: SPATIAL-TEMPORAL LOCALIZATION TRANSFORMER FOR MULTI-LABEL VIDEO CLASSIFICATION

Hongjun Wu[1] *Mengzhu Li*[1] *Yongcheng Liu*[2] *Hongzhe Liu*[1⋆] *Cheng Xu*[1] *Xuewei Li*[1]

[1] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Multi-label video classification (MLVC) is a long-standing and challenging research problem in video signal analysis. Generally, there exist many complex action labels in real-world videos and these actions are with inherent dependencies at both **spatial and temporal** domains. Motivated by this observation, we propose TranSTL, a spatial-temporal localization Transformer framework for MLVC task. In addition to leverage global action label co-occurrence, we also propose a novel plug-and-play Spatial Temporal Label Dependency (STLD) layer in TranSTL. STLD not only dynamically models the label co-occurrence in a video by self-attention mechanism, but also fully captures spatial-temporal label dependencies using cross-attention strategy. As a result, our TranSTL is able to **explicitly and accurately** grasp the diverse action labels at both spatial and temporal domains. Extensive evaluation and empirical analysis show that TranSTL achieves superior performance over the state of the arts on two challenging benchmarks, Charades and Multi-Thumos.

Index Terms— Multi-label Video Classification, Label Co-occurrence Dependency, Spatial Temporal Label Dependency, Transformer

1. INTRODUCTION

Multi-label video classification (MLVC) is a practical but challenging task in video signal processing. It aims to recognize multiple labels from a long and untrimmed video. As shown in Fig. 1, multiple action labels distribute in different spatial locations (a) and they often change over time (b). This indicates that labels are strongly related to the spatial-temporal context. Besides, there are also dependencies among labels. For example, 'sit on a sofa' and 'watch television' always happen at the same time. Therefore, the key challenges of MLVC are to not only capture the spatial and temporal label dependencies but also discover the label co-occurrence dependencies.

This work was supported, the National Natural Science Foundation of China (Grant No. 61871039, 62171042), the Academic Research Projects of Beijing Union University(No. ZB10202003, ZK40202101, ZK120202104).
*corresponding author: liuhongzhe@buu.edu.cn

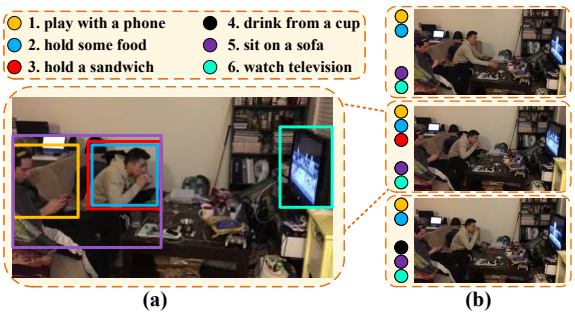

Fig. 1. Diverse actions are spatially distributed in different locations (a) and the action labels vary over time (b).

Graph convolution networks [1] have achieved great success in modeling category correlation, on account of the better representation of multivariate relationships. Hence many works are proposed to apply GCN for multi-label classification. ML-GCN [2] and MS-CMA [3] learn the static dependencies among labels of the target dataset with graph neural network. However, such static dependencies may lead to deviation in partial application. ADD-GCN [4] uses an updatable dynamic graph to achieve a more robust representation. PS-GCN [5] models the correlation among actions and attributes with a two-stream GCN. However, statistics-based label correlation may learn spurious correlation in the case of inaccurate statistics.

In order to capture spatial or temporal label dependencies, Sigurdsson et al. apply CRF on the feature map extracted by CNN to model the temporal label dependencies [6]. TRN studied the connections between multiple video clips [7]. To capture a longer range of time dependence, wang et al. [8] inserts non-local modules into 3D CNN. Timeception [9] builds multi-scale time convolution on the CNN feature map. For its preeminent performance in relation modeling, GCN is used to model the temporal and spatial label dependencies. STRG [10] finds regions of interest in time and space via GCN. To obtain higher-order relationships of actions, GHRM [11] designs temporal and semantic branches to fuse the basic relation information from different graphs. However, the methods above only exploit the local spatial-temporal label dependencies in the video.

