IIEEJ Transactions on

Image Electronics and Visual Computing

Special Issue on
Image Electronics Technologies Related to VR/AR/MR/XR

Vol. 13, No. 1 2025

The Institute of Image Electronics Engineers of Japan

Editorial Committee of IIEEJ

Contact Information

The Institute of Image Electronics Engineers of Japan (IIEEJ)
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan
Tel : +81-3-5615-2893　Fax : +81-3-5615-2894
E-mail : hensyu@iieej.org
http://www.iieej.org/ (in Japanese)
http://www.iieej.org/en/ (in English)
http://www.facebook.com/IIEEJ (in Japanese)
http://www.facebook.com/IIEEJ.E (in English)

IIEEJ Transactions on
Image Electronics and Visual Computing
Vol.13 No.1 June 2025

CONTENTS

Special Issue on Image Electronics Technologies Related to VR/AR/MR/XR

1 Upon the Special Issue on Image Electronics Technologies Related to VR/AR/MR/XR — Takafumi KOIKE

Contributed Papers

2 W-Attention Net for Lung Airway Precise Segmentation — Liang LYU, Jiaqing LIU, Shurong CHAI, Fang WANG, Tomoko TATEYAMA, Xu QIAO, Yen-Wei CHEN

14 Material Appearance Reproduction via GAN-Based Novel View Synthesis Considering the Naturalness of Intensity Changes in Viewpoint Transitions — Taishi IRIYAMA, Chihiro HOSHIZAWA, Takashi KOMURO

System development paper

25 Crossing over Virtual and Real Experiences through Digital Twin Partner Robot — Ryota SUZUKI, Yuto ISHIYAMA, Yoshinori KOBAYASHI

Regular Section

Contributed Papers

33 Basic Design and its Performance of Multiplication-Free Multi-Alphabet Arithmetic Code for Markov-Model Sources — Fumitaka ONO, Kazuto KAMIKURA, Yousun KANG

42 Online Hand Drawing Pattern Classification Using Sketch-RNN — Shione ISHIDA, Kyoko SUDO

Short Paper

50 Interactive Bayesian Optimization of Level of Abstraction for Stylized Image Composition — Ryoma HASHIMOTO, Yoshinori DOBASHI

Announcements

55 Call for Papers : Special Issue on Image Electronics Technologies Related to AI

56 First Call for Papers：The 9th IIEEJ International Conference on Image Electronics and Visual Computing 2026 (IEVC2026)

Guide for Authors

58 Guidance for Paper Submission

Published two times a year by the Institute of Image Electronics Engineers of Japan（IIEEJ）
3-35-4-101，Arakawa，Arakawa-ku，Tokyo 116-0002，Japan
Tel：+81-5615-2893 Fax：+81-5615-2894 E-mail：hensyu@iieej.org http://www.iieej.org/

Upon the Special Issue on Image Electronics Technologies Related to VR/AR/MR/XR

Editor: Takafumi KOIKE
(Hosei University)

In recent years, Virtual Reality (VR), Augmented Reality (AR), Mixed Reality (MR), Extended/Cross Reality (XR) have been applied to systems of various fields, and have brought revolutionary innovation in them. The applied area includes Entertainment, Medical & Health Care, Education, Practice & Disciplining, Manufacturing, Remote Work & Meeting, and so on, and has been extended even for the purposes of Art and Environment Preservation. Especially, VR technology will provide fully virtual environment and invite the users to another world in order to experience unfeasible experiments. Also, AR technology will improve the efficiency of daily task by shortening the access time to the desired information, and will empower the process to understand the individual utilized context at the same time. On the other hand, the long-time effect of such technologies on human senses and mental states has been gathering wide interest, and how to coexist with such technologies will be the future big theme for human beings.

Based on the above observation, Trans. on IE and VC has planned the special issue on “Image Electronics Technologies Related to VR/AR/MR/XR ”, targeting on June 2025 issue.

