HR-STAN: High-Resolution Spatio-Temporal Attention Network for 3D Human Motion Prediction

Omar Medjaouri
The University of Texas at San Antonio
omar.medjaouri@my.utsa.edu

Kevin Desai
The University of Texas at San Antonio
kevin.desai@utsa.edu

Abstract

3D human motion prediction requires making sense of the complex spatio-temporal dynamics which underpin human motion to make highly accurate predictions. Part of this complexity is due to the trade-off between long-term (>400ms) and short-term predictions (<400ms) which require different levels of granularity to observe patterns. Several works have explored methods of improving long-term prediction performance by utilizing longer motion histories but this typically comes at the cost of very short-term (<200ms) performance. Inspired by high-resolution network architectures, we propose a novel high-resolution spatio-temporal attention network (HR-STAN) which leverages parallel feature branches and dilated convolutions to observe human motion at different scales. Furthermore, we augment this architecture with split spatial and temporal attention mechanisms to efficiently capture spatio-temporal dependencies within a given motion. We evaluate the ability of our HR-STAN architecture at incorporating long-term motion histories while producing short-term predictions and show that it improves over several state-of-the-art methods on both the AMASS and Human3.6M benchmarks.

1. Introduction

As humans one of our strongest talents is our ability to predict the future. While driving, we may see a pedestrian approaching the street or a vehicle inching forward and instinctively predict their future motion. Similarly when making a pass towards a teammate we estimate their future motion given their body language and trajectory to ensure the pass connects. Thus, 3D human motion prediction is the task of taking an individual's pose history and using it to derive meaningful predictions of their future motion. This can be represented as a multivariate sequential modeling task where the relationships between elements within the kinematic tree must be successfully understood. Furthermore, a predictive model must be able to observe and understand the underlying motion dynamics of a given pose history and use them to generate plausible future poses.

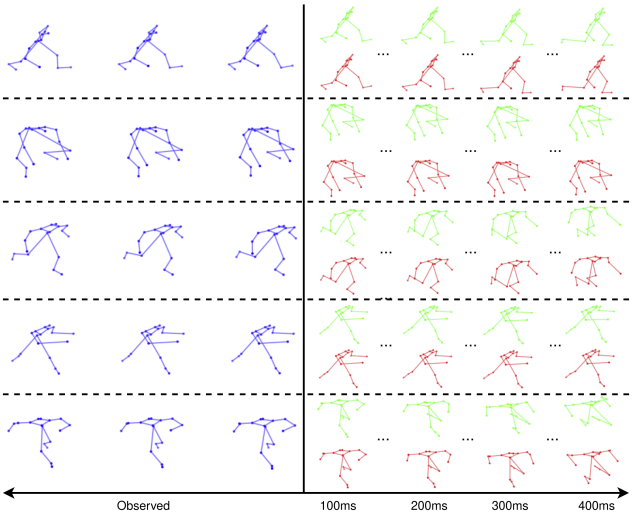

Figure 1. **AMASS Qualitative Results:** Overall performance of our HR-STAN approach compared to ground truth predictions. The featured pose histories and predictions are downsampled by 3x to better highlight the motion variety. Blue wireframes represent a pose history that is used as input into our approach while the green and red wireframes refer to ground truth future motion and predicted future motion respectively.

Following trends in sequential modeling, early works on 3D human motion prediction focused on utilizing recurrent neural networks (RNNs) to differing degrees of success [1, 5, 10, 12, 17, 24, 30]. As shown in [24], a zero-velocity baseline in which the last observed pose was used for all predictions proved to be more accurate than some of these early methods [10, 17]. These methods are typically described as auto-regressive [1] as they make predictions on a frame-by-frame basis, incorporating their own predictions with the observed pose history as they predict further out. Due to their auto-regressive design, these methods suffer from accumulated error when transitioning from making predictions entirely from observed pose history to a mixture of observed and predicted poses and often make pre-

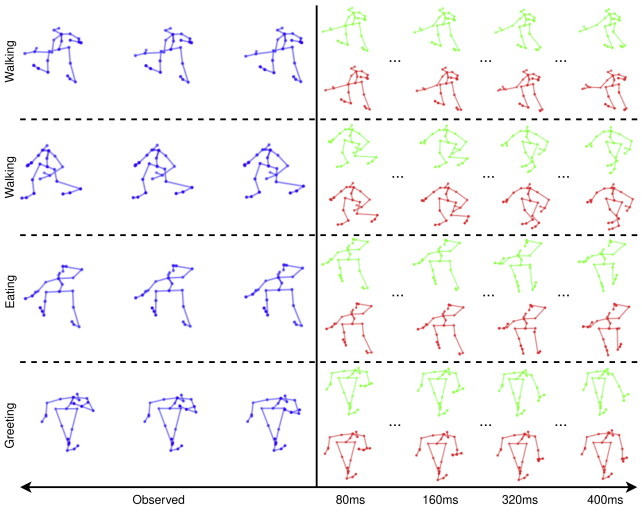

Figure 2. **Human3.6M Qualitative Results:** Overall performance of our HR-STAN approach compared to ground truth predictions. Green and red wireframes refer to ground truth future poses and predicted future poses respectively.

dictions which converge to a mean pose when conditioning on longer pose histories due to their short attention spans. Based on innovations in natural language processing, several works aimed to incorporate attention-based architectures [1, 5] to improve the predictions of auto-regressive methods. These attention-based architectures marginally improved performance in the short-term but substantially improved the ability for auto-regressive models to predict into the long-term without converging to a mean pose.