Recently, Transformer [12] from NLP field has achieved great success in vision area [13][14][15], due to its pow-

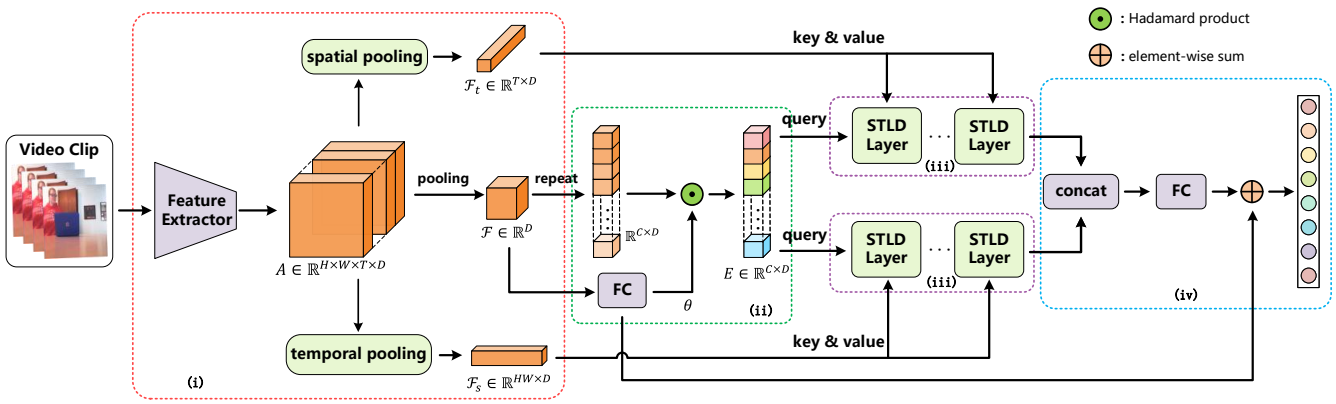

Fig. 2. The overall framework of our approach. Given a video clip, the proposed framework processes it with four steps. First, feature map A is extracted from a 3D CNN backbone and pooling functions are applied to A to obtain feature $\mathcal{F}_t$, $\mathcal{F}_s$, and $\mathcal{F}$ (shown in block (i)). Then $\mathcal{F}$ is decomposed into category-aware embedding E (shown in block (ii)). After that, E is refined using one or more attention-based Spatial-Temporal Label Dependency (STLD) layers in temporal branch and spatial branch (shown in block (iii)). Finally, outputs from both branches are concatenated for the final classification (shown in block (iv)).

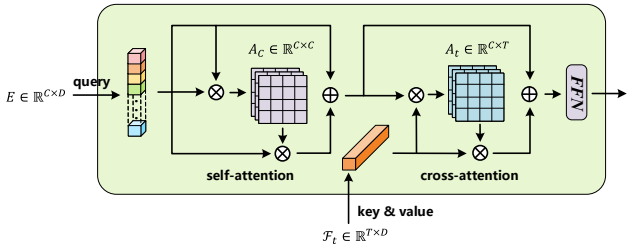

Fig. 3. An STLD layer in temporal branch of our framework. Given a category-aware embedding E as query, STLD layer models label co-occurrence dependencies with self-attention. After that, with feature $\mathcal{F}_t$ as key and value, cross-attention models dependency between specific category and each time step. STLD layers in spatial branch are similar to those in temporal branch.

erful capability to model longer range of relationships than CNN. Accordingly, we exploit Transformer in MLVC task for capturing complicated spatial-temporal label dependencies. Specifically, on the one hand, we directly mine the potential correspondences among labels implicitly from category-aware embedding. On the other, we encourage model to focus on temporal and spatial regions of interest to categories and learn the global temporal and spatial label dependencies.

