STIT: Spatio-Temporal Interaction Transformers for Human-Object Interaction Recognition in Videos

Muna Almushyti[1,2] and Frederick W. B. Li[1]
[1]Department of Computer Science, Durham University, UK
[2]Deanship of Educational Services, Qassim University, SA

***Abstract*—Recognizing human-object interactions is challenging due to their spatio-temporal changes. We propose the Spatio-Temporal Interaction Transformer-based (STIT) network to reason such changes. Specifically, spatial transformers learn humans and objects context at specific frame time. Temporal transformer then learns the relations at a higher level between spatial context representations at different time steps, capturing long-term dependencies across frames. We further investigate multiple hierarchy designs in learning human interactions. We achieved superior performance on Charades, Something-Something v1 and CAD-120 datasets, comparing to baseline models without learning human-object relations, or with prior graph-based networks. We also achieved state-of-the-art accuracy of 95.93% on CAD-120 dataset [1] by employing RGB data only.**

I. Introduction

Some human-object interactions (HOIs) are difficult to recognize, e.g., human is cleaning an oven, or taking food from the oven, where the oven affords to different interactions including open, clean and close. Also, the presence of various items in the scene at the same time could affect model learning.

Early action recognition works, such as ConvNet [2], [3], recurrent neural networks (RNNs) [4], [5] and 3D convolution models [6], [7], learn a global representation of an action without considering human-object interactions. However, contextual information about an interaction, including human-object and object-object relationships, is critical and discriminative at specific times and throughout a video.

Recent work explored graph-based techniques for action recognition in videos [8], [9], [10], [11], [12], using spatio-temporal graphs to learn objects and human relations. Transformers [13], [14] also learn spatio-temporal relations in videos, e.g., [15] focuses on object layout relationships with global video representation and [16] considers spatial and semantic embeddings of objects, yet hierarchical spatio-temporal relations for HOI recognization remains unexplored.

Since discriminative cues about an interaction can be intensive at specific moments across video frames [9], we propose to learn interactions in a hierarchical manner. Inspired by Transformers in vision tasks, we exploit them forming our spatial and temporal learning network. Through the spatial transformer, relationship between human and objects is learned, revealing the local context even in case of objects not being close to human within a frame. Later, long temporal dependency between interactions at different frames is captured via the temporal transformer, where it receives compact representations of interactions at each frame across a video.

Unlike other works, e.g., [17], where different transformer-based architectures was proposed for video classification, we investigate the use of hierarchical structures in modeling human and object interactions through transformers. To our best knowledge, we are the first to study hierarchical modeling in human-object interactions with transformers based solely on visual appearance features. Our main contributions include:

- Developing a novel transformer-based framework to learn spatio-temporal interrelations between humans and objects in videos, which captures both long-term and non-local dependencies in HOIs across video frames.
- Investigating how different hierarchical organizations in network design impact HOI learning.
- Evaluating our model on three datasets, namely Charades [18], Something-Something v1 [19] and CAD-120 [1]. STIT is flexible in adapting any backbone without end-to-end training. It outperforms all counterpart approaches and achieves state-of-the-art result in CAD-120 dataset [1] with 95.93% accuracy using RGB data only.

II. Related Work

Video action recognition models. Using all-video-frame features is required by standard methods for action recognition in videos. 2D/3D convolution neural networks (CNNs) [20], [21], [22], [23], [3] are used to extract video features, e.g., I3D [6] inflates 2D convolutions to 3D, yielding good results on the Kinetics action dataset. Sequence modelling networks, e.g., RNNs and LSTMs [4], can also be used. Other approaches, e.g., [24], focus on long-term dependencies and learn pixel-wise [25] or interframe interactions at various time scales [24]. Optical flow and depth data can also improve action recognition on top of the visual information retrieved from RGB images [2], [26], [27], [28]. The above methods rely on complete video features (e.g., global descriptors) rather than discriminative indicators of an action (e.g., spatial and temporal interactions between objects and humans).