In this special issue, two ordinary papers and one system development paper, accepted within the publication schedule of June issue, are contained. The first topic is “W-Attention Net for Lung Airway Segmentation”, which proposes an accurate segmentation method of pulmonary airways including bronchi, and develops a mixed-reality system to enhance clinical and educational visualization of airways. The second topic is “Material Appearance Reproduction via GAN-Based Novel View Synthesis Considering the Naturalness of Intensity Changes in Viewpoint Transitions”, which proposes a GAN-based novel view synthesis method that focuses on generating naturally varying viewpoint-dependent material appearances during viewpoint transitions. The last topic is “Crossing over Virtual and Real Experiences through Digital Twin Partner Robot”, which aims to enhance the user experience through collaboration with a digital twin partner robot to connect between VR experience and real experience.

We believe this special issue will contribute to stimulate the eagerness of this field and also the related activities. By the way, five papers on this special issue written in Japanese were already published in April 2025 issue of the Journal of IIEEJ.

Last but not least, we would like to thank all the reviewers and editors for their contribution to improving the quality of papers. We would also like to express our deepest gratitude to the members of the editorial committee of IIEEJ and the staff at the IIEEJ office for various kinds of support.

IIEEJ **Paper** -- *Special Issue on Image Electronics Technologies Related to VR/AR/MR/XR* --

W-Attention Net for Lung Airway Precise Segmentation

Liang LYU†, Jiaqing LIU†, Shurong CHAI†, Fang WANG††,
Tomoko TATEYAMA†††, Xu QIAO††††, Yen-Wei CHEN† (*Member*)

† Graduate School of Information Science and Engineering, Ritsumeikan University,
†† Department of Radiology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University,
††† Faculty of Engineering, University of the Ryukyus,
†††† School of Control Science and Engineering, Shandong University

<**Summary**> The lungs are essential to human health and respiration, yet lung diseases cause significant morbidity and mortality worldwide. Accurate segmentation of pulmonary airways, including bronchi, is crucial for diagnosis—especially underscored by the COVID-19 pandemic. Automated segmentation from CT scans is challenging due to the complex, tree-like airway structure and fine details at terminal bronchioles. To address this, we propose W-Attention Net, a dual-encoder model that integrates CNNs and Swin Transformers via cross-attention, capturing both local and global features for fine-grained airway segmentation. Our method outperforms existing approaches on a private dataset from Shandong University and the public LIDC-IDRI dataset. In addition, we developed a mixed-reality system with Microsoft HoloLens 2 to enhance clinical and educational visualization of airways.

Keywords: medical image, 3D segmentation, lung airway

1. Introduction

As an important respiratory organ, lung diseases such as pulmonary fibrosis and acute respiratory distress syndrome cause symptoms such as dyspnea, posing a potential threat to human health. Meanwhile, as a diagnostic tool widely used in hospitals, computed tomography (CT) has been widely adopted to aid diagnosis. However, for lung diseases, a prerequisite is to extract the lung airways from the CT images. Because most lung diseases are inextricably linked to the lung airways, the shape of the lung airways can also be used to help diagnose the disease. To be more specific, structural alterations of the airways often correlate with a range of chronic lung diseases, including asthma, cystic fibrosis (CF), chronic obstructive pulmonary disease (COPD), and coronavirus disease (COVID-19). So modelling of the airway tree facilitates the quantification of its morphological changes for the diagnosis of bronchial stenosis, acute respiratory distress syndrome, idiopathic pulmonary fibrosis, chronic obstructive pulmonary disease (COPD), occlusive bronchiectasis and pulmonary contusions.[1)–5)] Therefore, a method for accurate extraction of the tracheal and bronchial tree of the lungs from CT images, which is also known as accurate segmentation of the airway tree, is of great significance in helping physicians to diagnose lung diseases. Unfortunately, the airways of the lungs have a unique, tree-like structure that is thick in the main trunk portion and thin in the marginal portion. Airway segmentation remains a difficult task due to the high complexity of the airway structure and the inconsistency of the borders. In addition to this, intuitive and vivid visualization of the lung airways is equally important. Nowadays, deep neural networks have become a common solution for biomedical image segmentation, such as the well-known U-Net[6)]and 3D U-Net[7)]. Based on the high modeling capability of Convolutional Neural Networks (CNNs), researchers have made significant progress in airway segmentation. However, these methods still face challenges in achieving high continuity and integrity in airway prediction. They may not extract spatial and structural information of the entire bronchial tree well, and fail to preserve detailed structural information of the airway. Further, traditional CNNs lack the ability to extract global features, leading to poor performance. Besides, self-attention-based Transformer[8)]networks have shown great potential in the field of computer vision (CV) and have achieved the state-of-the-art results in image segmentation. One example is Swinney[9)], which replaces the convolution and pooling operations in the 3D U-Net[7)]