Our main contributions can be summarized as follows, (1) A novel framework TranSTL based on Transformer decoder is proposed to model spatial temporal label dependencies and label co-occurrence dependencies. (2) A flexible plug-and-play STLD module is designed to realize the proposed framework. (3) In addition to the common label co-occurrence dependencies, we show the significance of spatial temporal label dependencies for MLVC.

2. METHOD

In this section, we describe our proposed framework. It consists of four main parts: (i) feature extraction, (ii) a feature decomposition module to generate category-aware embedding, (iii) feature refinement with our STLD layers which models spatial temporal label dependencies and label co-occurrence dependencies, (iv) final classification to transform the refined features into class probabilities. An overview of our approach is presented in Fig. 2.

Feature extraction. Given a video clip $I \in \mathbb{R}^{T_0 \times H_0 \times W_0 \times 3}$, where T_0 is the length of the video clip and H_0, W_0 are the height and width of each frame. The feature map $A \in \mathbb{R}^{T \times H \times W \times D'}$ is extracted from the backbone network. Then a convolution layer transforms the dimension of A from D' to D to match with the desired dimension of STLD layers.

The previous methods all regard the time T and space H, W of the video as a whole. Unlike them, we propose to decouple the video features from the two dimensions of time and space, and model the temporal label dependency and the spatial label dependency in two branches respectively. Therefore, three pooling functions: the temporal max pooling function, the spatial max pooling function and the global average pooling function are used on the feature map $A \in \mathbb{R}^{T \times H \times W \times D}$ to obtain the spatial feature $\mathcal{F}^s \in \mathbb{R}^{H \times W \times D}$, the temporal feature $\mathcal{F}^t \in \mathbb{R}^{T \times D}$, and the video level feature $\mathcal{F} \in \mathbb{R}^D$. After that, $\mathcal{F}^s \in \mathbb{R}^{H \times W \times D}$ is flattened to $\mathcal{F}^s \in \mathbb{R}^{HW \times D}$ to match the input of STLD.

Category-aware embedding. For multi-label classification, we treat each label prediction as a binary classification task (1-existence, 0-absence). With video level feature $\mathcal{F}$, we implement the binary classifier with a fully-connected layer as follow:

$$y_m = W^\top \mathcal{F} + b, \tag{1}$$

where $W \in \mathbb{R}^{C \times D}$ and $b \in \mathbb{R}^C$ is the parameter of classifier and $y_m = [y_m^1, y_m^2, \cdots, y_m^C]$ are the prediction scores which are used to generate the final scores. Then, we repeat $\mathcal{F} \in \mathbb{R}^D$ C times to generate the matrix $G = [\mathcal{F}, \mathcal{F}, \cdots, \mathcal{F}] \in \mathbb{R}^{C \times D}$. After that, the category-aware embedding $E \in \mathbb{R}^{C \times D}$ can be obtained by element-wise multiplication as follow:

$$E_{ij} = G_{ij} W_{ij}. \quad (2)$$

STLD layer. We propose a multi-layer relationship modeling architecture. It consists of a spatial branch and a temporal branch to capture spatial and temporal label dependencies as well as label co-occurrence dependencies. Each branch has one or more Spatial-temporal Label Dependency (STLD) layers. STLD is implemented with standard Transformer decoder architecture, where each layer contains a self-attention module, a cross-attention module, and a position-wise feed-forward network(FFN). In each branch, category-aware embedding E is regarded as query, and both $\mathcal{F}_s$ and $\mathcal{F}_t$ are used as key and value in cross-attention respectively.

The details in a STLD layer is shown in Fig. 3. In self-attention module, category-aware embedding $E_0 \in \mathbb{R}^{C\times D}$ is regarded as query, key and value. Then an attention matrix $A^c \in \mathbb{R}^{C\times C}$ can be obtained as follow:

$$A^c = softmax(E_0 E_0^T / \sqrt{D}). \quad (3)$$

The attention matrix A^c contains the action label co-occurrence relevance for classification. With this attention matrix, a refined category-aware embedding E_1 can be obtained as $E_1 = A^c E_0$.

