GestFormer: Multiscale Wavelet Pooling Transformer Network for Dynamic Hand Gesture Recognition

Mallika Garg, Debashis Ghosh, Pyari Mohan Pradhan
Department of Electronics and Communication Engineering,
Indian Institute of Technology, Roorkee, India
mallika@ec.iitr.ac.in, debashis.ghosh@ece.iitr.ac.in, pyarimohan.pradhan@gmail.com

Abstract

Transformer model have achieved state-of-the-art results in many applications like NLP, classification, etc. But their exploration in gesture recognition task is still limited. So, we propose a novel GestFormer architecture for dynamic hand gesture recognition. The motivation behind this design is to propose a resource efficient transformer model, since transformers are computationally expensive and very complex. So, we propose to use a pooling based token mixer named PoolFormer, since it uses only pooling layer which is a non-parametric layer instead of quadratic attention. The proposed model also leverages the space-invariant features of the wavelet transform and also the multiscale features are selected using multi-scale pooling. Further, a gated mechanism helps to focus on fine details of the gesture with the contextual information. This enhances the performance of the proposed model compared to the traditional transformer with fewer parameters, when evaluated on dynamic hand gesture datasets, NVidia Dynamic Hand Gesture and Briareo datasets. To prove the efficacy of the proposed model, we have experimented on single as well multimodal inputs such as infrared, normals, depth, optical flow and color images. We have also compared the proposed GestFormer in terms of resource efficiency and number of operations. The source code is available at https://github.com/mallikagarg/GestFormer.

1. Introduction

Hand gesture recognition is an active and rapidly evolving area of research that involves various applications like sign gesture communication, human-computer interactions, gesture control appliances, autonomous vehicles, virtual reality, gaming etc. This is a challenging task since it involves variations in the pose, hand shape, position, directions and size of hand. There are also challenges due to variability of the image background, color differences, shadows and other lightening illumination which can be handled using depth sensors such as Leap Motion [44] and Microsoft Kinect sensor [12]. Gestures can be static or dynamic depending on the movement of hands. Static hand gestures are those where the hand remains relatively stationary and doesn't involve significant movement while dynamic hand gestures involve movement of the hands or fingers to convey meaning. In this work, we will focus on designing a model that recognizes dynamic hand gestures which are characterized by changes in hand position, orientation, or movement trajectories over time.

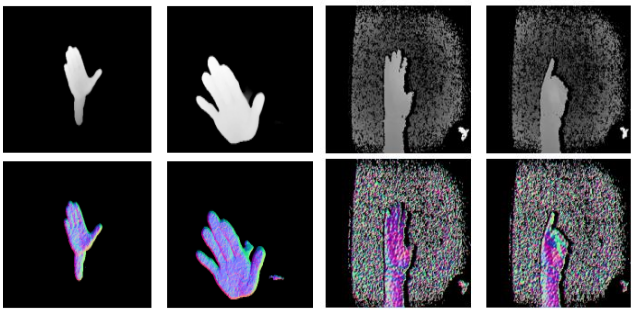

Figure 1. Some samples of depth (first row) and surface normals (last row) from the NVGesture and Briareo dataset.These samples are taken from [20].

With recent advancements in the deep learning algorithms, attention-based models have become popular in focusing on a certain portion of gesture image or video sequence. These attention-based models [3] have replaced traditional Recurrent Neural Networks (RNNs) [29], Long Short-Term Memory (LSTMs) [6] and various deep learning methods [23, 38] for hand gesture recognition. The recently introduced transformers are one such model that uses attention to focus on a certain portion of image or video sequence. A transformer-based model that classifies dynamic Hand Gesture Recognition is proposed in [20]. This method uses vanilla transformer [51], which comprises of two op-

eration. First, the attention operation is performed which is followed by the multi-layer perceptron (MLP). The attention operation models the relations between elements while the MLP is employed to model the relation within each individual element. Despite their effectiveness in these domains, their application to visual data, especially in dynamic hand gesture recognition tasks, remains relatively very limited.

So, we explore a transformer-based approach for dynamic gesture recognition. Some samples of the Dynamic gestures from NVGesture [40] and Briareo [39] dataset are shown in Fig. 1. Traditional Transformer [51] takes the advantage of quadratic attention which is computationally expensive, $O(n^2)$. This problem was addresses by Linformer [53], which uses linear attention with $O(n)$ complexity in both time and space. With advancements, the attention layer has been completely replaced by layers or modules that has no learnable parameters. PoolFormer [61] and FNet [34] proposes an attention-free network which uses average pooling and Fourier transform to mix the token of the input sequence. This helps reduces the complexity of the model to a great level. Inspired from PoolFormer, we also proposed a poolformer based technique that completely eliminates the attention mechanism and rely on token mixing. Pooling the input can aggregate token from input to learn contextual information and perform comparable to the Vision Transformer with very less complexity.

To further, enhance the performance of the poolformer for dynamic gesture recognition, we propose a novel **M**ultiscale **W**avelet **P**ooling **A**ttention (MWPA) mechanism which takes the advantage of wavelet transform [71] and can be used as an attention approximation mechanism. We also proposed a **G**ated **F**eed **F**orward **N**etwork (GFFN) to control the flow of the information through the different stages of the proposed **M**ultiscale **W**avelet **P**ooling **T**ransformer (MWPT).

