Master's Thesis

Transformer-Based Head Motion Prediction Algorithm Using Image Generation Model

Hyogeun Byun

The Graduate School

Sungkyunkwan University

Department of Computer Science and Engineering

Master's Thesis

Transformer-Based Head Motion Prediction Algorithm Using Image Generation Model

Hyogeun Byun

The Graduate School

Sungkyunkwan University

Department of Computer Science and Engineering

Transformer-Based Head Motion Prediction Algorithm Using Image Generation Model

Hyogeun Byun

A Master's Thesis Submitted to the Department of Computer Science and Engineering and the Graduate School of Sungkyunkwan University in partial fulfillment of the requirements for the degree of Master of Science in Engineering

April 2024

Supervised by

Sungkil Lee

Major Advisor

This certifies that the Master's Thesis
of Hyogeun Byun is approved.

심사위원장 signature

Committee Chair : Jaepil Heo

심사위원 signature

Committee Member : Jaemin Jo

지도교수 signature

Major Advisor : Sungkil Lee

The Graduate School
Sungkyunkwan University
June 2024

Table of Contents

List of Tables

List of Figures

Abstract

Transformer-Based Head Motion Prediction Algorithm Using Image Generation Model

In head-mounted display-based virtual reality, motion-to-photon latency refers to the time taken by the system to reflect sensor and user inputs in the rendering process. Prolonged latencies can lead to discrepancies between user movements and visual output, resulting in discomfort such as cyber sickness, thereby disrupting user immersion.

To minimize such delays, traditional virtual reality systems have employed methods like analyzing trends in head motion data or using deep learning models like Recurrent Neural Networks to learn the temporal characteristics of head motion data and predict subsequent movements. These predictions help prepare the rendering process, reducing latency. However, Recurrent Neural Networks suffer from issues with long-term dependencies, tending to forget past information over time, which can degrade prediction accuracy and limit parallel processing capabilities.

This paper proposes a model that uses a transformer-based head motion prediction model to address the long-term dependency issues and parallel processing limitations of Recurrent Neural Networks. The model proposed in this paper processes image frame data through a vision transformer-based model to extract image features focused on areas of interest within the image, assisting

in head motion prediction. During the decoding process, this model employs image generation techniques and demonstrates high scalability by utilizing deep learning models from natural language processing for predictive modeling. By leveraging additional data, the proposed model shows improved accuracy in predicting user head motions compared to existing RNN models.

Keywords: cyber sickness, virtual reality, rendering, AI

1. INTRODUCTION

The metaverse is an innovative concept that goes beyond mere extensions of virtual reality or augmented reality. It represents a unified virtual space where digitally enhanced physical reality converges with digital reality, creating an immersive experience of a new dimension. This concept focuses on digitizing the real world in unprecedented ways and adding elements of physical reality to the digital world, providing a shared virtual space for full immersion. The ultimate goal of the metaverse is to blur the line between reality and the virtual, allowing users to transcend the limitations of the physical world to explore and interact within this new digital ecosystem.

The metaverse promises to integrate reality and virtuality to create entirely new worlds, offering innovative experiences across diverse domains such as gaming, education, work, and social interaction.

In gaming, players can transcend merely watching a screen and enter the virtual world directly to engage in various activities. Leveraging virtual reality technology, these virtual worlds allow players to explore adventures and interactions that are impossible in the physical world while enjoying their own avatars with no physical limitations.

In education, virtual classrooms and laboratories enable students to have

learning experiences that closely mimic reality. The metaverse leverages virtual reality to allow students to move beyond traditional textbooks and flat screens to learn and experience in real-time within virtual environments. This transforms educational content into something more engaging, interactive, and practical.

In the realm of work, the metaverse offers virtual meeting rooms and collaboration spaces, enabling employees to work together beyond geographical limitations. It overcomes the challenges of remote work, making meetings and collaboration as lifelike as possible, bridging physical distance in the real world.

In social interaction, the metaverse enables people separated by physical distances to communicate through virtual gatherings and events. For instance, people can meet in virtual worlds through avatars or host and attend events in environments otherwise impossible in reality.

Despite the metaverse's boundless potential, significant technical challenges remain. A key enabling technology for the metaverse, head-mounted display (HMD)-based virtual reality, tends to induce side effects such as cyber sickness and visual fatigue after prolonged use[1,2]. Many of these issues are attributed to Motion To Photon (MTP) latency[3], a delay between user actions and corresponding changes in the virtual environment. Virtual reality inherently exhibits MTP latency due to the time required to render responses to sensor and user inputs. Increased latency complicates synchronization between physical movement and virtual reality perception, leading to discomfort, dizziness, and

nausea. Furthermore, high MTP latency causes visual fatigue as users continually adjust to slight delays between action and visual feedback. Thus, minimizing MTP latency is crucial to enhancing user comfort and experience in HMD-based virtual reality programs, crucial to developing the metaverse.

Previous research has attempted to address these issues by pre-rendering scenes with a margin beyond the user's field of view[4], rendering the center of the field of view at standard quality while peripheral areas at lower quality[5], or identifying user focus areas for pre-rendering[6]. Additionally, techniques have been developed to use deep learning to predict head motions[7] to facilitate pre-rendering. However, traditional approaches show low prediction rates due to inaccurate forecasts or limitations of Recurrent Neural Networks (RNNs).

This paper introduces a novel approach incorporating the latest research trends in RNNs into deep learning-based head motion prediction. Traditionally, deep learning has heavily relied on RNN models, known for capturing complex patterns and structures in sequential data. RNNs predict subsequent values by reflecting the values of previous sequences but face limitations in parallel processing and suffer from long-term dependency issues as the value of earlier data diminishes over time. This paper proposes using the Transformer model[8], which addresses RNNs' parallel processing issues and long-term dependencies, to base our predictions. We hypothesize that user head direction is influenced by the content viewed and verify this by extracting features from images and saliency data. These features are utilized to improve prediction accuracy, and an image generation model provides additional information to the decoder,

demonstrating high prediction accuracy.

The primary contributions of this paper are as follows:

- The application of the widely used Transformer model from natural language processing to overcome the limitations of RNNs in head motion prediction.
- The use of image and saliency data to enhance the accuracy of user head motion predictions.
- Improvement of overall prediction efficiency by utilizing image information during the decoding process with an image generation model.

2. Related work

This section introduces and analyzes the theoretical background and existing studies on head motion prediction. It explores the evolution of head motion prediction algorithms and major technical approaches, discussing how they have been utilized to predict user head movements. Additionally, the role and significance of the saliency map, commonly used in head motion prediction, is explored. The aim of this chapter is to provide readers with a comprehensive understanding of existing research in the field of head motion prediction and to identify the background and necessity of the novel approach proposed in this paper.

2.1 Head Motion Prediction

Head motion prediction, a technology for analyzing and forecasting user head movements in real time, is considered a critical research topic for enhancing user experience in Virtual Reality, Augmented Reality, and Mixed Reality environments. Higher accuracy in predictions allows users to interact naturally and intuitively in virtual worlds, maximizing immersion and enhancing the realism of simulations.

Early research in head motion prediction utilized linear models, such as Kalman filtering[9], to predict head movements. However, these traditional

approaches have limitations, especially when dealing with complex movements or extended prediction timelines, resulting in decreased accuracy.

