Body-part Tubelet Transformer for Human-Related Crime Classification

First Author[1][0000−1111−2222−3333] and Second Author[2,3][1111−2222−3333−4444]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

Abstract. Recent works show that vision transformers can demonstrate great capacity in solving Human Activity Recognition tasks based on skeletal trajectories. However, transformers typically lack inductive bias and are thus over-parameterized making them computationally complex. We propose to introduce inductive bias to pure transformer based approaches with the help of a 3D convolutional layer called Tubelet Embedder, which extracts spatio-temporal embeddings with the help of 3D convolution operation over the skeletal keypoints of a body part over sequential video frames. We conduct the experiments on the HR-Crime dataset and observe that this approach gives similar performance as the baseline approach while drastically reducing the computational complexity of the model. Thus, we demonstrate that Tubelet embeddings can replace the patch embeddings in pure transformers for Human Activity Recognition taks.

Keywords: Human Activity Recognition · Tubelet Embeddings · Transformers.

1 Introduction

In the past decade, there have been great advancements in the technological field in our world. Humans now live in a digital ecosystem. We are surrounded by motion sensors and cameras almost everywhere we go. In public places, governments and businesses have setup closed-circuit television cameras (CCTV) to monitor various activities like traffic, crowds and even to detect anomalous activities [6]. Even though humans have advanced a lot technologically, and complex surveillance systems are in place, we still observe a lot of criminal activities happening around us in plain sight. However, it's also a humongous task to manually inspect each video [1, 2] and there is a need to automate these tasks.

In the deep learning world, the task of identifying human activities from data like videos, skeletal trajectories, depth maps, etc. comes under the field of Human Activity Recognition (HAR). Most of the previous HAR works had focused on visual features extracted from videos like Spatio-Temporal Interesting Points (STIP) [7, 10], and recently, experiments with other kinds of data like those mentioned before have picked up. Deep learning methods also primarily focused on video data, however, they are computation intensive. Also, video features contain other kinds of information like background noises, lights, clothing etc. which could influence the results. Other features mentioned above have an advantage that they could be more descriptive compared to videos, like human skeletons which are compact, strongly structured and semantically rich.

Video Vision Transformer (ViViT) [3] introduced by the Google Research team used Tubelet embeddings which corresponds to 3D convolution capturing temporal and spatial information at once from videos. This introduced convolution to transformers. Typically, transformers are pretrained on large datasets since they lack inductive bias. However, convolutional networks have inductive bias and we investigate whether tubelet embeddings could help overcome the lack of inductive bias for transformers in HAR tasks. We propose an architecture which incorporates the tubelet embedding. For each body part, we pass the keypoints through a 3D convolution layer whose outputs are passed through a transformer encoder. This tubelet embedder is explored as an alternative for the patch embedding layer used in general transformers.

In [5], Boekhoudt et.al. analyzed several transformer architectures and explored different representations for human body movement. It was built on the works of Zheng et.al. [14]. While the architecture proposed in [14] lifted a 2D pose to a 3D pose by analyzing a sequence of skeletal trajectories, [5] modified the same architecture and used it to classify the skeletal trajectories by replacing the regression head with a classification head. Boekhoudt's work demonstrates that model architectures designed for pose estimation could also be modified for HAR tasks. We would be using this architecture as our baseline model.

We study how tubelet embeddings can be incorporated to transformer architectures for HAR tasks to investigate whether they could help encode information such as movement of different body parts. We also investigate if we can overcome the lack of inductive bias in transformers with the help of tubelet embeddings and thus reduce the computational complexity.

The contributions of this work are:

1. We incorporate a Tubelet Embedder to different transformer architectures and demonstrate that they can help match the performance of the baseline architecture with significantly lesser computational costs.
2. While obtaining similar results as to the baseline architectures, we improve the interpretability of the results by analyzing them visually using attention heatmaps and T-SNE plots.

The rest of the paper is divided as follows. In Section II, our proposed framework is presented and in Section III, we discuss the experimental setup. Section IV presents and discuss the obtained results. Section V will finally conclude the research work.

2 BPTubeFormer: Our proposed Body Part Tubelet Transformer

The proposed model has been illustrated in Figure. 1. We call this architecture, Body Part Tubelet Transformer (BPTubeFormer). We replace the patch embedding layer in the baseline model [5] with a Tubelet Embedder layer. The Tubelet Embedder is essentially a 3D convolution layer which takes as input, keypoints rearranged in the shape of a matrix and does 3D convolution over them using a 3D kernel.