Transformers for Computer Vision. Transformer-based networks are successful in natural language processing (NLP) [14] and computer vision tasks, e.g., image classification [13], [29], object detection [30], video segmentation [31] and action recognition [32], [33], [34]. Recently, vision transformer (ViT) [13] achieved state-of-the-art performance in image classification without applying convolution layers, and is extended to video action recognition where spatio-temporal tokens are extracted from videos and fed to transformer encoders [35],

[17]. Transformer encoder-decoder networks can detect human, object and their interactions in an image, benefited from self-attention to capture better contextual relationships [36], [37]. In contrast, we apply hierarchical transformer encoders and study their effectiveness in learning relations of human and object tokens in space and time of a video.

Human-object relations in videos. Several studies explored visual relationships to recognize HOIs in videos [38], [39], [11]. Graph neural networks, namely graph convolution network (GCN) [40] and graph attention networks (GAT) [41], can capture spatio-temporal relation between visual nodes, including humans and objects [42], [8], [43]. In [8], GCNs are used through space and time graphs to capture the evolution of objects and their context throughout a video. Also, spatial and temporal relationships between humans and objects in each video are modeled through a graph attention model, considering their spatial distance [12]. Herzig et al. [9] learned hierarchical context of actions by considering the relation between visual phrases at frame level via non-local operation, then aggregating the features to learn the temporal context of these relations. Transformers can learn action context by observing the visual relations among entire video features and the human in the center clip [44]. Transformers can also capture spatio-temporal contexts of objects by considering object spatial information (e.g., location) and object's category [15], [16]. Our research exploits the relationship between humans and objects in a hierarchical way through transformers, with visual appearances of humans and objects being used solely.

III. Methodology

A. Network Overview of STIT

The overall architecture of STIT is shown in Fig. 1. The inputs to the network are the extracted human and object region features from a backbone feature map through RoIAlign [45]. They can serve as tokens for spatial transformer encoders rather than dividing each frame into N patches as in [17], [13].

B. Transformer

A transformer [13], [14] mainly comprises multiple layers with multi-head self-attentions (MHA) in each layer and feed forward layers (MLPs). Layer normalization with residual connections is applied before MHA and MLPs. In self-attention, the input is transformed into three forms through linear transformation producing queries (Q), keys (K) and values (V). Self-attention is written as:

$$Attention(Q, K, V) = Softmax(\frac{Q.K^T}{\sqrt{d_k}})\ V \qquad (1)$$

where d_k is the key dimension. Moreover, position encoding (PE) is attached to input embeddings in order to maintain a track of the position of each input token.

C. Spatio-Temporal Transformers in STIT

As ViT [13] is flexible in learning token relations, we adapt it forming our spatial and temporal transformer encoders. In spatial encoder, local and non-local dependency relations between human and objects (i.e., tokens) in a frame can be captured. Non-local means when objects and humans are distant from each other within a frame. Spatial-level interactions imply capturing local contextual information where human and object relations at the same time step are learned. This can be done through multi-head attention in the spatial transformer layer where all pairwise interactions between tokens (i.e., humans/objects) in a frame are captured. Hence, each token representation will be refined with respect to all other object tokens appeared at the same moment via self-attention, which effectively captures each object context. Since we adapt ViT, we prepend a learnable class token to objects at each time step, which is proven by ViT that generates a compact representation for an image. Our STIT considers it as a representation for local context at each time step. The input of a spatial transformer at time t is human and objects that are embedded via linear projection to generate tokens of 1D dimension with the size of 2048 each. As in [46], [13], a 1D learned positional encoding are also added to tokens for retaining their positional information. We can write the input to the spatial transformer at time t as:

$$X_t = [\texttt{class}_t, \psi(h_t^1), \psi(o_t^1), \psi(o_t^2), ..., \psi(o_t^N)] + P_t \qquad (2)$$

$$z_t = \text{Spatial-Transformer}_t(X_t) \qquad (3)$$

where ψ stands for linear embeddings, h^i and o^i respectively represent a human and an object visual feature at time t, and N is the number of objects. P_t indicates the learned positional embedding with $N \times d$ size. $\texttt{class}_t$ is an extra token that is prepended to tokens at each time step t. This class token is randomly initialized and via spatial transformer layers, the token is attended and gathered information from all other tokens in a frame at time t. z_t is the updated version of $\texttt{class}_t$, and is the output of the spatial transformer at time t. Thus, z_t represents the local context of interactions at time t.