Alternatives to these recurrent models have also been explored [18, 22, 23, 36], which rather than producing predictions on a frame-by-frame basis produce fixed-length predictions using a fixed-length history. This style of approach has the advantage of producing the entire prediction from an observed set of poses, thus removing the potential for accumulated error. These approaches excel where pose histories are short and models can primarily focus on small movements but have trouble utilizing long-term motion histories [23, 36]. This is partially due to the fundamental limitation of convolution-based neural networks, where the performance is highly dependent on the overall size of their receptive field. Furthermore, many convolution-based networks compress their inputs into a latent space which destroys useful fine-grain details from the input. High-Resolution Networks (HR-Nets) [37, 38] have been proposed which maintain a high-resolution feature branch along with low-resolution branches, thus preserving fine-grain details while allowing the network to observe large-scale features simultaneously. In our approach, we utilize this style of convolution-based architecture to observe the dynamics of long-term motion as well as the fine-grain movements to make accurate short-term human motion predictions.

1.1. Proposed Approach

In this work, we propose a method for 3d human motion prediction, which leverages a high-resolution spatio-temporal attention network (HR-STAN) architecture to produce highly accurate short-term predictions. Our method encodes a given fixed-length pose history as a sequence of 1D pose vectors and maps directly to a fixed-length pose prediction sequence, thus removing the need for a pose encoding and decoding step [23, 36]. We demonstrate the state-of-the-art performance of this method using standard 3D human motion prediction benchmarks such as AMASS [21] and Human3.6M [16], comparing against several auto-regressive and fixed-length methods.

Our proposed HR-STAN builds upon the original high-resolution network [37] in several key ways. First, by decomposing convolutions into spatial and temporal components, our approach is able to more efficiently model spatio-temporal relationships throughout a given pose history. Second, rather than using strided convolutions in subsequent branches of the network, we utilize dilated convolutions to increase receptive field without feature compression. Finally, we introduce split spatial and temporal attention which more efficiently encourages the network to focus on the spatio-temporal relationships of a particular motion. The above contributions led to a novel high-resolution spatio-temporal attention network (HR-STAN) which achieves state-of-the-art performance on the 3D human motion prediction task on multiple benchmarks. We demonstrate the qualitative performance of this approach on the AMASS [21] and Human3.6M [16] datasets in Figure 1 and Figure 2 respectively, showing that our method generates accurate predictions of future human motion.

2. Related Work

2.1. Sequential Modeling

Traditionally, recurrent neural networks (RNNs) have dominated sequential modeling tasks and have achieved state of the art results on several tasks such as machine translation [4, 6, 7], language modeling [8, 25, 26], and even human motion prediction [10, 17, 24]. This is largely due to their ability to encode the history of a sequence and condition future predictions over the length of that history. However, in recent years their popularly has waned due to several drawbacks, most notable the exploding/vanishing gradient problem and their short attention spans [19]. Several works propose solutions for these problems such as incorporating same-layer neuron independence to allow for longer and deeper networks [19] or additional regularization strategies which aid in stability during training [25]. However, these works do not resolve some of the fundamental constraints of recurrent neural networks.

To address these constraints, several alternative architec-

tures have been proposed that have greater training stability and attention spans, such as the temporal convolutional network (TCN) and attention-based networks. Gehring et al. [11] introduced a novel architecture which used a TCN for language modeling. Bai et al. [3] explored the performance gap between recurrent and convolution-based models, demonstrating that TCNs were competitive with RNNs on many sequential modeling tasks due to their ability to capture long-term context. Attention-based models have also been proposed which have shown promising results on language modeling [2] and machine translation [35]. Specifically, [35] demonstrated that self-attention modules are able to capture complex temporal dependencies which allow them to achieve state-of-the-art performance. Finally, further works [13, 40] have bridged TCNs and attention-based networks together to leverage the receptive field of a convolution-based network with the flexibility of attention.

2.2. Human Motion Prediction

As seen in the more general field of sequential modeling, early deep learning approaches to human motion prediction heavily utilized recurrent neural networks [10, 17, 24, 30]. These methods are described as auto-regressive as they make predictions on a frame-by-frame basis rather than making predictions all at once. ERD [10] investigated simultaneous human pose estimation and prediction using a learned embedding of 3D human poses but found their approach was sensitive to hyperparameter tuning and had difficulty extrapolating long-term motions. Using a similar approach, Structural-RNN [17] proposed representing graph-based structural relationships using RNNs with nodes and edges being represented using LSTMs. However, Martinez et al. [24] demonstrated several issues inherent to these early recurrent methods. First, they found that a simple zero-velocity base line in which the last known pose is used as a prediction performed better on benchmarks than ERD or Structural-RNN. Futhermore, these RNN-based methods also suffer from first-frame discontinuities where they transition from conditioning predictions on ground truth data to conditioning on their own initial predictions [24]. To overcome these drawbacks, [24] proposed conditioning on frame-by-frame motion rather than positions directly and incorporating a sampling-based loss which trained the network using its own predictions as well as ground truth. These additions improved performance overall but the network had trouble with stationary motions. Similarly, Pavllo et al [30] investigated incorporating a forward kinematics-based loss to penalize the network for errors accumulated along the kinematic chain but did not make significant improvements to prediction accuracy.

