Received: February 19, 2024. Revised: May 17, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Improvement of Tradition Dance Classification Process Using Video Vision Transformer based on Tubelet Embedding

Edy Mulyanto[1,2] **Eko Mulyanto Yuniarno**[1,3] **Oddy Virgantara Putra**[1,4] **Isa Hafidz**[1, 5]
Ardyono Priyadi[1] **Mauridhi H. Purnomo**[2,3]*

[1]*Electrical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
[2]*Computer Engineering Department, Dian Nuswantoro University, Semarang, 50131, Indonesia*
[3]*Computer Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
[4]*Graduate Department of Informatics, Universitas Darussalam Gontor, Ponorogo, 63471, Indonesia*
[5]*Electrical Engineering Department, Institut Teknologi Telkom Surabaya, Surabaya, 60231, Indonesia*
* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Image processing has extensively addressed object detection, classification, clustering, and segmentation challenges. At the same time, the use of computers associated with complex video datasets spurred various strategies to classify videos automatically, particularly in detecting traditional dances. This research proposes advancement in classifying traditional dances by implementing a Video Vision Transformer (ViViT) that relies on tubelet embedding. The authors utilized IDEEH-10, a dataset of videos showcasing traditional dances. In addition, the ViViT artificial neural network model was used for video classification. The video representation is generated by projecting spatio-temporal tokens onto the transformer layer. Next, an embedding strategy is used to improve the classification accuracy of Traditional Dance Videos. The proposed concept treats video as a sequence of tubules mapped into tubule embeddings. Tubelet management has added TA (tubelet attention layer), CA (cross attention layer), and tubelet duration and scale management. From the test results, the proposed approach can better classify traditional dance videos compared to the LSTM, GRU, and RNN methods, with or without balancing data. Experimental results with 5 flods showed Loss between 0.003 to 0.011 with an average Lost of 0.0058. Experiments also produced an accuracy rate between 98.68 to 100 percent, resulting in an average accuracy of 99.216. This result is the best of several comparison methods. ViViT with tubeless embedding has a good level of accuracy with low losses, so that it can be used for dance video classification processes.

Keywords: Video vision transformer, Tubelet embedding, Video classification.

1. Introduction

Research in the field of computer vision, especially in human action detection, has increased in recent years. These works include the fields of transportation, health, security, and human interaction systems. In general [1, 2] action detection is more challenging than action recognition, especially for dealing with online video streams.

Dance is one area of study that can be studied further with the help of computer vision. Dance can be said to be a type of sporting event carried out by one or more people. Movement skills and techniques can influence the beauty of dance [3]. Indonesia is a large country with various tribes, customs, and cultures. One of their cultures is traditional dance, and almost every region and tribe has its own culture [4]. Tradition means the hereditary customs that occur from generation to generation and contain binding values or norms for the community. Traditional dances grow and develop in an area that becomes the people's cultural identity. However, research on dance, including its relationship to integrating the latest technology, has not yet been studied in more depth [5, 6]. Several variations of dance movements are known to be complicated and require technique. The application of traditional dance video classification through computer vision can help examine more deeply complex movements in the field of dance and reduce cultural diversity so that it can be preserved for educational purposes.

There are several methods of classifying a video dataset. Classify one frame at a time This method ignores the temporal video features and classifies each clip by looking at each frame, using convolutional neural network (CNN), also known as ConvNet. More specifically, using Inception V3, which has been previously trained on ImageNet and transfer learning to retrain the Inception on existing data, this takes two steps: perfecting the top solid layer with several epochs to maintain the possibility of previous research [7, 8]. Use CNN time-distributed and feature passing to recurrent neural network (RNN). This model considers the video's temporal features for the initial network using a Time Distributed wrapper, which allows distributing the CNN layer to an additional dimension, known as time. The ConvNet model uses a tiny network of VGG16 type, while for the RNN portion, long short-term memory (LSTM) or gated recurrent unit (GRU) is usually used [9, 10]. Extract features using CNN and passes the sequences on RNN. This model first runs each video frame via Inception, storing the output from the network end set layer. Then converts it into an extracted feature set for training on the RNN model, which uses the LSTM layer [11-13]. In short, classifying video datasets involves various techniques. One of them is frame-by-frame classification using CNNs such as Inception V3. Another approach combines CNNs and RNNs with time distributed processing to capture temporal features.

Transformer has experienced developments in the field of language processing (NLP), especially sequence-to-sequence models [14-16]. Initially, the transformer approach was used for machine translation as an alternative to natural language models based on RNN and CNN. Then, the Bidirectional Encoder Representations from Transformers (BERT) mechanism emerged with multiple NLP capabilities by pretraining transformers on unlabeled text. In [17] using video data and transformer method, applying Deep Video hashing on two separate modules 3DCNN and bidirectional encoder representation of transformer layer (BERT). Then appeared Generative Pre-trained Transformer 3 (GPT-3), which describes a transformer-based model with many parameters with different NLP and without fine-tuning. Besides being used in the NLP field, Transformers are also implemented in the computer vision field. The Vision Transformer (ViT) mechanism is used for image classification by using image patches as input to a transformer encoder [18]. Next is the Video Vision Transformer (ViViT) from ViT, which explores the application of using ViT in video classification [19]. ViViT can be used as an architecture for long-range spatiotemporal modeling to solve video sequence problems in 3D video signal extraction. Then, the self-attention mechanism can be used to combine features from various modalities automatically [20-22]. The next challenge is the use of transformer-based ViViT for Tubelet Embedding for dance video classification.

The primary contributions of this research are:

1. introduction of a unique classification technique by leveraging ViViT and tubelet embedding. This marks a departure from conventional methods such as LSTM, GRU, and RNN, showcasing innovation in applying advanced neural network models for improved accuracy in traditional dance video classification.

2. Utilization of the IDEEH-10 dataset, a collection of videos featuring traditional dances. This dataset becomes a crucial element in the research, providing a real-world and culturally relevant context for evaluating the proposed ViViT with the Tubelet embedding approach.

3. Incorporating the ViViT artificial neural network model for video classification, emphasizing the importance of using state-of-the-art models in handling complex video datasets. ViViT's capacity to project spatiotemporal tokens onto transformer layers is highlighted as a key feature in generating effective video representations.

4. Introduction of an embedding strategy that significantly improves the classification accuracy of traditional dance videos. Creating videos as sequences of tubelets and mapping them into tubule embeddings demonstrates a nuanced conventional dance understanding of sequences' temporal and spatial aspects.

5. Comparative evaluation asserting that the proposed ViViT with the Tubelet embedding outperforms traditional methods like LSTM, GRU, and RNN, irrespective of data balancing. This empirical evidence adds weight to the contribution by demonstrating the superior performance of the proposed approach.

6. The authors also propose the addition of TA (tubelet attention layer), CA (cross attention layer), and management of the scale and duration of tubelets. The management of tubelet scale and duration is important in video analysis because videos often have wide variations in object size, level of detail, and duration of object movement. In some cases, different tubelets in a video may have different scales from each other. For example, in a video featuring a scene

with distant objects and closely spaced objects, the scale ratio between the tubelets may vary. The development of techniques to adjust and normalize the scales between different tubelets can help obtain a consistent and reliable representation of the movement of objects in a video. We augment this scale and duration management with Relative Normalization between Tubelets. This addition will result in maintaining scale comparisons between different tubelets, relative normalization can be done by maintaining relative proportions between object sizes or dimensions within each tubelet.