To capture long-term HOI dependency, we add a second-level transformer for modeling temporal HOI evolution. The input tokens of the temporal transformer encoder are the updated class tokens outputted from the spatial transformers, that retain an abstract representation of interactions at each frame. The input to temporal transformer is then:

$$H = [\texttt{class}_{video}, \phi(z_1), \phi(z_2), \phi(z_3), ..., \phi(z_T)] + P_I \qquad (4)$$

$$Y_{interaction} = \text{Temporal-Transformer}(H) \qquad (5)$$

where ϕ is a linear transformation. Similar to spatial transformers, we prepended new class token to the token sequence which is $\texttt{class}_{video}$ in this level. z_i is the latent token generated by spatial transformers at temporal index i and T is the number of frames in a sequence. P_I is the positional encoding that learns and preserves the position of each token in a sequence. In the temporal transformer, the class token is attended to other tokens in the sequence, which are the compact representations of interactions at different time steps. Thus, high-level hierarchy of interactions is learned, providing discriminative cues of an action.

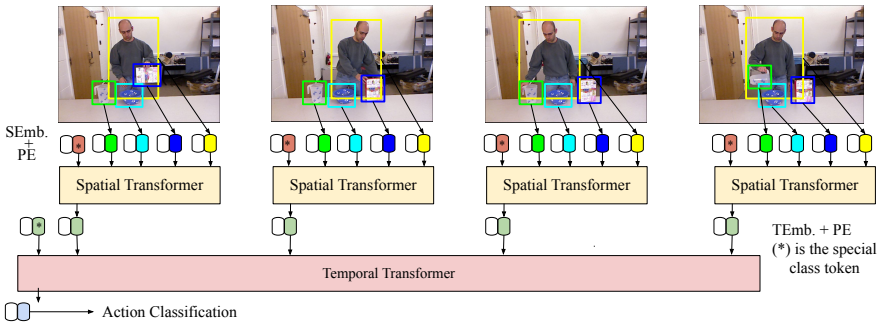

Fig. 1. Our proposed spatio-temporal transformer (STIT) model. SEmb. and TEmb. stand for spatial and temporal token embeddings, respectively.

IV. Experiments

We validate STIT on Charades [18], Something-Something v1 (SSv1) [19] and CAD-120 [1] datasets. Charades has 9,848 multi-label videos with indoor daily activities. During training phase, about 8K videos with action classes of 157 are used whereas 1.8K videos are used in the validation phase. We choose this due to its large-scale with most videos containing humans contacting with various objects. SSv1 contains 174 classes and 108,499 videos with single label each. Unlike Charades, most videos in SSv1 have clear background and actions that involve hands interacting with objects rather than involving whole human bodies. We also evaluate STIT on CAD-120 [1] which contains 120 videos with ten diverse human interactions performed by four different actors. Our implementation only utilized RGB images of CAD-120, despite it also contains depth images and skeleton information.

A. Experiments on the CAD-120 Dataset

1) Training details: To train STIT, we take 30 evenly sampled frames from each video. To extract human and objects features, we follow [47] where region of interest (RoIs) that indicate the bounding boxes of human and objects are cropped and reshaped to 224×224×3 for meeting the input size of 2D ResNet backbone [48]. Hence, 2048 features for each human and objects are extracted from ResNet-50 [48]. These human and object features are used as initial tokens for STIT. We present training hyper-parameters in Table I. The supplementary material contains further details.

