ARTICLE TEMPLATE

DSPT: Disassembly Sequence Planning Transformer for Interaction Guidance in VR

Sichun Huang[a], Ziteng Wang[a], Sio Kei Im [c] and Lili Wang[a ,b]

[a]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China;
[b]Peng Cheng Laboratory, Shengzhen, China;
[c]Faculty of Applied Sciences, Macao Polytechnic University, Macau, China.

ARTICLE HISTORY
Compiled December 9, 2025

ABSTRACT
The application of virtual reality technology in complex equipment disassembly training is widely used, and planning the disassembly sequence and interactively guiding the disassembly is an issue that requires in-depth research. Traditional methods based on physical collision detection are very accurate, but the computational efficiency is too low to meet the requirement of interactivity. In recent years, deep learning-based disassembly sequence prediction methods have emerged, which are fast in reasoning but suffer from inaccurate prediction of parts to be disassembled. In this paper, we propose a novel Transformer-based network, the Disassembly Sequence Planning Transformer (DSPT), to optimize the disassembly sequence for guiding users to disassemble objects in VR environments. First, we define Disassembly Sequence Features and Part History Features, along with their construction methods. Then, we introduce the parts-to-be-disassembled probability predictor based on a temporal-spatial score and propose a new loss function leveraging the temporal-spatial score to enhance the predictor's performance. Experimental results show that our method achieves higher sequence accuracy and stepwise accuracy, both outperforming the state-of-the-art method. The results of the user study demonstrate that our method significantly reduces the disassembly task completion time and improves the usability compared to comparison methods.

KEYWORDS
Virtual Reality; Disassemble Sequence Planning; Transformer; Interactive Guided Disassemble;

1. Introduction

With the rapid development of virtual reality (VR) technology, its application in industrial manufacturing, maintenance, and training is becoming more and more widespread. In VR disassembly tasks without disassembly guidance, users may frequently rotate objects and shift viewpoints, making the process cumbersome and potentially dizzying.

Lili Wang is corresponding author: wanglily@buaa.edu.cn, ZY2306429@buaa.edu.cn

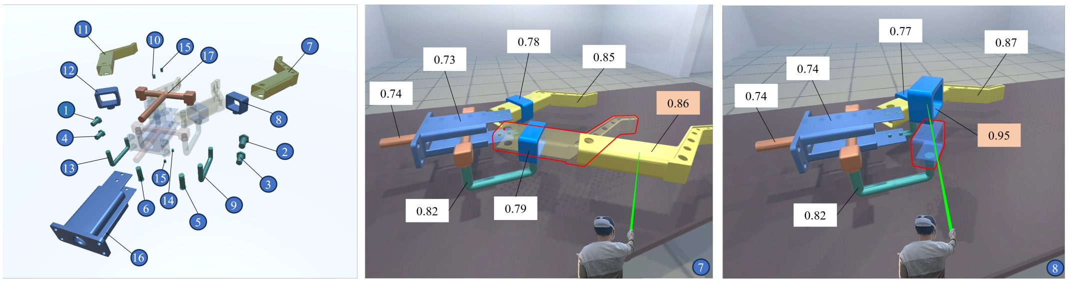


Figure 1. The left figure shows the DSPT-planned disassembly sequence, with numbers in blue circles indicating the predicted order. The middle and right figures display VR guidance process. At each step, DSPT predicts disassembly probabilities and highlights the most likely component to guide the user.

Disassembly Sequence Planning (DSP) determines the disassembly order of an object. It plays a critical role across multiple life-cycle stages, including repair, maintenance, recycling, and reuse. By determining the optimal order of part removal, DSP reduces labor time, minimizes component damage, and supports sustainable product recovery. Depending on the disassembly objective, DSP can be categorized as Selective, Partial, or Complete. Selective DSP targets specific components for maintenance or replacement; Partial DSP handles subassemblies for remanufacturing; and Complete DSP fully disassembles the product for material recycling.

When integrated with VR, DSP can provide significant support in industrial training and operational guidance, maintenance and repair assistance, as well as interactive decision-making and design verification. Specifically, by leveraging the optimal disassembly sequences generated by DSP within a VR environment, operators can engage in immersive training simulations. Trainees are able to familiarize themselves with disassembly steps, tools, and relevant precautions without the need to interact with physical equipment, thereby reducing operational risks. Furthermore, engineers can intuitively observe the disassembly process in VR, enabling verification and optimization of design schemes. The disassembly sequences produced by DSP can also be employed to assess the maintainability and disassemblability of assemblies, providing valuable insights for subsequent product design improvements.

Existing traditional 3D graphics-based methods for disassembly sequence planning mainly include physics simulation-based approaches such as Assembly Them All (ATA) method (Tian et al., 2022). However, since the method relies on physical collision tests and takes minutes to process individual parts and hours to generate a complete gravity-aware disassembly sequence, it is usually only suitable for offline planning. In real-time VR systems, each model upload requires a lengthy pre-planning phase before disassembly can begin. Automated Sequence Planning for Complex Robotic Assembly with Physical Feasibility (ASAP) method (Tian et al., 2024) addresses this by introducing a GNN, reducing per-part planning time to the second level and enabling interactive VR applications. However, its relatively high prediction error rate still results in a high disassembly failure rate in VR scenarios because they did not consider the influence of the disassembled parts on the next part to be disassembled.

To address these challenges, we propose the Disassembly Sequence Planning Transformer (DSPT) to provide guidance for users to disassemble objects in VR environments. This method not only considers the spatial relationships between object components but also incorporates the historical influence of disassembled parts on the remaining components. We define Disassembly Sequence Features and Part History Features, along with their construction methods. Based on them we get the temporal-spatial score for each part. Then we introduce a predictor based on a temporal-spatial score and propose a new loss function leveraging the temporal-spatial score as an optimization objective function to guide the updating direction of the model parameters to enhance the predictor's performance. In essence, we extend a disassembly sequence planning algorithm involving graphics and robotics to a VR disassembly sequence application. Experimental results show that our method outperforming the ASAP method in accuracy. We also designed a user study to evaluate the efficiency and usability of disassembly guidance based on the DSPT method. The results indicate that using our DSPT method for assisted disassembly significantly reduces the required time and improved the usability score compared to the ASAP-based approach.

Figure 1 illustrates a user disassembling an object using a disassembly guide based on the results predicted by DSPT. At each step, DSPT predicts disassembly probabilities and highlights the most likely component to guide the user effectively. This guidance mechanism effectively helps users intuitively understand the disassembly sequence in a VR environment. To see more examples of DSPT, please visit: `https://www.youtube.com/watch?v=WFlZnzhBhlo`.

In summary, the contributions of this paper are as follows: 1)We propose the Disassembly Sequence Planning Tranformer, DSPT, to give the guidance to users to disassemble the parts of objects in VR. 2)We define and construct Disassembly Sequence Features and Part History Features, based on which we compute temporal-spatial scores to quantify the influence of prior disassembly on the current timestep. 3)We propose a predictor that leverages temporal-spatial scores to estimate the disassembly probability of each part, along with a novel loss function designed to improve prediction performance using these scores.

The remainder of this paper is organized as follows: Section 2 reviews the related work on disassembly sequence planning and transformer-based method for sequence prediction. Section 3 introduces the proposed DSPT framework and its key components. Section 4 presents the experimental setup and evaluation results. Section 5 introduces the experimental design of the user study and the corresponding experimental results. Finally, Section 6 concludes the paper and outlines future research directions. Table 5 lists the acronyms appearing in this document and their corresponding full names.

2. Related Work

There has been a lot of work around DSP. We will provide a review of existing work in the following three dimensions: traditional DSP methods, deep learning-based DSP methods and Transformer-based sequence prediction methods. Table 1 summarizes the key characteristics of currently available disassembly sequence planning methods.

Table 1. Main characteristics of different DSP method type.

Method	Main Characteristics
Traditional method	Based on modeling, optimization and physic-simulation based approaches. Rely heavily on predefined disassembly rules tailored to specific products. Typically involve long perstep prediction times.
Learning method	The advantage of fast inference speed. But such models often suffer from low prediction accuracy and the generated sequences may not align with human operational habits in VR environments.
Ours	Our method is built upon a Transformer-based network architecture and is specifically designed to account for human operational habits and reduced cognitive load in VR environments. It achieves fast prediction speed while also providing improved accuracy compared with current state-of-the-art approaches.