VIVIT with Tubelet Embedding is a technique in video processing that combines the Vision Transform-er (ViT) model with the concept of tubelet embedding for a better comprehension of object motion in videos. Here are some key features of this method: (1) Vision Transformer (ViT): A neural network architecture initially designed for image processing, now adapted to handle video data. It distinguishes itself from traditional Convolutional Neural Networks (CNNs) by utilizing a self-attention-based method. ViT partitions the picture or video frame into blocks and then utilizes self-attention transformation on these blocks to enhance feature representations., (2) Tubelet embedding is the representation of a sequence of contiguous frames in a video as dense vector embeddings. The embeddings contain data regarding spatial and temporal variations inside the tubelet. Embedding tubelets enhances the representation of object motion in videos, allowing the model to grasp the temporal context of object movement more thoroughly., (3) The integration of ViT with tubelet embedding enables the model to gain a more comprehensive comprehension of object motion within videos. Tubelet embedding enhances the model's capacity to acquire more complex object motion representations, whereas ViT allows for the adaptive and flexible processing of visual information., (4) ViT architecture is highly scalable to input size, allowing it to be applied to films of various resolutions without requiring major adjustments. Tubelet embedding creates more condensed representations of tubelets, reducing computational burden and facilitating efficient video processing on a broad scale., and (5) The VIVIT with Tubelet Embedding model accounts for temporal dependencies in object motion by utilizing information from tubelet embeddings. This enables the model to identify intricate and changing motion patterns in videos. VIVIT with Tubelet Embedding provides a strong and efficient method for video processing, especially for assessing object motion and comprehending temporal contexts in videos.

VIVIT with Tubelet Embedding offers various ad-vantages in video processing compared to older approaches. Here are some of the primary benefits: (1) Comprehensive Representation of Object Motion: By employing tubelet embedding, which captures both spatial and temporal information inside a sequence of frames, VIVIT may provide a more comprehensive representation of object motion in films. This allows for a richer understanding of the dynamics and context of object movement throughout time., (2) Tubelet embedding allows VIVIT to properly capture temporal dependencies in object motion. The model can analyze the sequence of frames in a tubelet to comprehend the evolution of object movements over time, resulting in more precise and contextually detailed representations., (3) The Vision Transformer (ViT) design in VIVIT provides scalability and efficiency benefits, especially for processing large-scale video data. ViT's self-attention mechanism enables it to handle films of different resolutions effectively, without requiring significant adjustments, thereby making it well-suited for real-world applications with a wide range of video inputs., (4) VIVIT benefits from the flexibility of the ViT architecture in learning visual representations. ViT utilizes self-attention mechanisms to selectively focus on pertinent spatial and temporal characteristics within the input data, enhancing the acquisition of object motion patterns and visual context., (5) The utilization of tubelet embedding enhances the interpretability and explainability of the model's predictions. VIVIT enhances transparency in decision-making processes by embedding tubelets into dense vector representations to provide insights into how object motion is encoded and utilized by the model., and (6) VIVIT with Tubelet Embedding has shown cutting-edge performance in many video comprehension tasks, including action recognition, object detection, and video captioning. Its capacity to capture detailed temporal dynamics and context has resulted in better performance in comparison to conventional approaches. VIVIT with Tubelet Embedding provides a robust and flexible method for video processing, allowing for precise, contextually detailed, and understandable representations of object motion in videos.

In summary, this research advances the state-of-the-art in traditional dance video classification by introducing a cutting-edge ViViT with the Tubelet embedding approach, backed by a comprehensive evaluation using a relevant dataset. The innovative classification technique and improved accuracy showcase the potential for practical application in dance video classification processes.

Received: February 19, 2024. Revised: May 17, 2024.

The article is organized into several sections: Section 2 covers materials and methods, including dataset, preprocessing. Section 3 explain classification process, including ResNet101 architecture integrated with LSTM, GRU, and RNN layers, configured with and without weight class balancing. ViViT using Tubelet Embedding also explained with result analysis. Section 4 explains results and discussions, and section 5 is conclusion.

2. Related work

Advancements in image recognition have been reflected in the architectures designed for video understanding. In the early stages of video research, appearance and motion information were encoded using manually created features [23]. The initial success of AlexNet on ImageNet [24, 25] prompted the adaptation of 2D image convolutional networks (CNNs) for video applications, resulting in the development of two-stream networks [26-28]. These models independently analyzed RGB frames and optical flow images before merging them at the final stage. The availability of bigger video classification datasets, such as Kinetics [29], has made it easier to train spatio-temporal 3D CNNs [30, 31]. These models have a much higher number of parameters, which means they need larger training datasets. Due to the increased computational requirements of 3D convolutional networks compared to their image counterparts, several topologies employ convolution factorization techniques. The utilization of grouped convolutions [32-34] allows for the incorporation of both spatial and temporal aspects. In addition, we exploit the factorisation of the spatial and temporal dimensions of films to enhance efficiency, specifically inside transformer-based models.

Simultaneously, in the field of natural language processing (NLP), Vaswani et al. [35] attained the most advanced outcomes by substituting convolutions and recurrent networks with the transformer network, which just comprised self-attention, layer normalization, and multilayer perceptron (MLP) operations. The current cutting-edge designs in Natural Language Processing (NLP) [36, 37] continue to be built on transformers and have been expanded to handle large-scale datasets from the web [38]. Several modifications of the transformer model have been suggested to decrease the computational burden of self-attention when dealing with longer sequences [39-43], as well as to enhance parameter efficiency [44]. While self-attention has been widely used in computer vision, it is usually included as a layer towards the end or in the later phases of the network [45, 46]. Alternatively, it can be used to enhance residual blocks inside a ResNet design.

Prior studies have made efforts to substitute convolutions in vision architectures [47, 48]. However, it was only recently demonstrated by Dosovitisky et al. [49] that their ViT architecture, which utilizes pure-transformer networks like those used in NLP, can achieve cutting-edge performance in image classification as well. The authors demonstrated that these models are most effective when applied on a large scale. This is because transformers lack some inherent biases found in convolutional networks, such as translational equivariance. Consequently, training these models necessitates datasets larger than the commonly used ImageNet ILSRVC dataset. ViT has sparked significant further research in the community, and it is worth mentioning that there are several ongoing efforts to expand its application to other computer vision tasks [50, 51] and enhance its efficiency in handling data [52, 53]. Specifically, [54, 55] have also suggested transformer-based models for video.

Building upon the advancements in object detection using deep convolution neural networks, frame-level approaches have significantly enhanced action detection in videos [56, 57]. Some researchers utilize 3D convolution networks, such as [58, 59], to effectively collect temporal information in order to recognize activities. Feichtenhofer et al. [60] propose a slowfast network to more effectively gather spatio-temporal information. Both Tang et al. [61] and Pan et al. [62] suggest explicitly incorporating the modeling of relationships between actors and objects. In a recent study, Chen et al. [63] suggest training actor location and action classification simultaneously using a single backbone. Vaswani et al. [35] introduced the transformer model for machine translation, which quickly gained popularity as the primary framework for sequence-to-sequence tasks, such as [64, 65]. Recently, it has also made significant progress in object identification [66, 67], picture classification [49, 68], and video recognition [69-71]. Girdhar et al. [72] introduce a video action transformer network for the purpose of action detection. They utilize a region-proposal network to perform localization. The transformer is employed to enhance action recognition by consolidating features from the spatio-temporal context around actors. We present a comprehensive approach to concurrently determine the location and identify actions.