In addition to recurrent architectures, several works proposed other types of auto-regressive prediction models which used attention to focus on spatio-temporal dependencies over a given pose history. [5] leveraged a transformer architecture akin to [35] along with a progressive decoding strategy which first predicted the future locations of central joints and then expanded outwards to peripheral joints. While this strategy improved performance, it suffered when errors in base joints were propagated up the kinematic chain towards peripheral joints. Spatio-Temporal Transformer [1] expanded on the original transformer architecture by decomposing the attention mechanism into spatial and temporal components, reducing the size of the model and improving overall performance. This approach greatly improved on the ability for auto-regressive models to make long-term predictions but did not improve short-term prediction performance by a significant margin.

Alternatives to recurrent or auto-regressive methods have been explored, with several networks using feed-forward models [18, 22, 23]. These models use fixed-length input sequences and produce fixed-length predictions to alleviate the effect of accumulated error seen in auto-regressive methods. ConvSeq2Seq [18] utilized a long-term motion encoder and a short-term motion encoder/decoder to capture long-term motion features while preserving short-term fidelity, but was sensitive to convolution filter size. Mao and Liu [23] first proposed LTD which first encoded the motion using the discrete cosine transform (DCT) and then used a graph-based convolutional network (GCN) with learned connectivity to capture spatial dynamics. Further extending on this work, Mao and Liu [22] utilized motion attention to improve prediction accuracy on periodic actions. This approach proved beneficial towards long-term predictions, but did not have large impact on short-term predictions.

3. Method

We begin the explanation of our approach with a formal description of the problem of 3D human motion prediction. Let $P = \{p_1, p_2, ..., p_n\}$ represent a full description of a given human pose which is described using N joint positions $p_i \in \mathbb{R}^3$. Similarly, we can describe the segments which make up a given kinematic tree as $S = \{s_{1,2}, s_{2,3}, ..., s_{i,j}\}$ where a given segment is defined as $s_{i,j} = p_j - p_i$. Given a pose history $M = \{P_1, P_2, ..., P_t\}$ of length T where $t \in \mathbb{R}$ represents the individual pose in the sequence, the goal is to predict the future poses $M' = \{P_{t+1}, P_{t+2}, ..., P_{t+l}\}$ where $l \in \mathbb{R}$ represents the predicted pose l frames into the future.

3.1. Network Architecture

In order to make predictions using a given pose history M, we must first determine an encoding of M such that we can extract useful features from it using our network. Previous works have used a number of different methods to first encode individual poses such as the discrete cosine transform (DCT) [5,22,23,36] or a learned embedding [1,18,24].

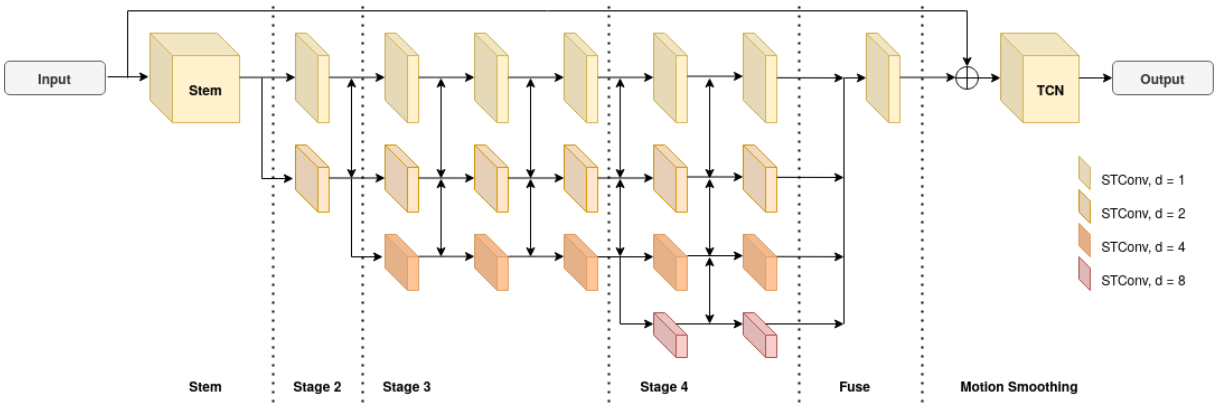

Figure 3. **HR-STAN Architecture Overview:** The input pose history is first mapped to 2D and processed via the stem to obtain initial motion features. The features are further processed by several novel spatial and temporal convolution modules arranged in a multi-stage high-resolution spatio-temporal architecture based on [37]. Different branches of the network operate at different dilation levels and several transformation layers combine features from different branches to encourage cross-talk between branches. Finally, the features of each branch are fused and the raw predictions are smoothed using a multi-layer temporal convolutional network (TCN) which produces the final motion predictions.

Due to the structure of our network, we are able to process a pose history directly by representing it as a sequence of pose vectors of size $3N$ such that the input to our network is a 2D matrix of size $(3N, T)$. By extension our network also produces predictions in the form of $(3N, L)$ and as such is directly able to map from pose history to future poses.