2) Model Variants and prior works: We analyse four variants of our model, namely `LSTM-Spatial-T`, `LSTM-Pool`, `LSTM-GAT` and `GAT-Temporal-T`, to investigate the importance of our STIT model components, with `T` standing for transformer. `LSTM-Spatial-T` investigates the impact of a temporal transformer on learning temporal dependencies across frames. We replace our model with two layers of Long-short Term memory (LSTMs) [49] which may be used in sequence learning of videos [4], [5]. The `LSTM-Pool` and `LSTM-GAT` models investigate the role of the spatial transformer in understanding the spatial context of HOIs. We hence use pooling and Graph Attention Networks (GAT) [41] to replace our model. Finally, `GAT-Temporal-T` investigates how the spatial transformer affects the temporal transformer when it is replaced by GAT [41]. We also train an additional model (`NO-Relation`) that ignores the spatial relationship between humans and objects, instead concatenating their features and pooling them across time. Table II shows our STIT model outperforms all other variants. We observe a 2.62% drop in accuracy when replacing the the temporal transformer with LSTM, indicating temporal modeling via transformers is superior to LSTMs. We further observe our spatial transformers outperform GAT when they work with either LSTMs or temporal transformer. Finally, disregarding the spatial relationship between HOIs degrades model performance by 9.24%.

Fig. 2 compares HOIs recognition by model variants. In the first example, the spatial transformer gives more discriminative context than other models, successfully identifying the human is having a meal. In the second example, the human is stacking objects, which is analogous to the reversed action of "unstacking objects". The spatial transformer cannot understand such an action on its own. Yet our STIT correctly recognizes the action via the temporal transformer. Similarly, GAT with a temporal transformer can also correctly predict the action. This confirms that the temporal modelling via transformer outperforms LSTMs to recognize these type of interactions. Some examples of failure are shown in the last example, where the person is picking an object and all models incorrectly identify it as arranging objects. This may be because that in some videos the arranging and picking action of the same object could be similar but the difference is based on the human pose. We leave this for future work by considering human skeleton information. We provide a confusion matrix of our prediction results on CAD-120 [1] in the supplementary. Notably, as in Table III, comparing to previous works, which mostly use depth and skeleton data, our STIT model still achieves the best results even with RGB data only. Even [50] has proposed a 3D model to leverages RGB data for action recognition, our STIT model achieves 2.33% of higher accuracy. This demonstrates the importance of performing HOI reasoning both at each frame and over a course of HOIs. Also, transformer properties, such as multi-head attention and learnable token placements, along with a two-level hierarchy of human-object relation modelling, help our STIT model achieve state-of-the-art accuracy on CAD-120 [1].

TABLE I
A SUMMARY OF TRAINING SETTINGS FOR OUR STIT MODEL ON CAD-120 [1] AND CHARADES[18].

Dataset	Optimizer	LR	Epochs	Decay	Training Strategy
CAD-120 [1]	Adam	2.e-6	100	each 50 steps	Leave-One-Out Cross-Validation
Charades [18]	SGD	0.018	60	each 40 steps	Two-stage Training

Fig. 2. Prediction results of some actions by applying four different models on CAD-120 [1]. For simplicity, bounding boxes are not shown.

3) Ablation studies: To validate the effectiveness of each component of our STIT model, we conduct two main experiments with STIT-Spatial and STIT-Temporal. In STIT-Spatial, the temporal transformer is replaced by average pooling (e.g., over time dimension) whereas in STIT-Temporal the spatial transformers are replaced by pooling (e.g., pooling over nodes at time t). As shown in Table IV, ignoring either spatial or temporal hierarchy leads to decreased model performance. Moreover, a 2.60% performance loss over our STIT model is observed when omitting the temporal transformer, because long term dependencies between HOIs over time is not explicitly modeled. We also notice that model performance decreases significantly by 13.5% when when replacing the spatial transformer with pooling. Because human and objects features are merely extracted from ResNet-50 [48] that is pre-trained on ImageNet [51]. In contrast, embeddings in spatial transformers enhance the token features besides learning the relations between human and objects at each frame, which lead to model accuracy improvement to 95.93%.

As shown in Table IV, we conduct additional experiments to explore the affect of using class token as a representation of the spatial context at time t and for the video which is used as the output of the temporal transformer. STIT-spatial-mean, STIT-Temporal-mean and STIT-mean indicate replacing the output of spatial, temporal and both spatial and temporal transformers in STIT with mean token instead of latent class token, respectively. Notably, using latent token as the output of spatial and temporal transformers leading to better results.