This paper compares several algorithms for dance video classification. The algorithms employed are VIVIT with Tubelet Embedding, ResNet101, MobileNetV2, LSTM, and GRU. ResNet101 demands significant processing resources because to its deep architecture with several layers, making it less efficient for real-time applications or devices with limited computational capacity. ResNet101 lacks explicit techniques to characterize temporal dependencies in video data, which may result in limits in capturing temporal context and dynamics. MobileNetV2 may not deliver as deep feature representations as ResNet101 due to its focus on speed and efficiency, perhaps resulting to inferior performance in tasks requiring thorough feature extraction. While MobileNetV2 is built for efficiency, it may be less efficient in circumstances where richer feature representations are required, as it may struggle to compress complicated information adequately. LSTM and GRU designs are computationally more complex compared to CNNs and may need greater computational resources, making them less suited for deployment on resource-constrained devices. Despite their capacity to manage long-term dependencies, LSTM and GRU architectures may still struggle with capturing very long-range temporal relationships efficiently, which might influence performance in certain sequence modeling applications. VIVIT with Tubelet Embedding gives a thorough depiction of object motion in videos by using tubelet embeddings. This provides for a detailed understanding of temporal dynamics in object movement. With tubelet embeddings collecting temporal information, VIVIT enables efficient processing of movies by efficiently adding temporal context into the analysis. The Vision Transformer (ViT) architecture employed in VIVIT allows scalability to changing input sizes, making it suited for processing videos with different resolutions. In summary, VIVIT with Tubelet Embedding excels in providing a comprehensive understanding of object motion in videos and efficient video processing, while ResNet101, MobileNetV2, LSTM, and GRU have their weaknesses in terms of computational cost, limited temporal context modeling, and feature representation capabilities.

3. Materials and methods

3.1 Dataset

Implementing the IDEEH-10 dataset marks a pioneering effort in algorithmic education, explicitly targeting the identification of traditional dances endowed with cultural importance. Recognizing the absence of appropriate datasets tailored for instructing algorithms in discerning culturally significant dances, a curated assortment of videos has been introduced.

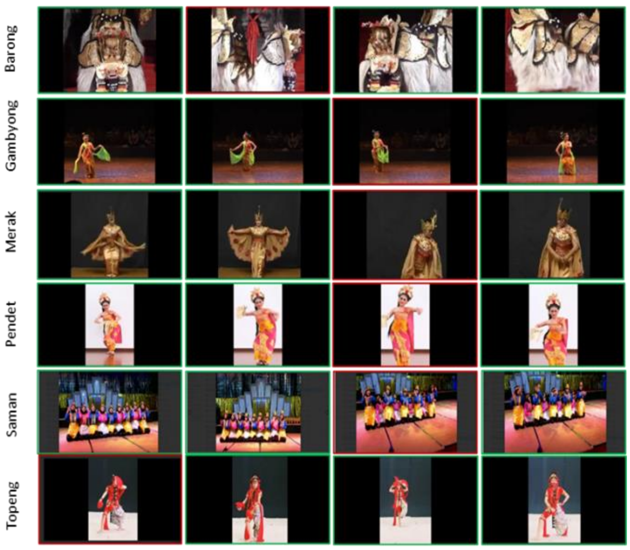

Figure. 1 Overview of traditional dance dataset. Anomalous frames are marked with red borders while frames with green borders are normal

Table 1. Dataset of Traditional Dance

No.	Type of Dance	Clip	Duration per clip (s)	Frame
1	Gambyong	194	10	93.120
2	Saman	189	10	90.720
3	Pendet	187	10	89.760
4	Topeng	191	10	91.680
5	Barong	192	10	92.160
6	Merak	186	10	89.280

Table 2. Label Each Frame

Gambyong	Class
frame1_1	1
frame1_2	1
……….	..
……….	..
frame1_93120	1

Saman	Class
frame2__1	2
frame2__2	2
……….	..
……….	..
frame2_90720	2

Pendet	Class
frame3_1	3
frame3_2	3
……….	..
……….	..
frame3_89760	3

Topeng	Class
frame4__1	4
frame4__2	4
……….	..
……….	..
frame4_91680	4

Barong	Class
frame5_1	5
frame5_2	5
……….	..
……….	..
frame5_92160	5

Merak	Class
frame6__1	6
frame6__2	6
……….	..
……….	..
frame6_89280	6

IDEEH-10, an acronym denoting "Indonesian Dances by Edy Eko Hery," encompasses a rich compilation of ten distinct traditional dances, each deeply rooted in the country's cultural tapestry. These dances serve as poignant expressions of Indonesia's cultural heritage, contributing to preserving and appreciating its artistic traditions. The dataset is meticulously structured, providing a comprehensive resource for algorithmic training, with detailed information on various dance forms and their corresponding frame counts elucidated in Fig. 1 and Table 1, respectively. This initiative not only fosters the advancement of machine learning in cultural recognition but also contributes to the broader goal of safeguarding and promoting diverse cultural legacies through technological innovation.

The acquisition of the dataset involved the meticulous recording of traditional cultural dances, capturing the nuanced movements and expressions inherent in each performance. An essential preprocessing phase was implemented to construct a robust dataset. The initial step in this process entailed segmenting each recorded video into distinct sub-clips, aligning with the various movements executed by different dancers. This granular dissection served a dual purpose: to articulate the dancers with specific labels and to isolate individual dance moves. Consequently, each video was systematically sectioned into fragments, with each segment corresponding to the particular number of traditional dances captured. This methodological approach facilitated the categorization of dancers and enabled the nuanced analysis of discrete dance elements. Through this meticulous preprocessing, the resulting dataset becomes a comprehensive repository, laying the groundwork for training algorithms to discern and appreciate the intricate details of traditional cultural dances with heightened accuracy and cultural sensitivity.

Following the initial segmentation of videos into sub-clips, the subsequent stage involves a refined cutting process to enhance the focus on the dancers within each sub-clip. This meticulous adjustment is executed on an individual basis, further honing the precision of the dataset. The subsequent prediction outcomes are bolstered in accuracy by narrowing the visual scope to showcase the dancers primarily. This targeted approach facilitates a more effective identification of the dancers and their intricate movements, thus refining the algorithm's capacity to recognize and interpret these cultural expressions.

The refinement continues; the process advances to a frame-based classification methodology. In this phase, a frame is extracted from each sub-clip, which has already undergone the meticulous trimming process. This frame is a crucial snapshot, encapsulating a pivotal moment within the dance sequence. Adopting a frame-centric approach, the algorithm can access a snapshot of the dancers' poses and expressions at a specific instance. This granular level of analysis enhances the accuracy of predictions and allows for a nuanced understanding of the diverse movements encapsulated in each sub-clip.

Therefore, the frame-based classification strategy underscores the significance of isolating key frames within the trimmed sub-clips, serving as the foundational step in the algorithm's journey toward discerning and classifying traditional dance forms with a heightened level of precision and cultural understanding.

The quantity of frames extracted from the meticulously curated sub-clips, tailored to accentuate the focal point of interest, is inherently flexible, accommodating the diverse requirements of the dataset. This adaptability ensures that the number of frames is aligned with the specific needs and intricacies of the traditional dances under consideration. Subsequently, the subsequent phase in this intricate process involves the assignment of labels to each frame within the sub-clips. This labeling initiative is fundamental in imparting a structured and categorical identity to the extracted frames, facilitating the seamless integration of these visual snippets into a coherent dataset.