In cross-attention module, E_1 inspects and selects spatial feature $\mathcal{F}_s$ or temporal feature $\mathcal{F}_t$ to combine. Take temporal branch as an example, we first generate an attention matrix $A^t \in \mathbb{R}^{C\times T}$ as follow:

$$A^t = softmax(E_1 \mathcal{F}_t^T / \sqrt{D}). \quad (4)$$

A^t denotes the presence (or absence) of specific action category in each time step. Then the refined embedding E_2 is obtained as $E_2 = A^t \mathcal{F}_t$. Intuitively, the embedding E continuously refines the temporal or spatial context information from the video via the cross-attention module layer by layer.

Final classification. Assuming we have l STLD layers in each branch, we can get the final category-aware embedding E_s^l and E_t^l output from the last layer. We concatenate the embedding from both branches to generate $E^l = [E_s, E_t] \in \mathbb{R}^{C\times 2D}$. To perform multi-label classification, we treat each label prediction as a binary classification task and project the feature of each class $E_c^l \in \mathbb{R}^{2D}$ to the logits $y_r = [y_r^1, y_r^2, \cdots, y_r^C]$ using a linear projection layer:

$$y_r^c = \sum_{i=1}^{c} (W_c^\top E_c^l)_i + b_c, \quad c = 1, 2, \cdots, C, \quad (5)$$

where $W_c \in \mathbb{R}^{2D}$, $W = [W_1, W_2, \cdots, W_C] \in \mathbb{R}^{C\times 2D}$ and $b_c \in \mathbb{R}$, $b = [b_1, b_2, \cdots, b_C] \in \mathbb{R}^C$. Then, we add up y_r and y_m with element-wise sum and obtain the final logits $y = [y_1, y_2, \cdots, y_C]$. Finally, we exploit a non-linear $Sigmoid$ function to map the logits $y = [y_1, y_2, \cdots, y_C]$ to the predicted probability scores of the range [0, 1].

Loss function. We use the optimized asymmetric loss [16], which is a variant of focal loss with focusing parameters γ to modulate positive and negative values. Given a video clip I, we get its category prediction score $y = [y_1, y_2, \cdots, y_C] \in \mathbb{R}^C$ and calculate the asymmetric loss via follow method:

$$\mathcal{L} = \frac{1}{C} \sum_{C=1}^{C} \begin{cases} (1-y_c)^{\gamma+} \log(y_c), & t_c = 1, \\ (y_c)^{\gamma-} \log(1-y_c), & t_c = 0, \end{cases} \quad (6)$$

where t is a binary ground truth to indicate if video clip has label c. The total loss is computed by averaging this loss over all samples in the training dataset.

Table 1. **Comparison with previous work on Charades.** IN means ImageNet-1K and K400 represents Kinetics-400.

method	pretrain	mAP
I3D[17]	K400	32.9
Timeception[9]	K400	37.2
STRG+NL[10]	IN+K400	37.5
VideoGraph[18]	K400	37.8
GHRM [11]	K400	38.3
SF-R50[19]	K400	38.0
VidTr-L[13]	K400	43.5
MViT-B,64×3[14]	K400	46.3
STLD-I3D(Ours)	K400	38.6(+5.7)
STLD-SF-R50(Ours)	K400	46.8(+8.8)

Table 2. **Comparison with previous work on Multi-Thumos**. K400 means Kinetics-400.

method	pretrain	mAP
I3D[17]	K400	72.43
Timeception[9]	K400	74.79
GHRM[11]	K400	79.89
STLD-I3D(Ours)	K400	80.31

3. EXPERIMENT

In this section, we first introduce datasets and implementation details. Then our method is compared with previous works. At last, we conduct ablation studies and visualizations.