2.1. Disassembly Sequence Planning

Traditional DSP methods can be divided into three categories: modeling approaches, optimization techniques, and physical simulation based approaches based on traditional 3D graphics. Modeling methods form the foundation of DSP, providing a structured representation of the product to be disassembled. Various modeling approaches have been proposed to capture the relationships between components, constraints, and disassembly operations(Behdad & Thurston, 2010, 2012; SONG, Hu, GAO, YANG, & Zhang, 2010). Graph-based models use nodes to represent components and edges to denote the precedence relationships between them. Kuo & Wang (2010) employed graph-based methods to describe the precedence relationships among components in a product. Min, Zhu, & Zhu (2010) proposed a weighted AND/OR graph to represent product structure and element constraints for disassembly planning. Guo, Liu, Zhou, & Tian (2017) used an AND/OR graph to represent product information for all disassembly sequences and optimized selective disassembly sequences using a scatter search algorithm. S.-e. Zhao, Li, Fu, & Yuan (2014) proposed a Disassembly Petri Net(DPN) to determine the optimal disassembly sequence based on cost and environmental benefits. Gunji et al. (2021) proposed a disassembly sequence planning method based on the stability graph cut set approach to address the problems of low disassembly efficiency and unclear component classification in end-of-life product recycling. Optimization techniques are essential for finding the best disassembly sequence that maximizes profit, minimizes cost, or achieves other objectives (Duta, Filip, & Popescu, 2008; Giudice & Fargione, 2007; Shimizu, Tsuji, & Nomura, 2007). Tripathi, Agrawal, Pandey, Shankar, & Tiwari (2009) proposed a fuzzy disassembly optimization model to maximize net revenue at the EOL disposal of a product, while Lambert & Gupta (2008) developed a heuristic algorithm for DSP with sequence-dependent costs. Xie, Huang, Zhong, & Kuang (2007) developed a hybrid GA and simulated annealing method to prevent premature convergence in DSP, while McGovern & Gupta (2007) presented an ACO-based approach to solve the DSP problem. W. Li, Xia, Gao, & Chao (2013) developed an SS algorithm for selective DSP, demonstrating its effectiveness in handling multi-resource constraints and optimizing disassembly profit. Bahubalendruni & Varupala (2021) proposed a two-level automatic disassembly sequence planning method based on CAD attributes and multi-matrix analysis, which addresses the issues of environmental pollution and resource waste caused by the lack of systematic disassembly schemes during the safe disposal of Waste Electrical and Electronic Equipment (WEEE). Anil Kumar, Bahubalendruni, Prasad, & Sankaranarayanasamy (2021) proposed a multi-level partial disassembly sequence planning (MDL) method to address the problems of excessive exposure to toxic substances, low disassembly efficiency, and the difficulty of balancing resource recovery with environmental protection during the disassembly of end-of-life products. Gulivindala, Bahubalendruni, P, & Eswaran (2023) proposed an environmental risk reduction-oriented disassembly sequence planning model (ECDSP), which enables the collaborative optimization of efficient recovery and harmless treatment within a predefined safety threshold. Bahubalendruni, Biswal, Kumar, & Nayak (2015) introduces an assembly predicate consideration method to optimize the Assembly Sequence Generation (ASG) process, addressing issues such as computational inefficiency and infeasible assembly sequences caused by traditional approaches that neglect critical assembly constraints. Physical simulation based disassembly sequence planning approach automatically plans a feasible disassembly sequence by attempting to disassemble the part from all directions through 3D graphical techniques and detecting collisions using a customized physical simulator (Tian et al., 2024, 2022; Zhu et al., 2024).

As noted in a number of reviews of DSP issues(Ghandi & Masehian, 2015; Guo, Zhou, Abusorrah, Alsokhiry, & Sedraoui, 2020; Ong, Chang, & Nee, 2021; Z. Zhou et al., 2019), traditional disassembly sequence planning methods usually rely on heuristic rules, search algorithms, optimization strategies, or physical simulation. In contrast, our method utilizes the powerful learning capabilities of neural networks to more accurately predict a reasonable disassembly sequence.

Traditional disassembly sequence planning methods based on heuristic rules, search strategies, or optimization algorithms often suffer from limited generalization capability and rely heavily on predefined disassembly rules tailored to specific products. Physics-based simulation approaches, while more realistic, typically involve long per-step prediction times, making them unsuitable for real-time interaction in virtual reality environments. In contrast, our method learns the underlying patterns of disassembly behavior directly from data, enabling it to generalize to previously unseen assemblies. Moreover, it offers fast single-step prediction, which makes it well suited for integration into VR environments where real-time guidance is essential.

2.2. Learning Method for DSP

In recent years, researchers have explored a variety of deep learning-based methods to improve the rationality of the disassembly sequence, reduce the disassembly cost, and enhance the adaptability to complex assemblies (Aslan et al., 2022; Sinanoğlu & Rıza Börklü, 2005). Reinforcement learning enables the system to gradually optimize the assembly strategy for a more efficient and flexible disassembly process by continuously trying different sequences and paths in the simulation environment. Parzeller, Koziol, Dagner, & Gerhard (2024) proposed an automated assembly sequence planning method based on reinforcement learning. The method automatically generates assembly sequences through a 'disassembly to assembly' strategy. Wang et al. Wang, Su, Sun, Chen, & Xie (2024) proposed the Object-Embodiment-Centric Imitation and Residual Reinforcement Learning (OEC-IRRL). M. Zhao, Guo, Zhang, Fang, & Ou (2020) proposed an assembly sequence planning system based on deep reinforcement learning (ASPW-DRL). Allagui et al. (2023) proposed a reinforcement learning approach based on the Q-Network (QN) algorithm to optimize the disassembly sequence planning.

Neural networks can help aid in establishing a deep mapping of part geometry to assembly relationships, laying the foundation for predictions in automated sequence planning. Zhu et al. (2024) proposed a multi-level reasoning framework incorporating the Part Assembly Sequence Transformer (PAST) for the automated assembly problem. PAST is a sequence-to-sequence neural network that recursively infers the assembly sequence from the target blueprint. Ma et al. (2023) proposed a Graph-Transformer framework based on heterogeneous graphs. This framework employs a heterogeneous graph attention network to encode LEGO models and utilizes an attention mechanism for decoding to generate the assembly sequence. Tian et al. (2024) proposed the ASAP method, which utilizes a graphical neural network (GNN) module to learn the physical feasibility of learning from a large amount of product assembly data and to predict the disassembly sequence of an assembly.

Compared with existing neural network-based approaches, our method explicitly accounts for the temporal dependencies in disassembly sequence planning. Additionally, our method incorporates spatial information about parts, guiding the network to select those that are easier for users to disassemble.

2.3. Transformer-based Method for Sequence Prediction

This section focuses on exploring the application of Transformer architectures in sequence prediction. Here, we cover not only sequence prediction involving disassembly and assembly but also general temporal forecasting tasks such as economic time series prediction.

In recent years, transformer-based sequence prediction models have made significant progress across multiple fields, becoming one of the mainstream methods for handling sequence data. The transformer architecture was first introduced by Vaswani et al. (2017), with its core innovation being the introduction of the self-attention mechanism. This mechanism allows the model to capture long-range dependencies within sequences without relying on recursive or convolutional structures. This characteristic gives the transformer a unique advantage in sequence prediction tasks.

In the field of sequence prediction, the transformer has demonstrated powerful modeling capabilities. Traditional time series forecasting model, such as Autoregressive Integrated Moving Average Model (ARIMA), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) , often face issues like low computational efficiency and gradient vanishing when handling long sequences. To address these problems, H. Zhou et al. (2021) proposed the Informer model, which significantly reduces the computational complexity of long-sequence predictions while maintaining high prediction accuracy. This is achieved through the design of a sparse self-attention mechanism which called ProbSparse Self-Attention and distillation operations. S. Li et al. (2019) introduced the Temporal Fusion Transformer (TFT), which combines temporal dynamic features with static features and uses a multi-level attention mechanism to efficiently model multivariate time series. TFT not only captures long-term dependencies in time series but also identifies important temporal patterns and external variables, achieving excellent performance on multiple real-world datasets. Kang & McAuley (2018) proposed the SASRec model, which uses a unidirectional transformer decoder to capture sequential dependencies in user behavior sequences, achieving state-of-the-art performance on several public datasets. Regarding automatic assembly and disassembly, a multi-level reasoning framework (Zhu et al., 2024) is proposed, which incorporates the Part Assembly Sequence Transformer (PAST) to address the automation of assembly tasks from parts to the target blueprint. A graph transformer framework based on heterogeneous graphs (Ma et al., 2023) is introduced to encode

LEGO models and generate assembly sequences using an attention mechanism for decoding.

Like these methods, our approach is also based on the architecture of transformer to utilize its powerful multi-attention mechanism. Building upon the Transformer architecture, we propose a definition and construction method for temporal-spatial scores, integrating them into our network to enhance its performance in disassembly sequence planning tasks.

3. Method

In interactive systems, high error rates in guidance undermine user confidence and increase cognitive load (Daronnat, Azzopardi, Halvey, & Dubiel, 2021). Frequent error feedback erodes trust in the system, causing users to question its capabilities during operation and leading to frustration. Secondly, high error rates compel users to expend additional cognitive resources on judgment, correction, and repetitive actions, significantly increasing external cognitive load. As cognitive load rises, users' task efficiency and accuracy decline, further compromising overall interaction quality. Therefore, we propose a Transformer-based approach capable of predicting object disassembly sequences with higher accuracy.

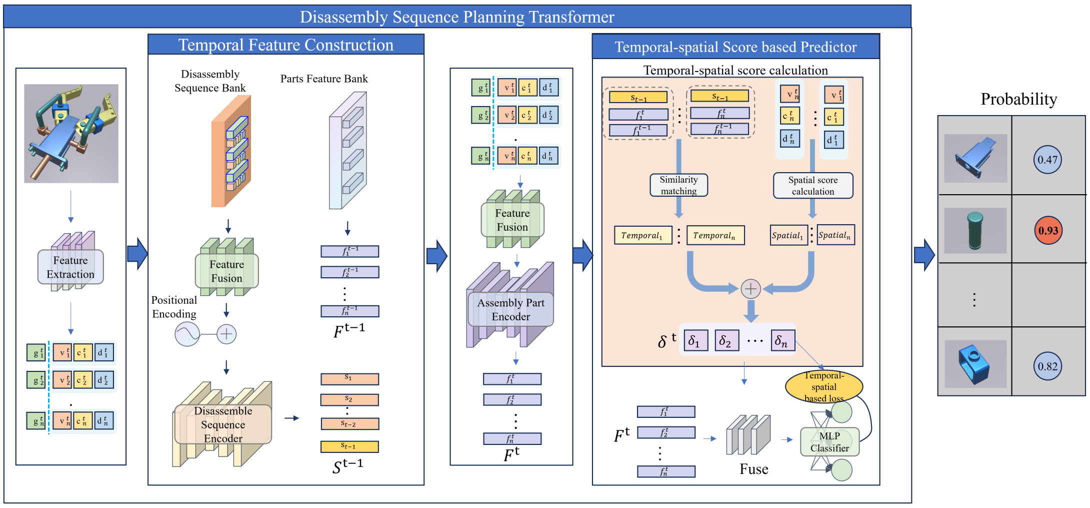


Figure 2. Network architecture diagram of DSPT. It primarily includes the construction of part features, the construction of temporal features, and a predictor based on temporal-spatial scores.