For each frame extracted from the sub-clips, a singular class label is attributed, establishing a cohesive link between the visual representation and the designated dance form. Table 2 is a visual reference, elucidating the correspondence between class labels (1 through 6) and specific traditional dances. To elaborate further, label 1 corresponds to the Gambyong dance class, label 2 denotes the Pendet dance, label 3 signifies the Saman dance, label 4 represents the Topeng dance, label 5 encapsulates the Barong dance, and label 6 is indicative of the Merak dance. This meticulous classification strategy streamlines the subsequent stages of analysis, enabling the algorithm to recognize individual frames and categorize them according to the distinct

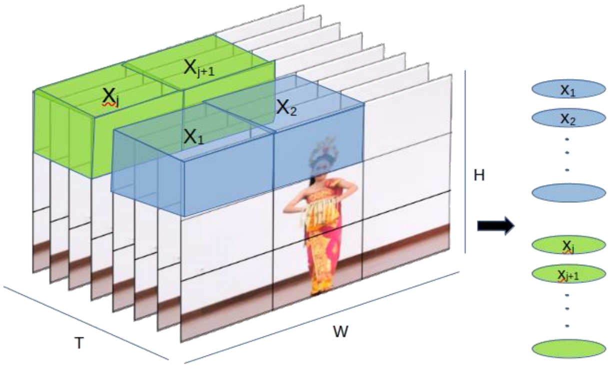

Figure. 2 Tubelet embedding

traditional dance forms they represent. The dataset attains a structured framework through this systematic labeling, laying the foundation for accurate and culturally informed predictions in standard dance classification.

Labelling process is carried out for all frames in the training process. The next process is to divide the data into two parts based on composition.

The dataset was prepared for the combination of Resnet101, LSTM, GRU and MobileNet methods. The dataset has also been prepared for embedding video clips, in order to map on $V \in \mathbb{R}^{\mathrm{T\times H\times W\times C}}$ to a sequence of tokens $\tilde{z} \in \mathbb{R}^{\mathrm{n}_t\times \mathrm{n}_h\times \mathrm{n}_w\times \mathrm{d}}$. It then adds position embedding and reshaping into $\mathbb{R}^{\mathrm{N\times d}}$ to obtain z, the input to the transformer.

Embedding of tubelets A different approach, illustrated in Fig. 2, involves extracting non-overlapping, spatio-temporal "tubes" from the input volume and projecting them linearly onto Rd. This approach is a continuation of Vit embedding technique into three dimensions, which aligns with a three-dimensional convolution. For a tubelet with dimensions t×h×w, $n_t = \left\lfloor \frac{T}{t} \right\rfloor, n_h = \left\lfloor \frac{H}{h} \right\rfloor$ and $n_w = \left\lfloor \frac{W}{w} \right\rfloor$, tokens are extracted from the temporal, height, and width dimensions, respectively. Reducing the diameters of the tubelets leads to a higher number of tokens, hence increasing the computational workload.

This method incorporates spatio-temporal information during tokenization, as opposed to the "Uniform frame sampling" approach where the transformer merges temporal information from distinct frames.

3.2 Methodology

In Fig. 3, we comprehensively depict the methodology employed to elevate the traditional dance classification process, utilizing a video visual transformer based on tubelet embedding. The initiation of this process unfolds with a meticulous video data preprocessing stage, where preparatory measures are taken for the video data corresponding to the Gambyong dance class, Pendet dance, Saman dance, Topeng dance, Barong dance, and Merak Dance. In this preparatory phase, preemptive actions are taken to address missing values and engage in feature selection for data segments that necessitate attention due to processing complexities.

The subsequent stage involves systematically inputting video data for each of the six dances: Gambyong, Pendet, Saman, Topeng, Barong, and Merak Dance. During this video input phase, frames are meticulously extracted from each batch, capturing pivotal moments and nuances within the dance sequences. This extraction process serves as a foundational step in generating a visual representation that encapsulates the essence of each traditional dance form. Through this comprehensive approach, the video vision transformer based on tubelet embedding is equipped with a rich and diverse dataset, laying the groundwork for accurate and nuanced classification of traditional dances. Fig. 3 acts as a visual guide, unraveling this transformative process's intricacies that enhance the classification accuracy and cultural understanding of the algorithmic model.

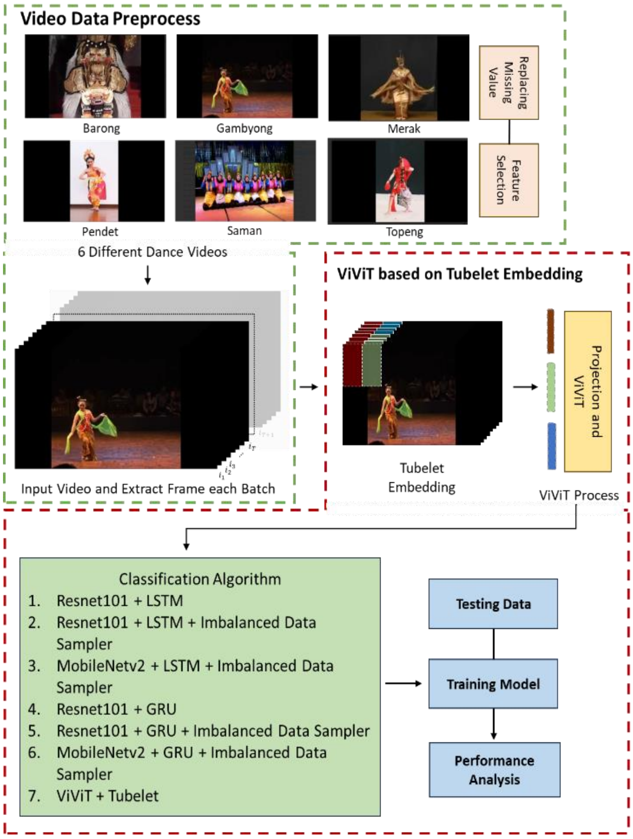

Figure. 3 Flowchart for Dance Classification

The video transformation process starts with applying transformative techniques at the Tubelet Embedding stage within the ViViT framework. ViViT, for Video Vision Transformer, is a neural network model designed explicitly for video classification. Its operational mechanism involves projecting each video segment into a feature space, a pivotal step that encapsulates the spatial and temporal dimensions of the input data. This transformative projection is executed spatio-temporally, and the resulting representation is denoted as a Tubelet.

The development of strategies to alter and normalize the scale between distinct tubelets can help produce a consistent and trustworthy portrayal of object movement in video. We supplement this scale and duration management using Relative

Normalization between Tubelets. This addition will result in maintaining scale comparisons across distinct tubelets, relative normalizing can be done by retaining relative proportions between the sizes or dimensions of objects within each tubelet.

Algorithm of Relative Normalization between Tubelets:

```
function relative_normalization(tubelets):
    normalized_tubelets = []
    // Iterate through each tubelet
    for each tubelet in tubelets:
        total_area = 0
        total_frames = length(tubelet)
        // Calculate total area of objects
        for each frame in tubelet:
            // Calculate area of object in the frame
            area = calculate_object_area(frame)
            total_area = total_area + area
        // Calculate average area of objects
        average_area = total_area / total_frames
        // Normalize each frame relative to the average
        // area
        normalized_tubelet = []
        for each frame in tubelet:
            // Calculate scaling factor based on ratio between
            // object area in frame and average area
            scaling_factor = calculate_object_area(frame) /
            average_area
            // Normalize size of object in frame using scaling
            // factor
            normalized_frame =
            normalize_object_size(frame, scaling_factor)
            // Append normalized frame to normalized tubelet
            append normalized_frame to normalized_tubelet
        // Append normalized tubelet to list of normalized
        // tubelets
        append normalized_tubelet to normalized_tubelets
    return normalized_tubelets
```

3.2.1 Tubelet encoder

Diverging from the conventional transformer encoder, the Tubelet encoder is specifically crafted to handle information in the 3D spatio-temporal domain. Each encoder layer comprises a self-attention layer (SA), two normalization layers, and a feed-forward network (FFN), as outlined in reference. Only the essential attention layers are included in the equations below.