Thus, we summarize our key innovations as:

1. We propose a novel GestFormer, a multiscale wavelet pooling transformer (MWPT) model for dynamic hand gesture recognition.
2. We propose a novel token mixer called Multiscale Wavelet Pooling Attention (MWPA) which uses multiscale pooling and a wavelet transform to map the input to wavelet space before passing it through the pooling layer. This helps boosts the long-range understanding capabilities of the model.
3. We also propose a Gated Feed forward network which helps to precisely filter the information forwarding to subsequent stages of the transformer block.
4. Experiments on NVGesture and Briareo dataset are done to prove the efficacy in terms of performance and resource utilisation of the proposed model.

2. Related Work

In the literature, there are several techniques that rely on traditional methods for hand gesture recognition which often involve manual engineering of features extraction and the use of classical machine learning algorithms. Earlier hand-crafting features were extracted from raw data, such as images or depth maps of hand gestures to train classical machine learning algorithms such as Support Vector Machines (SVM) [2], Bayesian-classifier [33], Hidden Markov Models (HMMs) [32], etc. With these traditional methods for hand gesture recognition, there are issues like robustness, scalability, and adaptability to diverse environments and user conditions that reduces the performance of the traditional methods.

Later, with the advent of deep learning, there has been a shift towards more data-driven approaches that automatically learn features from raw data, leading to significant improvements in performance and robustness of the system. With advanced deep learning technologies, Recurrent Neural Networks (RNNs) [29] and Long Short-Term Memory (LSTMs) [6] were developed for handling continuous sign gestures. Nowadays, transformer models are used for gesture recognition, which are designed for sequential data [13].

2.1. Transformer for Vision Tasks

Transformer-based networks have shown remarkable success in the field of natural language processing [51], computer vision tasks and modelling sequential data. Since transformers rely on attention mechanism, these models have shown huge progress in object detection [70], text generation [55], image classification [11, 27, 61], segmentation [8], recommendation systems [46], super-resolution [28], dialogue system [60], pose estimation [42], text understanding [5] and many more. Development of ViT [18] marked a significant milestone in the utilization of transformers for vision-based tasks. ViT is a pure transformer-based convolution-free approach which achieves competitive performance compared to CNNs. Later, transformers were used for video based tasks [41].

Inspired by ViT, a Video Vision Transformer (ViViT) [4] has been introduced that extracts token from video sequence. ViViT presents variants of the models to factorise the spatio-temporal dimensions of the input video: Spatio-temporal attention, Factorised encoder, Factorised dot-product encoder and Factorised dot-product attention. All these models factorise different components of the transformer model to factorise large spatio-temporal token in the video sequence. DeepViT [69] is another vision transformer which elaborates the issue that attention map goes similar as transformer digs deeper. This signifies that the self attention mechanism fails at deeper layers. So, DeepVit found a solution to this problem by re-generating the

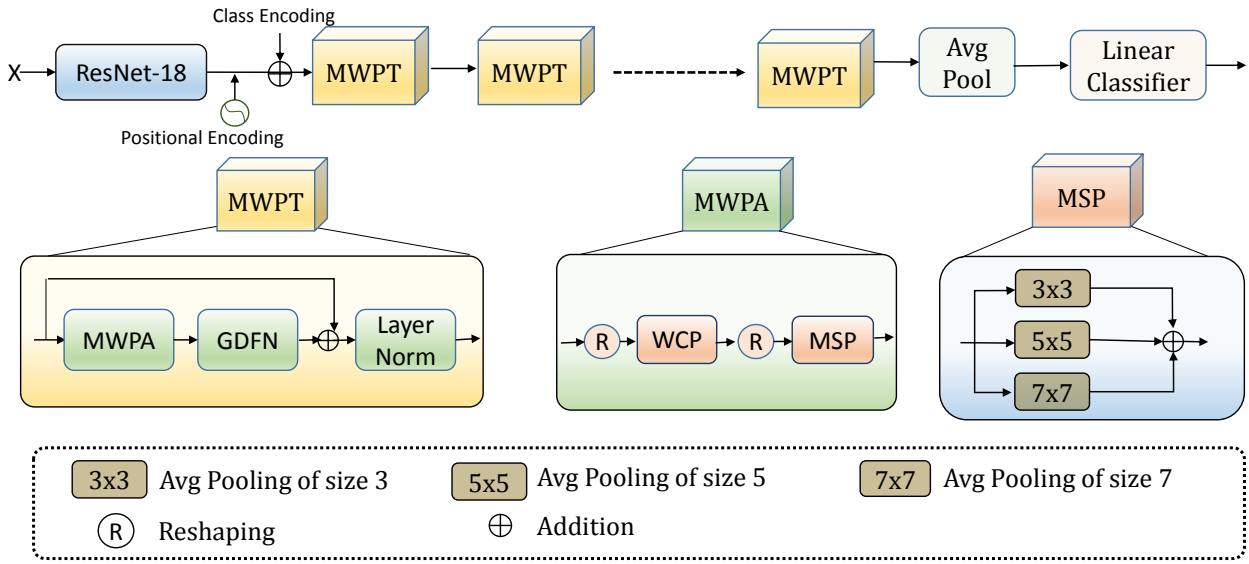

Figure 2. The overall architecture of the proposed GestFormer for dynamic hand gesture recognition. The proposed GestFormer consists of Multiscale Wavelet Pyramid Attention module which comprises of Wavelet Coefficient Processing (WCP) and Multi-scale Pooling architecture (MSP) to perform learning in the wavelet coefficient space with multiscaled pooling to capture the scaled attentive information. GestFormer also leverage the Gated Dconv FFN (GDFN) to control the forward flow of the information.

attention to get more diverse attention in the deeper layers.

