Dynamic Graph Warping Transformer for Video Alignment

Junyan Wang[1]
junyan.wang@unsw.edu.au
Yang Long[2]
yang.long@ieee.org
Maurice Pagnucco[1]
morri@cse.unsw.edu.au
Yang Song[1]
yang.song1@unsw.edu.au

[1] School of Computer Science and Engineering,
University of New South Wales,
Sydney, Australia

[2] Department of Computer Science,
Durham University,
Durham, UK

Abstract

Video alignment aims to match synchronised action information between multiple Video alignment aims to match synchronised action information between multiple video sequences. Existing methods are typically based on supervised learning to align video frames according to annotated action phases. However, such phase-level annotation cannot effectively guide frame-level alignment, since each phase can be completed at different speeds across individuals. In this paper, we introduce dynamic warping to take between-video information into account with a new Dynamic Graph Warping Transformer (DGWT) network model. Our approach is the first Graph Transformer framework designed for video analysis and alignment. In particular, a novel dynamic warping loss function is designed to align videos of arbitrary length using attention-level features. A Temporal Segment Graph (TSG) is proposed to enable the adjacency matrix to cope with temporal information in video data. Our experimental results on two public datasets (Penn Action and Pouring) demonstrate significant improvements over state-of-the-art approaches.

1 Introduction

The amount of video materials available through online platforms, *e.g.,* YouTube, has been growing rapidly. Research in video action understanding is a pressing need due to its wide application in video recognition, human-computer interaction, etc. However, action categories hardly capture the dynamic progression of an action. For example, during a push-up action, a video consists of body up and body down phases as shown in Figure 1. Therefore, the human action video alignment problem has received increasing attention in recent years, which aims to automatically synchronise actions between multiple videos.

The nature of the video alignment task initially encourages research focusing on supervised learning [9, 12, 29] where each frame is predicted to match one of the action phases. In addition, Song *et al*. [31] proposed unsupervised alignment with additional language resources, and NN-Viterbi [27] combines a neural network and a non-differentiable Viterbi

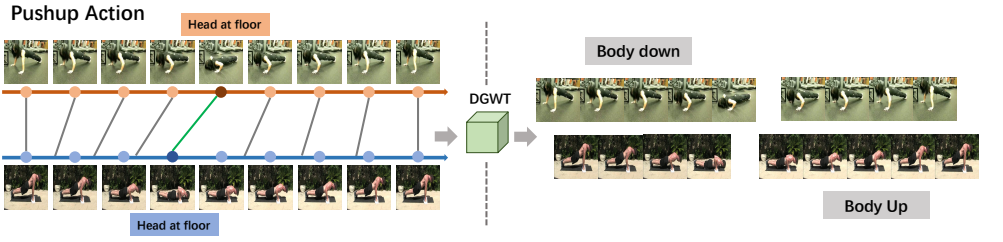

Figure 1: An illustration of the video alignment pipeline using our proposed DGWT model. The learned attention representations are used for fine-grained temporal understanding. Highlights of DGWT include a temporal segment graph and dynamic attention warping.

process to learn from ordering supervision iteratively. At the same time, direct manual annotation for visual correspondence across videos is challenging and infeasible to scale. Thus, we focus on designing a self-supervised approach. Recently, self-supervised learning methods have been utilised to deal with the problem that an action phase can have an arbitrary length and contain a varying number of frames in different videos. For example, Dwibedi *et al*. [12] introduced a differentiable cycle-consistency loss that can be used to find corresponding time points in multiple videos. Nevertheless, most of the existing research applies a simplified approach by assuming that the phases can be aligned through key events. However, predicting the key events does not guarantee perfect alignment. As illustrated in Figure 1, the push-up action can be broken into body down and body up phases but each phase can be completed in different time intervals with arbitrary speed of motion within that phase. Therefore, within an action phase, a direct frame-by-frame mapping (with interpolation) between videos would not produce a perfect alignment.