Given a video clip $V \in \mathbb{R}^{T\times H\times W\times C}$ where T, H, W, C denote the number of frames, height, width, and colour channels, Tube R first applies a 3D backbone to extract video feature $F_b \in \mathbb{R}^{T'\times H'\times W'\times C'}$, where T′ is the temporal dimension and C′ is the feature dimension.

$$F_{en} = Encoder(F_b) \tag{1}$$

Where F_b is the backbone feature and $F_{en} \in \mathbb{R}^{T'\times H'\times W'\times C'}$ denotes the C′ dimensional encoded feature embedding. F_b is the input of the encoder function and the result will be stored in F_{en}.

$$SA(F_b) = sfx \times \sigma_v(F_b) \tag{2}$$

$$sfx = softmax\left(\frac{\sigma_q(F_b)\times\sigma_k(F_b)^T}{\sqrt{C}}\right) \tag{3}$$

Where SA is self-attention layer. Each encoder layer comprises a self-attention layer. Sfx is softmax function. The σ(*) is the linear transformation plus positional embedding. Emb_{pos} is the 3D positional embedding.

$$\sigma(x) = Linear(x) + Emb_{pos} \tag{4}$$

3.2.2 Tubelet decoder

Drawing inspiration from [73], we employ tubelet queries denoted as Q={Q1, ..., QN} that are informed by the video data. Rather than manually designing 3D anchors, we learn a tubelet query to capture the inherent dynamics in a tubelet. The initialization of box embeddings is uniform across all tubelet queries. To encapsulate relations within these tubelet queries, we introduce a tubelet-attention (TA) module consisting of two self-attention layers.

$$F_q = TA(Q) \tag{5}$$

We provide a tubelet-attention (TA) module with two self-attention layers to describe relations in tubelet queries. We possess a spatial self-attention layer that manages the spatial connections among box query embeddings within a frame. The tube decoder utilizes the tubelet attention module to process tubelet inquiries Q in order to produce the tubelet query feature F_q.

The decoder component is a pivotal element within the system architecture. It comprises two crucial components: the tubelet-attention module and a cross-attention (CA) layer. These components play a vital role in decoding, especially in extracting and deciphering tubelet-specific features from the encoded information. The tubelet-attention module is responsible for capturing intricate relationships and patterns within the tubelet queries, employing self-attention mechanisms. On the other hand, the cross-attention layer facilitates extracting relevant information by considering interdependencies between the tubelet queries and other pertinent aspects of the input data. Together, these components contribute to the nuanced decoding of tubelet-specific features, providing a comprehensive and accurate representation of the dynamic information encapsulated in the original video data.

$$CA\left(F_q, F_{en}\right) = sfx_{CA} \times \sigma_v(F_{en}) \tag{6}$$

The tubelet-attention module and a cross-attention (CA) layer. These components play a significant role in decoding, notably in extracting and understanding tubelet-specific aspects from the encoded information. The tubelet-attention module is responsible for capturing detailed correlations and patterns inside the tubelet queries, leveraging self-attention techniques.

$$sfx_{CA} = softmax\left(\frac{F_q \times \sigma_k(F_{en})^T}{\sqrt{C}}\right) \tag{7}$$

$$F_{tub} = Decoder\left(F_q, F_{en}\right) \tag{8}$$

The Tubelet Embedding stage employs a series of transformer layers, strategically integrating the Tubelet projections into the feature space. This nuanced and multi-layered approach ensures that the video is accurately represented and encapsulates the dynamic and temporal intricacies inherent in the dance sequences. The Tubelet embedding process, characterized by these transformer layers, becomes instrumental in producing sophisticated video representations, serving as a robust foundation for subsequent stages in the ViViT model.

By combining spatial and temporal dimensions within the Tubelet projections, ViViT achieves a comprehensive understanding of the video content. This amalgamation of spatial and temporal features enhances the model's capacity for discerning intricate patterns and dynamic movements within the traditional dance sequences, thereby contributing to the overall effectiveness of the video classification process.

After the Tubelet Embedding stage in ViViT, the process advances to video classification, where several algorithms are systematically employed to facilitate a comprehensive comparison of the proposed method's application. This rigorous evaluation involves implementing seven distinct analysis methods, each meticulously chosen to provide a well-rounded assessment of the proposed approach's efficacy. The algorithms enlisted for comparison include Resnet101 and LSTM, Resnet101, LSTM, and Imbalanced Data Sampler, MobileNetv2, LSTM, and Imbalanced Data Sampler, Resnet101 and GRU, Resnet101, GRU, and Imbalanced Data Sampler, MobileNetv2, GRU, and Imbalanced Data Sampler, and finally, the proposed method ViViT with Tubelet Embedding.

The Resnet101 and LSTM combination leverages the powerful image recognition capabilities of Resnet101 and the sequential learning prowess of Long Short-Term Memory networks (LSTM). Introducing the Imbalanced Data Sampler into this mix addresses any potential class imbalance issues, enhancing the model's ability to handle diverse datasets effectively. The MobileNetv2 and LSTM tandem and the Imbalanced Data Sampler further explore the intersection of lightweight mobile architectures and sequential learning.

In parallel, the Resnet101 and GRU combination, accompanied by the Imbalanced Data Sampler, explores the fusion of Resnet101's deep feature extraction and the sequential processing capabilities of Gated Recurrent Units (GRU). Similarly, the MobileNetv2, GRU pairing, and Imbalanced Data Sampler delve into the potential synergy between mobile-friendly architectures and recurrent neural networks.

Finally, the proposed method, ViViT with Tubelet Embedding, is a unique and innovative approach. ViViT, being a Video Vision Transformer, and Tubelet Embedding, contributing spatial-temporal understanding through transformer layers, collaborate to offer a novel solution to video classification challenges. This comparative analysis aims to discern the strengths and weaknesses of each method, providing valuable insights into the applicability and performance of the proposed ViViT and Tubelet Embedding methodology within the context of traditional dance classification.

4. Results and discussion

In this section, we delve into the results and discussions derived from exploring traditional cultural dance classification within the context of video frames. Our study encompasses a comparative analysis with various alternative methods, revealing a notable enhancement in classification accuracy by utilizing a visual video transformer based on the Tubelet embedding.

The primary focus of our investigation lies in scrutinizing the effectiveness of different methodologies applied to the intricate task of classifying traditional cultural dances within individual video frames. Drawing comparisons with several other approaches, it becomes evident that incorporating a visual video transformer based on the Tubelet embedding leads to a substantial increase in classification accuracy.

This improvement can be attributed to the unique characteristics of the Tubelet embedding approach, which harnesses the power of transformer layers to capture spatial and temporal intricacies within the dance sequences. The spatial-temporal understanding enabled by the Tubelet embedding proves pivotal in discerning subtle nuances and dynamic movements inherent in traditional cultural dances. As a result, the classification accuracy achieves a significant boost, surpassing the performance of alternative methods.

This finding underscores the efficacy of leveraging advanced video transformation techniques, precisely the Tubelet embedding approach within a visual video transformer. The increased accuracy validates the relevance of this methodology for traditional cultural dance classification. It opens avenues for further exploration and refinement in video frame-based classification methodologies. The ensuing discussions delve into the nuanced aspects of these findings, shedding light on the implications and potential applications of this enhanced classification accuracy within the broader context of cultural preservation and technological advancements.