3.1. *Disassembly Sequence Planning Transformer*

We propose a novel network-based disassembly sequence prediction method, Disassembly Sequence Planning Transformer. This method leverages the powerful sequential learning capabilities of the transformer network architecture to model the historical and spatial relationships between parts during disassembly, optimizing the planning of the disassembly sequence. Our network takes the features of each part in the assembly as input and outputs the disassembly probability of each part in the current step. The part with the highest probability is selected and removed from the list of remaining parts. This iterative process continues until all parts are disassembled, ultimately generating a physically feasible and reasonable disassembly sequence. Figure 2 shows our network architecture.

First, we extract the features of each part under the current timestep. We adopt the same approach as ASAP (Tian et al., 2024) to extract geometric features (g^t), connection features (c^t) and distance features (d^t), which characterize the fundamental structure and constraints of the parts. We also introduce visibility features (v^t) to measure the visibility of parts from different viewpoints. Specifically, in a 3D space, we calculate the visibility ratio of each part from multiple viewpoints and take the average across all viewpoints to obtain the overall visibility feature. In this study, we select eight different viewpoints (front-top-left, front-top-right, back-top-left, back-top-right, front-bottom-left, front-bottom-right, back-bottom-left, and back-bottom-right) for visibility ratio computation. This multi-view visibility analysis enables the model to gain a more comprehensive understanding of part removability, thereby improving the rationality of the disassembly sequence planning. We denote the geometric feature of the part i to be disassembled at time t as g_i^t, the connection feature as c_i^t, the distance feature as d_i^t and the visibility feature as v_i^t.

After extracting each part's geometric features, connection features, distance features and visibility features at the current time step, our network first constructs the Disassembly Sequence Features and the Part History Features (Section 3.2). Next, the network fuses the geometric, connection, distance and visibility features of each part and feeds them into the Assembly Part Encoder to obtain the features of parts $F^t = \{f_1^t, f_2^t, ...f_n^t\}$ where f_i^t represents the feature vector of part i at moment t. Assembly Part Encoder is a conventional Transformer encoder designed to encode the integrated features of each component. During this process, we leverage the multi-head attention mechanism of the Transformer architecture, enabling parts to efficiently capture feature information from other parts. This allows for a more comprehensive understanding of the overall assembly structure and its internal dependencies. Subsequently, we employ the temporal-spatial score based predictor to estimate the disassembly probability of each remaining part at the current time step and select the part with the highest probability for removal (Section 3.3). Once the disassembly is executed, the system's memory module is updated by storing and adjusting the disassembly information from this step, along with the feature set of all remaining parts at the current time step (Section 3.2). This ensures that the model continuously learns and adapts to the evolving disassembly dynamics.

3.2. Temporal Feature Construction

This section primarily introduces the Temporal Feature Construction component in Figure 2. As shown in the Figure 2, this section has consists of 3 main parts: Disassembly Sequence Bank, Parts Feature Bank and Disassembly Sequence Encoder. These modules work in conjunction with each other to help the model effectively store and utilize the historical information of the disassembled parts to provide valuable guidance for the undisassembled parts in the subsequent disassembly steps, thus further optimizing the prediction process of the disassembly sequence.

The Disassembly Sequence Bank stores the geometric features, connection features, distance features and visibility features of the disassembled parts. These features reflect the historical information of the disassembled components. Meanwhile, the Parts Feature Bank stores the Part History Features F^{t-1}. (feature vectors of the parts to be disassembled from the previous time step). These feature vectors describe the state of each part before the current step, providing reference for the current disassembly decision.

We take the features of the disassembled parts from the Disassembly Sequence Bank and perform feature aggregation on these values to obtain a comprehensive feature for each disassembled part. After that, each part in the disassembled part sequence undergoes positional encoding to ensure the network can understand the order information of the parts in the disassembly sequence. The positional encoding is designed to encode each part's relative position or order in the disassembly process into the feature vector, enabling the network to capture the importance of the temporal sequence in the disassembly process. In this study, the classical positional coding formula proposed by Vaswani et al. (2017). was used:

$$PE_t^i = \begin{cases} \sin(\frac{1}{10000^{2k/d_{model}}}t) & \text{if } i = 2k \\ \cos(\frac{1}{10000^{2k/d_{model}}}t) & \text{if } i = 2k+1 \end{cases} \tag{1}$$

where t represents the position of the part in the disassembly sequence, which indicates the order of the current part in the disassembly process. d is the dimensionality of the positional encoding vector and i is the index of the element in the vector.

Then, the sequence of parts with positional encoding is passed through the Disassemble Sequence Encoder. The function of the Disassemble Sequence Encoder is to aggregate information from the historical disassembly steps and generate the Disassembly Sequence Feature $S_{t-1} = \{s_1, s_2, ..., s_{t-1}\}$ that contains the features of all previously disassembled parts. s_k represents the feature vector of part that was removed at the k-th time step. This Disassembly Sequence Feature (S_{t-1}) provides a global representation of the disassembled parts, capturing the features and historical information of all disassembled parts in the entire disassembly process.

After obtaining the Disassembly Sequence Features S_{t-1}, we select the feature vector (s_{t-1}) of the part disassembled at time $t-1$, and combine it with the Part History Features (F^{t-1}) stored in the Parts Feature Bank, as well as the feature vector of the part to be disassembled at time t calculated by the Assembly Part Encoder (F^t). These feature vectors are used as input to further compute the temporal-spatial score values for each part to be disassembled (detailed in Section 3.3). This process helps us to combine historical information with the current state to accurately predict the proper disassemble sequence.

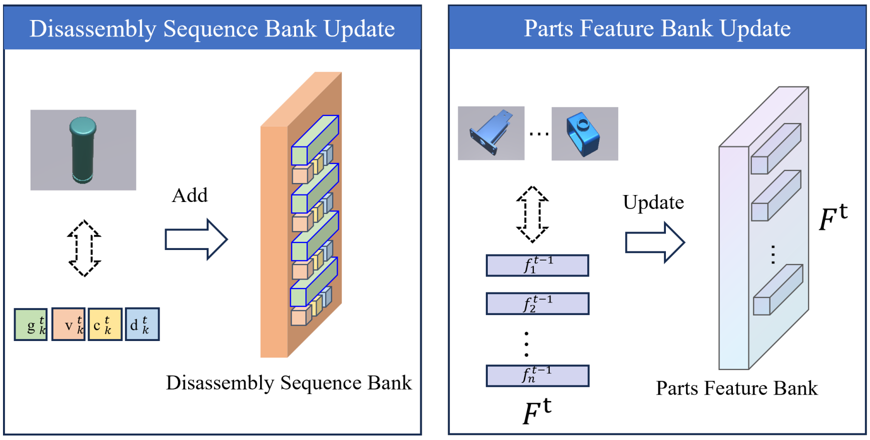

Disassembly Sequence Bank Update
Add
g_k^t v_k^t c_k^t d_k^t
Disassembly Sequence Bank
Parts Feature Bank Update
...
Update
f_1^{t-1}
f_2^{t-1}
f_n^{t-1}
F^t
F^t
Parts Feature Bank

Figure 3. Updated schematic of Disassembly Sequence Bank and Parts Feature Bank, which are updated at the end of each time step prediction

After each step of the disassembly prediction, we update the contents of the Parts Feature Bank and Disassembly Sequence Bank to provide new information for the next prediction. Specifically, as shown in Figure 3, after completing the current disassembly step, the features of the remaining components to be disassembled at time t (F^t) are updated in the Parts Feature Bank. At the same time, the geometry feature, connection feature, distance feature and visibility feature of the part selected for disassembly at time t are added to the Disassembly Sequence Bank. Through this updating process, the model continuously accumulates information from the disassembly process, enhancing its ability to remember historical disassembly steps. Through continuous updating, the Disassembly Sequence Bank and Parts Feature Bank provide a dynamic, evolving knowledge base for disassembly tasks, ensuring that the model will always be able to make more accurate predictions based on the latest historical data throughout the disassembly process.

3.3. Temporal-spatial Score Based Predictor

Temporal-spatial score based predictor is designed to predict the disassembly probability of each remaining part at the current time step based on the temporal-spatial score value of the part. This section primarily introduces the Temporal-spatial Score based Predictor component in Figure 2. The input to this module includes the previous-step disassembly feature s_{t-1} extracted from Disassembly Sequence Features, the Part History Features F_{t-1}, the current parts features F_t computed by the Assembly Part Encoder, the distance features d^t, the connection features c^t and the visibility features v^t. By integrating these inputs, the model can compute the disassembly probability of each part at the current time step.