With the advancement of deep learning, approaches to head motion prediction have significantly changed. Prediction models based on Long Short-Term Memory(LSTM)[10] have been proposed in various studies. LSTM's ability to model long-term dependencies in time-varying data makes it highly effective for learning patterns in complex sequence data. Research has been conducted on combining saliency networks with head motion data[11], simultaneously utilizing image features and head motion rotation speeds[12], and using multiple convolutional neural network models to integrate various saliency data and user gaze data[13].

Recent studies have explored various methods to resolve the long-term dependency issues of recurrent neural networks. These include using encoder-decoder-style recurrent networks to predict more complex patterns of head movements[14], and employing reinforcement learning to further enhance the performance of head motion prediction[15]. Additionally, Transformer-based prediction models[16] have also been proposed, utilizing multi-head self-attention architectures to predict future head directions from previous frames.

2.2 Saliency Detection

Saliency detection is a research field aimed at predicting users' visual

saliency in images or videos, with the fundamental goal of anticipating and understanding natural visual behaviors, thereby enabling more efficient user interface design and more accurate information delivery systems.

Initial research stages focused on manually extracting various features within images to enhance object detection accuracy. In particular, methods were explored to detect saliency using basic features such as intensity, color, and orientation of images[17]. Developments in these initial studies have led to proposals for image filtering[18] and various mathematical algorithms[19] to improve the accuracy of saliency detection.

With the advancement of deep learning technologies, saliency detection research has included more complex and sophisticated methodologies. Studies have been conducted applying convolutional neural network models developed for image classification to saliency detection using transfer learning[20] and developing specialized loss functions for saliency detection[21]. Additionally, deep learning models for detecting saliency in 360-degree images have also been actively researched, including methods utilizing spherical coordinate information[22], developing loss functions specialized for spherical coordinates[23], and extracting saliency directly using head motion data[24].

3. Algorithm

This chapter describes the head motion prediction algorithm proposed in this paper. The previous transformer-based algorithm[16] utilizes a transformer structure based on a multi-head self-attention mechanism. In this study, an image processing module is used to extract head direction information from images, and the features obtained from this module are applied to a transformer network based on multi-head self-attention and encoder-decoder attention. The overall structure of the algorithm is depicted in Figure 1. A key feature of this paper is the use of additional image features such as saliency in the encoder stage to enhance prediction accuracy. However, this paper goes beyond simply utilizing features; it uses an image generation model to extract and vectorize image features in the encoder, which are then converted back into images in the decoding process to be used for predicting the next frame. The predicted images are utilized in the decoder alongside head motion data to demonstrate a higher prediction probability.

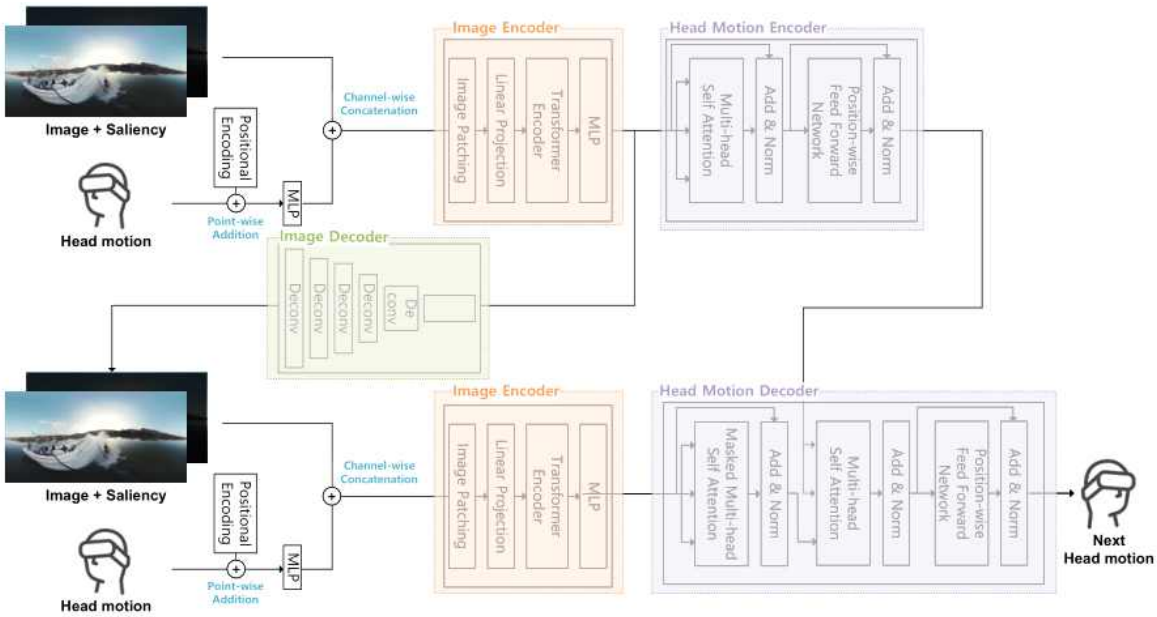

Figure 1. An overview of the model proposed in this paper, given images and head motion data. An image encoder is used to extract information about the image and head motion, which is then fed into the encoder of the head-motion network. The information extracted by the image encoder is also used as input for the image generation model, creating images to be used in the decoder. The generated image and head motion data are used in the decoder in a similar manner as in the encoder, extracting information to make predictions.

3.1 Head Motion Prediction Network

In this section, we elaborate on the architecture and principles of the Transformer, the core model of this paper, and discuss how its features have been applied to head motion prediction. As a model for head motion prediction, this paper adopts the Transformer as the base model to overcome the long-term dependency and parallel processing challenges inherent in Recurrent Neural Networks (RNNs).

3.1.1 Model Architecture

The encoder-decoder structure of the Transformer, as shown in Figure 2a, represents a significant innovation in deep learning-based sequence models. The encoder processes the input sequence through multiple layers using multi-head attention and a feed-forward network. Each encoder layer employs the multi-head attention mechanism to extract information from the input sequence and transforms the data using a feed-forward network to produce a latent space representation. This latent space representation serves as input for the decoder, which generates a new sequence based on the information extracted by the encoder and previous prediction results.

3.1.2 Multi-Head Attention

Figure 2b illustrates the core multi-head attention mechanism of the Transformer. Multi-head attention extracts crucial information from different parts of the data in parallel and combines this information to grasp the overall

context of the sequence.

The data features are divided into units called heads, which use queries, keys, and values. Each data point (query) evaluates its relationship with all other data points (keys) and learns the strength of these relationships (values). Each head focuses on different aspects of the data, enabling a richer interpretation through diverse perspectives.

The scaled dot-product attention mechanism computes the dot product between query and key vectors to measure similarity, and then scales it to improve computational stability. The scaled similarity is then converted to probabilities using a softmax function to represent the strength of relationships between data points. Finally, these probability distributions are multiplied by the value vectors to produce the attention output. Through this process, multi-head attention effectively analyzes the data from different perspectives to extract valuable information.

3.1.3 Feed-Forward Network

The feed-forward network is located after each attention mechanism in the encoder and decoder layers, and processes data through two consecutive linear transformations with a non-linear activation function in between. This helps further learn the features extracted by each attention mechanism and enhances the model's representational capacity.