Following the basic high resolution network architecture [37], our HR-STAN is structured in stages which progressively expand the receptive field of the network, enabling it to observe increasingly larger motion patterns. Figure 3 describes the architecture overview of the network. Each stage consists of parallel branches which operate at different dilation levels and each branch consists of several identical spatio-temporal convolution (STConv) modules which act as the main engines of the network. Let $F_{i,j}$ refer to a specific STConv module in the network such that it is the j^{th} module in a given branch and it operates at the dilation level $i \in \{1, 2, 4, 8\}$. Furthermore, let $h_{i,i'}$ refer to the convolution modules which serve as transforms between branches i and i'. These modules fuse features from parallel branches using stacked 3x3 and 1x1 convolutions. We can describe the output of a specific branch at a given stage as:

$$f_i' = h_{1,i}(F_1(f_1)) + h_{2,i}(F_2(f_2)) + ... + F_i(f_i) \quad (1)$$

where f_i and f_i' refer the input and output features of a given branch. Each stage can be stacked on itself an arbitrary number of times to increase the depth of the network. After several stages, the output of each branch is fused together in a fusion layer that provides an initial raw prediction. The fusion layer is of a similar structure to previous stages in the network; however, the additional dilation branches are no longer propagated. The output of the fusion layer can be described with equation (2). The initial prediction is then combined with the last observed pose in the pose history to obtain the raw prediction (3).

$$f' = F_1(f_1) + h_{2,1}(F_2(f_2)) + ... + h_{i,1}(F_i(f_i)) \quad (2)$$

$$M' = f' + M_T \quad (3)$$

The raw predictions are then concatenated with the pose history and smoothed using a final motion smoothing stage. This stage consists of a positional encoder [35], several temporal convolution layers, and spatio-temporal attention modules to produce the final prediction.

3.2. Spatio-Temporal Convolution Module

The core of the network is the spatio-temporal convolution (STConv) module which consists of separate spatial and temporal convolution branches. They are further processed using split spatio-temporal attention modules before being fused together. While the initial assumption would be that processing both spatial and temporal dimensions simultaneously would allow a network to exploit spatio-temporal relationships better, previous works have found that decomposing convolutions into their respective dimensions has proved beneficial [33]. To this end, the spatial and temporal branches of the STConv module are augmented with separate spatial and temporal attention mechanisms which modulate the output of each branch based on the importance given to spatio-temporal features. Figure 4 gives an outline of an STConv module, which is formally defined as below.

Given a feature vector f of shape (D, T) where D refers to the spatial dimension and T refers to the temporal dimension, the STConv module applies separate 1D dilated convolutions [39] with kernel size k and dilation d along

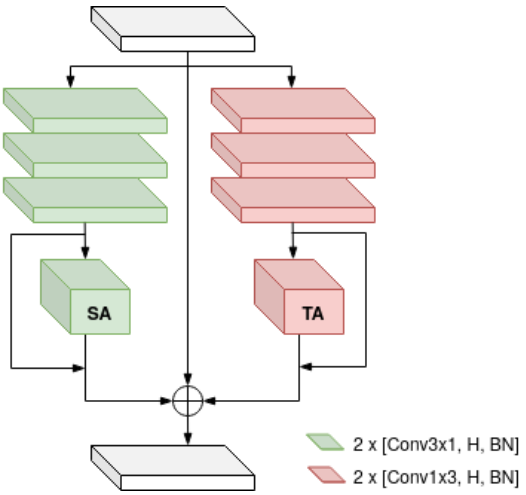

Figure 4. **STConv Module:** Each STConv module consists of separate *spatial* and *temporal* components which are fused with a residual connection to ensure stable gradient flow. Both branches feature three consecutive blocks of 1D convolutions with Hardswish activation [14] and Batch Normalization. Each branch also uses a *spatial* or *temporal* attention mechanism to direct network attention towards important spatio-temporal motion features.

the spatial and temporal axes to generate separate features as described in equations (4) and (5).

$$f_s[x,y] = \sum_{i=0}^{k-1} W[i] * f[x + d * i, y] \tag{4}$$

$$f_t[x,y] = \sum_{i=0}^{k-1} W[i] * f[x, y + d * i] \tag{5}$$

The output features of these convolution layers are normalized while training using a standard batch normalization process [15] and passed through the Hardswish activation function [14] before being processed by the spatial and temporal attention modules.

3.2.1 Efficient Spatio-Temporal Attention

Following [35] a lot of focus has been given to dot-product attention and its ability to capture complex relationships between features. However, one large drawback of this type of attention is its memory and computational complexity. As shown in [32], given the standard dot-product attention equation [35] with queries $\mathbf{Q} \in \mathbb{R}^{n*d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n*d_k}$, and values $\mathbf{V} \in \mathbb{R}^{n*d_k}$, the memory complexity is $\mathcal{O}(n^2)$ and the computational complexity is $\mathcal{O}(d_k n^2)$. While this is not necessarily a problem for smaller networks with fewer stacked attention modules, these complexities pose a significant problem for our proposed approach. However, [32] proposes an alternative attention mechanism with linear complexities, $\mathcal{O}(dn + d^2)$ memory complexity and $\mathcal{O}(dn^2)$ runtime complexity ,that is a better option for our approach due to the number of layers.

$$\mathbf{E}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\mathbf{Q}}{\sqrt{\mathbf{n}}}(\frac{\mathbf{K^T}}{\sqrt{\mathbf{n}}}\mathbf{V}) \tag{6}$$