Datasets and implementation. Charades[20] contains around 9.8K videos, among which about 8K for training and 1.8K for validation. It has 157 action labels and 66.5K annotated activities, about 6.8 labels per video. Multi-Thumos[21] dataset is an untrimmed dataset that contains 413 long videos, 200 for training and 213 for testing. There are a total of 65 actions in Multi-Thumos, with an average of 11 actions in each video. The metric mean average precision (mAP) is used for evaluation on both Charades and Multi-Thumos.

I3D[17] and SlowFast-R50[19] are chosen as our backbone. The hidden size D of STLD layer is 1024 and the number of total layers l is 2. In loss function, $\gamma+$ and $\gamma-$ is set to 1 and 4 respectively.

Comparison results. In table 1 we compare our TranSTL with three baseline networks, I3D and SlowFast-R50 as well as other state-of-the-art methods that work on Charades. Our method achieve significant improvement on mAP score over

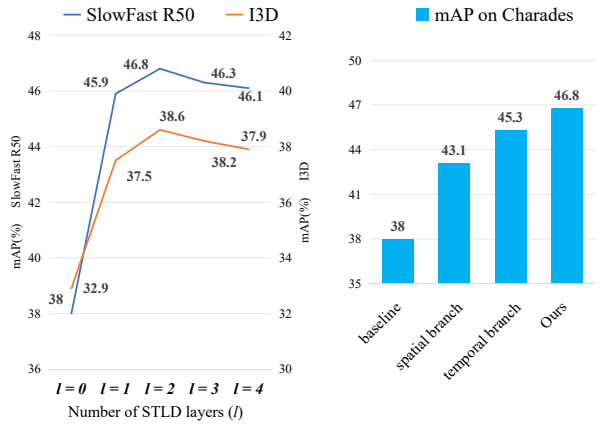

Fig. 4. Ablation studies on Charades. On the left is the comparison of using different numbers of STLD-layers (l). Our model achieves the best results on Charades with two STLD-layers. On the right is the effects of the temporal branch and spatial branch of our model.

all the baseline networks pre-trained with different datasets. Compared with baseline I3D and SlowFast-R50, TranSTL provides 5.7% and 8.8% higher mAP. With SlowFast-R50 as backbone, our method respectively outperforms MViT-B by 0.5% and VidTr-L by 3.3%. With the same backbone, I3D, our method outperforms GHRM, which captures high-order global semantic and local temporal relations. Moreover, we find that STLD improves SlowFast-R50 more than I3D. We speculate that the SlowFast framework is more helpful for our STLD to model spatial and temporal label dependencies, thanks to the higher temporal resolution feature in the fast pathway. These shows that modeling label co-occurrence dependencies and spatial-temporal label dependencies help the network to work better for multi-label video classification by plugging STLD layers into any video network.

We compare our method with other video classification methods on Multi-Thumos dataset in Table 2. Our model outperforms the I3D baseline and Timeception by 7.88% and 5.52% in mAP, respectively. Moreover, our model outperforms the state-of-the-art method GHRM, achieving state-of-the-art performance on this large dataset. It convincingly proves the effectiveness of our method.

Ablation studies. Since stacking multiple STLD layers is effective for modeling spatial-temporal label dependencies and co-occurrence label dependencies, we first test how performance changes as the number of STLD layers increases. on the left of Fig. 4, we show the effects of using different numbers of STLD layers on the Charades. The results show that stacking two STLD layers is sufficient for the Charades. By stacking more STLD layers, the performance on Charades declines continuously.

Our model contains a temporal branch and a spatial branch that aim to model spatial label dependencies and temporal label dependencies, respectively. We ablate these two branches by removing one of them. As shown on the right of Fig. 4, the results on Charades drop remarkably when either branch is removed, which demonstrates that the temporal branch and spatial branch are complementary.