B. Experiments on the Charades Dataset

1) Implementation details: To train our STIT, we employ two models as our backbones including Inflated 3D ConvNet (I3D) [6] with Resnet-50 and Slowfast-R50 [3]. We initialize I3D with pre-trained parameters on Kinetics-400 dataset [55] from [56]. For Slowfast-R50, we access the model via the Slowfast Github repository [56] where it has previously been trained on Charades. As input, we sample 32 (as in [8]) and 64 (as in [3]) frames from each video clip with 224×224 pixels for I3D and Slowfast-R50, respectively. We use a 2-stage training, which is different from [8], [9], [10], where we do not train the backbone and our model together for the third stage as end-to-end. This indicates the flexibility of our model to be integrated to any backbone with fewer number of training stages and with different settings of backbones including the one that is already trained on the same dataset as in Charades or using pretraind model as we used for training our model in CAD-120 [1].

Since Charades dataset does not provide human and object bounding boxes, we use Region Proposal Network (RPN) in Faster R-CNN [57] to produce object proposals. We use the top 15 proposals at each frame. We apply RoIAlign on the output feature maps of I3D model and the Slow path. Thus, human and object tokens are with the size of 2048 each after max pooling. Following [8], [9], [10], we concatenate the output of our STIT model with the output of the backbones (e.g., before FC), then fed the concatenated feature to fully connected layer with sigmoid activation for classification. Training hyper-parameters for Charades can be seen in Table I. We employ binary cross-entropy loss to train our STIT model with multi-label videos in charades. The supplementary material elaborates further implementation details.

From each video, we apply multi-view inference where 10 clips are sampled from a video as in [8], [3]. The evaluation metric is the mean average precision (mAP) where scores from different views are fused to report the results.

2) Comparison with state-of-the-art approaches: Table V shows the results of all prior methods that applied on the same dataset. The most close methods are those using the same backbone network as ours. It is observed that considering pose (P) information is not enough for correctly capturing HOIs. This indicates the importance of learning human-object relations in both space and time. Although we utilize fewer number of proposals (e.g., 15), our results are better than [8] where 50 proposals were used. Also, we achieve superior results comparing to STAG [9] that considers relations between a compact interactions, which include visual phrase (e.g., union box of both human and object). This indicates the power of learning the local context of human and objects through spatial transformers even without the visual phrases. Furthermore, learning the relation between visual tokens of human and objects gives more cues rather than considering the layout of human and objects as in STLT+I3D model. Also, our STIT

TABLE II
PERFORMANCE OF MODEL VARIANTS ON CAD-120 [1].

Model	Accuracy%
LSTM-Spatial-T	93.31
LSTM- Pool	90.26
LSTM-GAT,	92.47
GAT-Temporal-T	88.39
NO-Relation	86.69
STIT **(ours)**	**95.93**

TABLE III
RESULTS WITH CAD-120 [1]. NOTE THAT [52], [53], [1] AND [54] HAVE EMPLOYED ADDITIONAL SKELETON OR DEPTH INFORMATION.

Model	Accuracy%
Wang et al. [52]	81.2
Liu et al.[53]	93.3
koppula et al.[1]	80.6
Tayyub et al. [54]	95.2
Sanou et al. [50]	93.6
STIT **(ours)**	**95.93**

TABLE IV
ABLATION RESULTS ON CAD-120 [1].

Model	Accuracy%
Baseline (concat.)	86.69
STIT-Spatial	93.34
STIT-Temporal	82.43
STIT-spatial-mean	95.13
STIT-Temporal-mean	93.34
STIT-mean	95.04
STIT **(ours)**	**95.93**

TABLE V
COMPARISON WITH PRIOR APPROACHES ON CHARADES DATASET [18]. NOTE THAT SLOWFAST NETWORK ACHIEVED 45.2%MAP ON CHARADES USING R101 NETWORK BUT FOR FAIR COMPARISON WE REPORT SLOWFAST RESULTS WITH R50 NETWORK.