Given a skeletal trajectory extracted from a video, we divide it into different segments based on the user defined segment length and assign the action label of the video to all the segments. We reshape the keypoints to a matrix and feed them into a Tubelet Embedder to obtain different embeddings as output. These would replace the patch embeddings used in the original transformers. These Tubelet Embeddings are passed to a Transformer encoder to capture the temporal and spatial relationship between the different keypoints within a body part. Here we use five different Transformer Encoders, one for each body part. Thus, each encoder can exclusively learn about a body part. The output features are concatenated to obtain the Tubelet Encoding of that particular body part.

We concatenate a class embedding token to these five tubelet encodings along with their position embeddings. These are passed through a Body Part Transformer and the class token is passed through a classifier head to obtain the activity classification. With the help of this architecture, we learn and interpret the movements belonging to a body part.

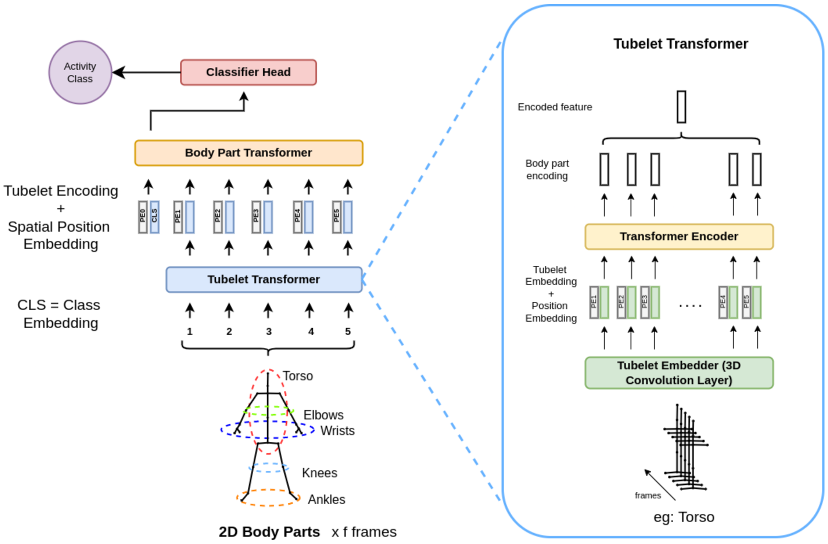


Fig. 1: Overview of the proposed BPTubeFormer.

Input representation: The input we have is a sequence of 1D keypoints while for dealing with 3D convolution, we need inputs in the shape of a square matrix. We take the body part with most number of keypoints as the standard and pad the keypoint data of all other body points to meet the shape of that body part representation. For example, consider a dataset which has a body part containing 9 body points. This means we have 18 values to represent that body part. We can rearrange them in the shape of 3 × 3 × 2. This is already in a convenient shape. For a smaller body part which has only 2 keypoints, we can reshape the 4 values to a 1 × 2 × 2 matrix but we also need to pad 2 rows and 1 column, as seen in Figure **??**, to match the matrix shape of the largest body part. This is necessary because different shapes would mean different number of embeddings as the result for 3D convolution and it would be impossible to concatenate them to the same dimension.

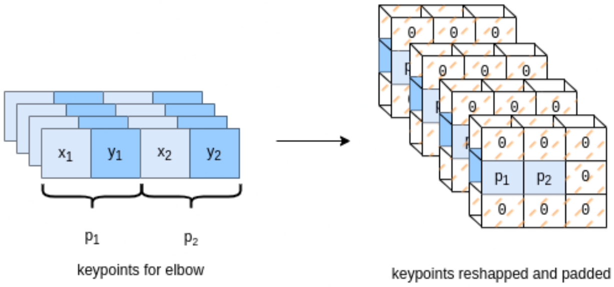


Fig. 2: Input representation: reshaping and padding of keypoints.

3 Experimental setup

The HR-Crime dataset [4] is a subset of UCF-Crime dataset [13] which is a surveillance video dataset consisting of 789 human-related anomaly videos and 782 human-related normal videos. It consists of anomaly videos from 13 categories: Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shooting, Shoplifting, Stealing and Vandalism [**?**]. The skeletons thus extracted had 17 keypoints.