Table 3 presents the classification results using Resnet101 and LSTM, considering variations in folds, architectural aspects, and data balance. In this analysis, the batch size (bs) value for the fold in the comparison method is consistently 64. Notably, the highest accuracy of 100 was achieved with Resnet101 and LSTM, considering fold variations (2, 3, and 4), architectural nuances, and the incorporation of an imbalanced data sampler. The loss rate for these folds remained impressively low, all below 0.042.

Following closely in performance was the combination of Resnet101 with LSTM, which yielded a minimum accuracy of 80.733. The least accurate performance in this comparison was observed in the Resnet101 and LSTM combination, with a maximum accuracy value of 65.137 on fold 1 and a minimum accuracy of 11.009 on fold 3,

Table 3. Classification using Resnet101 and LSTM by considering architectural variations and data balance

bs = 64	Resnet101 + LSTM		Resnet101 + LSTM + Imbalanced Data Sampler	
Fold	Loss	Acc	Loss	Acc
1	1.101	65.137	0.559	82.568
2	1.696	19.266	0.518	87.156
3	1.806	11.009	0.575	80.733
4	1.652	31.192	0.458	84.403
5	1.644	35.185	0.576	74.074

Table 4. Classification using Resnet101 and GRU by considering architectural variations and data balance

bs = 64	Resnet101 + GRU		Resnet101 + GRU + Imbalanced Data Sampler	
Fold	Loss	Acc	Loss	Acc
1	0.243	92.660	0.069	100
2	1.418	49.541	0.270	93.578
3	1.645	37.614	0.119	98.165
4	1.645	31.192	0.067	98.165
5	1.644	35.185	0.125	100

Table 5. Classification using Resnet101 and RNN considering architectural variations and data balance

bs = 64	Resnet101 + RNN		Resnet101 + RNN + Imbalanced Data Sampler	
Fold	Loss	Acc	Loss	Acc
1	0.129	98.165	0.039	100
2	1.445	47.706	0.201	96.330
3	1.615	37.614	0.063	99.082
4	1.638	31.192	0.114	98.165
5	1.644	35.185	0.270	95.370

Table 5. Classification using Resnet101 and RNN considering architectural variations and data balance

bs = 64	Resnet101 + RNN		Resnet101 + RNN + Imbalanced Data Sampler	
Fol d	Loss	Acc	Loss	Acc
1	0.129	98.165	0.039	100
2	1.445	47.706	0.201	96.330
3	1.615	37.614	0.063	99.082
4	1.638	31.192	0.114	98.165
5	1.644	35.185	0.270	95.370

Table 7. Classification results using ViViT and Tubelet

ViViT + Tubelet	Framesize=128, seq=25, lr=1e-4, PATCH_SIZE = (8,8,8)				
Fold	Loss	Acc	F1	Pre	Rec
1	0.003	98.7	0.809	0.798	0.831
2	0.004	100	0.826	0.818	0.844
3	0.011	100	0.837	0.831	0.857
4	0.003	98.7	0.815	0.808	0.831
5	0.008	98.68	0.815	0.808	0.831

corresponding to a loss of 1,101 and 1,806, respectively.

Moving on to Table 4, which details the classification outcomes with a batch size of 64 for the folds, the highest accuracy of 100 was achieved with MobileNetv2, GRU, and an imbalanced data sampler. Similar to the previous analysis, this result was observed across fold variations 2, 3, and 4, with a loss rate below 0.044. Resnet101, with the same combination, followed with high accuracy, particularly on folds 1 and 5.

In contrast, the least accurate combination was identified as Resnet101 and GRU, with an accuracy value of 31,192 on fold 4. These findings provide a detailed insight into the classification performance of different combinations, shedding light on the impact of fold variations, architectural choices, and the incorporation of data balancing techniques on the overall accuracy of the classification models.

From the simulation results in Table 5, the bs value used for the fold in the comparison method is also 64. The highest accuracy was obtained with a value of 100 and a loss of 0.039 using Resnet101, RNN, and Imbalanced Data Sampler. Almost all folds have a value above 95. The overall loss from the experiment is less than 0.27. Meanwhile, the Resnet101 and GRU methods have an accuracy value of more than 95 only on fold 1 with 98,165 and a loss of 0.129. The lowest value is on fold 4, with an accuracy value of 31,192 and a loss of 1,638.

This study investigates six models integrating the ResNet101 architecture with LSTM, GRU, and RNN layers, as illustrated in Fig. 3. Models 1, 3, and 5 are configured without weight class balancing, whereas models 2, 4, and 6 incorporate weight class balancing. When employing class weight balancing, the observed accuracy percentages consistently fall within a range of at least 60.

The primary objective of the class weight balancing process is to enhance each model's performance and predictive accuracy. This approach considers the imbalanced class distribution within the dataset by optimizing results. Using weight class balancing has a discernible impact, elevating the classification accuracy percentage from 80 to 100.

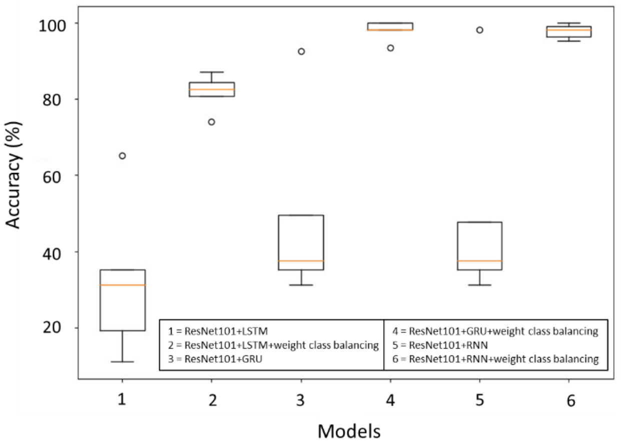

Figure. 4 The accuracy of the classification results of the ViViT and Tubelet Embedding methods is compared with other methods considering weight class balancing

Among the models, those incorporating the ResNet101 and RNN layers consistently achieved the highest accuracy results, showcasing the efficacy of this combination. Conversely, the models featuring ResNet101 and LSTM layers always recorded the lowest accuracy results. This comparative analysis sheds light on the significance of weight class balancing in improving model performance, providing valuable insights for optimizing traditional dance classification models.

Table 6 provides an overview of the results of employing the ViViT and the Tubelet architecture for traditional dance classification. The architectural configuration encompasses a frame size of 128, seq (sequence length) set to 25, a learning rate of 1e-4, and a patch size of (8, 8, 8). Notably, the achieved accuracy level reaches 100 in folds 2 and 3, with minimal losses of 0.004 and 0.011, respectively.

Examining fold 2, the model showcases exemplary performance with the highest F1 score, Precision, and Recall values observed on this particular fold, registering at 0.826, 0.818, and 0.844, respectively. These metrics align seamlessly with the high accuracy level achieved in fold 2. Overall, the incurred losses across all folds are consistently below 0.011, underscoring the robustness and efficiency of the ViViT and the Tubelet architecture.