network structure with Transformer networks, achieving competitive results by capturing long-range dependencies between global features through self-attention mechanism. However, one drawback of such Transformer-based networks is that they require a large amount of data or pre-trained weights to ensure result accuracy.

In this paper, we introduce a novel dual-encoder model with multi-cross attention for lung airway segmentation to address challenges in medical image segmentation. Utilizing two encoder modules, we analyze the contour shape, intensity distribution, and connectivity of bronchi and vessels in a data-driven manner. Our 3D segmentation network, inspired by 3D U-Net[7]and Swin-Unet[9] , effectively overcomes challenges in identifying long and thin tubes, demonstrating high sensitivity for bronchioles, small arteries, and veins. Using a CNN and transformer for feature extraction, our method successfully segments the airway and bronchial tree in lung CT datasets from Shandong University and LIDC-IDRI[10]. Experimental results show that the network with multi-cross attention for feature fusion effectively segments the airway, retaining both the detailed features at the ends and the overall structure. Our approach achieves the best results on both datasets. The main contributions of this paper are as follows:

1. We propose a novel 3D lung airway segmentation system based on dual encoders and cross-attention mechanisms, leveraging shape features of the lung airway, achieving high accuracy without pre-training.
2. We design a 3D medical image segmentation network based on 3D U-Net[7]and Swin-Unet[9], successfully addressing the challenges of identifying long, thin tubular structures, showing high sensitivity to bronchioles, small arteries, and veins.
3. To utilize dual encoders, we introduce a module using dual encoders and multi-cross attention for feature fusion, allowing interaction between CNN and transformer-extracted information, preserving detailed features at the bronchi ends and maintaining overall structure, achieving state-of-the-art performance on both datasets.
4. We additionally implemented a prototype 3D interactive visualization system based on Microsoft HoloLens to explore the potential application of our segmentation results. A brief description is provided in Appendix A.

A preliminary version of this work was presented as a four-page conference paper at the 2023 IEEE Global Conference on Consumer Electronics (GCCE)[11]. The present draft involves both substantial conceptual, method improvement, and experimental extensions including the following: (1) We add the novel multi-cross attention fusion model to get better feature exchanges with two encoder. (2) We integrated the new multi-cross attention fusion model with the previous model and performed ablation studies to demonstrate the effectiveness of our new model. (3) We add another dataset to evaluate our segmentation model. (4) We have created a system based on Microsoft Hololens to interact with 3D models. (5) We have added a new subjective experiment and proposed a method to evaluate our system.

The structure of this paper is as follows. Chapter 2 introduces the related work used in lung part segmentation. Chapter 3 introduces our proposed novel system, including a dual encoder based segmentation network and the dual encoder fusion module. Chapter 4 discusses the experiments conducted details in this study and their results. Chapter 5 gives conclusions and summarizes this paper.