Baseline: As the baseline model, we use the ST-Tran architecture proposed by Boekhoudt et.al. [5] which obtained the state of the art performance on the HRC dataset as can be seen in Table 1.

Experiments, exploring Tubelet Embeddings:

1. Experiment 1: BPTubeFormer with different kernel sizes for the tubelet embedder.

2. Experiment 2: BPTubeFormer with replication padding for the inputs.
3. Experiment 3: BPTubeFormer with different strides for the tubelet embedder.
4. Experiment 4: BPTubeFormer with different segment lengths for the tubelet embedder.

Implementation details: The experiments in this research work are carried out using the PyTorch machine learning framework which is based on the Torch library. The models are trained with a learning rate of 0.001. We also use an early stopping criteria which would stop training if there is no improvement in the validation loss for 3 epochs. The batch size was set to be 1000, however, few of the models took more memory and hence the batch size had to be reduced to 500 or 100. The weights were updated using the Adam optimizer and the loss was calculated using the cross entropy loss function.

Validation: We evaluate the models in two ways, qualitatively and quantitatively. Quantitative evaluation would help us determine which model/architecture performs the best while qualitative evaluation would help us understand how the transformer learns and interprets the problem.

Quantitative evaluation: To evaluate the models quantitatively, we use various metrics. We report the metrics with their mean and standard deviation, since we do 3 fold cross validation. We use the Balanced Accuracy of the model which is a weighted metric, to evaluate the models quantitatively since the HRC dataset exhibits class imbalance.

In the end, we summarize the results using a confusion matrix for the best performing model. To analyze the complexity of the models, we also compare the number of multiply–accumulate (MAC) operations as well as the total number of trainable parameters in the model. The MAC operation computes the product of two numbers and adds that product to an accumulator ($a \leftarrow a + (b \times c)$). It is the most common operation used in machine learning models due to its usage in matrix operations. These are calculated using the THOP[3] library.

Qualitative evaluation: For qualitative evaluation, we use the best performing model to predict the activity of a given trajectory. We can then generate heatmaps using the attentions scores of the Body Part Transformer and the Tubelet Transformer to explore which parts of the input the different heads attend to. In the body part transformer, we would consider the attention scores between different body parts, and in the tubelet transformer, we would consider the attention scores between different tubelet embeddings for each body part. The proper classifications and misclassifications would also be analysed.

We also plot the t-distributed Stochastic Neighbor Embedding (T-SNE) plots [8] and Silhouette plots [12] of the class token embeddings generated by the BPTubeFormer models for the test sets from both the datasets. T-SNE essentially helps us to understand high dimensional data and project it to lower dimensional space. T-SNE plots are generated using the Tensorboard[4] toolkit.

[3] THOP : https://github.com/Lyken17/pytorch-OpCounter

[4] Tensorboard : https://www.tensorflow.org/tensorboard

Silhouette plots demonstrate how well objects are classified as clusters of data using the Scikit-learn library [11].

4 Results and discussion

Quantitative Evaluation: The quantitative results of our experiments have been summarized in Table. 1. For Experiment 1, we experiment with different kernel-stride values. We follow the ViViT [3] approach by keeping the same values for both the kernel and the stride. The experiments are done with a segment length of 24 and we change the kernel-stride size from (1, 3, 3) until (8, 3, 3). The highest balanced accuracy was obtained with a kernel-stride size of (2, 3, 3) which was 0.4982 ± 0.0004. It was observed that the performance dropped constantly as we increased the kernel-stride size. The least balanced accuracy was obtained with the largest kernel-stride size of (8, 3, 3). A kernel-stride size of (1, 3, 3) which considers only one frame at a time also gave a lower balanced accuracy. Thus it can inferred that shorter tubelets (more than one frame) give rise to better performance. Shorter tubelets would also mean more number of tubelet embeddings generated. For example, in the case of this experiment, a kernel-stride value of (2, 3, 3) would give 12 tubelet embeddings while a kernel-stride value of (8, 3, 3) would give only 3 tubelet embeddings.