Specifically, we compute the temporal-spatial score for each part based on s_{t-1}, F_{t-1}, F_t, d^t, c^t and v^t which will be introduced in Section 3.3.1. This score quantifies the priority of a part for disassembly at the current time step by incorporating historical information, geometric relationships and spatial constraints. Next, we fuse each part's temporal-spatial score with its feature vector and process it through an MLP network, ultimately predicting the disassembly probability for the current step.

During training, we optimize the model parameters using a temporal-spatial score based loss function, which enhances prediction accuracy through gradient backpropagation.

3.3.1. Temporal-spatial score calculation

In this section, we will provide a detailed introduction of how the temporal-spatial score values of parts are calculated. The temporal-spatial score is a composite score formed by combining the sequence history information of the disassembly sequence, the history feature information of the part to be disassembled and the spatial location information of the part to be disassembled during the disassembly process, with the purpose of helping the model to identify and predict the optimal disassembly sequence.

During the execution of disassembly tasks, the disassembly process of parts exhibits pronounced characteristics of continuity and dependency. These characteristics persist throughout the entire lifecycle of the disassembly operation and directly influence the scientific soundness of disassembly decisions as well as the efficiency of task execution. Specifically, the disassembly decision at any given moment does not exist in isolation; rather, it depends not only on the intrinsic properties of the target part but also on the dynamic contextual environment shaped by the sequence of parts that have already been removed.

For example, in classic assemblies involving screws or bolts, when the previous disassembly step involves removing fasteners like screws or bolts, the next step typically prioritizes removing other fasteners of the same type. This batch removal of similar parts helps reduce operational switching costs. Only after all fasteners are removed can previously constrained parts that were inaccessible for disassembly be further disassembled.

To precisely capture and quantify the inherent time dependencies within the disassembly process, thereby providing effective temporal context for modeling disassembly decisions, this paper introduces the concept of temporal score. The core function of this metric lies in measuring the dynamic relationship between parts awaiting disassembly and the sequence of already disassembled components, as well as the state evolution of the parts themselves. By quantifying this relationship, the temporal score provides an objective temporal evaluation criterion, supporting each step of the disassembly decision-making process.

The temporal score of the part i to be disassembled at time t can be calculated using the following formula:

$$temporal_i^t = \left|\text{cosine}(f_i^{t-1}, f_i^t) + \text{cosine}(s_{t-1}, f_i^t) - 1\right| \tag{2}$$

Here, $\text{cosine}(f_i^{t-1}, f_i^t)$ computes the similarity of the disassembly object i between two consecutive time steps, while $\text{cosine}(s_{t-1}, f_i^t)$ measures the similarity between each currently disassemblable part and the part disassembled in the previous step.

This formula is designed to maximize the score values for two categories of parts. The first category comprises parts with high similarity to their own historical features and high similarity to the features of the disassembly sequence. The core characteristics of these parts undergo no significant changes during disassembly and align closely with the features of the previously disassembled parts. Typical examples include fasteners of the same type. The second category comprises parts with low similarity to both the part's historical characteristics and the features of the disassembly sequence. These parts undergo significant changes in their feature states due to prior disassembly operations (such as fastener removal) and exhibit substantial differences from the features of the previously disassembled part. Typical examples include parts constrained by fasteners.

In addition to considering the historical continuity of disassembly, from the perspective of the rationality of human operation in a VR environment, we also need to consider the spatial layout of the parts. Specifically, we believe that in a disassembly task, parts that are farther from the center of the object and have fewer connecting edges should be prioritized for disassembly. This is because these parts are usually more independent, less likely to be interfered with by other parts during disassembly. Furthermore, their minimal contact with other components often implies that their removal will have limited impact on the overall stability of the assembly. Moreover, parts farther from the center are typically easier to handle, making them more suitable for disassembly first. Therefore, we define our spatial feature of part i as:

$$spatial_i^t = \frac{v_i^t \times d_i^t}{c_i^t \times c_i^t} \tag{3}$$

where v_i^t represents the visibility feature value of part i at time t, d_i^t represents the distance feature value of part i at time t, c_i^t represents the connection value of part i at time t.

This formula is designed to maximize parts that are farther from the assembly's center, have fewer connecting edges to other parts, and exhibit higher visibility.

Thus, the calculation formula for our temporal-spatial feature is:

$$\delta_i^t = \alpha \times temporal_i^t + \beta \times spatial_i^t \tag{4}$$

Here, δ_i^t represents the temporal-spatial score value of part i at time t, which is used to measure the importance or disassemblability of the part at the current disassembly step. This score value combines temporal information (the sequence of disassembly) and spatial information (the part's position, visibility, etc.), effectively guiding the disassembly process and making the disassembly sequence more rational and coherent. α and β are balanced weight.

This scoring mechanism fuses these complementary signals to generate disassembly sequences that better align with human disassembly habits. The temporal component integrates cumulative dependencies from prior steps, reflecting how human operators rely on progressively accumulated context to plan disassembly actions. The spatial component captures geometric feasibility to measure the difficulty of part removal.

3.3.2. Temporal-spatial score based loss

During the training process of this task, we introduce a regularization term into the loss function to guide the model to prioritize parts with higher temporal-spatial score values for disassembly. Specifically, the purpose of the regularization term is to constrain the model's loss function, encouraging the model to favor parts with higher temporal-spatial score when making disassembly decisions. Our loss function is computed as:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(-\left[y_i \log\left(\frac{1}{1+e^{-\hat{y}_i}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-\hat{y}_i}}\right)\right] + \gamma y_i \delta_i\right) \tag{5}$$

Here, y_i represents the ground truth label, $\hat{y}_i$ denotes the model's predicted value, and γ is the balancing weight that controls the contribution of the regularization term.

4. Experiments

4.1. Dataset for Disassembly Sequence Planning

In this experiment, we used the dataset released by Tian et al. (2024), which is sourced from the Fusion360 Gallery Assembly Dataset (Willis et al., 2022) and a real mechanical assembly dataset (Lupinetti, Giannini, Monti, & Pernot, 2019). The dataset consists of 2,146 assemblies, each comprising between 3 and over 50 components, and covers common object types found in industrial domains, such as car engines, water pumps, radios, propellers, and game controllers. The dataset is already divided into a training set and a test set, with the training set containing 1906 parts and the test set containing 240 parts. We split the training set by 9:1 into training data and validation data like others. We used the given physical simulation code by Tian et al. (2024) to generate the disassembly data for all removable parts at each step of the process for each assembly. This physical simulation code accounts for the influence of gravitational forces on the stability of each component and the entire assembly, so the network can learn and model the influence of gravity on both individual parts and the overall assembly stability.

4.2. Comparison Methods

To evaluate the performance of our proposed DSPT method, we compare it with four methods as well as a SOTA method. The four method are random method which generate disassembly sequence randomly, Assembly Them All method (Tian et al., 2022) which designs a special data structure to store randomly generated disassembly sequences, Heuristic-Volumes method (Tian et al., 2024) which is an explicit disassembly sequence prediction method based on part volumes, Heuristic-Distance method (Tian et al., 2024) which is an explicit disassembly sequence prediction method based on the distance between parts and the assembly center. The SOTA method is ASAP method (Tian et al., 2024) which uses a Graph Neural Network (GNN) to predict the disassembly sequence of an assembly. To ensure the accuracy of the evaluation results, we used physical simulation to compare the correctness of different disassembly sequences. The physical simulation takes into account several factors, such as the number of parts to be held, the stability of the objects, and the impact of the gravitational environment on the disassembly process. In our simulation, the number of parts to be held is set to the default value of 3.

4.3. Metrics

Sequence accuracy. Sequence accuracy measures the consistency between the disassembly sequence generated by the model and the feasible disassembly sequence. It is calculated as the ratio of the number of assemblies with successfully predicted complete disassembly sequences to the total number of assemblies. This metric reflects the model's accuracy in global disassembly planning and provides an intuitive evaluation of its overall performance in complete disassembly tasks. Successful prediction of our sequence is defined as all steps being predicted correctly; whenever any of these steps is wrong, the sequence is considered a prediction failure.

Stepwise accuracy. Stepwise accuracy focuses on the accuracy of the model's decision making at each disassembly step. If the model predicts a part that belongs to the feasible disassembly set at a given step, the prediction is considered correct. Stepwise accuracy is calculated as the ratio of the total number of parts correctly predicted by the model to the total number of parts in all test samples. Compared to sequence accuracy, stepwise accuracy provides a finer-grained evaluation of the model's stability and reliability in the step-by-step reasoning process. This metric helps analyze the model's local decision-making capability under different assembly structures and disassembly constraints.

4.4. Experimental Setup

In this experiment, model training and disassembly sequence planning were conducted on an NVIDIA RTX 4080 GPU and Intel i9-13900F CPUs with 64GB RAM. All modules utilize 1024 hidden neurons. Assembly Part Encoder and Disassemble Sequence Encoder are two independent Transformer encoders respectively implemented in PyTorch. Each Transformer encoder comprises 8 hidden layers. The multi-head attention mechanism employs 8 attention heads, ensuring the model can capture the spatial relationships and historical information of components from multiple perspectives.

Our model is trained by Adam optimizer. Training takes 12 hours for 150 epochs with learning rate of 0.0005. During training, we monitor the model's performance on the validation set and adjust model parameters based on validation loss to prevent overfitting and improve generalization capability.

During the prediction process, we input the model-generated predictions into a physics simulation. This simulation system accounts for the impact of gravitational forces on the stability of each component and the entire assembly. By simulating the disassembly process, we verify the correctness of the predictions. This approach effectively evaluates the feasibility of the model's predictions, ensuring their applicability in real-world disassembly tasks.