Given this formulation of attention, we define our spatial and temporal attention using the following variables. For the spatial branch, we define our queries, keys, and values as $\mathbf{Q}_s \in \mathbb{R}^{D*T}$, $\mathbf{K} \in \mathbb{R}^{D*T}$, and $\mathbf{V} \in \mathbb{R}^{D*T}$. Conversely within the temporal attention branch we define them as $\mathbf{Q}_s \in \mathbb{R}^{T*D}$, $\mathbf{K} \in \mathbb{R}^{T*D}$, and $\mathbf{V} \in \mathbb{R}^{T*D}$. Note that in either case, the alternate dimension serves as the dimensionality of the feature space such that the temporal component of a specific joint serves as its feature space. Using the efficient attention mechanism described in Equation 6, we construct spatial and temporal attention blocks which compute attention vectors and combine them with features f_s and f_t using a residual mechanism. Finally, the separate spatial and temporal branches are combined with a residual to produce the output of the STConv module using the equation:

$$f' = f_s + f_t + f \tag{7}$$

3.3. Loss Functions

Our HR-STAN model is trained using a loss function consisting of two major terms. The first is based on the standard Mean Per Joint Position Error (MPJPE) proposed in [16] to measure the error between the predicted and ground truth joint positions and is described as (8):

$$L_{mpjpe} = \frac{1}{LP} \sum_{l}^{L} \sum_{p}^{P} \|p_{t+l} - \hat{p}_{t+l}\|_2 \tag{8}$$

where $p_{t+l}, \hat{p}_{t+l} \in P$ represent the 3D positions of both ground truth and predicted joints for the pose P on the l^{th} frame in predicted sequence M'. The second term acts as a regularization term which penalizes the network for predicting unnatural segment angles by computing the cosine similarity of each of the vectors which make up the human pose. The similarity regularization term can be described as (9):

$$L_{cs} = \frac{1}{S} \sum_{s}^{S} \frac{s \cdot \hat{s}}{\max(\|s\|_2 * \|\hat{s}\|_2, \epsilon)} \tag{9}$$

where $s, \hat{s} \in S$ refer to each associated joint segment in both the ground truth and predicted pose sequence respectively. With these terms defined, the final loss function is denoted as:

$$L = L_{mpjpe} + \lambda L_{cs} \tag{10}$$

where λ represents a configurable hyperparameter which affects the relative weighting of L_{cs} with L_{mpjpe}.

4. Experimental Setup

To benchmark against previous architectures [1, 18, 22–24, 30], we evaluate the short term performance of HR-STAN on both the Human 3.6M dataset [16] and the AMASS dataset [21] using standard protocols. While previous works have represented poses using both 3D joint positions and several angle-based representations, our approach targets the 3D joint position representation and thus we compare directly against previous works which report their results in this format. Following [1, 5, 18, 22, 23] we evaluate the performance of our approach using the MPJPE [16] metric on several predefined key-frames.

4.1. Datasets

AMASS [21]: this dataset consists of different motion capture datasets such as CMU [9], BioMotion Lab [34], and MPI-HDM05 [27, 28] which have been parameterized using the SMPL [20, 31] body mesh model. Using the SMPL model, we convert the captured sequences from the dense mesh SMPL representation to a sparse 3D joint position representation consisting of 22 body joints. Following previous work [22], we also remove the 4 static joints and obtain an 18-joint pose representation. AMASS [21] consists of captures at 60Hz and we process the data at this original rate instead of downsampling to 25Hz as with Human3.6M [16]. However, we maintain the same overall time duration, i.e., making 24 frame predictions rather than 10. Finally, we use the same training, validation, and test splits as [1, 18, 22, 23] to provide a fair performance comparison.

Human3.6M While Human3.6M has long been a standard benchmark for many tasks including 3D human motion prediction, it has been argued that it should be given less importance given its small size and limited variability in subjects [1]. However, because it has actions separated into categories it provides valuable insight into model performance and thus we report results using this benchmark as well as AMASS [21] It consists of seven actors performing 15 different action sequences such as smoking, sitting, and walking and each pose is represented using a set of 32 joints. Following previous work [1, 18, 22, 23] we split reserve subject 11 and 5 as our validation and test sets respectively. Additionally, we downsample the original 50 Hz sequences to 25 Hz to better compare against previous methods. While [18, 23] evaluate the performance of their approaches on 8 random sub-sequences per action from the test set, [22] argues that this leads to high variance in reported performance and thus evaluates using 256 sub-sequences instead. To compare against these works we report our results on 256 random sub-sequences per action.

4.2. Implementation Details

Our models were implemented using the PyTorch [29] library and trained as well as evaluated using an NVIDIA RTX 2080. Due to the difference in frame-rate between both datasets, we utilize a different network architecture when evaluating performance on AMASS versus Human3.6M. The version trained and evaluated on AMASS is depicted in Figure 3 while the version trained on Human3.6M forgoes the Stage 4 module entirely and only uses 3 stacked Stage 3 modules. The final temporal smoothing module consists of a positional encoder based on [35], several dilated temporal convolutions, and the split spatio-temporal attention modules described in Section 3.2.1. For all versions of the HR-STAN we used an initial learning rate of 1e-3 along with an AdamW optimizer. For the versions trained specifically on Human3.6M, we used a step-wise learning rate scheduler with a gamma of 0.5 which was applied every 50 epochs for 400 epochs. For the versions trained on AMASS we applied the same gamma every 25 epochs and only trained for 300 epochs due to the size of the dataset in comparison to Human3.6M. All versions of the model utilized a λ of 0.1 for the cosine similarity loss descried in Equation 10.