To this end, this paper proposes a novel self-supervised video alignment model, Dynamic Graph Warping Transformer (DGWT), which predicts the action phase labels using a Transformer-based spatio-temporal feature extraction method, with a temporal graphical operation inserted in between the spatial and temporal transformers to further enhance the learning of frame dependencies. Meanwhile, end-to-end learning is guided by a dynamic attention warping loss function to optimise the model explicitly for frame alignment in addition to frame-wise phase label prediction.

Our design is motivated by the following factors. 1) Spatio-temporal feature learning remains a challenging issue in the video analysis task. In contrast to previous state-of-the-art architectures with sophisticated convolutions, *e.g.,* 3D CNN or memory design, we embrace the recent development of Vision Transformer (ViT [1]) as the *Spatial Transformer* and another *Temporal Transformer* to encode spatial and temporal features. 2) Compared to many successful graph neural networks on images, temporal data does not exhibit a straightforward graphical structure. How the adjacency matrix can be built in the context of video alignment remains unexplored. Thus, to enhance the ability of the temporal transformer for identifying key frames and matching the global action progress, we design a *segment-level graph operation* to formulate the given sequence into a graphical model. 3) In order to achieve explicit optimisation of frame-level alignment, we introduce the *warping* problem into the self-supervised paradigm. In line with the spirit of self-supervised learning, we explore frame-level guidance between different videos via warping. Most dynamic warping methods are imposed on low-level signals while action videos often contain extensive visual

features and complex content. We thus designed a *warping loss* that is customised for videos with multiple stages and various causal relations between actions.

Given these design considerations, our contributions are summarised as follows: 1) To the best of our knowledge, we present the first approach that successfully introduces graph neural networks and warping to learn representations for video alignment that demonstrates state-of-the-art performance. 2) The spatio-temporal transformers are applied to encode local temporal dependency into an arbitrary-length high-level attention representation. 3) Within the spatio-temporal transformers, a Temporal Segment Graph (TSG) is designed to convert temporal data into a graphical structure. With TSG, frames within a segment can be smoothed and the distinction between segments is enhanced to effectively identify the transition of adjacent action phases. 4) We propose a Dynamic Attention Warping (DAW) loss function that can compare videos of variable lengths based on their attention representations to find the optimal frame alignment without extra supervision.

2 Related Work

Video Alignment. Human action recognition is a fundamental and well studied problem in computer vision, and various standard benchmarks span across still images [6, 28, 37] through to videos [17, 19, 30, 32]. Human action alignment as a branch of video recognition task has recently received growing attention in the community. Early studies have presented various solutions to this problem, including unsupervised learning [31], weakly-supervised learning [5], and self-supervised learning [12, 29]. In particular, self-supervised learning methods can effectively deal with arbitrary length videos. Time-Contrastive Networks (TCN) [29] is a self-supervised approach for learning representations and robotic behaviors entirely from unlabelled videos recorded from multiple viewpoints. Temporal cycle-consistency (TCC) [12] loss is used to find correspondences across time in multiple videos. Recently, Purushwalkam *et al*. [26] enhanced TCC loss to learn correspondence in space and time via cross video cycle consistency. Besides, Cao *et al*. [4] proposed a few-shot learning framework (Ordered Temporal Alignment Module) that can learn to classify a previously unseen video. In contrast, we apply graph neural networks and Dynamic Programming loss to help the model learn the representation without label supervision.

Graph Attention Networks. Neural network algorithms for processing graphical data have become one of the most important machine learning areas [36]. Specifically, Graph Convolutional Networks (GCNs) [7, 8, 9, 14] can learn local and global structural patterns of graphs with convolutional functions. However, typically the graph node neighbourhoods are aggregated with equal or pre-defined weights which can vary greatly. Thus, other methods [33, 34, 35] have applied the attention mechanism into graph neural networks. Graph attention networks (GATs) [34] utilise self-attention to enhance node features. The attention-based graph neural network (AGNN) [33] is designed to replace all the intermediate fully-connected layers with propagation layers and attention mechanisms. In our work, in contrast to other graph attention networks, DGWT restructures sequential data into a segment-based graphical structure data to capture relationships in videos.