4.5. Comparison Result

4.5.1. Quantitative comparison

Our experimental results demonstrate that, compared to the ASAP method and other comparison methods, our proposed DSPT network model achieves significant performance improvements in the disassembly sequence planning task.

Table 2. Accuracy rate (%) comparison of DSPT on the test dataset against several baseline methods.

Method	Sequence Accuracy(%)	Stepwise Accuracy(%)
random	16.67%	74.84%
ATA	13.72%	79.21%
Heuristic-Volumes	47.29%	81.45%
Heuristic-Distance	49.01%	84.67%
ASAP	50.49%	82.30%
DSPT	**60.29%**	**87.68%**

As Table 2 shows, on the test set, our method achieved a sequence accuracy of 60.29% and a stepwise accuracy of 87.68%.

From the experimental results, it is evident that our method achieves significant improvements in disassembly sequence planning compared to the random and ATA methods. Specifically, compared to the random method, our sequence accuracy improves by 43.62%, and stepwise accuracy improves by 12.84%. Compared to the ATA method, our sequence accuracy improves by 46.57%, and stepwise accuracy improves by 8.47%. This improvement is primarily due to the fact that the random and ATA methods still rely on random selection and trial-based approaches, failing to fully utilize the structural information of the assembly to make reasonable disassembly plans. As a result, these methods exhibit clear shortcomings in maintaining a globally rational disassembly order.

When compared with Heuristic-Volumes, Heuristic-Distance and ASAP methods, in terms of stepwise accuracy, our method achieves improvements of 6.23%, 3.01%, and 5.38%, respectively. Compared with the Heuristic-Volumes and Heuristic-Distance methods, our network can learn more complex part constraint relationships and is more accurate in predicting disassembly at each step, resulting in a higher stepwise accuracy rate. Compared with the ASAP method, our method additionally considers the influence of historical disassembly information on the current prediction. It makes the prediction of disassembly sequences more reasonable, and better adapts to the structure of complex assemblies, providing clearer guidance for the disassembly process.

Although the stepwise improvement is relatively modest compared with the Heuristic-Volumes, Heuristic-Distance and ASAP methods, our method shows a significant advantage in sequence accuracy, with improvements of 13.00%, 11.28% and 9.8% over these methods. This is because the disassembly process usually involves a multi-step decision-making process where many of the disassembly steps are flexible. There may be more than one feasible sequence for disassembly of an assembly. However, overall sequence accuracy is often determined by a few key steps, incorrect decisions at these crucial steps can lead to cascading errors in subsequent steps, ultimately reducing sequence accuracy. Our model excels at predicting these critical steps, ensuring that the final disassembly sequence remains both reasonable and feasible.

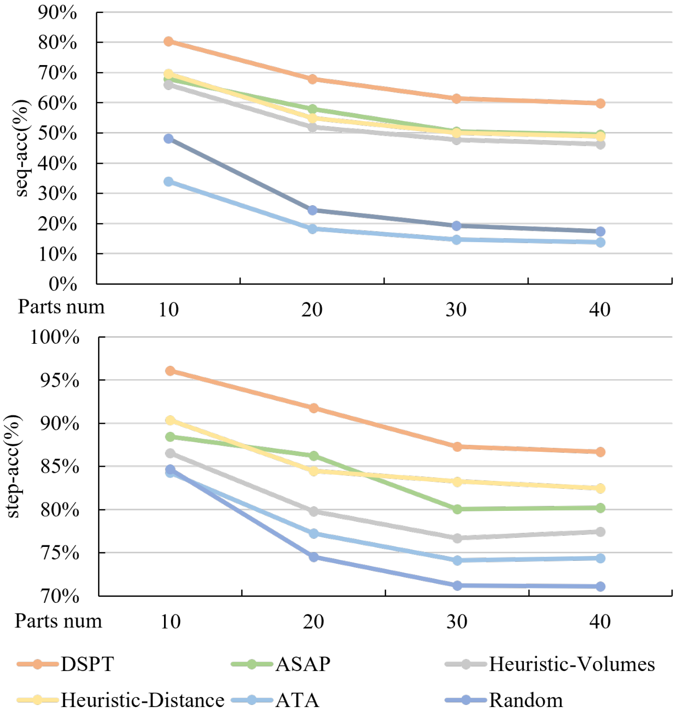


Figure 4. Line graphs of sequence accuracy and stepwise accuracy for assemblies with different numbers of parts.

Our experimental results also reveal the impact of the number of parts on the accuracy of the disassembly sequence and stepwise accuracy as Figure 4 shows. Specifically, although the disassembly sequence accuracy and stepwise accuracy tend to decrease as the number of assembled parts increases, both accuracy decreases tend to level off as the number of parts increases. The main reason for this trend is that complex assemblies have more connections and structural constraints, which increases the difficulty of inferring the disassembly sequence. Although the accuracy of the stepwise disassembly decisions remains relatively high, disassembly is a process of cumulative error, and small deviations in early decisions may be amplified in subsequent steps, thus affecting the final sequence accuracy. Moreover, regardless of the number of parts, our method always outperforms all the comparison methods in terms of sequence accuracy and stepwise accuracy.

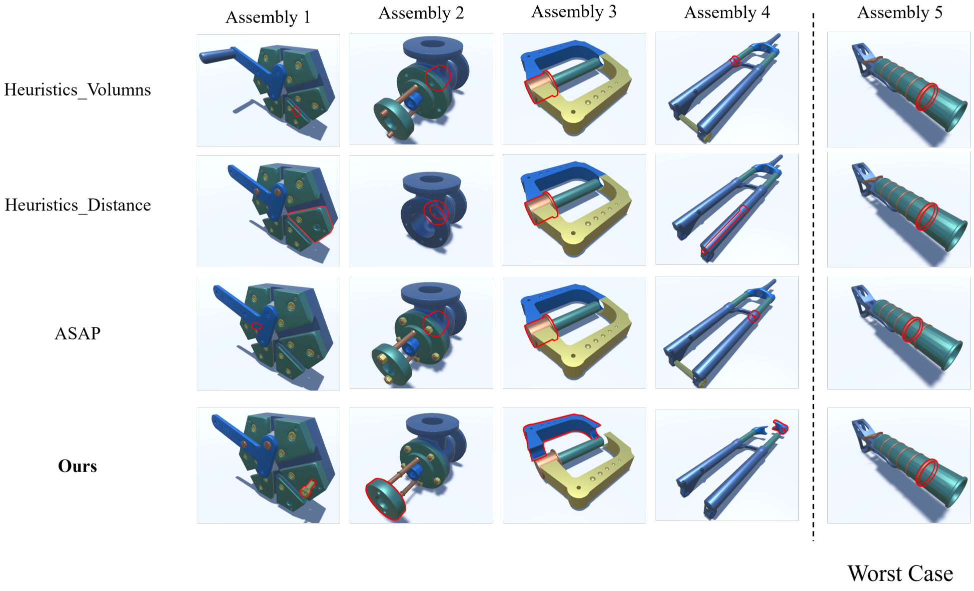


Figure 5. Disassembly process comparison between different methods. We highlighted the predicted part when the comparison methods made their first incorrect prediction during the disassembly process. For Assembly 1 to 4, the disassembly steps predicted by DSPT were all feasible. However, for Assembly 5, none of the methods were able to fully predict a feasible disassembly sequence correctly.

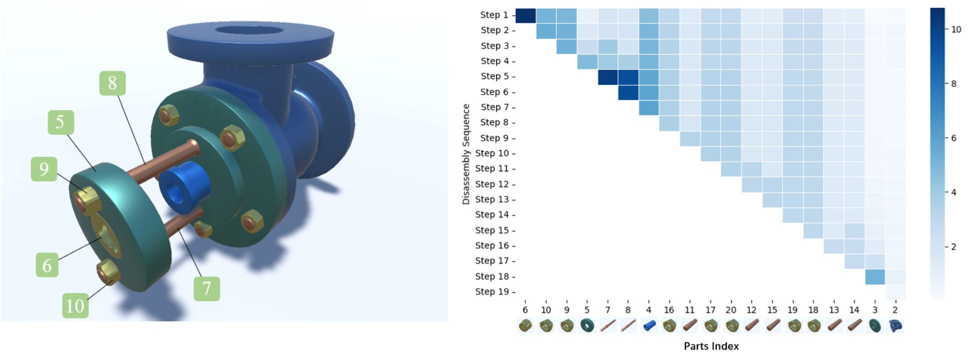
4.5.2. Visual comparison

We selected 5 assemblies for a visual analysis of the disassembly process as Figure 5 shows. Among them, 4 assemblies are cases where the comparison methods performed poorly on these assemblies, and the remaining assembly is a case where all methods (including ours) failed to predict accurately. We selected the moments in the prediction process where the comparison method made the first prediction error and the similar moments in the prediction sequence of our method for visualization.