5. Results

Whenever possible, we relied on reported results for other methods in our experiments; thus, we compare against different methods when evaluating the performance of our proposed approach on AMASS [21] and Human 3.6M [16]. More specifically, [22] adapted previously released code for Res. Sup [24] and convSeq2Seq [18] to produce 3D position metrics. The authors also reported results for their proposed method [22] and previous method [23] which are used as a comparison to our method. For results on AMASS [21], we relied on reported results of [1], which reports results for Res. Sup [24], QuaterNet [30], convSeq2Seq [18], LTD [23], and LTD Attention [22].

We evaluate our network architecture with several permutations to gain a better understanding of it's benefits. When reporting results, we use the nomenclature HR-STAN-X to indicate the length of the pose history the model uses. On the Human3.6M dataset we evaluated the performance of our method using pose histories of length 10, 25, and 50. Similarly, for the AMASS dataset we used pose histories of length 30, 60, and 90, after accounting for the difference in sampling rates of the two datasets.

5.1. AMASS

Table 1 presents the results of our method on the AMASS dataset compared to several previous state-of-the-art methods. Overall our method improves prediction accuracy across the entire prediction length and does so with a shorter pose history than ST-Transformer [1], which uses 120 frames. Through evaluating different versions of the model trained with different pose history lengths (30, 60, 90 frames), we found that increasing the length of the pose

Method	MPJPE			
	100ms	200ms	300ms	400ms
Zero-Velocity [24]	23.6	39.6	53.1	64.2
convSeq2Seq [18]	23.6	37.4	48.6	57.9
QuaterNet [30]	16.7	28.7	39.6	49.0
LTD [23]	15.6	25.7	35.3	44.1
LTD Attention [22]	13.4	23.2	32.1	39.7
ST-Transformer [1]	12.8	21.8	30.9	39.5
HR-STAN-30	10.7	**19.5**	27.9	35.8
HR-STAN-60	**10.6**	**19.5**	**27.7**	**35.3**
HR-STAN-90	10.9	19.8	28.0	35.7

Table 1. **Results on AMASS:** Results are reported using the test set (*BMLrub*) and error is represented in millimeters. Overall, our approach is better across the entire prediction length.

history did not have a significant impact on network performance. This indicates that short term predictions are largely dominated by features which appear in the final second of the pose history and that increasing pose history length does little to improve short term predictions while increasing computational costs. Using the shortest pose history our HR-STAN approach improved upon state of the art performance while our best version, HR-STAN-60, which utilized a pose history of 60 frames or 1 second yielded an 11% reduction in sum of the MPJPE over the entire prediction length when compared to the leading alternative approach. A sample of the qualitative results are presented in Figure 1 which demonstrate that our HR-STAN predicts realistic poses while predicting 3D joint positions directly.

5.2. Human 3.6M

Table 2 highlights the average performance of our HR-STAN compared to an auto-regressive method (Res. Sup. [24] and several fixed-length methods (convSeq2Seq, LTD, LTD Attention) [18, 22, 23]. Through our evaluation of different input pose history lengths, we found that HR-STAN-10 produced the best results with HR-STAN-25 still performing better than previous state of the art methods. Using the average frame-wise MPJPE on the Human3.6M test set, HR-STAN-10 performed 27.6% better on very short term ($<$ 200ms) and 11.3% better on ($<$ 400ms) predictions than the best alternative method.

Method	MPJPE			
	100ms	200ms	300ms	400ms
Res. Sup. [24]	25.0	46.2	77.0	88.3
convSeq2Seq [18]	16.6	33.3	61.4	72.7
LTD [23]	11.2	23.4	47.9	58.9
LTD Attention [22]	10.4	22.6	47.1	58.3
HR-STAN-10	**6.9**	**17.0**	**42.5**	**56.4**
HR-STAN-25	7.5	17.9	43.8	57.7
HR-STAN-50	8.2	19.4	46.8	61.6

Table 2. **Results on Human3.6M:** Results are reported using the test set and error is represented in millimeters. Overall our approach performs better across the entire prediction length.

As mentioned in Section 4.1, the Human3.6M [16] dataset consists of several different action sequences with some being highly periodic such as walking or eating and others being aperiodic such as posing or greeting. To further investigate the performance of our method on different action types, we captured the frame-wise MPJPE of our best performing model (HR-STAN-10) over each action type and present them in Table 3. Notably there are a few action types in which our performance lags behind LTD Attention [22] at the upper end of the prediction length such as Walking, Eating, Greeting, and Walking Together. These actions feature a higher degree of periodicity and as such the longer pose history that LTD Attention uses is very beneficial to predicting future motions. Additionally, both LTD [23] and LTD Attention [22] rely on first transforming the pose history into the frequency domain using the DCT and thus periodic features are naturally easier to capture than aperiodic features. For largely aperiodic actions such as Posing or Purchases our HR-STAN greatly outperforms previous work, indicating that our method is better able to handle abrupt changes in motion.