2. Related Work

Over the past decades, there have been various methods proposed for lung segmentation in medical imaging, especially in computed tomography (CT) scans. Lung segmentation plays a crucial role in many applications, including computer-aided diagnosis, lung disease analysis, and treatment planning. Here is a review of some commonly used lung segmentation methods:

1. Thresholding-based methods: These methods utilize intensity thresholding techniques to separate lung tissue from the surrounding structures based on Hounsfield unit (HU) values.[12]The most straightforward approach is global thresholding, where a fixed threshold is applied to the CT image. However, this method may not be robust in the presence of intensity variations and lesions. To address this, adaptive thresholding techniques, such as Otsu's method or local thresholding[13], can be employed to handle intensity variations within the lung region. 2. Region-growing methods: Region-growing algorithms start from a seed point within the lung region and iteratively expand the region by incorporating neighbouring voxels based on certain similarity criteria. These methods typically use intensity or gradient-based similarity measures to determine the lung boundaries.[14]
3. Graph-based methods: Graph-cut algorithms formu-

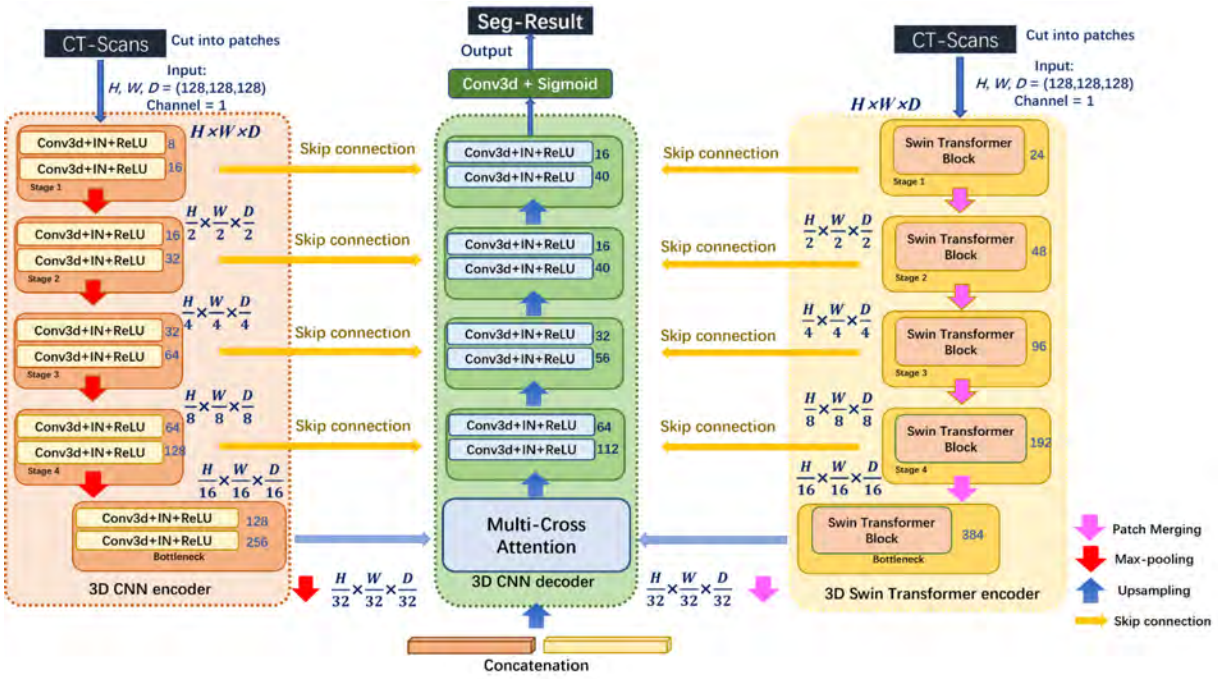


Fig. 1 The detailed structure of w-Attention net

late lung segmentation as an energy minimization problem. They construct a graph, where nodes represent image elements, and edges encode the similarity between adjacent elements. The energy function combines both data fidelity terms and smoothness priors. By solving the graph-cut optimization, the lung region can be accurately segmented. Graph-cut-based methods can handle intensity variations and leaking issues, but they may be computationally expensive[15)–18)]. 4. Machine learning-based methods: With the advancement of deep learning techniques, convolutional neural networks (CNNs)[19)–21)] and transformer-based network[9),22)–24)]have been widely applied to lung segmentation. These methods train models on large annotated datasets to learn the lung appearance and spatial context. Once trained, the models can segment lungs in unseen images. CNN-based methods have shown promising results in terms of accuracy and robustness but require a substantial amount of marked data for training. Inspired by the achievements of sequential dual attention (Ding et al., 2022)[25)]and channel cross-attention (Wang et al., 2022)[26)], we introduce Multi Cross-Attention, an attention module designed to efficiently extract both local and global information. By integrating features from both 3D CNNs encoder and 3D Swin Transformer encoder, our module aims to address intricate details such as the bronchi distal notch. Our primary objective is to devise a fusion mechanism that effectively connects the dual encoder and decoder, achieving enhanced performance with only a minimal increase in parameters.

3. Proposed Method

3.1 Overview of W-Attention net

The primary goal of our method is to achieve high degree of accuracy and robust segmentation of lung airways, especially the small and peripheral branches that are often missed by conventional approaches. To this end, we design a novel architecture, W-Attention Net, that integrates local and global features through a dual-encoders structure and a bidirectional cross-attention mechanism.