3.1.4 Positional Encoding

Unlike RNNs, the Transformer inherently lacks the ability to understand the order of sequential data. To address this, positional encoding is used to add positional information to each element in the sequence. This enables the model to recognize the temporal or spatial order of the sequence, allowing it to comprehend sequential relationships.

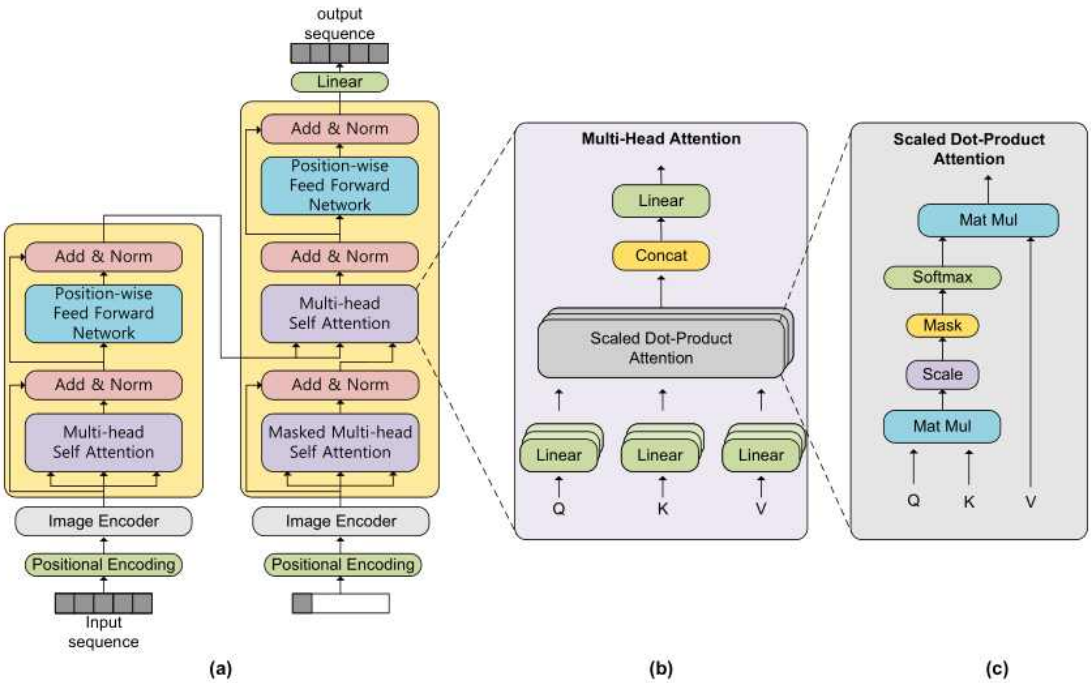

Figure 2. This figure illustrates the transformer-based model used in this paper. (a) shows the architecture of the head motion prediction model, (b) depicts the multi-head attention, and (c) presents the scaled dot-product attention from (b).

3.2 Image Encoder

Considering the significant influence of visual content on head movements in HMD-based virtual reality environments, this paper proposes a novel input data structure that integrates head motion data with image features. Frames from 360-degree videos are segmented, and at each frame, user head motion data is concatenated channel-wise with images, providing the input for the transformer module. This configuration effectively facilitates the learning of interactions between visual tendencies and images.

Figure 3 visualizes the vision transformer[25] used in this paper. Figure 3b illustrates the model architecture of the vision transformer. The vision Transformer is based on the transformer architecture that has demonstrated exceptional results in natural language processing, treating each image as a sequence of grid-like patches. However, this approach requires extensive data for effective training. To address this limitation, as shown in Figure 3a, the image encoder model developed in this research employs convolutional layers in the initial phase of image patch segmentation to extract features. These features are subsequently enhanced through attention mechanisms and residual learning, thereby increasing the efficiency of network learning and improving overall performance.

The multi-head attention mechanism of the Vision Transformer divides input data into several parts to process and integrate information from various scales and perspectives simultaneously. The feedforward layer processes data through two consecutive linear transformations, with a nonlinear activation function applied in between to enhance the model's expressive power. Through this configuration, the Vision Transformer precisely extracts various image features and accurately learns the interactions with user head movements.

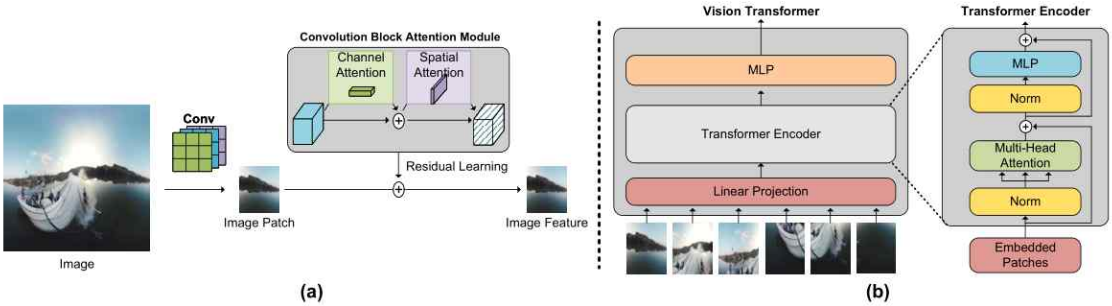

Figure 3. The image encoder used in this paper is depicted. (a) shows image patches extracted via convolutional layers passing through CBAM[26]. (b) illustrates the ViT[25] module used in this paper, where image features are extracted through an MLP in the final stage and used as input to the head motion encoder proposed in this paper.

3.3 Image Decoder

Differing from traditional methods that only utilize images or saliency during the encoding process, this paper also utilizes them in the decoding process. Features of head motion data combined with image and saliency data extracted by the image encoder are used as latent vectors to generate images. This process involves using an inverse convolutional neural network to create a 4-dimensional image combining the next frame's image and saliency.

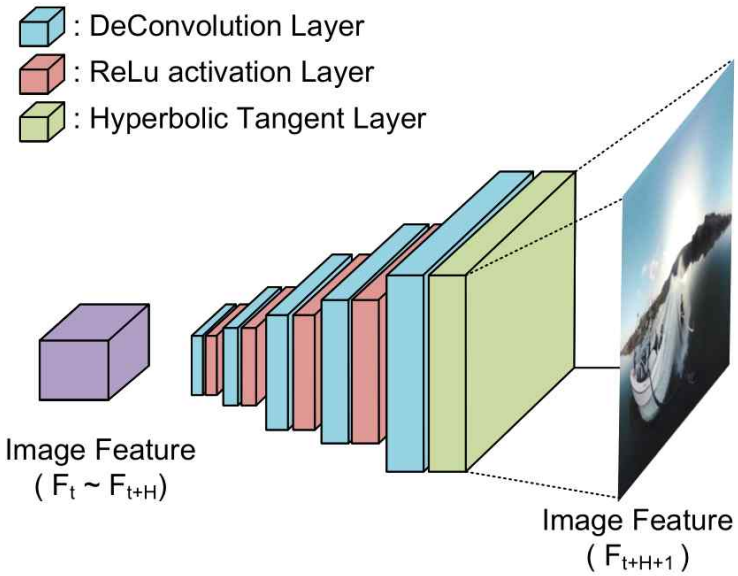

Figure 4. This figure displays the Image Decoder used in this paper. Five conventional deconvolution layers were employed to expand the features to the same size as the image, with ReLU layers used in all but the final layer, which utilizes a Tanh function.