As can be seen from Figure 5, these comparison methods have their own shortcomings in predicting the disassembly sequence. Heuristic-Volumes method prioritizes the smallest parts, leading to the prediction of the sequence of parts to be disassembled sorted by volume. However, in cases where small parts are constrained by large parts, the method predicts an unreasonable disassembly sequence. For Assembly 1, 2, and 4 in the figure, Heuristic-Volumes method incorrectly predicts parts that are surrounded by other large parts. For Assembly 3, Heuristic-Volumes method incorrectly predicts parts that are constrained by other large parts. Heuristic-Distance method always prioritizes the parts farthest from the center point. However, not all peripheral parts are prioritized for disassembly. The method predicts unreasonable parts to be disassembled in the case where parts far from the center are constrained by parts close to the center. For Assembly 2 and 4 in figure, Heuristic-Distance method incorrectly predicts parts that are surrounded by other parts. For assembly 1 and 3, Heuristic-Distance method incorrectly predicts parts that are constrained by other parts. ASAP method does not take into account the effect of visibility on disassembly. As a result, the predicted disassembled parts may be surrounded by other parts and cannot be disassembled during the prediction process of this method. For Assembly 1, 2, 3 and 4 in the figure, the ASAP method incorrectly predicts parts from being bounded by other parts.

In contrast, the DSPT method combines geometric features, distance features, connection features, and visibility features. At the same time, it takes into account the guidance of the disassembled sequence to the current disassembly, so that its predicted disassembly sequence is more consecutive and reasonable. For assembly 1, since the screws were mainly removed during the previous disassembly, DSPT prioritizes the continued removal of the screws based on the continuation of the disassembly pattern. For assembly 2, the DSPT predicts that the green disk will be preferred for disassembly because the screws holding it in place were removed in previous steps, and the part is farther away from the center of the assembly and has fewer connected edges, resulting in a higher temporal-spatial score. For assembly 3, the DSPT prefers to predict the blue part as the next disassembly target because the small ball placed on the blue part has been removed, while the part is farther away from the center of the assembly and has fewer connected edges. For assembly 4, the DSPT preferentially predicts the part for disassembly since the screws restricting the movement of the part have been removed in the previous time step, while the part is farther away from the center of the assembly and has fewer connecting edges, giving it a higher temporal-spatial score.

However, some assemblies remain challenging for all methods to predict, such as the Assembly 5 in the diagram. In this case, all methods incorrectly predicted a non-removable circular ring. But due to a small raised edge at the front of the enclosed object, the ring could not be directly removed. For all methods, it is difficult to accurately understand such fine geometric details, which highlights the limitations of the current methods in dealing with complex local geometric constraints.



(a)

(b)

Figure 6. (a) is the object we disassembled and some of its part indexes. (b) is visualization figure of temporal-spatial score in the disassembly process of the object. The y-axis represents the disassembly order, the x-axis represents part indexes, arranged from left to right according to the disassembly sequence.

Figure 6 is a visualization of the temporal-spatial scores during the disassembly of Assembly 2, where each cell corresponds to the temporal-spatial score of each part in each step (calculated from Equation 4). The depth of the color represents the size of the temporal-spatial score values, with darker colors indicating higher temporal-spatial score values. At each time step, the part selected for disassembly is usually the part with the higher temporal-spatial score for that step. In this disassembly process, part 6, part 10, and part 9 are adjacent to part 5, and they act as fasteners that limit the disassembly of part 5. As these three parts are gradually disassembled, we find that the temporal-spatial score of part 5 shows a significant upward trend. This is due to the gradual increase of its temporal score, indicating that the feasibility of disassembling this component increases with time. After the removal of part 10, the temporal-spatial score of part 7 also shows a significant increase. This is because part 10 is also a fastener that restricts part 7, and once removed, the temporal score of part 7 rises, leading to an increase in its overall temporal-spatial score. Part 9 and part 8 are similar.

4.6. Ablation Experiments

To verify the effectiveness of temporal-spatial score based prediction and the temporal-spatial score based loss function in disassembly sequence prediction, we perform ablation experiments by adding these two components step by step. First, we established a baseline model, Transformer + MLP (TM), where the Transformer serves as the feature encoder, and the MLP predicts the next part to be disassembled. In this setting, we trained the model using the Binary Cross Entropy Loss (BCELoss). Building upon this, we introduced the temporal-spatial score based predictor (TSS), allowing the model to incorporate the influence of previously disassembled components and thereby improving its ability to model disassembly sequences. In this setting, BCELoss remained the loss function. Finally, we incorporated the temporal-spatial score based loss function (TSSL), resulting in our complete model.

Table 3. Comparison of sequence accuracy and stepwise accuracy (%) for different model setups under the test set. △ denotes accuracy improvement compared to the baseline model.

Method	Sequence Accuracy(%)	Stepwise Accuracy(%)	△
TM	46.76%	79.50%	
$TM + TSS$	49.75%	85.54%	+2.99% / +5.64
$TM + TSS + TSS$L	60.29%	87.68%	+14.13% / +8.18%

As Table 3 shows, experimental results demonstrate that the baseline model achieves a certain level of effectiveness in disassembly sequence prediction but exhibits lower sequence accuracy due to the lack of historical information utilization. With the addition of the temporal-spatial score, both the stepwise accuracy and the overall sequence accuracy improved, indicating that incorporating historical information helps the model infer the next disassembly step more accurately. Further introducing the temporal-spatial score based loss function led to additional improvements in both sequence and stepwise accuracy, confirming the effectiveness of the proposed loss function. These findings suggest that this loss function effectively guides the model in optimizing disassembly sequences, making them more consistent with physical constraints and assembly logic, thereby enhancing the rationality and feasibility of the disassembly process.

5. User Study

We designed a user study to evaluate the effectiveness of the DSPT method for disassembly guidance in a VR environment.

We have developed a disassembly guidance system based on DSPT and ASAP methods (Figure 7).

In the disassembly guidance process, our disassembly sequence prediction model will predict the next probability of all parts of the object to be disassembled, and the system will highlight the one with the highest probability. Then, users can select it using the controller. Users can attempt to disassemble it using the controller, with the disassembly direction determined by the user. Once the part is successfully removed, the system highlights the next recommended part to be removed. If the user finds that the recommended part cannot be disassembled, the system allows them to independently select another removable part. After disassembling the component, the

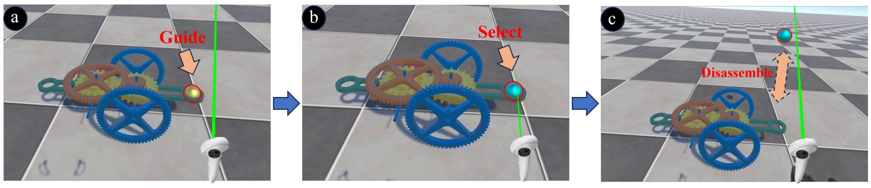


Figure 7. The virtual scenario of our disassembly assistance platform. The platform will highlight the components predicted by the model. Users can perform disassembly operations on this platform.

system recalculates the optimal disassembly sequence based on the updated assembly status and dynamically adjusts the subsequent disassembly guidance. This process continues until the user has successfully disassembled all components and completed the disassembly task.

In terms of prediction time per step, our method has an average prediction time of 1.73 seconds per step compared to 0.54 seconds for the ASAP method. Obviously, our method requires a longer prediction time compared to ASAP. The main reason for this difference is the image matching and alignment process required for the visibility feature computation, which increases the computational cost and leads to an increased time overhead. Since after the system highlight the guidance, the user needs to select and remove components. These operations take an estimated 3-6 seconds to complete, so we estimate that the average prediction time of our method is still within an acceptable range and does not have a significant impact on the overall user experience.

5.1. Participants and Apparatus

Sixteen participants (9 males, 7 females), aged 21 to 26 (mean 22.88, variance 1.36), took part in this experiment. None of them had experience with VR equipment. All participants had normal vision (or were corrected to normal vision by wearing glasses). The experimental system used a Unity 2022 and a Pico 4 headset. The system ran on the Pico 4 headset.

5.2. Task and Procedure

We randomly selected 8 assembly models and imported them into our self-developed disassembly assistance platform. Each model consists of more than 30 parts. Users must complete the disassembly of these eight assemblies according to the system's instructions. During the experiment, we conducted a comparative analysis between our proposed method and the ASAP method to evaluate their disassembly performance and user experience.

The average duration of the experiment was approximately 40 minutes per participant. Before starting, participants were required to complete a questionnaire covering personal information and experience with virtual reality head-mounted displays (VR HMDs) to assess their background. Next, we provided a detailed explanation of the experimental procedure and guided participants to stand at the center of the experimental area while wearing the VR HMD device. To ensure familiarity with the VR system and interaction operations, we allocated sufficient adaptation time before the formal experiment. Once the experiment began, participants followed the system's guidance to disassemble the current assembly model. After completing each disassembly task, a 1-minute break was given to reduce fatigue effects. To minimize order effects, we adopted a Latin square design to balance the order of Technique conditions, ensuring fairness and objectivity. During the experiment, we collected a total of 16 participants × 2 methods × 8 assembly models = 256 experimental datasets, providing a robust sample size for subsequent analysis. After completing the disassembly tasks for each method, participants filled out Usability questionnaire, followed by a brief rest period. At the end of the experiment, we conducted short interviews to gather participants' subjective feedback on both methods.

5.3. Metrics

Total disassembly time. The total time taken by users to complete all assembly disassembly tasks on the platform under different disassembly sequence planning methods. Different methods provide users with varying disassembly guidance based on their planned disassembly sequences, which in turn affects the overall efficiency of the disassembly process. By comparing the total time taken by these methods, we can assess the effectiveness of each method in actual disassembly tasks and analyze their impact on user operation efficiency.

Usability score. A usability questionnaire (Kim, Lee, & Billinghurst, 2015) to evaluate users' perceptions of technology usability, focusing on intuition, efficiency, accuracy, naturalness, satisfaction, and ease of use, scored from 1 to 10. The six questions are: is this method intuitive (Q1), is the method efficient (Q2), is the method accurate (Q3), is the method natural (Q4), is the method satisfied (Q5), is the method easy to use (Q6).