Model	Backbone	Modality	mAP%
2-Stream [58]	VGG-16	RGB+Flow	18.6
2-Stream+LSTM [58]	VGG-16	RGB+Flow	17.8
Async-TF [58]	VGG-16	RGB+Flow	22.4
Multiscale TRN [24]	Inception	RGB	25.2
I3D [6]	Inception	RGB	32.9
I3D [8]	R50-I3D	RGB	31.8
STRG [8]	R50-I3D	RGB	36.2
STAG [9]	R50-I3D	RGB	37.2
Pose and Joint-Aware [59]	R50-I3D	Pose+RGB	32.81
LFB Max [60]	R50-I3D-NL	RGB	38.6
STLT+I3D [15]	R50-I3D	RGB	38.5
I3D+STIT **(ours)**	R50-I3D	RGB	**39.62**
Slowfast 16 x 8 [3]	R50-3D	RGB	38.9
Slowfast 16 x 8+STIT **(ours)**	R50-3D	RGB	**42.49**

TABLE VI
ABLATION RESULTS ON CHARADES [18] USING I3D-R50 BACKBONE.

Model	mAP%
I3D	34.23
STIT-Spatial	36.60
STIT-Temporal	38.78
STIT-spatial-mean	38.94
STIT-Temporal-mean	38.64
STIT-mean	37.06
STIT **(ours)**	**39.62**

Fig. 3. Comparison between I3D and our STIT on Charades [18].

model can be incorporated with any backbone model rather than I3D without end-to-end training. As a result, our STIT model with Slowfast 16 x 8 surpasses its baseline. Thus, the results show that our STIT outperforms all other counterpart approaches which reflects the power of structure learning of HOIs through our two-level hierarchy of transformers.

3) Ablation studies: To evaluate STIT, we conduct ablation studies to demonstrate the impact of each part of STIT on learning HOIs. To study the impact of each hierarchy in our model on learning discrminative representation of HOIs, we conduct the same ablation experiments as in Sec. IV-A3, including STIT-Spatial, STIT-Temporal, STIT-spatial-mean, STIT-Temporal-mean and STIT-mean. Table VI shows the model performance after applying these settings. We find that removing the temporal transformer leads to a 3% decline in model performance whereas the performance loses only 0.84% when replacing the spatial transformers with pooling. This indicates the importance of temporal dependencies between interactions that can be captured via temporal transformer.

We apply different settings in using latent class token over the mean of transformer tokens. We observe that latent token provides better compact representation for spatial and temporal contexts, which are learned via spatial and temporal transformers, respectively. Also, learning human-object relations via our STIT outperforms the I3D baseline, achieving 5.39% mAP improvement. Fig. 3 shows examples of HOIs that our STIT performs better than I3D. Our STIT model can distinguish between different interactions with the same objects, such as taking, holding, and placing a laptop, whereas I3D cannot. Furthermore, our model can discriminate between how the same HOI can be performed with various objects, such as holding a towel versus holding a box. More importantly, interactions that occur simultaneously can be recognized. For example, it can be seen in Fig. 3, the human in the third example is washing a window and this interaction involved another interaction, which is holding towel at the same time.

C. Experiments on the Something-Something v1 Dataset

As in [8], we sample 32 frames and use 10 object proposals that are generated as in Charades experiment from each frame. We train our model on the top of fixed I3D backbone where we extract the tokens features from. We train our model for 50 epochs with batch size of 8 videos. We start with a 0.02 learning rate and it is reduced by a factor of 10 at 35,45 epochs. Supplementary material provides further details.

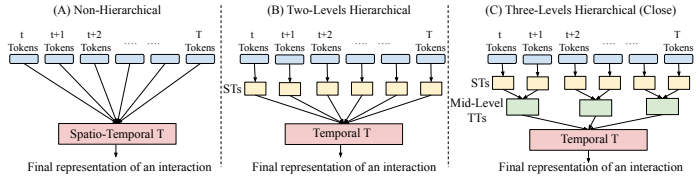

Fig. 4. Different network designs for modeling HOIs. STs and TTs stand for spatial transformers and temporal transformers, respectively. For simplicity, we use six frames as an example.