Dynamic Programming. Dynamic Time Warping (DTW) [22] is one of the most popular self-supervised algorithms for measuring similarity between two temporal sequences, and computes the best possible alignment between two time series of different lengths. This method has been widely applied to analyse time series applications, such as speech recognition [25, 39]. Recently, DTW approaches have been applied to a few video analysis tasks.

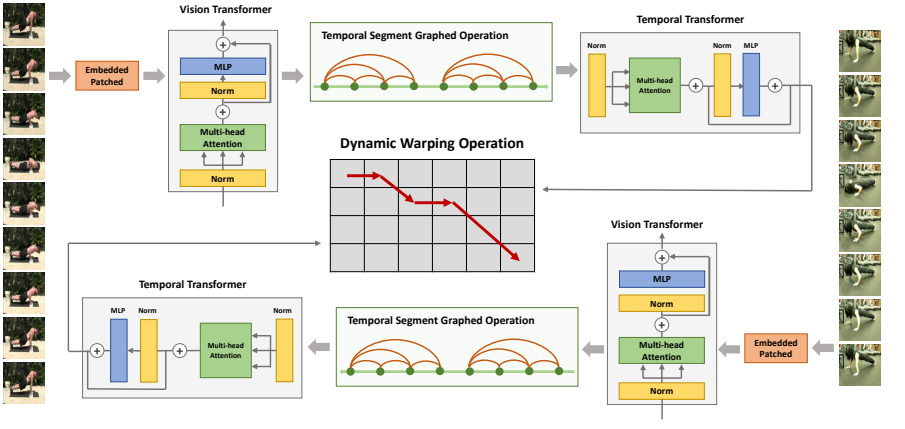

Figure 2: Overall architecture of our proposed DGWT model. This end-to-end framework consists of Spatio-temporal Transformer, graphical operation and warping optimisation.

Chang *et al*. [5] proposed Discriminative Differentiable Dynamic Time Warping (D3TW) that solves sequence alignment with discriminative modelling and end-to-end training. The Time Shift Dynamic Time Warping (TS-DTW) model [20] is derived for performing automatic alignment whilst achieving data selection and matching between inherently inaccurate and incomplete sequences. For video alignment, Haresh *et al*. [15] leverage the combination of temporal alignment loss using Soft-DTW and temporal regularization terms for aligning video sequences, and Hadji *et al*. [13] combined DTW loss with global cycle-consistency loss to enforce the temporal alignment. Our work focuses on aligning action attention and obtaining an effective graphed attention representation by utilising dynamic programming.

3 Dynamic Graph Warping Transformer

In this section, we present our Dynamic Graph Warping Transformer (DGWT) network for self-supervised feature learning. The problem is essentially modelled as a sequential labelling process. Formally, given a sequence of T frames from a video $X = \{x_1, \ldots x_t, \ldots x_T\}$, we aim to build an end-to-end model to predict the action phase labels $Y = \{y_1, \ldots, y_t, \ldots y_T\}$. In our proposed DGWT, a Spatio-temporal Transformer is designed to extract spatial and temporal features from videos. We also design a Temporal Segment Graph (TSG) that can divide a video into several fixed-length segments to learn frame dependencies to better identify key events. In addition, a key component of our method is to extract *between-video* information for video alignment. To find the optimal frame alignment without extra supervision, we introduce Dynamic Attention Warping into the loss function so that the attention representation can be constrained by action phases and progress information. An overview of our DGWT model is illustrated in Figure 2.