The 3D CNN encoder is responsible for capturing fine-grained local details, while the 3D Swin Transformer encoder models long-range anatomical dependencies. These complementary representations are fused using a multi-cross attention fusion module, which enables effective information exchange between the two encoders. This design is tailored to handle the complex, tree-like morphology of the bronchial structure in volumetric CT data.

Inspired by the 3D U-Net[7)], Swin-UNet[9)], and Swin UNETR[27)], we propose a novel 3D segmentation network with two encoders and one decoder to address the aforementioned challenges. We employ a CNN and a Transformer in the encoder stage to extract local and global features, respectively.

To obtain better segmentation of lung airways, we propose a new network called W-Attention Net. As shown in **Fig. 1**, our method takes 3D CT images as input. The ar-

chitecture resembles the shape of the letter "W", consisting of two encoders and one decoder. The 3D CNN-based encoder performs two convolution operations followed by a max pooling layer at each stage to extract local features. The 3D Swin Transformer[28]-based encoder captures global features through a self-attention mechanism.

In the decoder stage, feature maps generated by the two encoders are concatenated and upsampled using transposed convolutions. Skip connections are added between corresponding levels of the encoders and the decoder to preserve spatial information. Finally, a $1 \times 1 \times 1$ convolution followed by a sigmoid activation function is applied to generate the final segmentation output.

After determining the basic structure of the network, we introduce a Multi-Cross Attention Module to allow bidirectional information fusion between the two encoder branches. This mechanism enhances the overall representational power of the network and contributes significantly to its high segmentation accuracy.

3.2 3D CNN encoder

We first use CNN to extract local features. The CNN encoder stacks multiple layers and each layer contains two consecutive blocks. Each block includes three operations. Firstly, a 3D convolution operation is applied to extract local features and expand the number of channels. Then, a normalization operation is performed. Finally, the ReLU activation function is utilized to enhance the network representation capacity. For the convolution operation, the first convolution layer that increases the number of channels by 8 times, the subsequent convolutions double the number of channels. In each layer, a maximum pooling operation is applied to reduce the spatial size of the feature maps while preserving important feature information.

3.3 3D Swin Transformer encoder

We employ the Swin Transformer to extract global and structural information. Because we want to directly process 3D data, we adopt the 3D Swin Transformer, which is also the Video Swin Transformer[28]as the backbone to capture the long-range dependencies for 3D CT volume data. The video spin transformer also helps reduce computational and time complexity, making the network more efficient in processing the 3D input data.

The structure of the 3D Swin Transformer block is shown in **Fig. 2**. Here, 3D W-MSA and 3D SW-MSA are regular multihead self-attention modules with window partitioning, respectively, and LN is linear norm operation. MLP and LN denote layer normalization and multilayer perception, respectively. The 3D W-MSA and 3D SW-MSA operations are shown in **Fig. 3**.