Although, transformers have marked incredible progress in vision based tasks, they face certain difficulty when these models deal with large sequential data. Since, transformers use quadratic attention, and vision transformers used large sequence length of image tokens, transformers used in visions are computationally expensive and space complexity is also high. Along with this, the vanilla transformer, outputs a feature map of same dimension at each transformer stage. To tackle these issue, various models that reduce the dimension of the input sequence progressively in the transformer stages are introduced recently. There have been 2 ways to reduce the dimensionality, Convolution based reduction and pooling based reduction. Pyramid Vision Transformer (PvT) [54], Pyramid Pooling Transformer (P2T) [57], MsMHA-VTN [24], MViT [22], Improved MViT [36], PSViT [9], POSTER [67] are some methods which use pooling to reduce the sequence length and reduce the computation cost of the entire system. There are some other methods that use pyramid hierarchy but they incorporate convolutional layers instead of pooling e.g. POTTER [66], Convolutional Vision Transformer (CvT) [56], Swin Transformer[37], CSwin [17], CeiT [63], Unifying CNNs [35], CoFormer [15], etc.

2.2. Token Mixing

Since, the computation cost of quadratic attention is very high, researchers are now more inclined to replace this attention with some low computational token mixing. PoolFormer [61] exploits a general pooling non-parameteric operator to help in basic token mixing. It is the MetaFormer which is actually a generalised mixer for token in computer vision tasks. Another model that mixes the input token by linear transformations (Fourier transform) is FNet [34]. Convolution can also be used to mix tokens as in ConvMixer [49]. Wavemix [28] uses wavelet transformer and convolution. Similarly, MLP-Mixer [47] presents a method that uses MLP for mixing tokens. It separates the channel-mixing and token-mixing task and both tasks use MLP in this architecture. All these token mixing architectures have comparable performance when compared with the transformer model with less computational requirements.

2.3. Transformer for Gesture Recognition

Transformers have nowadays been used in gesture recognition. In [20], RGBD data is used to predict the class of dynamic gesture using color image , depth maps. It particularly shows that depth maps and the normals which are derived from depth map outperforms other modalities. This method also leverages single and multimodal inputs using basic transformer model. To give the model the order of sequence, positional embedding is employed. An advancement over sinusoidal positional encoding is proposed using a new positioning scheme based on Gated Recurrent Unit (GRU) into Transformer networks [3].

Earlier, multimodal output was taken by fusing the output probability of single modal inputs using decision level fusion, but multimodal fusion at inputs can also be done at

the feature fusion stage. One such method [25] which uses convolutional transformer blocks to fuse at the input level is called early fusion. It also performs experiments with mid fusion, late fusion and multi-level fusion. Spatio-temporal features can also be extracted using transformer models using transformations to canonical maps from both spatial and temporal information [7]. Transformer uses columnar structure to map input to same dimensional features. MsMHA-VTN [24] maps the input to multidimensional subspace using pyramid attention networks. This also helps in the reduction of the computational cost of the model. A combined spatiotemporal vision and spatiotemporal channel attention mechanisms can extract context information from the input feature using self attention [10] on multimodal RGBD data.

3. Method

3.1. Overview

We propose a transformer-based gesture recognition framework that is designed for dynamic sequence of hand gesture. An overview of the proposed GestFormer is shown in Figure 2. GestFormer takes a sequence of m frames as an input which can be represented as $X = \{x_1, x_2, .., x_m\}$, $X \in \mathbb{R}^{m \times w \times h \times c}$, where $w \times h$ is the size of each frame with c channels. First, the features, F are extracted from each frame using a ResNet-18 [26] model which outputs a map of $\mathbb{R}^{m \times k}$. These features are then fed to the proposed GestFormer block to learn the wavelet of multiscale features. Our proposed GestFormer consists of 6 stages of Multiscale Wavelet Pooling Transformer (MWPT) blocks to get the refined features which finally helps to predict the probability distribution of n classes using a linear classifier.

3.2. Multiscale Wavelet Pooling Transformer (MWPT)

In traditional transformers [51], input is projected into three different vectors, Query, Key, and Value using linear transformation. The attention from these 3 vectors is computed using scaled-dot product of the Query and Key, normalising it and applying softmax to obtain the weights of the value. Computation of the attention in this transformer has quadratic complexity which increase with long sequences. To deal with this issue, we use PoolFormer [61] as the core architecture of our proposed MWPT model. PoolFormer replaces the attention mechanism with pooling based token mixing which is a simple non-parametric operation and it has fewer parameters compared to the traditional transformer.