Model	Model	Kernel	Stride	Segment Length	Balanced Accuracy
Baseline	ST-Tran [5]	-	-	60	0.4926 ± 0.0043
	M1.2 [9]	-	-	-	0.3820 ± 0.0050
Exp. 1	BPT-V1	(1, 3, 3)	(1, 3, 3)	24	0.4797 ± 0.0065
	BPT-V2	(2, 3, 3)	(2, 3, 3)	24	**0.4982 ± 0.0004**
	BPT-V3	(4, 3, 3)	(4, 3, 3)	24	0.4873 ± 0.0048
	BPT-V4	(5, 3, 3)	(5, 3, 3)	24	0.4768 ± 0.0017
Exp. 2	BPT-8-V1	(2, 2, 2)	(2, 1, 1)	24	0.4890 ± 0.0025
	BPT-8-V2	(2, 2, 2)	(1, 1, 1)	24	0.4917 ± 0.0022
Exp. 3	BPT-V4	(2, 3, 3)	(2, 3, 3)	60	0.4899 ± 0.0023

Table 1: The performance of the proposed BPTubeFormer (BPT) model for the different implemented experiments on the HR-Crime [4] dataset. The first two rows present the performance of the baseline experiments [5] and [9].

For Experiment 2, we kept the kernel and stride values as different values. For example, kernel-stride values of (2, 2, 2)-(2, 1, 1) and (2, 2, 2)-(1, 1, 1). This would mean that some convolutions would overlap. It was observed that these kernel-stride values slightly lowered the balanced accuracy values by 1%. Thus, we infer that keeping the same value for kernel and stride (as proposed in ViViT [3]) extracts more information with the help of disjoint convolutions.

With Experiment 3, while padding the square matrices, we used the *replicate* padding mode which pads the cells with the values on the boundary of the input. By default, we had been using the *constant* padding mode which pads with 0. We observe that using the *replicate* mode, the performance lowered by nearly 2% compared to the *constant* padding mode. This might be suggesting that, with the *replicate* padding mode, we might be adding noise to the input and *constant* padding with 0 keeps the actual information. In Experiment 4, we experiment with a longer segment length of 60. In the other experiments, we had been using a segment length of 24. We observe that with a segment length of 60, the performance dropped by nearly 1%. This suggests that longer segment lengths cannot guarantee better performance.

We use a confusion matrix to analyze the classification performance of BPTubeFormer which gave the best performance. This can be seen in Figure 3. It can be seen that all models perform better than random guessing ($\frac{1}{13} \approx 0.0769$). We observe that out of the 13 classes, 10 classes exhibit an accuracy above 50% (*Robbery* has nearly 50% accuracy). The least performance was seen for the class *Arson*, 22%. The best performance was seen for the class *Shoplifting*, similar to the finding in [5]. However, the accuracy for the class Vandalism saw great improvement, rising to 50% from 0.04% in the baseline model . Accuracies of some other classes also improved, namely, Shooting, Fighting and Explosion. Accuracy for the class Arson got reduced by 3 times.

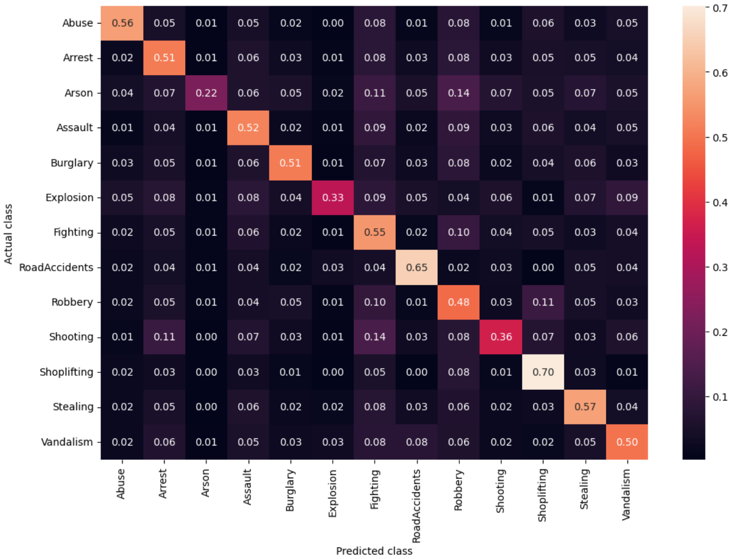


Fig. 3: Confusion matrix for the BPTubeFormer-2 model.