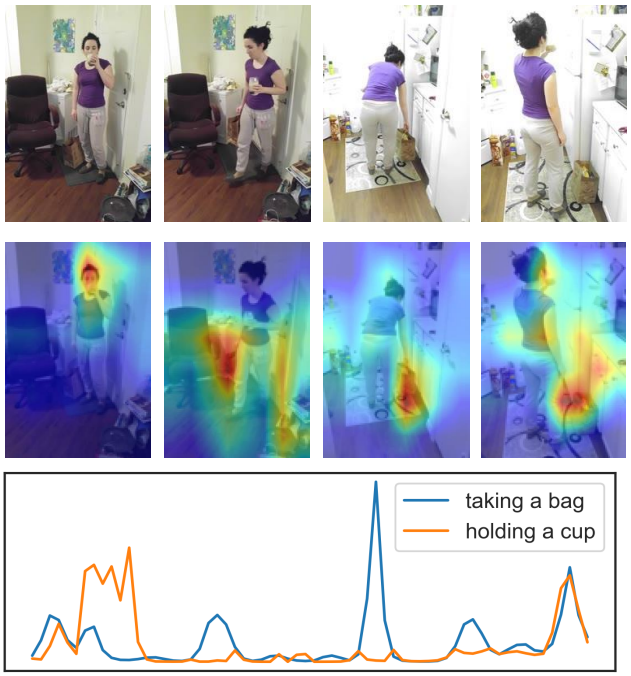

Fig. 5. Visualization of spatial and temporal attention in STLD. On the top are the original frames. In the middle are the heat maps of spatial attention. At the bottom is the visualization of the attention score of temporal attention.

Visualization. To further demonstrate the effectiveness of the proposed method, we visualize the cross attention maps in the last STLD layer. Partial visualization results are shown in Fig. 5. We find that our model is capable of locating the spatial important regions of actions or related objects and select relevant temporal instances. It indicates that our STLD layer is able to refine the feature to obtain more accurate spatial and temporal information and make the correct prediction.

4. CONCLUSION

In this paper, a novel spatial-temporal localization Transformer framework named TranSTL has been proposed to tackle the task of multi-label video classification (MLVC). In addition to the potential correspondence among labels, there are inherent dependencies at both spatial and temporal domains in a video. In TranSTL, our proposed Spatial Temporal Label Dependency(STLD) layer not only dynamically models the label co-occurrence in a video by self-attention mechanism, but also fully captures spatial-temporal label dependencies using cross-attention strategy. Therefore, our TranSTL is able to obtain more accurate semantic information from the video for MLVC task. The effectiveness of our method is evidenced by its favorable performances compared with other state-of-the-arts on two challenging benchmarks, Charades and Multi-Thumos. In the future, we will develop a lightweight MLVC model for higher efficiency.

5. REFERENCES

[1] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.

[2] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[3] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12709–12716.

[4] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 649–665.

[5] Junyu Gao, Tianzhu Zhang, and Changsheng Xu, "Learning to model relationships for zero-shot video classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3476–3491, 2021.

[6] 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*, 2017, pp. 585–594.

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

[8] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[9] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders, "Timeception for complex action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 254–263.

[10] Xiaolong Wang and Abhinav Gupta, "Videos as space-time region graphs," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 399–417.

[11] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng, "Graph-based high-order relation modeling for long-term action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8984–8993.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] Xinyu Li, Yanyi Zhang, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe, "Vidtr: Video transformer without convolutions," *arXiv preprint arXiv:2104.11746*, 2021.

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer, "Multiscale vision transformers," *arXiv preprint arXiv:2104.11227*, 2021.

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," *arXiv preprint arXiv:2106.13230*, 2021.

[16] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor, "Asymmetric loss for multi-label classification," *ICCV*, 2021.

[17] 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*, 2017, pp. 6299–6308.

[18] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders, "Videograph: Recognizing minutes-long human activities in videos," in *ICCV Workshop on Scene Graph Representation and Learning*, 2019.

[19] 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*, 2019, pp. 6202–6211.

[20] 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*. Springer, 2016, pp. 510–526.

[21] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 375–389, 2018.