3.1 Pure Transformer Architecture Overview

The first step in video analysis tasks is generally spatio-temporal feature extraction. The success of attention-based models in NLP has recently inspired approaches in computer vision

to integrate transformers into vision tasks, such as action recognition [1, 2, 24]. Recently, a pure-transformer based architecture has outperformed its convolutional counterparts in image classification, which is Vision Transformer (ViT) [11]. Inspired by recent action recognition networks [1, 2, 24], we propose a spatio-temporal transformer model for video alignment as shown in Figure 2. Therefore, we propose a spatio-temporal Transformer model for self-supervised feature learning.

As the architecture overview shown in Figure 2, the input video $X \in \mathbb{R}^{T\times h\times w\times c}$ is firstly mapped to a sequence of tokens $Z \in \mathbb{R}^{n_t\times n_h\times n_w\times d}$ by using the same patch embedding method as ViT, and then the tokens are reshaped into $\mathbb{R}^{N\times d}$ after adding position embedding. The spatial vision transformer forwards all tokens extracted from the video through the transformer encoder to extract visual features of each frame $H^s = \{h_t^s\}_{t=1}^T \in \mathbb{R}^{T\times d}$. The temporal transformer then applies attention mechanisms to encode global dependencies for the extracted visual feature sequence H^s and outputs the temporal attention representation of the video sequence $H^a = \{h_t^s\}_{t=1}^T \in \mathbb{R}^{T\times d}$.

In a self-attention block of transformers, the queries, keys and values $Q = XW_q$, $K = XW_k$ and $V = XW_v$, are linear projections of the input X with Q, K, $V \in \mathbb{R}^{N\times d}$. The process of a self-attention block is defined as:

$$A^l = LN(Softmax(\frac{QK^T}{\sqrt{d_k}})V) + A^{l-1}\ , \tag{1}$$

$$H^{a^l} = MLP(LN(A^l)) + H^{a^{l-1}}, \tag{2}$$

where H^{a^l} denotes l^{th} self-attention map and d_k represents the attention feature dimension. LN and MLP represent the Layer Normalisation and Muti-Layer Perception, respectively.

Even though Transformers can extract both spatial and temporal representations, they cannot perform sufficiently well for the video alignment task. Specially, the temporal Transformer has difficulty identifying key frames and matching the global action progress. Therefore, we further propose the TSG component to enhance the ability of the temporal Transformer and DAW loss to find the optimal frame alignment without extra supervision.

3.2 Temporal Segment Graph

There is significant research investigating dependencies between complex information in video data. Since the emergence of the deep learning paradigm, spatial dependency has been encoded via convolution operations. However, the key challenge of video alignment is to learn dependencies between frames in a local range and identify key frames and match them with global actions. The temporal attention mechanism has the ability to learn these dependencies and, thus we propose a temporal segment graph to help the temporal transformer distinguish key frames among global actions.

Compared to spatial temporal convolution and recursive neural networks, graphical models [34] provide a new solution to model frame dependencies. However, graphical models are normally applied on images that can be naturally divided into spatial partitions. For video analysis, we need to construct graphs to represent the sequential information in videos effectively. To address this problem, we propose a novel solution to build fixed-size graphs using temporal segments, and the operation is conducted in between the spatial and temporal transformers. In this graph $\mathcal{G} = \{V, E\}$, nodes V capture the spatial representation of nodes V capture the, and the edges E represent the relationship between nodes in a segment. To

formulate this step, we use D to denote the diagonal matrix of node degrees, and A denotes the adjacency matrix. Thus with the normalized graph Laplacian matrix [36], the temporal graph operation can be denoted as

$$H^{graph} = \sigma(D^{-\frac{1}{2}}AD^{\frac{1}{2}}H^{s}W) \,, \tag{3}$$

where W is a layer-specific trainable weight matrix and $\sigma()$ denotes an activation function. Given the extracted spatial features H^s of all frames and the adjacency matrix A, the output H^{graph} represents the graph enhanced feature. Note that the length T of each video is different. Therefore, it is not trivial to create a universal graphical model across all videos.