TABLE VII
PERFORMANCE OF STIT MODEL ON SOMETHING-SOMETHING V1 DATASET [19] COMPARED WITH PRIOR WORKS. TOP-1 ACCURACY IS REPORTED ON THE VALIDATION SET.

Model	Backbone	Top-1% Val
MultiScale TRN [24]	Inception	34.4
I3D [8]	R50-I3D	41.6
I3D+similarity graph [8]	R50-I3D	42.7
STRG [8]	R50-I3D	43.4
ECO [61]	BNInception+3D ResNet-18	46.4
TSM [62]	R50	44.8
TSN [63]	R50	19.9
STM [64]	R50	47.7
STIT **(ours)**	R50-I3D	47.92

As shown in Table VII, our STIT model outperforms other approaches, which confirms the importance of our proposed hierarchical learning of human-object interactions, even when learning different nature of interactions, i.e., hand-object interaction in Something-Something v1. Our hierarchical representation of actions outperforms other relation approaches without hierarchical representation, such as similarity graph [8], and other models relying on global representation of actions.

D. Structure Learning of HOIs via Hierarchical Designs

We now justify our network design in employing two-level hierarchies including spatial and temporal. We also consider different time windows (e.g., number of frames) for aggregating the local contexts with different number of temporal transformers, which are referred as close to a small time window (Close) and wide to a large window (Wide). For simplicity, in Fig. 4, we show example of close window with two frames. Note that for the Charades [18] experiments in Table VIII, the total number of frames is 16 and we use windows of 4 and 8 frames for Close and Wide windows, respectively. For CAD-120 Dataset [1], we choose 5 and 15 frames for Close and Wide windows, respectively. We run experiments with different designs of hierarchical modeling of HOIs including our model as shown in Fig. 4. Explanations of these designs are as follows: (A) Hierarchical learning is not considered. The pairwise relations between all tokens from different time steps are learned via the spatio-temporal transformer. (B) This is our STIT design where two-level of hierarchy is used to get the latent representation of HOIs. (C) We use small window of 2 where three mid-level temporal transformers are used to learn the relation between compact representations of HOIs with two frames range. Then, a higher-level temporal transformer models the relations between the mid-level representations of HOIs to produce the final representation of HOIs. Thus, we have three-level hierarchies of transformers including spatial, mid-level and high-level transformers. The last design amends (C) with a larger window frames (e.g., Wide).

TABLE VIII
RESULTS OF APPLYING DIFFERENT HIERARCHICAL DESIGNS IN MODELING HOIS. H STANDS FOR HIERARCHICAL

Architecture	Charades[18] (I3D)	Charades[18] (Slowfast)	CAD-120[1]
Three-Levels H (Close)	35.15	40.66	89.02
Three-Levels H (Wide)	35.23	40.86	84.44
Non-Hierarchical	38.91	41.24	94.21
Two-Levels H **(our STIT)**	**39.62**	**42.49**	**95.93**

As in Table VIII, we find that using more than two levels of transformers leads to model overfit where deeper levels of transformers can affect model generalization. Also, without hierarchical learning and using only one level of spatio-temporal transformer as in Fig. 4 (A), the model produces better results than three levels of hierarchy with specific temporal range because it captures the whole relations from different time steps. Hence, long-term temporal relations are captured well. Among all these architectures, we verify that our STIT with two-level hierarchy is the best for modelling HOIs and for capturing discriminative cues of action context.

Due to the different natures of actions and how they are being performed by human, some actions can be recognized with no-hierarchy, while others may require deeper-hierarchies. For example, recognizing a picking object action requires a deeper hierarchy in STIT while no-hierarchy fails to identify the action. In contrast, without a hierarchy in STIT, stacking objects actions in some videos are easier to be recognized. Because in picking objects actions, the spatial reasoning for objects at specific time is critical. However, we believe that in stacking objects, recognizing such action requires information about how the status of each object changes across time.