Qualitative Evaluation: In the figure below we present an example of attention of our BPTubeFormer. **(a, b, c, d) Attention heatmap**: Attention matrix of the last layer for heads 1 to 8. (a) represents the attention weight between the different body part tubelet encodings. Here index 0 corresponds to the class token. (b), (c) and (d) shows the attention weights between the different tubelet embeddings within the Tubelet transformer for the body parts Torso, Wrists and Elbows respectively. **(e) Skeleton body parts and attention**: The importance of each body part for the class token is illustrated. Each blob represents a body part - blue for torso, green for elbows, red for wrists, purple for knees, and yellow for ankles. The bigger the blob, the higher the attention score.

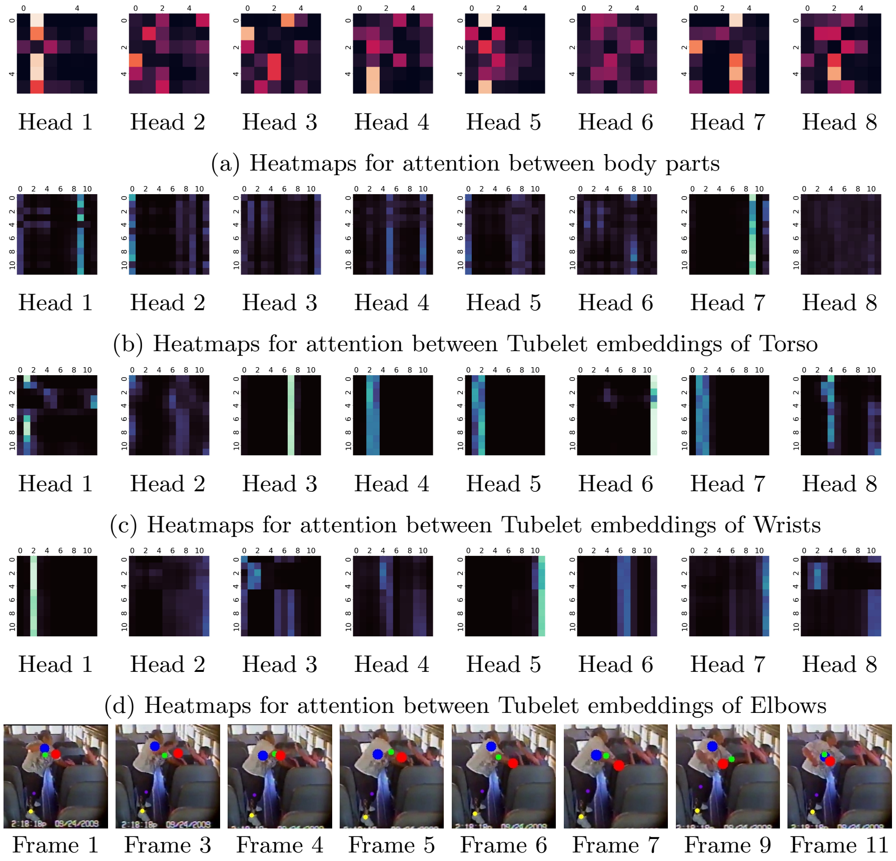
We take a segment from the *Abuse* class in which a woman is hitting a child on a running bus. Some important frames from this video can be seen in figure 4e. From Figure 4a, it can be observed that the first, fourth and fifth heads focus most on the second column (Torso). The seventh head focuses on the fourth column (Wrists). The second, third and eighth heads focus on the third column (Elbows). The last two columns (Knees, Ankle) are focused on none of the heads. This makes sense since it can be observed in Figure 4e that in all the frames, the lower body part remains constant and only the upper body part is under action. The wrists and torso blobs are the biggest and the knee and ankle blobs are the smallest, as expected.

Figures 4b, 4d and 4c show the attention weights between different tubelet embeddings within the Tubelet transformer for the body parts torso, elbows and wrists. The segment length is 24 and the kernel size is (2,3,3). Hence, the number of tubelet embeddings is 12, with each of them representing two frames. This is why the attention matrix is in the shape 12×12. It can be seen that for the tubelet embeddings of the Wrists, Head 1 focuses on the frames 3 and 4. Heads 5 and 7 collectively focuses on frames 3 to 6. Head 3 completely focuses on frames 15 and 16. Head 4 focuses on frames 5 to 8.

For the Torso, Heads 1 and 9 focus on frames 19 and 20. Also, Head 4 focuses on frames 11, 12, 21, 22. For the Elbows, Head 1 focuses on frames 3 and 4, Head 5 and 7 on frames 23 and 24, Head 6 on frames 13-16. It can be seen that wrists being the most active body part in this activity, it is attending to the frames 3 to 6 where it shows the most movement, as can be seen in Figure 4e.