3.4 Image Attention Module

The transformer's attention mechanism plays a pivotal role in improving the accuracy of the decoder by emphasizing significant parts of the input data. Similarly, in the field of image processing, the Convolution Block Attention Module (CBAM) [26] employs a specialized attention mechanism to highlight meaningful features in images while eliminating redundant information.

CBAM is integrated into the intermediate layers of a convolutional neural network and employs two primary attention mechanisms. First, channel-wise attention assesses the significance and characteristics of each channel to highlight meaningful channels. By performing average pooling and max pooling on the input feature map, followed by a linear layer, it extracts vectors that represent the importance of each channel. Second, spatial attention identifies and emphasizes crucial regions within the image. It generates a map that highlights important regions by performing average pooling and max pooling on the feature map, which has already undergone channel-wise attention.

This paper addresses the limitation of ViT that require extensive data for training by employing CNNs to patchify images and using the CBAM module to extract critical feature values. The extracted attention values are added to the existing patch values through residual learning. This approach enhances the important features in the patch images obtained from the image encoder, thereby improving the accuracy of head movement prediction.

4. Experiments

This section describes the data used in the experiments and various details regarding the experimental setup.

4.1 Dataset

The data used in this paper was derived from the PVS-HM dataset[15], which includes 76 pieces of 360-degree video and head motion data(roll, pitch, yaw) collected from 58 participants using HTC Vive devices. A subset of this data was used for training and validation. To utilize the image data effectively, only the pitch and yaw angle data were converted into spherical coordinates (x, y, z), where the sum of the squares of the components (x, y, z) always equals one. This coordinate system is suitable for representing head motion as a vector in three-dimensional space.

Saliency was extracted using the deep learning-based network Panosalnet[11], which measures user attention and employs a pre-trained model for extraction. Saliency, image, and head motion data were segmented frame-by-frame and concatenated in a set window size to form the input data.

User data is inherently variable, as experimental conditions, sensor errors, or user errors can alter results even when the same user views the same video

repeatedly. Outliers were present for these reasons, and anomalous data were identified by setting viewing angles and using the IQR (Interquartile Range) method. This method utilized the differences between two consecutive head motion data points to represent the data distribution, using only the data within the 75th percentile. Data points above this threshold indicated large variations.

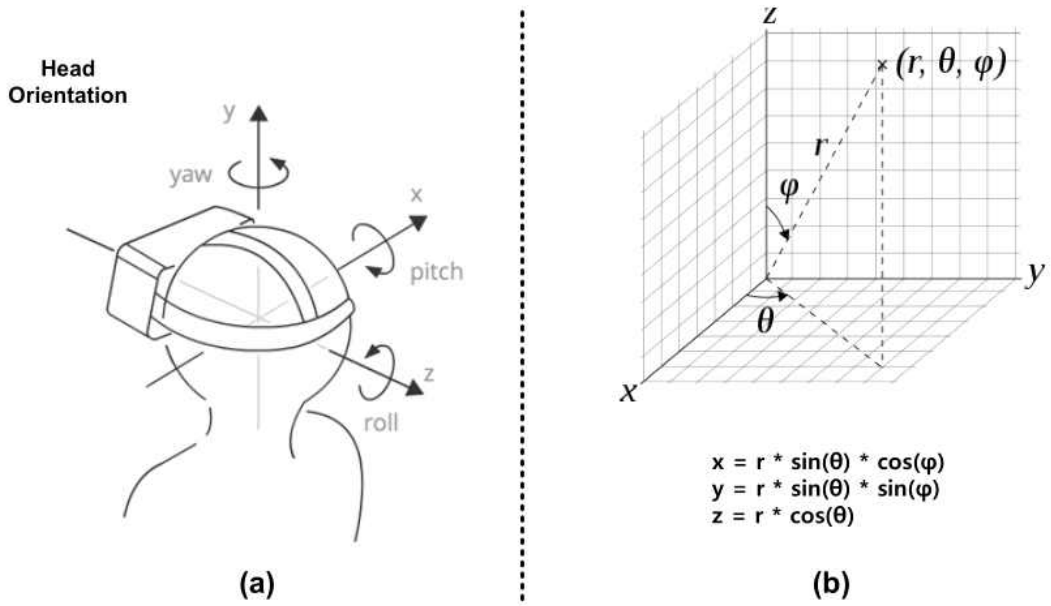

Figure 5. (a) displays the Euler angles—roll, pitch, yaw—used in virtual reality, while (b) demonstrates how these angles are transformed into coordinates within a spherical coordinate system. Here, R is set to 1 for the transformation into the spherical coordinates.

4.2 Loss Function

In TRACK[14], the Mean Square Error (MSE) loss is used, and in VPT360[16], MSE loss is augmented with the change between previous and current sequences. In this study, as shown in Equation 1, the absolute distance of each predicted coordinate is used as the loss. Given that the real values and predicted values are typically smaller than those used in general regression problems, Smooth L1 Loss was adopted for model training. Smooth L1 Loss combines the robustness to large values characteristic of L1 Loss with the sensitivity to small errors of L2 Loss by integrating the absolute and squared differences between predicted and actual values. A threshold is set where values below this threshold use L2 Loss, and values above it use L1 Loss. During model training, the differences between actual values and predicted values are calculated, applying L1 Loss for larger discrepancies and L2 Loss for smaller ones. The threshold was set at 0.05, based on histogram analysis of sequence data differences, which showed most values concentrated within a specific range.

$$diff = \frac{1}{3}((\hat{x}-x)^2 + (\hat{y}-y)^2 + (\hat{z}-z)^2) \tag{1}$$

$$Loss_n = \begin{cases} 0.5 * diff^2/\beta & \text{if } diff < \beta \\ diff - 0.5*\beta & otherwise \end{cases} \tag{2}$$

4.3 Experimental Implementation Details

The model training centered on iterative learning with the image encoder and transformer. Considering training duration and memory efficiency, the image encoder and transformer were each trained through three iterations. Experiments were conducted using an NVIDIA GTX 3090 GPU, and the model was optimized using the Adam optimizer. The initial learning rate was set at 0.0001 due to the small range of head motion data values, with b1 and b2 parameters set at 0.9 and 0.999, respectively. The total training epochs were set at 30, and to optimize memory and the sequential nature of the data, the batch size was set at 8. Each epoch took approximately 35 minutes.

5. Results

This section presents the experimental results to validate the performance of the model proposed in this research. Table 1 compares the techniques presented in this paper with those suggested by the TRACK model across various models and datasets, focusing on predictions segmented from 360-degree videos. In these experiments, five frames were used as input to predict the subsequent five frames. The TRACK model utilizes an LSTM-based Seq2seq network, employing head direction data for prediction, and additionally uses saliency data extracted from images as input values. This saliency is extracted through a neural network, and the features are further input into the recurrent neural network-based Seq2seq model for prediction.

In this paper, experiments were conducted using both image and saliency data, and similar experiments were also carried out using the TRACK model. The results shown in Table 1 utilize the Mean Overlap[15] metric. Mean Overlap is a performance metric that calculates the field of view of a person using head motion and shows the degree of overlap between the predicted and actual field of view as a percentage. Higher values indicate more overlap, with the maximum value being 100%.