In the video alignment task, the key objective is to detect key event frames that can distinguish two adjacent action phases. Because a video may contain multiple repetitive action phases, we suggest to keep each segment short so that each segment can distinguish the fine changes between frames. For example, in push-up actions, the up and down phases are almost visually reversible. With short segments, the segment that contains key event frames, *e.g*, the highest and lowest body positions, can be distinguished. To this end, we build a local graph partition with an aggregation operation:

$$D_{\tau}(i,i) = \sum_{t=j}^{L}(A_{\tau}(i,j)) \,, \tag{4}$$

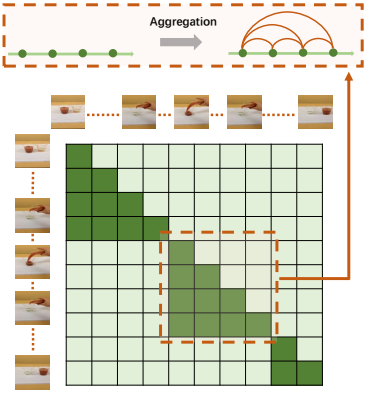

Figure 3: The process of building a graph with aggregation operation in a segment. The green matrix represents the diagonal degree matrix.

where D_{τ} and A_{τ} represent the τ^{th} segment of D and A respectively. The aggregation operation of one segment is shown in Figure 3. The $T \times T$ adjacency matrix indicates T/L graph partitions, each of which denotes a segment. If T is not a multiple of L, the remaining frames would contain a smaller segment. After the graphical operation, the transition matrix will smooth the frame-level spatial features in each segment and increase the distinction between adjacent segments, to the benefit of the temporal transformer for better understanding the progress of the actions in an action sequence.

3.3 Dynamic Attention Warping

Different individuals may complete an action in variable time intervals and speed. In other words, a perfect prediction model does not guarantee accurate alignment. For example, a down-phase of 1.5 seconds cannot be uniformly extended (interpolated) to match another one with 3 seconds because the pace may be changing during the course of action. Therefore, our solution here is to utilise self-supervised learning to tackle this challenge and introduce the dynamic warping algorithm so that the temporal attention representation learning can be guided by more accurate alignment information.

We locate this self-supervised alignment as a dynamic time warping problem, which has been widely used in alignment tasks. We denote two attention representation sequences

$H^{a^\beta} = \{H_1^{a^\beta}, H_2^{a^\beta}, \ldots, H_l^{a^\beta}\} \in \mathbb{R}^{l \times d_a}$ and $H^{a^\gamma} = \{H_1^{a^\gamma}, H_2^{a^\gamma}, \ldots, H_m^{a^\gamma}\} \in \mathbb{R}^{m \times d_a}$ of length l and m corresponding to the video X^β and X^γ. The goal of the Dynamic Attention Warping (DAW) is to find the best alignment automatically. We impose rigid constraints on eligible warping paths based on the observation that each video frame can only be aligned to a single action phase label. Thus, we first build the cosine matrix $\Delta(H^{a^\beta}, H^{a^\gamma}) := [\cos(H_i^{a^\beta}, H_j^{a^\gamma})]$. Then, we calculate the cost $r_{l,m}$ of aligning frame x_i^β of video β to frame x_j^γ of video γ as in Algorithm 1. An example of the warping path is shown in Figure 4.

Algorithm 1: Forward recursion to compute the alignment cost.

```
/* l : the length of video X^β;                    */
/* m : the length of video X^γ;                    */
/* r : the distance;                               */
```

Input : $H^{a^\beta}, H^{a^\gamma}$

$r_{0,0} = 0; r_{i,0} = \infty; r_{j,0} = \infty$

for j = *1 to m* **do**

  for i = *1 to l* **do**

    $r_{i,j} = \cos(H_i^{a^\beta}, H_j^{a^\gamma}) + min\{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\}$

Output: $r_{l,m}$

As shown in Figure 4, all paths that connect the upper left entry Δ_{11} to the lower right entry Δ_{lm} using only $\rightarrow\searrow$ moves. Thus, the DAW loss can be defined as $\mathcal{F}_{DAW} = r_{l,m}$ that minimises the alignment cost between the two attention representations as the optimal alignment. In this case, we can obtain the best alignment. Besides, DAW requires $(l \times m)$ operations and $(l \times m)$ storage cost. With the DAW loss, we can measure the similarity between two video sequences, which may vary in speed. By minimising the loss, we intend to make the temporal Transformer understand the aligned attention information.