Thus, while comparing the attention maps of the Tubelet transformer, we observe that different heads attended to the tubelet embeddings corresponding to the frames which represented important movements in the activity. For example, the movement of wrists while Abusing. This proves that tubelet embeddings can detect movements of various body parts and helps in better interpretability of the models.

Silhouette plots. Figure **??** shows the silhouette plot for the BPTubeFormer model. We observe an average silhouette score of 0.0413 with generally low values for the silhouette coefficients for all the classes. Most embeddings in the Abuse, Road Accidents, Stealing and Vandalism classes are cohesive indicated by the long right tail and the height of each silhouettes in Figure **??**. Separation value (indicated by the left tail of the silhouettes) is high for the classes Assault, Burglary, Explosion, Fighting, Robbery, Shooting and Shoplifting. This indicates that they have embeddings which are largely separated and overlaps with other clusters. There are also some classes which have most of their embeddings separated. These are Assault, Burglary and Robbery.



(a) Heatmaps for attention between body parts

(b) Heatmaps for attention between Tubelet embeddings of Torso

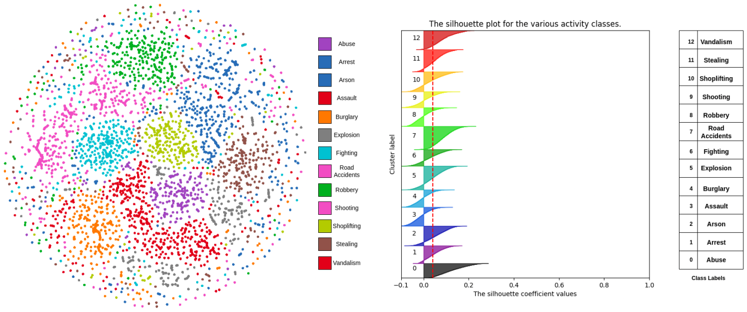
(c) Heatmaps for attention between Tubelet embeddings of Wrists

(d) Heatmaps for attention between Tubelet embeddings of Elbows

(e) Visualization for the attention scores for each body part

Fig. 4: Visualization of self-attention in BPTubeFormer on a sequence of 24 frames from a video in the *Abuse* category of the HR-Crime dataset [4]. **(a, b, c, d) Attention heatmap**: Attention matrix of the last layer for heads 1 to 8. (a) represents the attention weight between the different body parts. Index 0 corresponds to the class token. (b), (c) and (d) shows the attention weights between the different tubelet embeddings within the Tubelet transformer for the body parts Torso, wrists and elbows respectively. **(e) Skeleton body parts and attention**: The importance of each body part for the class token is illustrated. The size of the blob represents the relevance of the body part - blue for torso, green for elbows, red for wrists, purple for knees, and yellow for ankles.

T-SNE plots. We visualize the T-SNE plot for Experiment 1 in Figure 5. The plot for the HRC dataset can be seen in Figure 5a. We can observe that the embeddings are tightly packed and some embeddings are loosely scattered around the edges. Tensorboard displays embeddings with non-unique colours and hence, some colours are shared by two different classes. For example, red is shared by *Assault* and *Vandalism*. We can see that the embeddings for Robbery and Shoplifting are close to each other with some embeddings of Shooting placed in between them. Overall, most of the embeddings are well clustered.



(a) T-SNE.

(b) Silheoutte score.

Fig. 5: (a)T-SNE plots for the BPTubeFormer models on the HRC dataset. (b) Silhouette plot for the classifications of the BPTubeFormer model on the HRC dataset. The y axis labels indicate the activity class labels. The red line indicates the average silhouette score for all the classes.

Computational Complexity of the models: The comparison of the models based on the complexity is given in Table 2. We observe that while the baseline ST-Tran [5] model uses 613.5×10^9 MACS operations, our BPTubeFormer model uses just 30.5×10^9 MACS operations which is nearly 20 times less. In terms of the number of parameters, the ST-Tran model uses 9.5 Million parameters and our BPTubeFormer model uses almost half of it, 4.9 Million parameters. The ST-Tran architecure has a Spatial Transformer which deals with 17 (keypoints) 32-dimensional embeddings and a Temporal Transformer Encoder which deals with 60 (frames) 544-dimensional embeddings. In the case of the BPTubeFormer architecture, it has a Tubelet Transformer which deals with 12 (tubelet embeddings) 32-dimensional embeddings and a Body Part transformer which deals with 6 (body parts) 384-dimensional embeddings.