5.4. Result

In the analysis of the total disassembly time metric, we first identify and remove outliers for each condition where the selection time exceeds M ± 3 · SD, ensuring the validity and robustness of the data. Next, we perform a paired t-test to examine the significance of the difference in disassembly efficiency between the two methods. Additionally, we use Cohen's d (Sawilowsky, 2009) to calculate the effect size, quantifying the actual degree of difference between the two methods. The effect size d value is converted into qualitative categories, including: Huge ($d > 2.0$), Very Large ($2.0 > d > 1.2$), Large ($1.2 > d > 0.8$), Medium ($0.8 > d > 0.5$), Small ($0.5 > d > 0.2$), and Very Small ($0.2 > d > 0.01$). All statistical analyses are performed using SPSS software.

Table 4. Total disassembly time (s) comparison of DSPT against ASAP method. In this table, EC refers to the DSPT method and CC refers to the ASAP method.

Technique	Avg ± std. dev.	$(CC\text{-}EC)$ $/ CC$	p	Cohen's d	Effect size
$DSPT(EC)$	1038.50 ± 55.67				
$ASAP(CC)$	1246.46 ± 62.42	16.68%	$< 0.0001^*$	3.516	Huge

As Table 4 shows, the experimental results indicate a significant difference between the DSPT method and the ASAP method in terms of the total disassembly time metric ($p < 0.001^*$). Specifically, the DSPT method demonstrates a significant improvement over the ASAP method, indicating that the DSPT method has a clear advantage in reducing the time required for disassembly tasks.

Figure 8 shows Usability Scores. The t-test results indicate that the user ratings for each metric are significantly higher compared to the ASAP method ($p < 0.001^*$).

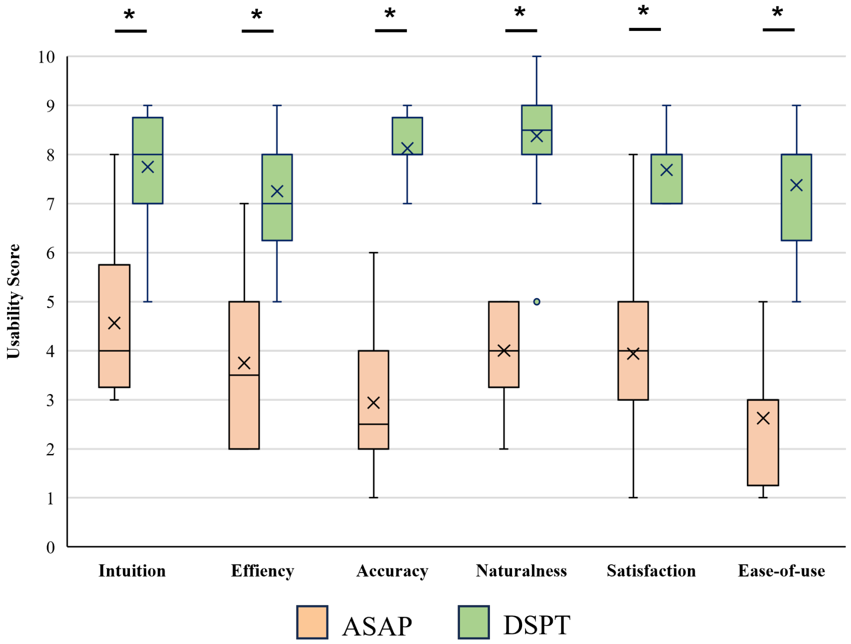


Figure 8. Usability scores for individual questions. Significant difference are denoted with the asterisk and line.

The data shows that users believe our system not only performs better in terms of task resolution but also provides a better user experience. This is due to the high prediction accuracy of the entire sequence, which provides a better user experience by reducing user frustration when the prediction fails. This is because our method achieves higher prediction accuracy, preventing users from repeatedly attempting to disassemble non-removable parts. This reduces cognitive load and enhances the disassembly experience. Furthermore, by incorporating spatial information about parts into sequence planning, our method prioritizes components located on the outermost layer of assemblies that have the least impact on overall structural stability. The resulting prediction sequences more closely align with the natural human disassembly patterns observed in VR environments.

5.5. Discussion

After compiling and analyzing the results of the user interviews, we found that the majority of users indicated that the disassembly sequence of the DSPT plan was more ergonomic and practical. This may be due to the fact that our model takes into account not only the temporal scores but also the spatial scores, combining them to get our temporal-spatial score values. Moreover, we use a loss function constrained based on temporal-spatial score during model training, which makes the model more inclined to select the outermost parts with fewer connections for disassembly. In contrast, the disassembly sequence planned by the ASAP method tends to involve parts that are located deep within the assembly and surrounded by multiple other parts. While disassembly can still be performed by the user, the process is often inconvenient and requires more maneuvering space and time. Additionally, the disassembly sequences planned by DSPT showed higher accuracy, allowing users to follow the planned sequence in most cases, thus reducing cognitive load. In comparison, the lower prediction accuracy of the ASAP method led users to spend more time and effort thinking about the next disassembly target, thereby increasing the complexity and inconvenience of the task. When we asked users if they felt inconvenienced by the DSPT method's average prediction time for the operational process, all users indicated that they didn't feel a significant delay. As we estimated, the average operation time of the user is about 7.65 seconds, which is much longer than 1.73 second, so the prediction time has less impact on the overall disassembly process and will not interrupt the user's interaction experience.

6. Conclusion

This paper proposes a novel transformer-based disassembly sequence planning network, DSPT (Disassembly Sequence Planning Transformer), designed to optimize interactive disassembly processes in VR environments. Unlike traditional rule-based methods, DSPT fully leverages the powerful modeling capabilities of deep learning and utilizes the multi-head attention mechanism to capture complex internal relationships within assemblies, enabling a more intelligent and dynamic disassembly sequence prediction. DSPT optimizes disassembly planning through feature encoding, historical information storage and retrieval and multi-head attention mechanism. Experimental results demonstrate that DSPT outperforms existing methods, such as the GNN-based ASAP approach, in terms of sequence accuracy and stepwise accuracy. Specifically, DSPT achieves a highest sequence accuracy of 60.29% and a highest stepwise accuracy of 87.68%, significantly surpassing other benchmark methods. Moreover, in VR interaction experiments, DSPT effectively reduces users' cognitive load and enhances the smoothness and operability of the disassembly process. The experimental results also validate that in the design of human-computer interaction systems, for guidance-oriented tasks, the accuracy of guidance significantly impacts the user's interaction experience. Incorrect guidance imposes greater cognitive load on users and diminishes their interaction experience.

The proposed DSPT framework has several practical implications for real-world applications. In industrial VR environments, it can be used to guide operators through complex disassembly tasks, providing step-by-step, context-aware visual instructions that reduce cognitive load and improve task efficiency. It also supports training scenarios, where users can practice disassembly procedures on virtual assemblies of varying complexity before performing real-world operations.

Our method still has certain limitations. First, the generalization ability of our model needs further improvement. We only test on objects with fewer parts. For highly intricate or irregular assemblies, the current feature encoding approach may not fully capture the complex relationships between parts. In the future, we aim to enhance DSPT's generalization ability by optimizing network structure to learn more complex connectivity relationships and constructing a more diverse and extensive dataset to provide richer information for model training.

Second, the inference time of the model needs optimization. Currently, DSPT calculates part visibility features using image matching methods, which incur significant computational overhead. In future work, we plan to refine the visual feature extraction method and optimize the computational process to accelerate feature extraction, thereby improving the model's inference efficiency in real-time applications.

Third, our approach does not consider disassembly path planning, so in the future we also plan to integrate disassembly path planning to consider not only the part disassembly order, but also kinematic constraints and environmental factors to generate more accurate disassembly paths. This improvement will provide users with more intuitive and efficient operation guidance in the virtual environment.

Fourth, the current method does not account for continuous oblique disassembly directions, which may limit the system's applicability in more complex geometric structures. In future work, we can extend the framework to support more refined or even continuous directional spaces, enabling this method to handle the planning of more complex assembly disassembly sequences.

Table 5. Full names and their corresponding acronyms appearing in this document.

Full Name	Acronyms
Disassembly Sequence Planning Transformer	DSPT
Automated Sequence Planning for Complex Robotic Assembly with Physical Feasibility	ASAP
Virtual Reality	VR
Assembly Them All	ATA
Disassembly Petri Net	DPN
Waste Electrical and Electronic Equipment	WEEE
Multi-level Partial Disassembly Sequence Planning	MDL
Environmental Risk Reduction-oriented Disassembly Sequence Planning Model	ECDSP
Assembly Sequence Generation	ASG
Object-Embodiment-Centric Imitation and Residual Reinforcement Learning	OEC-IRRL
Assembly Sequence Planning System Based on Deep Reinforcement Learning	ASPW-DRL
Q-Network	QN
Part Assembly Sequence Transformer	PAST
Graphical Neural Network	GNN
Autoregressive Integrated Moving Average Model	ARIMA
Recurrent Neural Network	RNN
Long Short-Term Memory	LSTM
Temporal Fusion Transformer	TFT

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China through Projects U25A20441, 62507002 and 62372026, the National Key R&D plan 2025YFE0203800, and the Fundamental Research Funds for the Central Universities.

References

Allagui, A., Belhadj, I., Plateaux, R., Hammadi, M., Penas, O., & Aifaoui, N. (2023). Reinforcement learning for disassembly sequence planning optimization. *Computers in Industry*, *151*, 103992.