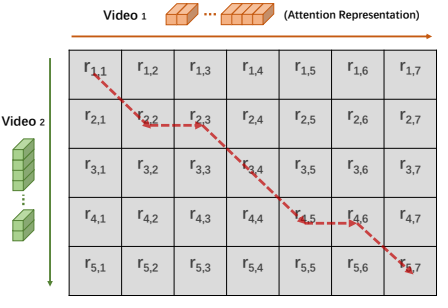

Figure 4: Dynamic Attention Warping formulation for video alignment. The $\rightarrow\searrow$ moves through the matrix showing the process of computing the distance.

4 Experiments

4.1 Experimental Setup

Datasets. The Penn Action [38] and Pouring [29] datasets provide collections of human action videos which can be used for dense alignment. The Penn Action dataset contains actions from different sports and exercises, and the Pouring dataset focuses on the interaction between hands and a drinking glass. We follow the annotation rule provided by Dwibedi *et al.* [12] for ground truth annotation. For the Penn Action dataset, we add key events and

Table 1: Action phase classification (%) results on both Penn Action and Pouring datasets.

Method	Penn Action	Pouring
ResNet [16]	44.96	43.85
SaL [21]	74.87	85.68
TCN [29]	81.99	89.19
TCC [12]	81.26	89.23
DGWT	**83.16**	**90.76**

Table 2: Action phase progression results on both Penn Action and Pouring datasets.

Method	Penn Action	Pouring
ResNet [16]	0.6267	0.6986
SaL [21]	0.5943	0.7451
TCN [29]	0.6762	0.8057
TCC [12]	0.6726	0.8030
DGWT	**0.6856**	**0.8183**

phases labels where a phase is the period between two key events and all frames in the period have the same phase label. As densely labelling each video frame is time-consuming and challenging work, we adopt the views of most annotators on key event frame annotation [10, 18]. Besides, we exclude the *strumming guitar* and *jumping rope* actions because the key event is difficult to define, following the same process as in [12].

Evaluation Metrics. Following the evaluation protocol of [12], DWGT is first trained on the training set and then frozen. An SVM classifier is trained on the learned features from DWGT to output the phase labels for each frame of the training data, with no additional fine-tuning. We use **phase classification accuracy** which is the accuracy of each frame, and **phase progression** measures how well the progress of a process or action is captured which is computed as the the average R-squared measure (coefficient of determination) [23]. Both metrics are implemented by the SVM classifier to evaluate the learned video representation. In addition, the training and validation splits of both Penn Action and Pouring datasets follow the setting in [12].

Implementation Details. Before training, all frames from each video sequence are resized to 224×224 pixels. we first extracted the spatial features from the pretrained ViT-Base [11] last attention layer. In the temporal transformer network, we stack six self-attention blocks with 768 attention dimensions per layer which is the same as the ViT network. For training, we apply ADAM as the optimiser of our model and set the learning rate as 1×10^{-4}. All hyperparameters are optimised via cross-validation.

4.2 Comparison with SOTA methods

Our main comparison to state-of-the-art methods (training-from-scratch results) is summarised in Table 1 and Table 2. We also add the results ResNet-50 features pre-trained on ImageNet dataset. Results of compared approaches are obtained from [12]. For both classification metric (Table 1) and progression metric (Table 2), it can be seen that our proposed DGWT model outperforms other approaches on both datasets consistently.