V. CONCLUSION

The structural learning of HOIs captures crucial cues about how human interacts with different objects. Our STIT explicitly uses hierarchical learning of the context of humans and objects to capture their interactions both at specific time and across a video. We show STIT has outperformed existing approaches on both Charades and CAD-120 datasets. By studying different levels of hierarchy for modeling HOIs, we find that two levels of hierarchy is enough for capturing local and global context of interactions via spatial and temporal transformers, respectively. In future work, we will investigate techniques to distill human objects contexts from various relation views to recognize HOIs.

Acknowledgement: This work made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is co-ordinated by the Universities of Durham, Manchester and York.

REFERENCES

[1] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202–6211.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[5] F. Li, C. Gan, X. Liu, Y. Bian, X. Long, Y. Li, Z. Li, J. Zhou, and S. Wen, "Temporal modeling approaches for large-scale youtube-8m video understanding," *arXiv preprint arXiv:1707.04555*, 2017.

[6] J. Carreira and A. 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.

[7] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

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

[9] R. Herzig, E. Levi, H. Xu, H. Gao, E. Brosh, X. Wang, A. Globerson, and T. Darrell, "Spatio-temporal action graph networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[10] H. Tan, L. Wang, Q. Zhang, Z. Gao, N. Zheng, and G. Hua, "Object affordances graph network for action recognition." BMVC, 2019.

[11] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell, "Something-else: Compositional action recognition with spatial-temporal interaction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1049–1059.

[12] M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, and R. Cucchiara, "Video action detection by learning graph-based spatio-temporal interactions," *Computer Vision and Image Understanding*, vol. 206, p. 103187, 2021.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] G. Radevski, M.-F. Moens, and T. Tuytelaars, "Revisiting spatio-temporal layouts for compositional action recognition," *The British Machine Vision Conference (BMVC)*, 2021.

[16] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang, "Spatial-temporal transformer for dynamic scene graph generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 372–16 382.

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021.

[18] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

[19] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[20] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[23] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

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

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[26] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 536–553.

[28] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–118.

[29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[31] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[32] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," *arXiv preprint arXiv:2104.11227*, 2021.

[33] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *arXiv preprint arXiv:2106.13230*, 2021.

[34] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," *arXiv preprint arXiv:2102.00719*, 2021.

[35] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" *arXiv preprint arXiv:2102.05095*, 2021.

[36] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 74–83.

[37] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei *et al.*, "End-to-end human object interaction detection with hoi transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 825–11 834.

[38] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, "Object level visual reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 105–121.

[39] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, and C. Schmid, "Actor-centric relation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 318–334.

[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[41] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[42] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2016, pp. 5308–5317.

[43] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Visual-semantic graph attention networks for human-object interaction detection," *arXiv e-prints*, pp. arXiv–2001, 2020.

[44] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[45] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[47] S. P. R. Sunkesula, R. Dabral, and G. Ramakrishnan, "Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 691–699.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[50] I. Sanou, D. Conte, and H. Cardot, "An extensible deep architecture for action recognition problem," in *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2019)*, 2019.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[52] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 97–106.

[53] Z. Liu, Y. Yao, Y. Liu, Y. Zhu, Z. Tao, L. Wang, and Y. Feng, "Learning dynamic spatio-temporal relations for human activity recognition," *IEEE Access*, vol. 8, pp. 130 340–130 352, 2020.

[54] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. G. Cohn, and D. C. Hogg, "Qualitative and quantitative spatio-temporal relations in daily living activity recognition," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 115–130.

[55] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[56] H. Fan, Y. Li, B. Xiong, W.-Y. Lo, and C. Feichtenhofer, "Pyslowfast," https://github.com/facebookresearch/slowfast, 2020.

[57] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[58] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta, "Asynchronous temporal fields for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 585–594.

[59] A. Shah, S. Mishra, A. Bansal, J.-C. Chen, R. Chellappa, and A. Shrivastava, "Pose and joint-aware action recognition," *arXiv preprint arXiv:2010.08164*, 2020.

[60] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 284–293.

[61] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 695–712.

[62] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[63] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[64] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "Stm: Spatiotemporal and motion encoding for action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2000–2009.