The goal is to develop a model that is computationally less expensive and at the same time, performance of the model is also comparable. PoolFormer achieves competitive results on dynamic gesture recognition when initial experiments were performed. To further enhance the performance, we explore various techniques built on the core PoolFormer structure. The features obtained from the ResNet are first embedded using spatial embedding [28]. We also use positional embedding to make the model know the order of the sequence [50]. This encoded input with positional embedding is fed to the proposed MWPT blocks. We propose a novel token mixer called Multiscale Wavelet Pooling Attention (MWPA) which uses multiscale pooling and a wavelet transform before passing the input through the pooling layer. Our MWPA is purely convolution based architecture. After the tokens are mixed in the pool token mixer, we fed the features to the Gated Depthwise Feed Forward Network (GDFN) block, which helps in selectively passing the fine details in addition with the skip connection to the next stage after layer normalisation. A stack of 6 MWPA stages is used in the proposed MWPT.

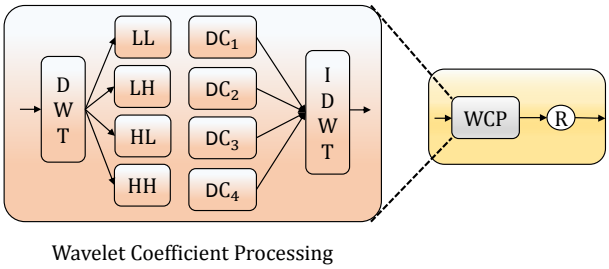

Figure 3. Detailed Wavelet transform Processing (WCP) block. The proposed WCP requires only linear time complexity. It first maps the input into its wavelet coefficients which decomposes the input into multipple sinusoidal waves. The wavelet coefficients is the magnitude of the sinusoidal. After enhancing these magnitudes using Dconv (DC_x, x= 1,2,3,4), the coefficients are re-mapped in input space via backward wavelet transform.

3.2.1 Multiscale Wavelet Pooling Attention (MWPA)

The PoolFormer uses a single input, unlike the vanilla transformer which uses 3 attention vectors. Since the input is fed to the pooling layer directly, it plays an important role in the full transformer block. Pooling helps to select the important features from the input. Further, providing enhanced features as input to the pooling layer can help the model to improve the performance. The enhanced features are calculated by using a wavelet-based forward and backward paradigm [71]. This facilitates the pooling layer to aggregate the enhanced features in wavelet coefficient space. We follow [43], which uses wavelet-based query for image inpainting to reduce the noise forwarding to the attention block. Applying wavelet transform has linear complexity in contrast to transformers which has quadratic complexity. Our model is still less complex.

We calculate the wavelet coefficient of the input features, F as:

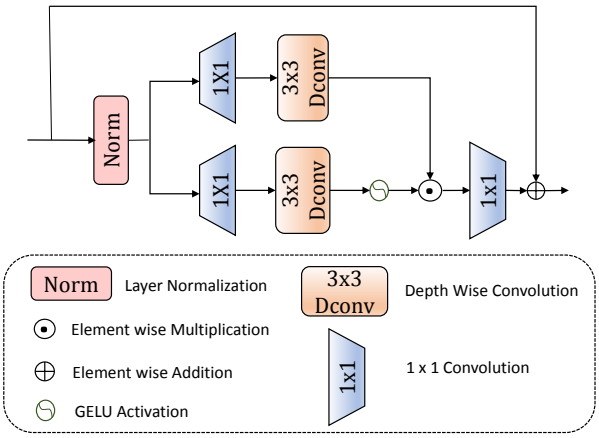

Figure 4. The detailed Gated Depth-wise Feed Forward Network (GDFN) structure. GDFN facilitates subsequent layers within the network hierarchy to concentrate on more detailed image attributes, thereby resulting in better performance of the complete model.

$$LL, LH, HL, HH = DWT(F), \tag{1}$$

where, the input feature is divided into 4 subspace, which are approximation (LL) and details in 3 orientations as horizontal (LH), vertical (HL) and diagonal (HH). These coefficients are the magnitude of the corresponding sinusoidal wave decomposed after wavelet transform. After the extraction of these coefficients, we separately enhance these features using Depth-wise separable convolution as shown in Fig. 3. Further, inverse wavelet transform is calculated from the processed output features, which are given as input to the pooling layer of the PoolFormer.

In order to extract the important features using pooling, we propose a Multiscale Pooling (MSP) mechanism which helps to aggregate the multiscale information (shown in MSP block in Fig. 2). A multiscale pooling can accurately capture the hand shape and size variations and recognise the hand with different scales. We propose to apply 3 filters for pooling the input features from the WCP block, (3×3, 5×5 7×7). Output from these 3 pooling layers are then averaged to get a strong aggregated enhanced feature. This enhanced feature is the overall output of the proposed MWPA block.