According to the results in Table 1, the TRACK model performs better when predicting solely based on head motion data without using images or saliency. However, when saliency data is additionally used, the model proposed in this study generally shows superior performance. Notably, while the TRACK model performed better with the Waterskiing video, in other videos, the proposed model shows better results when saliency is used in addition to the predictions. Moreover, when image and saliency data are combined channel-wise, the model proposed in this paper outperforms the TRACK model in all cases. Particularly in the case of the StarryPolar video, whereas the TRACK model shows lower accuracy compared to other videos, the proposed model, through its image processing module, predicts areas of interest within the image, generally exhibiting better performance.

		Waterskiing	F5Fighter	Mercedes Benz	Parasailing	StarryPolar
TRACK [14]	POS	94.8	93.3	95.3	94.9	**88.1**
	SAL	**95.4**	93.3	95.1	95.3	87.0
	IMG	94.8	**93.7**	**95.4**	**95.3**	86.0
	IMG+SAL	95.3	93.4	95.0	95.3	86.8
Ours	POS	92.6	93.2	94.3	94.3	89.7
	SAL	95.3	95.2	95.6	96.4	94.0
	IMG	95.5	95.3	95.7	96.5	94.2
	IMG+SAL	**95.5**	**95.4**	**95.8**	**96.6**	**94.3**

Table 1. Comparison of results for the model and dataset proposed in the paper. The top row lists the names of the 360-degree videos. Results are evaluated the individual frames of the input videos. Our model proposed in this paper outperforms the TRACK model for most cases.

Table 2 elaborately describes the structural features and functional roles of core modules in the model proposed by this study. This research explores an artificial neural network-based method of extracting image features utilizing image generation, analyzing its effects comprehensively. In the image generation process, critical features are extracted from image data through a deep learning model, and these features are used to generate new images. This methodology significantly enhances prediction accuracy by allowing the decoder to utilize additional image features.

In cases without Image Generation, the image encoding process remains intact, but the generation phase is omitted. The decoder does not use any additional image features but relies solely on the information extracted by the encoder for making predictions. This setup allows for the analysis of the impact that the image generation process has on the final prediction outcomes.

In the case known as 'Without CBAM', both CBAM and residual learning are omitted. Only convolutional layers are used when the image is segmented into patches. This configuration assesses the role of CBAM in the process of image feature extraction and recognition. The experiments indicate that CBAM significantly contributes to the effective extraction of information about human focal points.

When image generation is not utilized, there is an observable improvement in the decoder's prediction performance by approximately 1%. This improvement suggests that the image generation process positively influences prediction

accuracy. Additionally, employing CBAM results in a further improvement of about 0.1 to 0.2%. This finding substantiates that effectively capturing and utilizing information about human focal points in image feature extraction enables more precise predictions.

	Waterskiing	F5Fighter	Mercedes Benz	Parasailing	StarryPolar
Without Image Generation	94.1	93.8	94.5	95.3	91.0
Without CBAM	95.4	95.2	95.8	96.4	94.0
Ours	**95.5**	**95.4**	**95.8**	**96.6**	**94.3**

Table 2. Differences of results estimated when only using additional data in the encoder, when excluding the image generation model used in this paper, and when encoding images without the image attention mechanisms and the residual learning in the image encoder. The first row represents names of the videos, as in Table 1.

Figure 6 is an indicator used in this paper to visualize how accurately the proposed model predicts ground truth. In this study, we proceed with predictions using only the pitch and yaw values, excluding the roll value from the head movement data. The pitch and yaw values are converted into x, y, z coordinates, which are then used as the basis for predicting positions. The results are subsequently transformed back into pitch and yaw values and visualized on a two-dimensional graph with x and y axes. The dots on the graph represent the points looked at, and the lines indicate the paths of movement. A graph where the position and direction of the dots and lines are similar demonstrates better prediction accuracy.

The x-axis represents the yaw value, ranging from -180 to 180 degrees, while the y-axis represents the pitch value, ranging from -90 to 90 degrees. However, since the ground truth changes are minimal, the axes of the graph have been appropriately scaled to display the data clearly. This graph was generated using different videos, visualizing data from the start up to the 60th frame of each video. Additionally, the model's generalization capability was tested using data from new users not included in the training dataset.

In videos such as Flight and MercedesBenz, the model proposed in this study shows an overlay almost identical to the ground truth, suggesting that the model can operate with high accuracy in real-world applications. Although the TRACK model also displays similar performance, minor discrepancies between predicted values and ground truth occur in specific frames. While the TRACK model's predictions may appear to be significantly off, the very small variance

means that in reality, the differences are not substantial. Nevertheless, the model proposed in this paper demonstrates more detailed and stable predictions compared to the TRACK model.

For videos like F5Fight and LoopUniverse, the proposed model does not perfectly overlay but follows the general trend of the ground truth well. This indicates that the model effectively captures the primary patterns of the data. In contrast, the TRACK model shows more significant fluctuations and lacks consistency compared to the model proposed in this paper, suggesting lower stability in its predictions.