While maintaining exceptional accuracy across various folds, the least accuracy is noted in fold 5, recording a still-impressive accuracy of 98.68 with a minimal loss of 0.008. These findings affirm the reliability and generalizability of the ViViT and the Tubelet approach in traditional dance classification, highlighting its ability to consistently achieve high

Table 8 Best Loss and Accuracy, Average Loss and Accuracy from 5 Fold Experiment

Method	Best Fold		Average of 5 Folds	
	Loss	Loss	Loss	Acc
ViViT + Tubelet	0.004	0.004	0.0058	99.216
Resnet101 + LSTM	1.101	1.101	1.5798	32.3578
Resnet101 + LSTM+weight class balancing	0.518	0.518	0.5372	81.7868
Resnet101 + GRU	0.243	0.243	1.319	49.2384
Resnet101 + GRU +weight class balancing	0.069	0.069	0.13	97.9816
Resnet101 + RNN	0.129	0.129	1.2942	49.9724
Resnet101 + RNN + weight class balancing	0.039	0.039	0.1374	97.7894
MobileNetv2 + LSTM + Imbalanced Data Sampler	0.021	0.021	0.0542	99.6312
MobileNetv2 + GRU + Imbalanced Data Sampler	0.029	0.029	0.0468	99.2626

accuracy levels across different folds while also emphasizing the model's resilience even in scenarios where accuracy experiences a slight dip.

5. Conclusion

Image processing capabilities have improved greatly to address challenges such as object detection, classification, clustering, and segmentation. This proposed research offers advances in classifying traditional dances through the use of ViViT with tubelet embedding. The ViViT model, combined with Tubelet Embedding shows good performance in five evaluation processes. This model achieves high accuracy, reaching 100% on the second and third folds, with minimal loss. Additionally, the F1 score remains consistently above 0.8 at all levels, indicating a strong balance between precision and recall. These results show that the ViViT model with Tubelet Embedding can work to classify traditional dances accurately and shows its potential for video classification. In dance video research, six models were explored using the ResNet101 architecture integrated with LSTM, GRU, and RNN layers. Comparisons without and with weight class balancing were also carried out. Without weight class balancing, the highest results are in the minimum accuracy percentage range of 60, but when using balancing, accuracy can increase to 80 to 100. The highest accuracy results are obtained from the ResNet101 and RNN models, while the lowest results are found in the combination of ResNet101 and LSTM.

In Table 7, the classification of dance videos using Vivit and Tubelet with 5 flod experiments shows a loss between 0.003 to 0.011 with an average loss of 0.0058. The experiments also produced accuracy rates between 98.68 to 100 percent, resulting in an average accuracy of 99.216. These results are the best of the comparison methods shown in Tables 3 to 5.

Further research can be done to extend this strategy to various dance video genre contexts, refine the optimal frame extraction method, explore video data scalability, and evaluate real-world applications for the proposed video classification.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Edy Mulyanto, Eko Mulyanto Yuniarno and Mauridhi Hery Purnomo; methodology, Edy Mulyanto, Oddy Virgantara Putra, Isa Hafizh and Eko Mulyanto Yuniarno; validation, Edy Mulyanto, Oddy Virgantara Putra, Isa Hafizh and Eko Mulyanto Yuniarno; software and resources, Edy Mulyanto, Ardyono Priyadi ; investigation, Edy Mulyanto, Eko Mulyanto Yuniarno and Mauridhi Hery Purnomo; data curation, Edy Mulyanto, Oddy Virgantara Putra and Isa Hafizh; writing-original draft preparation, Edy Mulyanto; writing-review and editing, Edy Mulyanto, Eko Mulyanto Yuniarno, Ardyono Priyadi and Mauridhi Hery Purnomo; visualization, Edy Mulyanto, Oddy Virgantara Putra and Isa Hafizh ; supervision, Eko Mulyanto Yuniarno, Ardyono Priyadi; All authors read and approved the final manuscript.

References

[1] M. S. Hutchinson and V. N. Gadepally, “Video Action Understanding”, *IEEE Access*, Vol. 9, pp. 134611–134637, 2021, doi: 10.1109/ACCESS.2021.3115476.

[2] Y. Liu, F. Yang, and D. Ginhac, “TEDdet: Temporal Feature Exchange and Difference Network for Online Real-Time Action

Detection", *IEEE Access*, Vol. 10, pp. 37870–37881, 2022, doi: 10.1109/ACCESS.2022.3164730.

[3] R. Zhang, "Analyzing body changes of high-level dance movements through biological image visualization technology by convolutional neural network", *J Supercomput*, Vol. 78, No. 8, pp. 10521–10541, 2022, doi: 10.1007/s11227-021-04298-y.

[4] J M D, Sinag. "Dance Ethnography: An Analysis on Aeta Ambala Tribe of Barangay Tubo-tubo, Bataan", *Universal Journal of Educational Research*, Vol. 1, No. 4, pp. 218-231, 2022, doi: 10.5281/ZENODO.7269393.

[5] Y. Zhao, "Teaching traditional Yao dance in the digital environment: Forms of managing subcultural forms of cultural capital in the practice of local creative industries", *Technology in Society*, Vol. 69, p. 101943, 2022, doi: 10.1016/j.techsoc.2022.101943.

[6] R. E. Cisneros, K. Stamp, S. Whatley, and K. Wood, "WhoLoDancE: digital tools and the dance learning environment", *Research in Dance Education*, Vol. 20, No. 1, pp. 54–72, 2019, doi: 10.1080/14647893.2019.1566305.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks", In: *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA: IEEE, 2014, pp. 1725–1732. doi: 10.1109/CVPR.2014.223.

[8] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", *arXiv*, 2014, Accessed: Aug. 14, 2023, [Online], Available: http://arxiv.org/abs/1406.2199

[9] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond Short Snippets: Deep Networks for Video Classification", In: *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 4694-4702, 2015. doi: 10.1109/CVPR.2015.7299101.

[10] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis", In: *Proc. of 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, New York City, NY, pp. 540–546, 2017, doi: 10.1109/UEMCON.2017.8249013.

[11] E. Ergun, F. Gurkan, O. Kaplan, and B. Gunsel, "Video action classification by deep learning", In: *Proc. of 2017 25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey, pp. 1–4, 2017, doi: 10.1109/SIU.2017.7960446.

[12] M. Abdullah, M. Ahmad, and D. Han, "Facial Expression Recognition in Videos: An CNN-LSTM based Model for Video Classification", In: *Proc. of 2020 International Conference on Electronics, Information, and Communication (ICEIC)*, Barcelona, Spain, pp. 1–3, 2020, doi: 10.1109/ICEIC49074.2020.9051332.

[13] M. A. Russo, A. Filonenko, and K.-H. Jo, "Sports Classification in Sequential Frames Using CNN and RNN", In: *Proc. of 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT)*, Busan, pp. 1–3, 2018, doi: 10.1109/ICT-ROBOT.2018.8549884.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, ukasz, and I. Polosukhin, "Attention is All you Need", *presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, pp. 5598–6008,2017.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In: *Proc. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186, 2019.

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, … and D. Amodei, "*Language Models are Few-Shot Learners*", *arXiv*, 2020, doi: 10.48550/arXiv.2005.14165.

[17] S. Mostafa, E. Abyad, M. M. Soliman, K. Mostafa, and E. Sayed, "Deep Video Hashing Using 3DCNN with BERT", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 5, pp. 113–127, 2022, doi: 10.22266/ijies2022.1031.11.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", *arXiv*, 2021, doi: 10.48550/arXiv.2010.11929.

[19] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision

Transformer", *arXiv,* 2021, doi: 10.48550/arXiv.2103.15691.

[20] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A Survey of Transformers", *arXiv,* 2021, doi: 10.48550/arXiv.2106.04554.