3.2.2 Gated Depthwise Feed Forward Network (GDFN)

To transform the features from MWPA block, we follow [64] to apply two modifications in FFN: gating mechanism and depth-wise convolutions. The architecture of GDFN is shown in Fig. 4, which helps control the flow of important feature or fine information to the next stage of the transformer blocks. This is formulated by linearly transforming input using depth-wise convolution and performing element-wise product of two parallel features, of which one is Gelu activated represented as.

$$\mathbf{P'} = W_p^0 Gating(\mathbf{P}) + \mathbf{P}, \tag{2}$$

$$Gating(\mathbf{P}) = \phi(W_d^1 W_p^1(\mathbf{P})) \odot W_d^2 W_p^2(\mathbf{P})) \tag{3}$$

here, $\odot$ denotes the element-wise multiplication and ϕ represents the GELU activation.

3.3. Multi-Modal Late Fusion

Multi-modal methods have gained the popularity among research community and have been used in numerous application. RGB-D sensors provides RGB images, depth images, infrared images and it has been used to acquire the NVGestures and Briareo dataset for dynamic hand gesture recognition. Following [20], we also adapt late fusion technique to predict the multimodal accuracy of the inputs. We have simply averages the output probability score from each input modality trained separately which is given as

$$y = \arg\max_j \sum_i^n P(\omega_j|x_i), \tag{4}$$

where n is the number of modalities over which the results are to be aggregated, and $P(\omega_j|x_i)$ is the probability distribution of the i^{th} frames of a given input, which belongs to class ω_j.

4. Experiments and Discussion

Experiments are performed on single as well as multimodal inputs on NVGesture and Briareo. We also analyse the number of learnable parameters and the MACs along with the ablation on each component of the model.

4.1. Datasets

NVGesture: NVGesture [40] is a dynamic hand gesture dataset containing 1582 images in total from 25 different classes. Dataset is divided into two parts, having 1050 samples in training and rest in test dataset. Dataset samples were collected in three different modalities (RGB, IR, and depth) by a group of 20 subjects.

Briareo: Briareo dataset [39] is collected for dynamic hand gesture recognition. The dataset samples are collected using a RGB camera, depth sensor, and an infrared stereo camera, under natural lighting conditions. Since images are captured in natural lighting, images are dark and of low contrast. The dataset contains 12 different dynamic gestures which were performed by 40 subjects among them 33 were males and 7 were females. Each gesture is performed 3

Table 1. Results for different modalities on NVGesture [40] and Briareo [39] dataset. # is the number of input modalities used. Transformer results are the one reported in [20].

#	Input data					Accuracy			
						NVGesture		Briareo	
	Color	Depth	IR	Normals	Optical flow	Transformer [20]	GestFormer	Transformer [20]	GestFormer
1	✓					76.50%	75.41%	90.60%	94.44%
		✓				83.00%	80.21%	92.40%	96.18%
			✓			64.70%	63.54%	95.10%	**98.13**%
				✓		82.40%	**81.66**%	95.80%	97.22%
2	✓	✓				84.60%	82.57%	94.10%	96.78%
	✓		✓			79.00%	77.19%	95.50%	95.87%
		✓	✓			81.70%	79.88%	95.10%	96.57%
	✓			✓		84.60%	82.75%	96.50%	97.44%
		✓		✓		87.30%	**82.78**%	96.20%	96.33%
			✓	✓		83.60%	82.18%	97.20%	**97.57**%
	✓				✓	72.00%	72.61%	-	96.57%
3	✓	✓	✓			85.30%	**84.24**%	95.10%	96.78%
	✓	✓		✓		86.10%	83.81%	95.80%	**97.42**%
	✓		✓	✓		85.30%	83.40%	96.90%	96.88%
		✓	✓	✓		87.10%	83.61%	97.20%	96.79%
4	✓	✓	✓	✓		87.60%	**85.62**%	96.20%	96.33%
	✓	✓		✓	✓	-	85.85%	-	96.79%
	✓		✓	✓	✓	-	85.31%	-	**97.42**%
	✓		✓	✓	✓	-	84.55%	-	96.79%
		✓	✓	✓	✓	-	85.96%	-	96.79%
5	✓	✓	✓	✓	✓	-	**85.85**%	-	96.88%

times by every subject. Thus a total of 120 (40×3) sequences of each gesture is collected of at least 40 frames. Randomly, 32 subjects are placed in the train and the validation set and 8 subjects in the test set.

4.2. Implementation Details

The proposed GestFormer model was implemented, trained and tested using Torch=1.7.1 with 12 GB Nvidia GeForce GTX 1080 Ti GPU, CUDA 10.1 with cuDNN 8.1.1. 40 frames of a gesture are given as input to the model to optimise the loss using Adam optimizer over categorical cross entropy loss. The model is trained with a batch size of 8 at $1e^{-4}$ learning rate which decays after 50^{th} and 75^{th} epoch. Following [20], we use ResNet-18 model as feature extractor which is pre-trained on the ImageNet dataset [14]. Each modality was separately trained, and probability score for each modality is calculated. Late fusion was used to combine different modalities for integration of diverse sources of information for improved performance.

4.3. Results and Discussion

NVGesture: We follow [20] to performed experiments with single as well as multi-modality. The result compared with the traditional transformer are compared for single and multimodal combinations for NVGesture in Table 1. The proposed GestFormer achieves the state-of-the-art results with lesser number of parameters. Lesser parameters are the results of the pooling layers used to replace the attention mechanism. From the table, we can observe that GestFormer obtained best result on normals with an accuracy of 81.66% and nearly similar result is obtained in depth maps. This is because normals are derived from the depth images.