Finally, we also explore the qualitative performance on different Human3.6M action types, depicted in Figure 2. Despite the large variety in motion types, our method is able to accurately predict future motion.

5.3. Ablation Study

To further investigate the utility of several contributions of this work, we performed an ablation study using the AMASS dataset. For the purposes of the study, we used our HR-STAN-30 model as the base and measured the performance of different configurations on the AMASS [21] dataset as well as the size of each model in terms of number of parameters. Specifically we evaluated the impact of the split spatial and temporal convolutions vs a combined 3x3 convolution block baseline and the impact of our spatio-temporal attention modules. We present the frame-wise MPJPE and number of parameters of each configuration in Table 4 While the overall performance of the combined architecture is better than the split version by a small amount without attention modules active, the spatio-temporal split version is able to produce comparable performance with 57% of the parameters. Additionally, the attention modules also improve the overall mode performance with only a modest increase in the number of parameters. Interestingly, while the attention module has a large impact on accuracy when incorporated into the base split model, it mildly reduced performance in both the baseline model and the expanded split model. This may indicate that while the attention module enables a network to focus on important elements in the pose history, it's impact is minor when used in a network which already has the capacity to incorporate the entire pose history.

	Walking				Eating				Smoking				Discussion				Sitting			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res. Sup. [24]	23.2	40.9	61.0	66.1	16.8	31.5	53.5	61.7	18.9	34.7	57.5	65.4	25.7	47.8	80.0	91.3	23.8	44.7	78.0	91.2
convSeq2Seq [18]	17.7	33.5	56.3	63.6	11.0	22.4	40.7	48.4	11.6	22.8	41.3	48.9	17.1	34.5	64.8	77.6	13.5	27.0	52.0	63.1
LTD [23]	11.1	21.4	37.3	42.9	7.0	14.8	29.8	37.3	7.5	15.5	30.7	37.5	10.8	24.0	52.7	65.8	9.8	20.5	44.2	55.9
LTD Attention [22]	10.0	19.5	**34.2**	**39.8**	6.4	14.0	**28.7**	**36.2**	7.0	14.9	29.9	36.4	10.2	23.4	52.1	65.4	9.3	20.1	44.3	56.0
HR-STAN-10	**7.5**	**17.0**	39.4	51.9	**5.1**	**12.1**	29.2	38.5	**4.5**	**10.4**	**24.7**	**33.0**	**7.1**	**18.1**	**46.4**	**61.9**	**5.7**	**13.7**	**33.6**	**44.2**

	Directions				Greeting				Phoning				Posing				Purchases			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res. Sup. [24]	21.6	41.3	72.1	84.1	31.2	58.4	96.3	108.8	21.1	38.9	66.0	76.4	29.3	56.1	98.3	114.3	28.7	52.4	86.9	100.7
convSeq2Seq [18]	13.5	29.0	57.6	69.7	22.0	45.0	82.0	96.0	13.5	26.6	49.9	59.9	16.9	36.7	75.7	92.9	20.3	41.8	76.5	89.9
LTD [23]	8.0	18.8	43.7	54.9	14.8	31.4	65.3	79.7	9.3	19.1	39.8	49.7	10.9	25.1	59.1	75.9	13.9	30.3	62.2	75.9
LTD Attention [22]	7.4	18.4	44.5	56.5	13.7	30.1	**63.8**	**78.1**	8.6	18.3	39.0	49.2	10.2	24.2	58.5	75.8	**13.0**	29.2	60.4	73.9
HR-STAN-10	**5.4**	**13.9**	**36.7**	**49.1**	**10.3**	**26.2**	65.0	83.9	**6.1**	**14.5**	**35.0**	**46.2**	**7.1**	**18.6**	**49.9**	**67.2**	**8.4**	**21.3**	**53.4**	**69.4**

	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res. Sup. [24]	31.7	58.3	96.7	112.0	21.9	41.4	74.0	87.6	23.8	44.2	75.8	87.7	36.4	64.8	99.1	110.6	20.4	37.1	59.4	67.3
convSeq2Seq [18]	20.7	40.6	70.4	82.7	12.7	26.0	52.1	63.6	14.6	29.7	58.1	69.7	27.7	53.6	90.7	103.3	15.3	30.4	53.1	61.2
LTD [23]	15.6	31.4	59.1	71.7	8.9	18.9	41.0	51.7	9.2	19.5	43.3	54.4	20.9	40.7	73.6	86.6	9.6	19.4	36.5	44.0
LTD Attention [22]	14.9	30.7	59.1	72.0	8.3	18.4	40.7	51.5	8.7	19.2	43.4	54.9	20.1	40.3	73.3	**86.3**	8.9	18.4	35.1	**41.9**
HR-STAN-10	**7.4**	**17.5**	**43.5**	**58.1**	**5.0**	**12.7**	**35.0**	**48.6**	**6.6**	**15.9**	**39.7**	**53.0**	**12.2**	**30.4**	**73.0**	94.5	**5.9**	**13.6**	**33.3**	45.8

Table 3. **Results on Human3.6M:** Comparison between different methods across the different action types included in the Human3.6M test set. We use our best performing network configuration HR-STAN-10 which utilizes pose histories of 10 frames to make predictions. Our method greatly outperforms previous methods on aperiodic actions and on very short term <160ms predictions. **Bold** indicates the best results for that specific action frame. Underline indicates second best results.