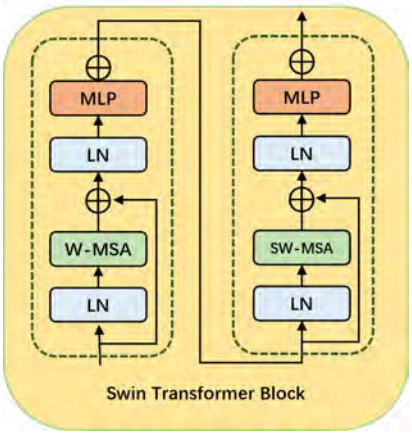


Fig. 2 The structure of 3D swin transformer block

3D W-MSA (Window-based Multi-head Self-Attention) reduces the computational complexity by dividing the input into multiple small windows, with self-attention calculations carried out within each window. If the window size is $M \times M \times M$, the number of positions within each window is M^3. For cases where M is much smaller than the image size, this significantly reduces computational burden. 3D W-MSA computes Q,K,V separately within each window, and then calculates the attention matrix and multiplies it with V to obtain the output. Although 3D W-MSA effectively reduces computational complexity, it is limited in that positions within each window can only attend to other positions within the same window, and cannot capture the correlation between different windows. So, 3D SW-MSA (Shifted Window-based Multi-head Self-Attention) is needed.

3D SW-MSA allows positions within each window to attend to information from surrounding windows by shifting the window positions in different layers. Specifically, if the initial window layout is referred to as a "regular window", in the next layer, the window is shifted by half of the window size, which is called a "shifted window". This allows a position to attend not only to information within its own window but also from its surrounding windows.

In terms of implementation, the window partitioning method in 3D SW-MSA is to first cyclically shift the input, then divide it as a regular window, calculate the self-attention, and finally cyclically shift it back.