Further, the accuracy increases when more than one modality is used as input. The results in multimodal approach are obtained using late fusion. When RGB images are fused with normals or depth maps, an increment in the accuracy is seen. It further increases when normals and depth inputs are fused. Among all the combination of 2 modalities, best performance is obtained when normal and depth is fused which is 82.78%. From the table, it can be clearly seen that adding a modality shows an increment in the accuracy. 3 modality reaches an accuracy of 84.24% with RGB, depth and IR fusion. Accuracy further improves to 85.62% with 4 modal input and the best accuracy is obtained with all the 5 modalities which is 85.85% on the proposed GestFormer. However, it is still less compared to the traditional transformer [20] on single as well as multimodal inputs.

Table 2. Comparison results for single modality on NVGesture dataset [40]. All results are taken from the respective papers.

Input modality	Method	Accuracy
Color	Spat. st. CNN [45]	54.60%
	iDT-HOG [52]	59.10%
	Res3ATN [16]	62.70%
	C3D [48]	69.30%
	R3D-CNN [40]	74.10%
	GPM [21]	75.90%
	PreRNN [59]	76.50%
	Transformer [20]	76.50%
	I3D [52]	78.40%
	ResNeXt-101 [31]	78.63%
	MTUT [1]	81.33%
	NAS1 [62]	83.61%
	Human [40]	88.40%
	MotionRGBD [68]	89.57%
	GestFormer	**75.41%**
Depth	SNV [58]	70.70%
	C3D [48]	78.80%
	R3D-CNN [40]	80.30%
	I3D [52]	82.30%
	Transformer [20]	83.00%
	ResNeXt-101 [31]	83.82%
	PreRNN [59]	84.40%
	MTUT [1]	84.85%
	GPM [21]	85.50%
	NAS1 [62]	86.10%
	MotionRGBD [68]	90.62%
	GestFormer	**80.21%**
Optical flow	iDT-HOF [50]	61.80%
	Temp. st. CNN [45]	68.00%
	Transformer [20]	72.00%
	iDT-MBH [50]	76.80%
	R3D-CNN [40]	77.80%
	I3D [52]	83.40%
	GestFormer	**72.61%**
Normals	Transformer [20]	82.40%
	GestFormer	**81.66%**
Infrared	R3D-CNN [40]	63.50%
	Transformer [20]	64.70%
	GestFormer	**63.54%**

We also compare the performance of the proposed GestFormer with other methods on single modality in Table 2, and on multimodal inputs in Table 3 and observe that GestFormer achieves state-of-the-art results. We can also observe from the Table 2 that our model is able to outperform Transformer model [20] when optical flow input is given to the model.

Briareo: Similar to NVGesture, we performed experiments on Briareo dataset with single and multimodal inputs as shown in table 1. A comparison is also shown with the basic transformer architecture [20]. From the comparison,

Table 3. Comparison results for multi-modalities on NVGestures dataset [40].

Input modality	Method	Accuracy
iDT [50]	color + flow	73.00%
R3D-CNN [40]	color + flow	79.30%
R3D-CNN [40]	color + depth + flow	81.50%
R3D-CNN [40]	color + depth + ir	82.00%
R3D-CNN [40]	depth + flow	82.40%
R3D-CNN [40]	all	83.80%
MSD-2DCNN [21]	color+depth	84.00%
8-MFFs-3f1c[30]	color + flow	84.70%
STSNN [65]	color+flow	85.13%
PreRNN [59]	color + depth	85.00%
I3D [52]	color + depth	83.80%
I3D [52]	color + flow	84.40%
I3D [52]	color + depth + flow	85.70%
GPM [21]	color + depth	86.10%
$MTUT_{RGB-D}$ [1]	color + depth	85.50%
$MTUT_{RGB-D+flow}$ [1]	color + depth	86.10%
$MTUT_{RGB-D+flow}$ [1]	color + depth + flow	86.90%
Transformer [20]	depth + normals	87.30%
Transformer [20]	color + depth +ir + normals	87.60%
NAS2 [62]	color + depth	86.93%
NAS1+NAS2 [62]	color + depth	88.38%
MotionRGBD [68]	RGB + Depth	91.70%
GestFormer	**depth + normals**	**82.78%**
GestFormer	**depth + color + ir**	**84.24%**
GestFormer	**depth + color + ir normal**	**85.62%**
GestFormer	**depth + color + ir normal + op**	**85.85%**

Table 4. Comparison of the results obtained for different modalities on Briareo dataset [39].

Method	Tensor sizes	
C3D-HG [39]	color	72.20%
C3D-HG [39]	depth	76.00%
C3D-HG [39]	ir	87.50%
LSTM-HG [39]	3D joint features	94.40%
NUI-CNN [19]	depth + ir	92.00%
NUI-CNN [19]	color + depth + ir	90.90%
Transformer [20]	normals	95.80%
Transformer [20]	depth + normals	96.20%
Transformer [20]	ir + normals	97.20%
GestFormer	**ir**	**98.13** %
GestFormer	**ir + normals**	**97.57**%

we can conclude that GestFormer performs better on Briareo dataset compared to [20] with approx. 2-4% rise in accuracy with each modality. It can also be observed that our modal has better results on all the modalities and also on all the experiments individually, except the 2 experiments with 3 modalities. Best performance is observed when infra-red input is used, obtaining an 98.13% accuracy. Combining modalities did not lead to a notable improvement in GestFormer's performance.