Anil Kumar, G., Bahubalendruni, M. R., Prasad, V., & Sankaranarayanasamy, K. (2021). A multi-layered disassembly sequence planning method to support decision making in demanufacturing. *Sādhanā*, *46*(2), 102.

Aslan, O., Bolat, B., Bal, B., Tumer, T., Sahin, E., & Kalkan, S. (2022). Assemblerl: Learning to assemble furniture from their point clouds. In *2022 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 2748–2753).

Bahubalendruni, M. R., Biswal, B. B., Kumar, M., & Nayak, R. (2015). Influence of assembly predicate consideration on optimal assembly sequence generation. *Assembly Automation*, *35*(4), 309–316.

Bahubalendruni, M. R., & Varupala, V. P. (2021). Disassembly sequence planning for safe disposal of end-of-life waste electric and electronic equipment. *National Academy Science Letters*, *44*(3), 243–247.

Behdad, S., & Thurston, D. (2010). Disassembly process planning tradeoffs for product maintenance. In *International design engineering technical conferences and computers and information in engineering conference* (Vol. 44144, pp. 427–434).

Behdad, S., & Thurston, D. (2012). Disassembly and reassembly sequence planning tradeoffs under uncertainty for product maintenance.

Daronnat, S., Azzopardi, L., Halvey, M., & Dubiel, M. (2021). Inferring trust from users' behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration. *Frontiers in Robotics and AI*, *8*, 642201.

Duta, L., Filip, F. G., & Popescu, C. (2008). Evolutionary programming in disassembly decision making. *International journal of computers, communications & control*, *3*(3), 282–286.

Ghandi, S., & Masehian, E. (2015). Review and taxonomies of assembly and disassembly path planning problems and approaches. *Computer-Aided Design*, *67*, 58–86.

Giudice, F., & Fargione, G. (2007). Disassembly planning of mechanical systems for service and recovery: a genetic algorithms based approach. *Journal of Intelligent Manufacturing*, *18*, 313–329.

Gulivindala, A. K., Bahubalendruni, M. R., P, M. B., & Eswaran, M. (2023). Mechanical disassembly sequence planning for end-of-life products to maximize recyclability. *Sādhanā*, *48*(3), 109.

Gunji, B. M., Pabba, S. K., Rajaram, I. R. S., Sorakayala, P. S., Dubey, A., Deepak, B., ... Bahubalendruni, M. R. (2021). Optimal disassembly sequence generation and disposal of parts using stability graph cut-set method for end of life product. *Sādhanā*, *46*(1), 21.

Guo, X., Liu, S., Zhou, M., & Tian, G. (2017). Dual-objective program and scatter search for the optimization of disassembly sequences subject to multiresource constraints. *IEEE Transactions on Automation Science and Engineering*, *15*(3), 1091–1103.

Guo, X., Zhou, M., Abusorrah, A., Alsokhiry, F., & Sedraoui, K. (2020). Disassembly sequence planning: a survey. *IEEE/CAA Journal of Automatica Sinica*, *8*(7), 1308–1324.

Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. In *2018 ieee international conference on data mining (icdm)* (pp. 197–206).

Kim, H., Lee, G., & Billinghurst, M. (2015). A non-linear mapping technique for bare-hand interaction in large virtual environments. In *Proceedings of the annual meeting of the australian special interest group for computer human interaction* (pp. 53–61).

Kuo, T. C., & Wang, M.-L. (2010). Waste electronics and electrical equipment disassembly and recycling using petri net analysis. In *The 40th international conference on computers & indutrial engineering* (pp. 1–6).

Lambert, A., & Gupta, S. M. (2008). Methods for optimum and near optimum disassembly sequencing. *International Journal of Production Research*, *46*(11), 2845–2865.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, *32*.

Li, W., Xia, K., Gao, L., & Chao, K.-M. (2013). Selective disassembly planning for waste electrical and electronic equipment with case studies on liquid crystaldisplays. *Robotics and Computer-Integrated Manufacturing*, *29*(4), 248–260.

Lupinetti, K., Giannini, F., Monti, M., & Pernot, J.-P. (2019). Content-based multi-criteria similarity assessment of cad assembly models. *Computers in Industry*, *112*, 103111.

Ma, L., Gong, J., Xu, H., Chen, H., Zhao, H., Huang, W., & Zhou, G. (2023). Planning assembly sequence with graph transformer. In *2023 ieee international conference on robotics and automation (icra)* (pp. 12395–12401).

McGovern, S. M., & Gupta, S. M. (2007). Benchmark data set for evaluation of line balancing algorithms. *IFAC Proceedings Volumes*, *40*(2), 48–53.

Min, S., Zhu, X., & Zhu, X. (2010). Research on disassembly and/or graph construction and uncertain weight. *Chinese J. Eng. Des*, *17*, 19–24.

Ong, S.-K., Chang, M. M. L., & Nee, A. Y. (2021). Product disassembly sequence planning: state-of-the-art, challenges, opportunities and future directions. *International Journal of Production Research*, *59*(11), 3493–3508.

Parzeller, R., Koziol, D., Dagner, T., & Gerhard, D. (2024). Automating the assembly planning process to enable design for assembly using reinforcement learning. *Proceedings of the Design Society*, *4*, 2179–2186.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of modern applied statistical methods*, *8*(2), 26.

Shimizu, Y., Tsuji, K., & Nomura, M. (2007). Optimal disassembly sequence generation using a genetic programming. *International Journal of Production Research*, *45*(18-19), 4537–4554.

Sinanoğlu, C., & Rıza Börklü, H. (2005). An assembly sequence-planning system for mechanical parts using neural network. *Assembly Automation*, *25*(1), 38–52.

SONG, S.-H., Hu, D., GAO, X., YANG, M., & Zhang, L. (2010). Product disassembly sequence planning based on constraint satisfaction problems. *China Mechanical Engineering*, *21*(17), 2058.

Tian, Y., Willis, K. D., Al Omari, B., Luo, J., Ma, P., Li, Y., ... others (2024). Asap: Automated sequence planning for complex robotic assembly with physical feasibility. In *2024 ieee international conference on robotics and automation (icra)* (pp. 4380–4386).

Tian, Y., Xu, J., Li, Y., Luo, J., Sueda, S., Li, H., ... Matusik, W. (2022). Assemble them all: Physics-based planning for generalizable assembly by disassembly. *ACM Transactions on Graphics (TOG)*, *41*(6), 1–11.

Tripathi, M., Agrawal, S., Pandey, M. K., Shankar, R., & Tiwari, M. (2009). Real world disassembly modeling and sequencing problem: Optimization by algorithm of self-guided ants (asga). *Robotics and Computer-Integrated Manufacturing*, *25*(3), 483–496.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, C., Su, C., Sun, B., Chen, G., & Xie, L. (2024). Extended residual learning with one-shot imitation learning for robotic assembly in semi-structured environment. *Frontiers in Neurorobotics*, *18*, 1355170.

Willis, K. D., Jayaraman, P. K., Chu, H., Tian, Y., Li, Y., Grandi, D., ... others (2022). Joinable: Learning bottom-up assembly of parametric cad joints. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 15849–15860).

Xie, Y., Huang, M., Zhong, Y., & Kuang, B. (2007). Disassembly sequence planning based on the simulated annealing and genetic algorithm. *Mechanical Engineer*, *1*, 44.

Zhao, M., Guo, X., Zhang, X., Fang, Y., & Ou, Y. (2020). Aspw-drl: assembly sequence planning for workpieces via a deep reinforcement learning approach. *Assembly Automation*, *40*(1), 65–75.

Zhao, S.-e., Li, Y.-l., Fu, R., & Yuan, W. (2014). Fuzzy reasoning petri nets and its application to disassembly sequence decision-making for the end-of-life product recycling and remanufacturing. *International Journal of Computer Integrated Manufacturing*, *27*(5), 415–421.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 11106–11115).

Zhou, Z., Liu, J., Pham, D. T., Xu, W., Ramirez, F. J., Ji, C., & Liu, Q. (2019). Disassembly sequence planning: Recent developments and future trends. *Proceedings of the Institution* *of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *233*(5), 1450–1471.

Zhu, X., Jha, D. K., Romeres, D., Sun, L., Tomizuka, M., & Cherian, A. (2024). Multi-level reasoning for robotic assembly: From sequence inference to contact selection. In *2024 ieee international conference on robotics and automation (icra)* (pp. 816–823).

About the Authors

Sichun Huang is a master student in the School of Computer Science and Engineering of Beihang University, China. His current research focuses on virtual reality, augmented reality, and HCI.

Ziteng Wang is a Ph.D student at the School of Computer Science and Engineering, Beihang University, China. His current research focuses on virtual reality and augmented reality.

Sio Kei Im received his Ph.D. degree in Electronic Engineering from Queen Mary University of London (QMUL), United Kingdom. He is a professor at the Faculty of Applied Sciences, Macau Polytechnic University, and a researcher at the Engineering Research Center of Applied Technology on Machine Translation and Artificial Intelligence, Ministry of Education. His research interests include video coding, image processing, machine learning for NLP and multimedia.

Lili Wang received her Ph.D. degree from the Beihang University, Beijing, China. She is a professor with the School of Computer Science and Engineering of Beihang University, and a researcher with the State Key Laboratory of Virtual Reality Technology and Systems. Her interests include virtual reality, augmented reality, mixed reality, real-time rendering and realistic rendering.