[21] M. Yasuda, Y. Ohishi, S. Saito, and N. Harada, "Multi-view and Multi-modal Event Detection Utilizing Transformer-based Multi-sensor fusion", *arXiv*, 2022, doi: 10.48550/arXiv.2202.09124.

[22] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers", *arXiv,* 2019, doi: 10.48550/arXiv.1908.07490.

[23] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", *Int J Comput Vis*, Vol. 103, No. 1, pp. 60–79, 2013, doi: 10.1007/s11263-012-0594-8.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *Commun. ACM*, Vol. 60, No. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database", In: *Proc. of 2009 IEEE Conference on Computer Vision and Pattern Recognition,* Miami, FL, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.

[26] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks", In: *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1725–1732, 2014, doi: 10.1109/CVPR.2014.223.

[27] A. Abbas, A. Chadha, Y. Andreopoulos and M. Jubran, "Rate-Accuracy Trade-Off in Video Classification with Deep Convolutional Neural Networks", In: *Proc. of 2018 25th IEEE International Conference on Image Processing (ICIP),* Athens, Greece, pp. 793-797, 2018.

[28] D. Zhang, H. Gao, H. Dai and X. Shi, "Two-stream Graph Attention Convolutional for Video Action Recognition", In: *Proc. of 2021 IEEE 15th International Conference on Big Data Science and Engineering (BigDataSE),* Shenyang, China, pp. 23-27, 2021.

[29] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C.Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset", *Computing Research Repository (CoRR),* CoRR abs/1705.06950, 2017.

[30] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, pp. 4724-4733, 2017, doi: 10.1109/CVPR.2017.502.

[31] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal Residual Networks for Video Action Recognition", In: *Proc. of Advances In Neural Information Processing Systems (NIPS),* pp. 3468-3476, 2016.

[32] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video Classification with Channel-Separated Convolutional Networks", In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV),* Seoul, Korea (South), 2019, pp. 5551-5560, 2019, doi: 10.1109/ICCV.2019.00565.

[33] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6450-6459, 2018, doi: 10.1109/CVPR.2018.00675.

[34] C. Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition", *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, WA, USA, pp. 200-210, 2020, doi: 10.1109/CVPR42600.2020.00028.

[35] J. You and J. Korhonen, "Attention Boosted Deep Networks For Video Classification", In: *Proc. of 2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, pp. 1761-1765, 2020.

[36] Z. Zhou, V. W. L. Tam and E. Y. Lam, "SignBERT: A BERT-Based Deep Learning Framework for Continuous Sign Language Recognition", *IEEE Access*, Vol. 9, pp. 161669-161682, 2021.

[37] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research*, Vol. 21, pp. 1-67, 2020.

[38] Y. Ouali, A. Bulat, B. Matinez and G. Tzimiropoulos, "Black Box Few-Shot Adaptation for Vision-Language models", In: *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV),* Paris, France, pp. 15488-15500, 2023.

[39] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating Long Sequences with Sparse

Transformers", *Computing Research Repository (CoRR)*, 2019.

[40] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. J. Colwell, and A. Weller, "Rethinking Attention with Performers", In: *Proc. of International Conference on Learning Representations*, 2020.

[41] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The Efficient Transformer", In: *Proc. of 2020 International Conference on Learning Representations*, 2020.

[42] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey", *ACM Computing Survey*, Vol. 55, pp. 1-28 2020.

[43] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long Range Arena: A Benchmark for Efficient Transformers", In: *Proc. of International Conference on Learning Representations*, 2020.

[44] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal Transformers", In: *Proc. of International Conference on Learning Representations*, 2018.

[45] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng,H. Shen, and H. Xia, "End-to-End Video Instance Segmentation with Transformers", *CVPR, Computer Vision Foundation / IEEE*, pp. 8741-8750, 2020.

[46] Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, and T. S. Huang, "CCNet: Criss-Cross Attention for Semantic Segmentation", *ICCV, IEEE*, pp. 603-612, 2019.

[47] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-Alone Self-Attention in Vision Models", *NeurIPS*, pp. 68-80, 2019.

[48] Z. Shen, I. Bello, R. Vemulapalli, X. Jia, and C.-H. Chen, "Global Self-Attention Networks for Image Recognition", In: *Proc. of 9th International Conference on Learning Representations*, 2020.

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", In: *Proc. of 9th* International Conference on Learning Representations, *ICLR*, 2020.

[50] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point Transformer", *IEEE Access*, Vol. 9, pp. 134826-13480, 2020.

[51] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers", *CVPR*, pp. 6881-6890, Computer Vision Foundation / IEEE, 2021.

[52] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, "Scalable Vision Transformers with Hierarchical Pooling", *ICCV*, pp. 367-376, IEEE, 2021.

[53] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers &amp; distillation through attention", In: *Proc. of 38th International Conference on Machine Learning*, Vol. 139, pp. 10347-10357, 2021.

[54] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?", In: *Proc. of 38th International Conference on Machine Learning*, Vol. 139, pp. 813-824, 2021.

[55] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video Transformer Network", *ICCVW*, pp. 3156-3165, IEEE, 2021.

[56] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzzolin, "Deep Learning for Detecting Multiple Space-Time Action Tubes in Videos", *The British Machine Vision Conference (BMVC)*, pp. 58.1-58.13, BMVA, 2016.

[57] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, "Online Real-time Multiple Spatiotemporal Action Localisation and Prediction", In: *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 3657-3666, 2016.

[58] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6047-6056, 2018.

[59] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, and C. Schmid, "Actor-Centric Relation Network", *ECCV 2018, Lecture Notes in Computer Science*, Vol. 11215, Springer, pp. 335-351, 2018.

[60] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition", In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 6201-6210, 2019.

[61] J. Tang, J. Xia, X. Mu, B. Pang, and C. Lu, "Asynchronous Interaction Aggregation for

Action Detection", *ECCV (15), volume 12360 of Lecture Notes in Computer Science*, pp. 71-87. Springer, 2020.

[62] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, "Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 464-474. 2021.

[63] S. Chen, P. Sun, E. Xie, C. Ge, J. Wu, L. Ma, J. Shen, and P. Luo, "Watch Only Once: An End-to-End Video Action Detection Framework", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 8158–8167. doi: 10.1109/ICCV48922.2021.00807.

[64] A. U. Khan, A. Mazaheri, N. da V. Lobo, and M. Shah, "MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering", *EMNLP (Findings)*, Vol. EMNLP 2020 of Findings of ACL, pp. 4648-4660, Association for Computational Linguistics, 2020.

[65] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled Transformer for Image Captioning", In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 8927–8936, 2019, doi: 10.1109/ICCV.2019.00902.

[66] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers", *ECCV (1)*, Vol. 12346 of Lecture Notes in Computer Science, pp. 213-229, Springer, 2020.

[67] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection", In: *Proc. of International Conference on Learning Representations*, 2020.

[68] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F.E. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 538-547, 2021.

[69] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale Vision Transformers", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 6804-6815, 2021.

[70] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. M. Snoek, "Actor-Transformers for Group Activity Recognition", *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 836-845, 2020.

[71] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe, "VidTr: Video Transformer Without Convolutions", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 13557-13567, 2021.

[72] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video Action Transformer Network", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 244-253, 2019.

[73] J. Zhao, Y. Zhang, X. Li, H. Chen, S. Bing, M.Xu, C. Liu, K. Kundu, Y. Xiong, D. Modolo, I.Marsic, C. G. Snoek, and J. Tighe, "TubeR: Tubelet Transformer for Video Action Detection", *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 13588-13597, 2021.