The performance of SaL is significantly better than using only spatial features (ResNet-50) in both the Pouring dataset and the Penn Action dataset, but worse than other self-supervised learning methods. We thus consider shuffling the order of the sequence can learn temporal information but might not be able to learn the alignment dependencies in a video sequence. Both of TCC and TCN are applied on the normal timeline and hence the results verify that shuffling the sequence order is not a good choice for the alignment task. In addition, TCN focuses on constrastive learning of different videos and TCC focuses on consistency learning among sequences. Finally, the improved performance of our model shows

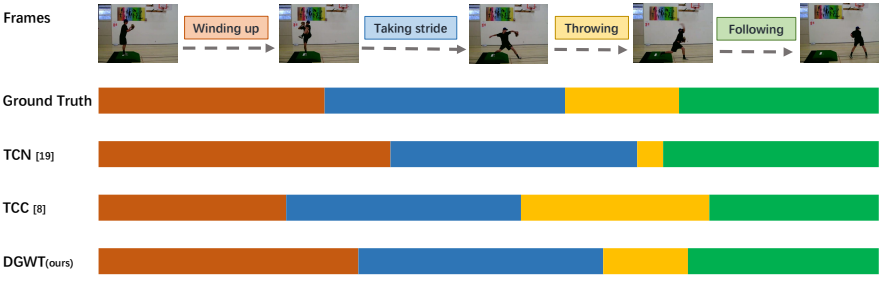

Figure 5: Qualitative results of different approaches for *Baseball Pitch* video 67 in the Penn Action dataset. The coloured area represents each action phase sequence.

that it is more effective to use segment-level attention representation learning which minimises the distance between multiple sequences, and our dynamic attention loss can better predict action progress.

Qualitative evaluation. Example qualitative results for an action video in the Penn Action dataset are shown in Figure 5. Overall, we can observe that the predicted action phase timeline by DGWT is more similar to the ground truth. This means that our proposed model can effectively learn frame dependencies within video sequences and the relationship between key events and action progress.

4.3 Ablation Study

The success of our DGWT can be attributed to both the framework design and technical improvement in each component. To analyse the effect of each component in DGWT, we construct ablation study models including DGWT$_{vit}$ that uses only pretrained vision transformer without any optimisation and DGWT$_{tet}$ that uses the spatio-temporal transformer representations and DAW loss but without TSG. In addition, models DGWT$_{sup}$ and DGWT$_{sup+daw}$ are trained in a supervised learning setting, with the former using cross entropy loss, and the latter using a combined cross entropy and DAW loss; and both models contain the complete network (spatio-temporal transformers with TSG). Results are summarised in Table 3, from which we can infer the following aspects.

Different vision backbones. In our pure transformer design (DGWT$_{vit}$), we apply vision transformer as our vision backbone. Compared to the ResNet backbone (shown in Table 1), the performance of vision transformer features is slightly better. We consider that the vision attention mechanism might be able to capture the main object information. However, without any temporal information, the result is relatively low. Thus, we think with spatial features only, the model cannot learn the alignment information.

Table 3: Action phase classification (%) results of ablation study.

Method	Penn Action	Pouring
DGWT$_{vit}$	46.22	48.34
DGWT$_{tet}$	81.56	89.68
DGWT	**83.16**	**90.76**
DGWT$_{sup}$	**84.42**	**91.32**
DGWT$_{sup+daw}$	**85.23**	**92.21**

Temporal transformer. When adding the attention mechanism to learn the temporal information ($DGWT_{tet}$), we observe that performance is significantly better than only applying spatial features (ResNet in Table 1 and $DGWT_{vit}$ in Table 3), which also proves that temporal information is important for video understanding. In addition, the performance is close to TCC and better than TCN and SaL. This means that by using pure transformers (spatial and temporal), the model can learn alignment information quite effectively.

Segment-level Graph. By adding the segment-level graph refinement (DGWT), we obtained better performance. This proves our hypothesis that TSG can benefit the temporal transformer for learning the transition of adjacent action phases.