The 3D Swin Transformer encoder relies entirely on an

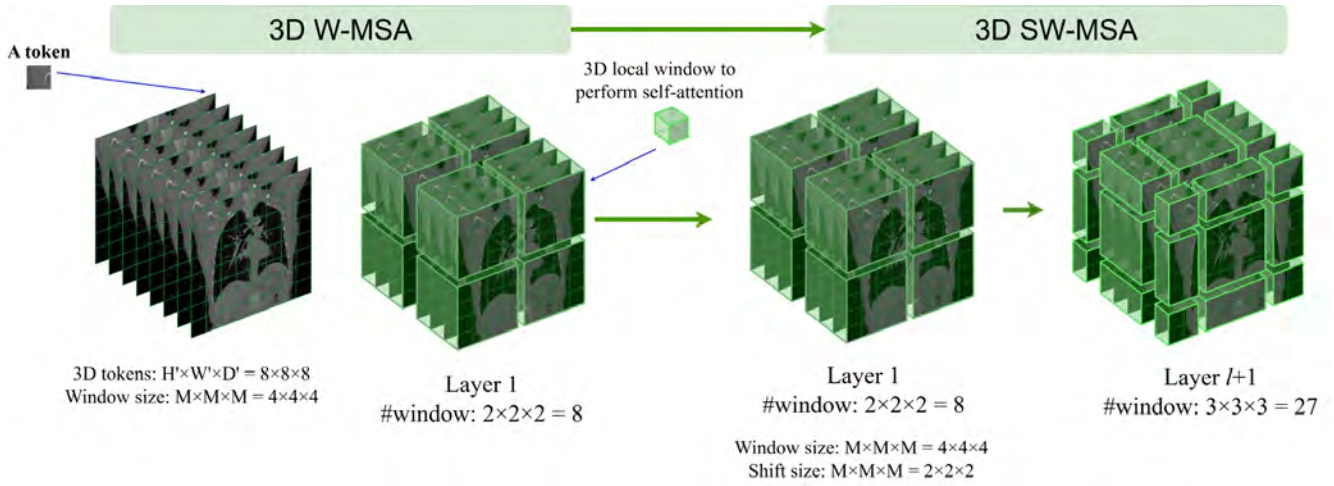


Fig. 3 3D W-MSA and 3D SW-MSA operation

attention mechanism to capture the long-range. The self-attention in transformer encoder is defined as follow:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where, $Q = W_q x, K = \cdot W_k x, V = W_v x$, d_k is the dimension of X. The input to the Swin Transformer encoder $x \in X$ is a token with a patch resolution of (H', W', D') and dimensions of $H' \times W' \times D' \times S$. We first create 3D tokens using the patch partitioning layer and project them into an embedding space of dimension C. The self-attention is computed in non-overlapping windows that are created during the partitioning stage, enabling efficient modeling of token interactions. The size of the embedding space C is set to 24 in our encoder.

The Swin Transformer encoder consists of 4 stages, each composed of 2 transformer blocks. In each stage, a linear embedding layer is used to create 3D tokens of size $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$. Fig. 3 shows the details of the shifted windowing mechanism for subsequent layers. Specifically, we utilize windows of size $M \times M \times M$ to evenly partition a 3D token into $\frac{H'}{M} \times \frac{W'}{M} \times \frac{D'}{M}$ regions at a given layer l in the transformer encoder. Subsequently, in layer $l+1$, the partitioned window regions are shifted by $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ voxels. In subsequent layers of l and $l+1$ in the encoder, the output is calculated as

$$\begin{aligned} \hat{z}^l &= W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \\ z^l &= MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \\ \hat{z}^{l+1} &= SW - MSA\left(LN\left(z^l\right)\right) + z^l \\ z^{l+1} &= MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \end{aligned} \quad (2)$$

Here, W-MSA and SW-MSA are regular and window partitioning multi-head self-attention modules, respectively; $\hat{z}^l$ and $\hat{z}^{(l+1)}$ denote the output of W-MSA and SW-MSA; MLP and LN denote layer normalization and Multi-Layer Perceptron, respectively.

To maintain the hierarchical structure of the encoder, the feature resolution is reduced by half using patch merging at the end of each stage. Additionally, the patch merging layer groups and concatenates patches of size $2 \times 2 \times 2$ to form a feature embedding of size $4C$. Then, a linear layer is used to reduce the dimensionality of the feature representation to $2C$. After bottleneck, the feature with reduced resolution using patch merging is connected to the CNN encoder part.

3.4 3D CNN decoder

We concatenate the feature maps that are extracted by the CNN and Swin Transformer encoder and feed them into the CNN decoder. The 3D CNN decoder consists of 5 layers. We also add skip connections between two encoders and decoder. Each layer in the decoder contains two blocks. Similarly to the CNN encoder, each block includes three steps. Firstly, a 3D convolution operation is performed to reduce the number of feature channels. Then normalization is applied, followed by ReLU activation. In the last layer of the decoder, a $1 \times 1 \times 1$ 3D convolution operation is used, followed by a sigmoid function to output the predicted segmentation probabilities. Finally, a Softmax operation is applied to generate the final results.

3.5 Fusion module based on cross attention

In order to enable effective information exchange between the dual encoders, we propose a multi-cross attention fusion module, which integrates local features from the 3D CNN encoder and global features from the 3D Swin Transformer encoder. This mechanism enhances

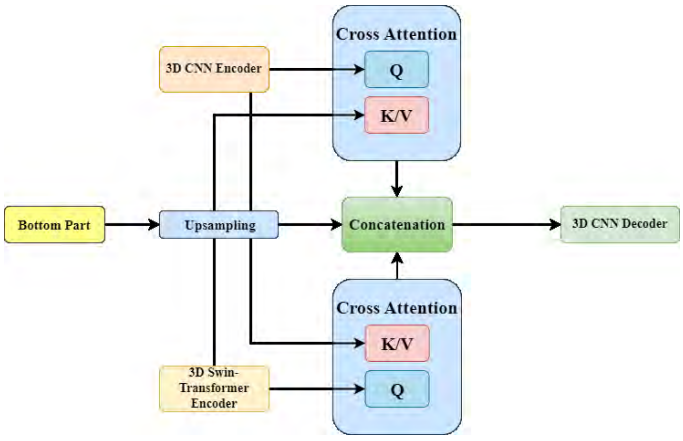


Fig. 4 Multi cross attention module

the network's sensitivity to fine-grained and structurally complex airway regions.