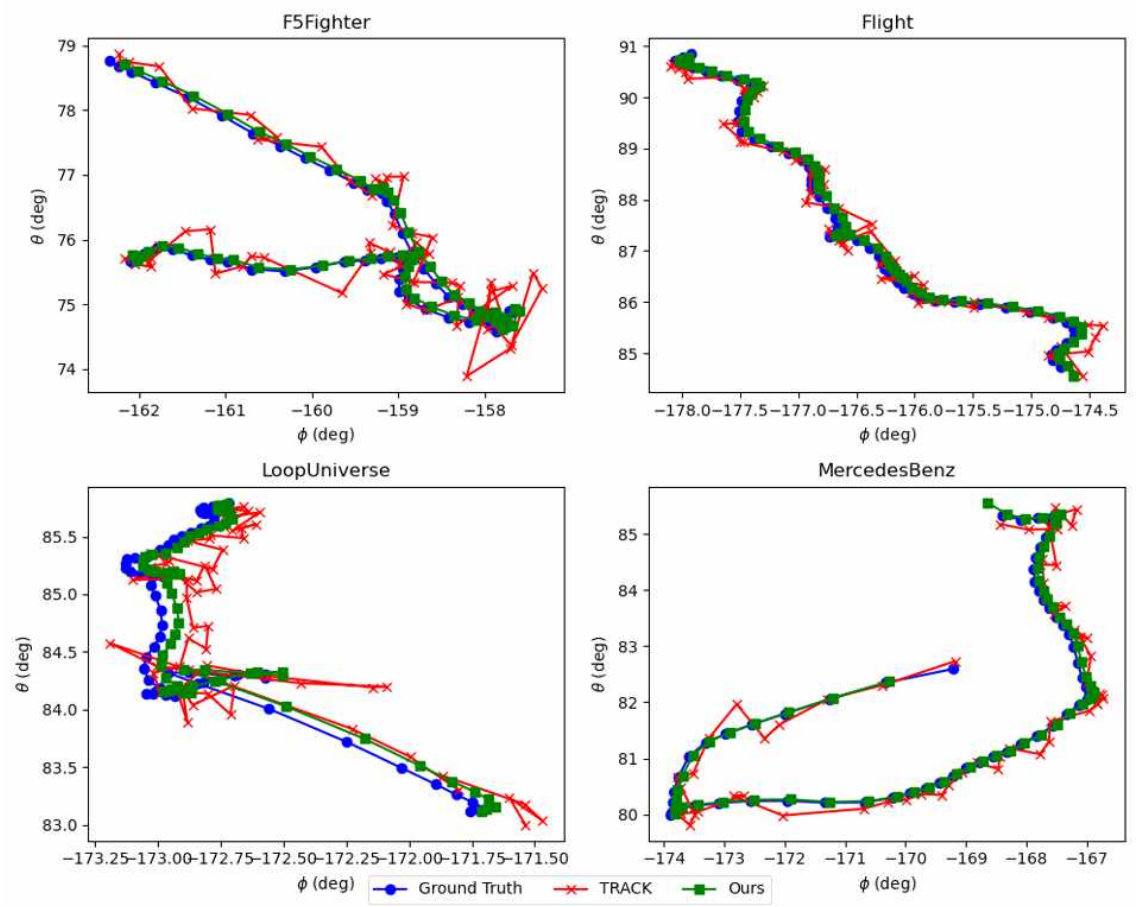

Figure 6. Demonstrates how closely the predicted values of the proposed model align with the ground truth. Each graph represents different videos analyzed using the same user data. Generally, the model proposed in this paper either closely matches or follows the trend of the ground truth, in contrast to the TRACK model, which often fails to do so.

Figure 7 represents the distances between the actual and predicted values over time. In virtual reality applications that utilize 360-degree videos, specific points are represented as if they are on the surface of a sphere. Due to this, the 'Great Circle Distance' is employed to accurately measure the distances between the actual and predicted values. The Great Circle Distance is a method used to calculate the shortest distance between two points on the surface of a sphere, such as Earth, where the line connecting these two points forms a circle with a center that coincides with the center of the sphere.

$$d = R \cdot \arccos(\sin(\phi_1) \cdot \sin(\phi_2) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \cos(\Delta\lambda)) \quad (3)$$

ϕ represents the yaw value, and λ signifies the difference in pitch values. This measurement technique is very useful in assessing the accuracy of location predictions in 360-degree videos, as it precisely captures changes in position on the sphere. The closer the distance between the two points on the sphere, the better the prediction, and the values on the graph are accumulated over time to represent these differences.

In cases like F5Fighter and Loop Universe, the graphs are nearly similar, yet the model proposed in this paper shows slightly better performance. For Flight and MercedesBenz, while the trends in predictions are similar, resulting in similarly trending graphs, the difference in the values indicates that the model proposed in this paper provides a closer match to the actual values.

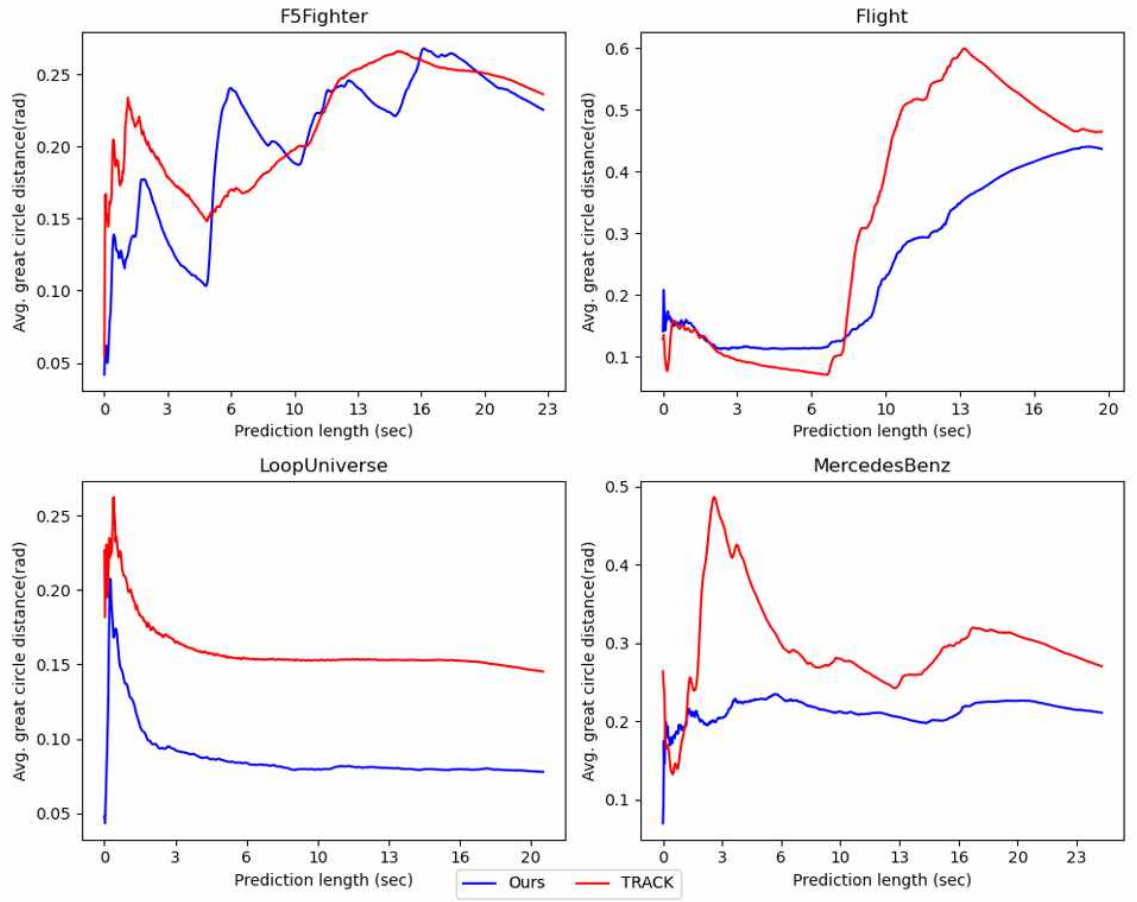

Figure 7. This figure displays the averaged Great Circle Distance between two points over time. Smaller values indicate more accurate predictions, and except for the F5Fighter video, the model proposed in this paper demonstrates better performance than the TRACK model from start to finish.

Figure 8 displays the angular error over time between the ground truth and predicted values, as calculated using Equation 2. Angular error is a metric that quantifies the difference in angles between two vectors or directions, serving as an indicator of system performance or prediction accuracy. The angular error calculated at each time point is graphically represented to visually analyze the model's performance variations over time.

$$d = \sqrt{\arctan(\sin(\Delta(\theta)/\cos(\Delta(\theta))^{2+}(\phi_1 - \phi_2)^2} \quad (4)$$

For Flight and LoopUniverse, the graph forms are almost identical, with the model proposed in this paper consistently showing superior performance. This consistency arises from similar prediction tendencies, which are reflected in the graph's trend. In the case of F5Fighter and MercedesBenz, although there are specific parts where the performance of the proposed model drops slightly, it generally exhibits better overall performance