The high number of MACS operations for ST-Tran model is because it takes in a segment length of 60, thereby significantly increasing the number of embeddings. The less computational complexity of the architectures which use Tubelet Embeddings is due to the fact that Tubelet embedders introduce 3D convolutional learning and 3D convolutions can extract spatial and temporal dependencies at the same time, with low computational cost. This avoids the

Model	MACS ($\times 10^9$)	# Params (Millions)
ST-Tran [4]	**613.5**	**9.5**
BPT-V2	**30.5**	**4.9**

Table 2: Comparison of the computational complexity of different models in terms of MACS and the number of trainable parameters. MACS is expressed in the order of 10^9 and the number of parameters in millions.

need to have separate Spatial and Temporal encoders with transformers. Convolutional learning also has the advantage that it introduces inductive bias to the architecture which transformers typically lack. This asserts the fact Tubelet Embedders are a promising method to improve the performance of Transformers used in HAR tasks. The above result prove that the proposed models which use Tubelet embeddings can match the performance of the architectures proposed by Boekhoudt et.al. with much lesser computational costs.

Limitations It was observed that Tubelet embeddings helped transformers to nearly match the performance of patch embeddings on both the datasets. A simple TTubeFormer model which replaced the patch embedding layer in T-Tran model [5] with a Tubelet Embedder layer matched the performance of the baseline model on HRC [**?**] dataset,

5 Conclusions

In this work, we present a case study of our proposed BPTubeFormer, a tubelet-based transformer model for human-related crime recognition tasks. We observe that they matched the existing baseline performance but with significantly lesser computational costs. We demonstrated that Tubelet Embedders could encode the movement of different body parts with better interpretability which is relevant for HAR tasks. With the help of Tubelet Embedders, we were able to introduce inductive bias to pure transformer based approaches. Thus we conclude that incorporating tubelet embeddings to transformers could benefit HAR tasks and the performance could be further improved with more complex model architectures.

References

1. Cctv camera market size, trends, growth and overview – 2030: Mrfr (Feb 2020), https://www.marketresearchfuture.com/reports/cctv-camera-market-8160
2. Video surveillance storage market segmentation by enterprise type (large enterprise, and small amp; medium enterprise); by end-user (residential,

commercial, defense, industrial, and others); by deployment type (on-premise, and cloud)-global demand analysis amp; opportunity outlook 2031 (Oct 2022), https://www.kennethresearch.com/report-details/video-surveillance-storage-market/10154361

3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer (2021). https://doi.org/10.48550/ARXIV.2103.15691, https://arxiv.org/abs/2103.15691
4. Boekhoudt, K., Matei, A., Aghaei, M., Talavera, E.: HR-Crime: Human-Related Anomaly Detection in Surveillance Videos (2021). https://doi.org/10.34894/IRRDJE, https://doi.org/10.34894/IRRDJE
5. Boekhoudt, K., Talavera, E.: Spatial-temporal transformer for crime recognition in surveillance videos. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (2022)
6. Cosgrove, E.: One billion surveillance cameras will be watching around the world in 2021, a new study says (2019)
7. Laptev, I.: On space-time interest points. International Journal of Computer Vision pp. 107–123 (2005). https://doi.org/10.1007/s11263-005-1838-7, https://doi.org/10.1007/s11263-005-1838-7
8. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
9. Matei, A.D., Talavera, E., Aghaei, M.: Crime scene classification from skeletal trajectory analysis in surveillance settings (2022). https://doi.org/10.48550/ARXIV.2207.01687, https://arxiv.org/abs/2207.01687
10. Mohana, D., U M, M.: Human action recognition using stip techniques. International Journal of Innovative Technology and Exploring Engineering (2020). https://doi.org/10.35940/ijitee.G5482.059720
11. Pedregosa, F., V.G.e.a.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
12. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics pp. 53–65 (1987)
13. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos (2018). https://doi.org/10.48550/ARXIV.1801.04264, https://arxiv.org/abs/1801.04264
14. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers (2021). https://doi.org/10.48550/ARXIV.2103.10455, https://arxiv.org/abs/2103.10455