DAW optimisation. The results also show that supervised learning with cross entropy loss does show higher performance than using only self-supervised learning, which is expected. On the other hand, by comparing $DGWT_{sup}$ and $DGWT_{sup+daw}$, we observe that when adding our proposed DAW loss, the model can obtain further improvement, demonstrating the benefit of attention warping even in a supervised setting. This demonstrates the advantage of our end-to-end framework that incorporates spatio-temporal transformers, graph attention based feature enhancement and explicit alignment optimisation with warping.

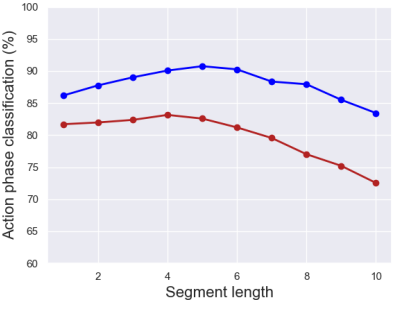

Figure 6: Action phase classification results (%) with segments of different lengths in DGWT on the Penn Action (red line) and Pouring (blue line) datasets.

Segment length in Learning. We observe that when the segment length is 4 for the Penn action dataset and 5 for the Pouring dataset, the action phase classification achieves the best results as seen in Figure 6. The results show that DGWT might not learn the alignment mechanism well when the number is too low or too high. When the segment is too long, our temporal segment graph might lose the key event information as the action progresses.

5 Conclusion

The key contribution of our work is to introduce warping as an essential task to the video alignment problem. We have presented the first approach to introduce spatio-temporal transformers for representation learning in video alignment tasks. Our Temporal Segment Graph converts temporal data into a graphical structure by dividing a video into small segments. Frames in each segment are non-reversible so can be encoded by aggregation using a graph. In this way, frames within a segment are smoothed and the distinction between segments is enhanced. With the proposed Dynamic Attention Warping loss, the model is able to optimise the alignment between videos of arbitrary lengths using self-supervised learning. Our experimental evaluation shows that on the Penn Action and Pouring datasets, our proposed DGWT model provides state-of-the-art performance.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015.

[4] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020.

[5] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019.

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015.

[7] Jianfei Chen, Jun Zhu, and Le Song. Stochastic training of graph convolutional networks with variance reduction. *arXiv preprint arXiv:1710.10568*, 2017.

[8] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: Fast learning with graph convolutional networks via importance sampling. In *International Conference on Learning Representations*, 2018.

[9] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 257–266, 2019.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019.

[13] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11068–11077, 2021.

[14] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[15] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[18] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[20] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[21] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[22] Meinard Müller. Dynamic time warping. In *Information retrieval for music and motion*, pages 69–84. Springer, Berlin, Heidelberg, 2007.

[23] Nico JD Nagelkerke et al. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.

[24] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

[25] Thomas Prätzlich, Jonathan Driedger, and Meinard Müller. Memory-restricted multiscale dynamic time warping. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 569–573. IEEE, 2016.

[26] Senthil Purushwalkam, Tian Ye, Saurabh Gupta, and Abhinav Gupta. Aligning videos in space and time. In *European Conference on Computer Vision*, pages 262–278. Springer, 2020.

[27] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7386–7395, 2018.

[28] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*, pages 52–1. BMVA Press, 2015.

[29] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141. IEEE, 2018.

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

[31] Young Chol Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea, and Henry A Kautz. Unsupervised alignment of actions in video with text descriptions. In *IJCAI*, pages 2025–2031, 2016.

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[33] Kiran K Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. Attention-based graph neural network for semi-supervised learning. *arXiv preprint arXiv:1803.03735*, 2018.

[34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[35] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The World Wide Web Conference*, pages 2022–2032, 2019.

[36] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[37] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, pages 1331–1338. IEEE, 2011.

[38] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013.

[39] Yaodong Zhang, Kiarash Adl, and James Glass. Fast spoken query detection using lower-bound dynamic time warping on graphical processing units. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5173–5176. IEEE, 2012.