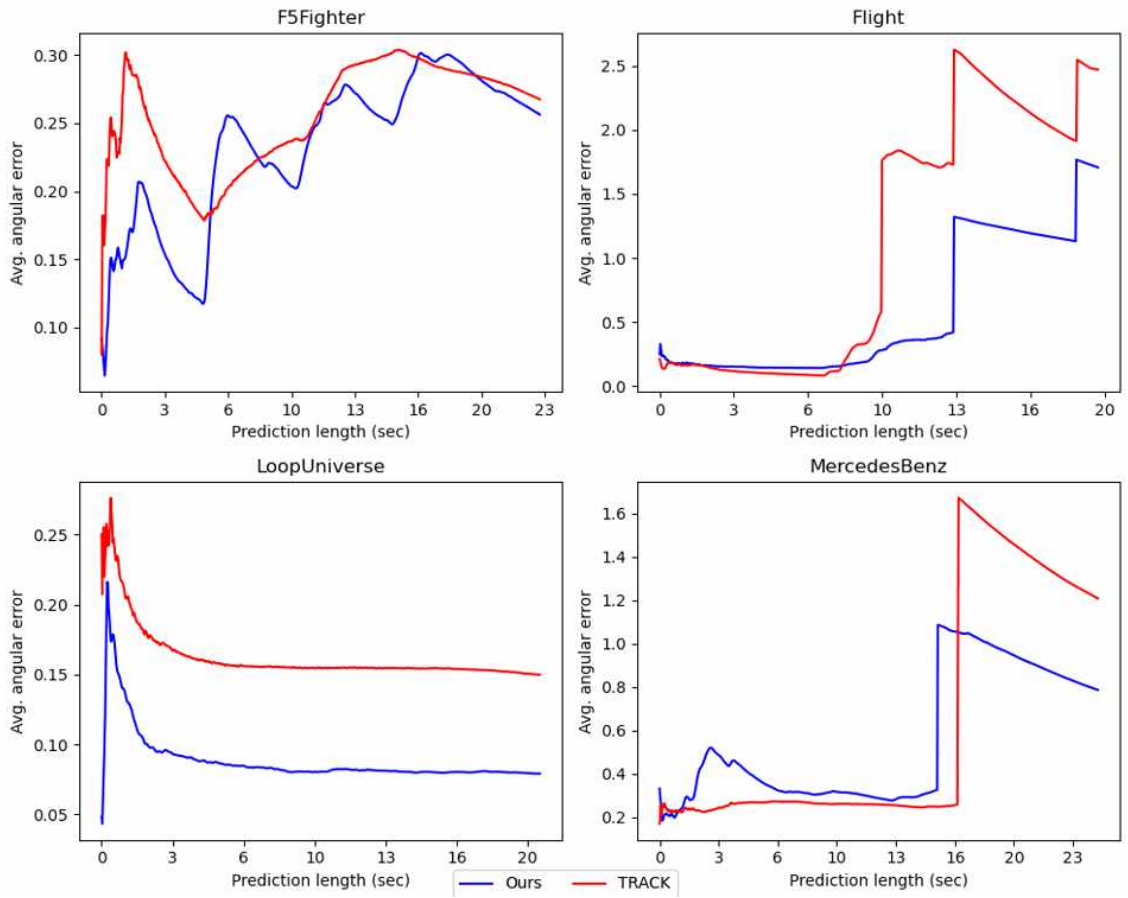

Figure 8. This figure displays the averaged angular error over time. Smaller values indicate more accurate predictions, and across all videos, the model proposed in this paper generally demonstrates better performance than the TRACK model.

Figure 9 displays the Mean Overlap values over time between the ground truth and predicted values. Mean Overlap is an indicator similar to the MIoU, measuring the degree of overlap in the field of view between predicted and ground truth values, calculated based on human visual perception. Unlike other error metrics, a higher Mean Overlap indicates greater accuracy, as more overlap correlates with better prediction accuracy.

For the F5 Fighter, although the graph fluctuates like other metrics, it generally exhibits superior performance most of the time. For Flight and MercedesBenz, although the TRACK model may perform better initially, over time, the model proposed in this paper demonstrates increasingly superior performance. In the case of LoopUniver, the trend is similar, and the model proposed in this paper consistently outperforms the TRACK model across all sections.

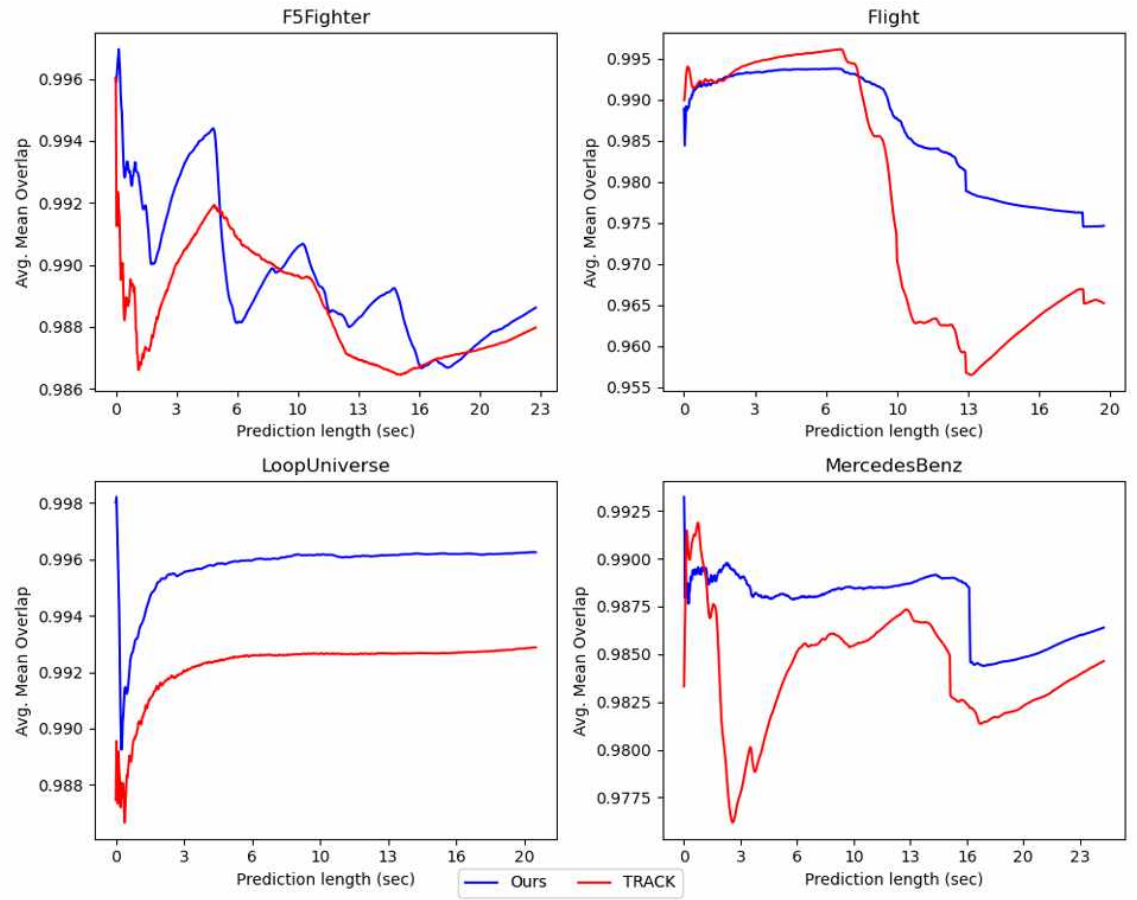

Figure 9. This figure displays the time-averaged Mean Overlap values. Higher values indicate more accurate predictions, with the y-axis representing the percentage of overlap. Although the differences are slight, the model proposed in this paper demonstrates better performance than the TRACK model across all videos.