Model		**Parameters**	**MPJPE**			
Split	*Attn*		100ms	200ms	300ms	400ms
x	x	7.38M	10.5	19.5	28.0	36.6
x	✓	7.69M	11.0	20.1	28.6	36.7
✓	x	4.21M	10.9	20.3	29.1	38.1
✓	x	7.42M	10.5	19.4	27.7	36.1
✓	✓	4.52M	10.7	19.5	27.9	35.8
✓	✓	7.73M	11.0	20.2	28.6	36.4

Table 4. **Ablation Study on AMASS:** Results are reported for several different network configurations trained and evaluated on AMASS train/test sets. *Split* refers to network configurations with split spatial and temporal convolutions in the STConv module rather than combined spatio-temporal convolutions. *Attn* refers to whether or not the STConv module is using the proposed split attention mechanism. Parameter numbers are in units of millions.

6. Limitations and Future Work

While the method presented in this paper was able to produce state-of-the-art results on multiple benchmarks, there are both limitations with this work and potential future research opportunities. First and foremost, this work mainly explored the performance of our method on short-term predictions but further work could explore the impact of this approach on long-term predictions. Futhermore, as seen in the results on the Human3.6M [16] benchmark, our approach generally performs better on aperiodic actions than periodic ones. One of the advantages of the approach described in Mao et al. [22,23] is that it inherently captures the periodicity of motions using DCT but this comes at the cost of performance on aperiodic actions. A future work could explore a combination of both methods which is able to capture periodic motion features while producing accurate predictions on aperiodic actions. Finally, while we decided to use a 2D representation for pose histories, it may not be the most efficient representation and several works [22,23] have successfully utilized graph representations. Further adapting the multi-branch architecture to a graph convolution-based network could allow for high prediction accuracy while drastically reducing the overall network size. Finally, our ablation study highlights an interesting result in which the impact of the spatio-temporal attention modules are minor when applied to larger models with a fixed pose history length of 30 frames. Further work could investigate the impact of the attention on different lengths of pose history as this may provide better analysis of its impact.

7. Conclusion

We proposed a novel *high-resolution spatio-temporal attention network (HR-STAN)* for 3D human motion prediction which stacks spatio-temporally separable dilated convolutions to observe multi-scale motion features. By maintaining branches at multiple spatio-temporal scales and incorporating separate spatial and temporal self-attention, the model is better able to capture long and short-term motion features and make accurate short-term predictions. Comparison of our approach with previous state-of-the-art auto-regressive and fixed-length methods shows 11% reduction in average MPJPE over the entire prediction on the AMASS dataset. For Human3.6M, it shows 11.3% reduction over the entire prediction and 27% reduction on very short-term (<160ms). Furthermore, our ablation study demonstrates the impact of the proposed modules in HR-STAN, showing an improvement in performance over the baseline using just 57% of the parameters.

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *2021 International Conference on 3D Vision (3DV)*, Dec 2021. 1, 2, 3, 6, 7

[2] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3159–3166, Jul 2019. 3

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018. 3

[4] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. 2

[5] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, Ding Liu, Jing Liu, and Nadia Magnenat Thalmann. Learning progressive joint propagation for human motion prediction. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII*, page 226–242, Berlin, Heidelberg, 2020. Springer-Verlag. 1, 2, 3, 6

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. 2

[8] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2

[9] Adam Bargteil Xavier Martin Justin Macey Alex Collado Fernando de la Torre, Jessica Hodgins and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, April 2008. 6

[10] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4346–4354, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. 1, 2, 3

[11] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning, 2017. 3

[12] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and Jose M. F. Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[13] Hongyan Hao, Yan Wang, Yudi Xia, Jian Zhao, and Furao Shen. Temporal convolutional attention-based network for sequence modeling. *CoRR*, abs/2002.12530, 2020. 3

[14] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 5

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 5

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339, 2014. 2, 5, 6, 7, 8

[17] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 1, 2, 3

[18] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018. 2, 3, 6, 7, 8

[19] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 6

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 2, 6, 7

[22] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. *Lecture Notes in Computer Science*, page 474–489, 2020. 2, 3, 6, 7, 8

[23] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2, 3, 6, 7, 8

[24] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4674–4683, 2017. 1, 2, 3, 6, 7, 8

[25] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models, 2017. 2

[26] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*, 2018. 2

[27] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. Efficient and robust annotation of motion capture data. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '09, page 17–26, New York, NY, USA, 2009. Association for Computing Machinery. 6

[28] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database hdm05. 06 2007. 6

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 6

[30] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion, 2018. 1, 3, 6, 7

[31] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands. *ACM Transactions on Graphics*, 36(6):1–17, Nov 2017. 6

[32] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities, 2018. 5

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 4

[34] Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, 09 2002. 6

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3, 4, 5, 6

[36] Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct 2021. 2, 3

[37] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, Oct 2021. 2, 4

[38] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. 2

[39] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. 4

[40] Jinliang Zang, Le Wang, Ziyi Liu, Qilin Zhang, Gang Hua, and Nanning Zheng. Attention-based temporal weighted convolutional neural network for action recognition. *Artificial Intelligence Applications and Innovations*, page 97–108, 2018. 3