Additionally, we also compared the results obtained by the proposed model with other methods in Table 4. It is evident that GestFormer achieves superior performance with an accuracy of 98.13%. Finally, we can also conclude from the results that GestFormer is able to achieve better results on single modalities, leading to a conclusion that even without using multimodal inputs for our methods, we are able to achieve better results than other state-of-the-art methods.

4.4. Ablation Study

We perform the ablation study on NVGesture depth modality. The proposed GestFormer has 8 baselines (BL1, BL2, BL3, BL4, BL5, BL6, BL7 and BL8) as shown in Table 5. Baseline BL1 is the transformer model with pooling layer similar to PoolFormer (A). Baseline BL2 explores the pooling transformer with the multi-scaling pooling network (B) where 3 types of filters are used for each scale. Baseline BL3 uses encoding of input using spatial embedding (C) with A as discussed in Section 3.2. Baseline BL4 and BL5 is the Wavelet transform (WCP) (D) and Gated Dconv FFN (GDFN) (E) used with A.

An initial experiment that shows the performance of PoolFormer is 76.04% which increases to 76.67% by using multi-scale pooling network. Further, addition of different modules to the poolformer aims to enhance the performance of the proposed model. From the table, we can conclude that addition of each baseline on BL1 has enhanced the performance of the model, giving a clear motivation of designing the proposed GestFormer model.

We have also compared the number of learnable parameters and the number of MAC of our model with other models and the traditional transformer model in Table 6. The numbers of parameters and MACs are comparatively less for GestFormer from other methods.

5. Conclusion

We proposed a novel GestFormer model for dynamic hand gesture recognition build on PoolFormer which is a computationally efficient model since it uses non parametric layer. We further enhance the performance by extracting wavelet coefficients and enhancing the features in wavelet space. We also leverage the multiscale contextual information by using multiscale pooling and a gated network to process the refined features. This helps the model to learn significant features with fewer parameters compared to the traditional transformer. Evaluating the proposed GestFormer on NVGesture and Briareo datasets shows our model achieves state-of-the-art results. For Briareo dataset, we can conclude that our GestFormer model is so efficient that it performs better with single input compared to other single and multimodal methods as well.

Table 5. Ablation study on the proposed GestFormer model.

Baseline	Module	Accuracy
BL1	PoolFormer (A)	76.04
BL2	A + MSP(B)	76.67
BL3	A + embedding(C)	77.29
BL4	A + WCP(D)	78.95
BL5	A + GDFN(E)	79.12
BL6	A + C + D	79.58
BL7	A + B + C + D	79.97
BL8	A + B + C + D + E	**80.21**

Table 6. Comparison in terms of the number of parameters (M)) and MACs. The numbers of MACs are counted by fvcore library.

Methods	Params (M)	MACs (G)
R3D-CNN [40]	38.00	-
C3D-HG [39]	26.70	-
Transformer[20]	24.30	62.92
GestFormer	**24.08**	**60.40**

References

[1] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1165–1174, 2019. 7

[2] Rajat Agarwal, Balasubramanian Raman, and Ankush Mittal. Hand gesture recognition using discrete wavelet transform and support vector machine. In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 489–493. IEEE, 2015. 2

[3] Neena Aloysius, M Geetha, and Prema Nedungadi. Incorporating relative position information in transformer-based sign language recognition and translation. *IEEE Access*, 9: 145929–145942, 2021. 1, 3

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2

[5] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International Conference on Document Analysis and Recognition*, pages 319–334. Springer, 2021. 2

[6] Danilo Avola, Marco Bernardi, Luigi Cinque, Gian Luca Foresti, and Cristiano Massaroni. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Transactions on Multimedia*, 21(1):234–245, 2018. 1, 2

[7] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[8] Mathilde Caron, Neil Houlsby, and Cordelia Schmid. Location-aware self-supervised transformers for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 117–127, 2024. 2

[9] Boyu Chen, Peixia Li, Baopu Li, Chuming Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, and Wanli Ouyang. Psvit: Better vision transformer via token pooling and attention sharing. *arXiv preprint arXiv:2108.03428*, 2021. 3

[10] Huizhou Chen, Yunan Li, Huijuan Fang, Wentian Xin, Zixiang Lu, and Qiguang Miao. Multi-scale attention 3d convolutional network for multimodal gesture recognition. *Sensors*, 22, 2022. 4

[11] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2

[12] Carlo Dal Mutto, Pietro Zanuttigh, and Guido M Cortelazzo. *Time-of-flight cameras and Microsoft KinectTM*. Springer Science & Business Media, 2012. 1

[13] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Sign language recognition with transformer networks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[15] Gayatri Deshmukh, Onkar Susladkar, Dhruv Makwana, Sparsh Mittal, et al. Textual alchemy: Coformer for scene text understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2931–2941, 2024. 3