6. Conclusion and Future Research

This paper targets the reduction of Motion To Photon (MTP) latency in head-mounted display (HMD)-based virtual reality. To achieve this, we leverage deep learning to understand trends in head motion data and propose a new model that predicts head motion. Our model overcomes the limitations of recurrent neural networks by incorporating a transformer structure. While transformers did not perform as well as recurrent neural networks when using only head motion data, the introduction of additional data and the generation of images in the decoder for predicting and utilizing image features in the decoding process have led to the development of a flexible model that performs well across all videos, not just specific ones.

Experimental results demonstrate that head motion varies depending on the displayed content. Our research adopts the encoder-decoder structure commonly used in natural language processing to compress data and minimize information loss that may occur during the decoding process. In the field of natural language processing, attention mechanisms developed to prevent information loss have proven effective. We have applied these concepts to the image generation and decoding processes, developing a model capable of predicting head motion data without information loss. This allows for effective modeling of the complex interactions between visual data and head motion, potentially enhancing user experiences in various virtual reality environments. This approach is one of the major contributions of our paper and is expected to significantly influence the

advancement of virtual reality technology and the design of efficient user interfaces.

However, the transformers used in this study increase computational complexity and processing time. The dataset used involved video frames and saliency extracted from convolutional layers, which added to the time required. Future research directions will focus on reducing the computational complexity of transformers and enhancing data processing efficiency to shorten processing times. Additionally, based on the achievements of this paper, we aim to pursue optimizations that can be practically applied in HMD environments. Through this research, we have confirmed the usefulness and potential of using deep learning to reduce MTP latency, and we will continue to explore effective ways to apply these findings in HMD environments.

References

[1] H. Han, H. Park, J. Kim, and J. Park, "Survey on Cyber Motion Sickness," Proc. of the KIISE Korea Computer Congress 2015, pp. 1401-1403, 2015.

[2] E. Chang, D. Seo, H. Kim, and B. Yoo, "An Integrated Model of Cybersickness: Understanding User's Discomfort in Virtual Reality," Journal of KIISE, Vol. 45, No. 3, pp. 251-279, 2018.

[3] J. Zhao, R. S. Allison, M. Vinnikov, and S. Jennings, "Estimating the motion to photon latency in head mounted displays," 2017 IEEE Virtual Reality (VR), pp. 313-314, 2017.

[4] S. Shi, V. Gupta, M. Hwang, and R. Jana, "Mobile VR on edge cloud: a latency-driven design," Proceedings of the 10th ACM multimedia systems conference, pp. 222-231, 2019.

[5] J. Kim, S. Choi, J. Choi, S. Ahn, and C. Park, "Trend in Head Motion Based Rendering Technology in Augmented Reality," Communications of the Korean Institute of Information Scientists and Engineers, Vol. 34, No. 12, pp. 23-28, 2016.

[6] S. Lee, et al., "Motion-constrained tile set based 360-degree video streaming using saliency map prediction," Proceedings of the 29th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 20-24, 2019.

[7] Y. Xu, et al., "Gaze prediction in dynamic 360 Immersive videos," proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5333-5342, 2018.

[8] A. Vaswani, et al., "Attention is all you need," Advances in neural information processing systems 30, 2017.

[9] R. T. Azuma, "Predictive tracking for augmented reality," University of North Carolina at Chapel Hill, 1995.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

[11] A. Nguyen, Z. Yan, and K. Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," Proceedings of the 26th ACM international conference on Multimedia, pp. 1190-1198, 2018.

[12] Y. Li, et al., "Two-layer fov prediction model for viewport dependent streaming of 360-degree videos," Communications and Networking: 13th EAI International Conference, ChinaCom 2018, Chengdu, China, October 23-25, 2018, Proceedings 13, pp. 501-509, 2019.

[13] Y. Xu, et al., "Gaze prediction in dynamic 360 immersive videos," proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5333-5342, 2018.

[14] M. F. R. Rondón, et al., "TRACK: A New Method From a Re-Examination of Deep Architectures for Head Motion Prediction in 360 Videos," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, No. 9, pp. 5681-5699, 2021.

[15] M. Xu, et al., "Predicting head movement in panoramic video: A deep reinforcement learning approach," IEEE transactions on pattern analysis and machine intelligence, Vol. 41, No. 11, pp. 2693-2708, 2018.

[16] F. Y. Chao, C. Ozcinar, and A. Smolic, "Transformer-based Long-Term

Viewport Prediction in 360° Video: Scanpath is All You Need," MMSP, pp. 1-6, 2021.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on pattern analysis and machine intelligence, Vol. 20, No. 11, pp. 1254-1259, 1998.

[18] Y. Fang, et al., "Video saliency detection in the compressed domain," Proceedings of the 20th ACM international conference on Multimedia, pp. 697-700 2012.

[19] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," Journal of vision, Vol. 13, No. 4, pp. 11-11, 2013.

[20] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on ImageNet," arXiv preprint arXiv:1411.1045, 2014.

[21] X. Huang, et al., "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," Proceedings of the IEEE international conference on computer vision, pp. 262-270, 2015.

[22] R. Monroy, et al., "Salnet360: Saliency maps for omni-directional images with CNN," Signal Processing: Image Communication, Vol. 69, pp. 26-34, 2018.

[23] Z. Zhang, et al., "Saliency detection in 360 videos," Proceedings of the European conference on computer vision (ECCV), pp. 488-503, 2018.

[24] A. Nguyen, Z. Yan, and K. Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," Proceedings of the 26th ACM international conference on

Multimedia, pp. 1190-1198, 2018.

[25] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[26] S. Woo, et al., "Cbam: Convolutional block attention module," Proceedings of the European conference on computer vision (ECCV), pp. 3-19, 2018.

논문요약

이미지 생성 모델을 이용한 트랜스포머 기반 헤드모션 예측 알고리즘

변효근
소프트웨어학과
성균관대학교

모션 투 포톤 지연은 헤드 마운트 디스플레이 기반 가상현실에서 사용자의 움직임과 영상 출력의 시차로 인한 사이버 멀미 같은 불편감을 줄 수 있고, 이러한 불편함이 지속되면 사용자의 몰입감을 방해할 수 있다. 기존의 모션 투 포톤 지연을 줄이는 방식은 직접 헤드 모션 데이터의 경향성을 파악하거나, 순환신경망 모델을 통해 헤드 모션을 예측하지만, 기존의 순환신경망 모델은 시퀀스 정보를 오랜 시간 동안 기억하지 못하는 장기 의존성 문제와 병렬 처리의 제약이 존재한다. 본 논문은 트랜스포머 기반 헤드 모션 예측 모델을 이용하여 이전의 영상 프레임의 데이터를 통해 이후의 프레임을 예측하는 기법을 제안한다. 본 논문에서는 이미지 생성모델을 통해 디코딩 과정에서도 이미지를 사용한다는 점과 자연어처리에서 사용되던 딥러닝 모델을 예측 모델로 사용함에 있어서 높은 확장성을 가진다. 또한 본 연구에서 제안한 모델은 데이터를 추가로 사용하여 기존 모델보다 사용자의 헤드 모션을 잘 예측함을 알 수 있다.

주제어: 사이버 멀미, 가상현실, 렌더링, AI

Master's Thesis

Transformer-Based Head Motion Prediction Algorithm Using Image Generation Model

2024

Hyogeun Byun