[16] Naina Dhingra and Andreas Kunz. Res3atn-deep 3d residual attention network for hand gesture recognition in videos. In *2019 international conference on 3D vision (3DV)*, pages 491–501. IEEE, 2019. 7

[17] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 3

[18] 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. 2

[19] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Multimodal hand gesture classification for the human–car interaction. In *Informatics*, page 31, 2020. 7

[20] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020. 1, 3, 5, 6, 7, 8

[21] Dinghao Fan, Hengjie Lu, Shugong Xu, and Shan Cao. Multi-task and multi-modal learning for rgb dynamic gesture recognition. *IEEE Sensors Journal*, 21(23):27026–27036, 2021. 7

[22] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 3

[23] Mallika Garg, Pyari Mohan Pradhan, and Debashis Ghosh. Multiview hand gesture recognition using deep learning. In *2021 IEEE 18th India Council International Conference (INDICON)*, 2021. 1

[24] Mallika Garg, Debashis Ghosh, and Pyari Mohan Pradhan. Multiscaled multi-head attention-based video transformer network for hand gesture recognition. *IEEE Signal Processing Letters*, 30:80–84, 2023. 3, 4

[25] Basavaraj Hampiholi, Christian Jarvers, Wolfgang Mader, and Heiko Neumann. Convolutional transformer fusion blocks for multi-modal gesture recognition. *IEEE Access*, 11:34094–34103, 2023. 4

[26] 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. 4

[27] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 2

[28] Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. Wavemixsr: Resource-efficient neural network for image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5884–5892, 2024. 2, 3, 4

[29] Ghazaleh Khodabandelou, Pyeong-Gook Jung, Yacine Amirat, and Samer Mohammed. Attention-based gated recurrent unit for gesture recognition. *IEEE Transactions on Automation Science and Engineering*, 18(2):495–507, 2020. 1, 2

[30] Okan Kopuklu, Neslihan Kose, and Gerhard Rigoll. Motion fused frames: Data level fusion strategy for hand gesture recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2103–2111, 2018. 7

[31] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE in-*

ternational conference on automatic face & gesture recognition (FG 2019), pages 1–8. IEEE, 2019. 7

[32] Pradeep Kumar, Himaanshu Gauba, Partha Pratim Roy, and Debi Prosad Dogra. Coupled hmm-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86:1–8, 2017. 2

[33] Pradeep Kumar, Partha Pratim Roy, and Debi Prosad Dogra. Independent bayesian classifier combination based sign language recognition using facial expression. *Information Sciences*, 428:30–48, 2018. 2

[34] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021. 2, 3

[35] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[36] Y Li, CY Wu, H Fan, K Mangalam, B Xiong, J Malik, and C Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021. 3

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[38] Garg Mallika, Debashis Ghosh, and Pyari Mohan Pradhan. A two-stage convolutional neural network for hand gesture recognition. In *Proceedings of the 6th International Conference on Advance Computing and Intelligent Engineering: ICACIE 2021*, 2022. 1

[39] Fabio Manganaro, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Hand gestures for the human-car interaction: The briareo dataset. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*, pages 560–571. Springer, 2019. 2, 5, 6, 7, 8

[40] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016. 2, 5, 6, 7, 8

[41] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021. 2

[42] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 2

[43] Shruti S Phutke, Ashutosh Kulkarni, Santosh Kumar Vipparthi, and Subrahmanyam Murala. Blind image inpainting via omni-dimensional gated attention and wavelet queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1260, 2023. 4

[44] Leigh Ellen Potter, Jake Araullo, and Lewis Carter. The leap motion controller: a view on sign language. In *Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*, pages 175–178, 2013. 1

[45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 7

[46] Kyle Dylan Spurlock, Cagla Acun, Esin Saka, and Olfa Nasraoui. Chatgpt for conversational recommendation: Refining recommendations by reprompting with feedback. *arXiv preprint arXiv:2401.03605*, 2024. 2

[47] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 3

[48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 7

[49] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 3

[50] Quo Vadis, Joao Carreira, and Andrew Zisserman. Action recognition? a new model and the kinetics dataset. *Joao Carreira, Andrew Zisserman*. 4, 7

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2, 4

[52] Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid. A robust and efficient video representation for action recognition. *International journal of computer vision*, 119:219–238, 2016. 7

[53] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2

[54] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 3

[55] Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2698–2703, 2022. 2

[56] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 3

[57] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[58] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 804–811, 2014. 7

[59] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Making convolutional networks recurrent for visual sequence learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2018. 7

[60] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14230–14238, 2021. 2

[61] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 2, 3, 4

[62] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 30:5626–5640, 2021. 7

[63] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021. 3

[64] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 5

[65] Wenjin Zhang, Jiacun Wang, and Fangping Lan. Dynamic hand gesture recognition based on short-term sampling neural networks. *IEEE/CAA Journal of Automatica Sinica*, 8(1): 110–120, 2020. 7

[66] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1620, 2023. 3

[67] Ce Zheng, Matias Mendieta, and Chen Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3155, 2023. 3

[68] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022. 7

[69] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 2

[70] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022. 2

[71] Yufan Zhuang, Zihan Wang, Fangbo Tao, and Jingbo Shang. Waveformer: Linear-time attention with forward and backward wavelet transform. 2022. 2, 4