As shown in **Fig. 4**, the module applies cross-attention in two directions:

CNN → Transformer: The feature map from the 3D CNN encoder is used as the query (Q), and the feature map from the 3D Swin Transformer encoder provides the keys (K) and values (V). Transformer → CNN:The feature map from the 3D Swin Transformer encoder is used as the query (Q), and the CNN feature map provides the keys (K) and values (V).

Each directional attention is computed using the standard scaled dot-product attention mechanism as defined in Equation (1).

Let F_{cnn} and F_{trans} be the input feature maps from the CNN and Swin Transformer encoders, respectively. Then, the two cross-attention outputs are computed as follows:

$$F_{\text{CA1}} = \text{Attention}(F_{\text{cnn}}W_1^Q,\ F_{\text{trans}}W_1^K,\ F_{\text{trans}}W_1^V) \quad (3)$$

$$F_{\text{CA2}} = \text{Attention}(F_{\text{trans}}W_2^Q,\ F_{\text{cnn}}W_2^K,\ F_{\text{cnn}}W_2^V) \quad (4)$$

where W_1^Q, W_1^K, W_1^V and W_2^Q, W_2^K, W_2^V are learnable projection matrices for the two directions.

The final fused representation is obtained by concatenating the two attention outputs along the channel dimension:

$$F_{\text{fused}} = \text{Concat}(F_{\text{CA1}},\ F_{\text{CA2}}) \quad (5)$$

Finally, the fused outputs from both attention directions, along with the upsampled bottleneck features, are concatenated and sent to the decoder and final predictor for output generation.

3.6 Loss functions

We utilize the combination of focal loss and dice loss as the loss functions in this work. For this task, given the label $y^\alpha(x)$ and prediction $p^\alpha(x)$ of each voxel x in the volume set X, the overall loss function is defined as follows:

$$Loss = L_{focal} + L_{Dice}. \quad (6)$$

In this context, focal loss is employed to address the issue of class imbalance in image segmentation tasks. The lung bronchus occupies a relatively small proportion within the entire CT image of the lung. For the lung airways represent a relatively small portion of the whole lung CT images, the sample categories are unbalanced, which leads to a training process that focuses more on samples from most categories and relatively neglects samples from a few categories.

Focal Loss addresses this by introducing an adjustment factor to reduce the weight of easy-to-classify samples, thus allowing the model to pay more attention to those challenging samples. This adjustment factor α based on the predicted probability of each sample. It decreases the weight of high-probability samples in the loss computation, while maintaining relatively higher weights for low-probability samples. On the other hand, Dice loss, based on the Dice coefficient, aims to minimize the disparity between the segmentation results generated by the model and the ground truth segmentation. The computation formulas for Focal Loss and Dice Loss are as follows:

$$L_{\text{focal}} = -\frac{1}{|X|}\sum_{x\in X}\left(1-p_t^\alpha(x)\right)^2\log\left(p_t^\alpha(x)\right), \quad (7)$$

$$L_{\text{Dice}} = \frac{2\sum_{x\in X}p^\alpha(x)y^\alpha(x)}{\sum_{x\in X}\left(p^\alpha(x)+y^\alpha(x)\right)+\epsilon}$$

where $p_t^\alpha(x) = p^\alpha(x)$ if $y^\alpha(x) = 1$, otherwise $p_t^\alpha(x) = 1 - p_t^\alpha(x)$. Parameter ϵ is used to avoid division by zero.

3.7 Advantages of our proposed network

The combination of dual encoders and a Multi-Cross Attention Fusion Module enables W-Attention Net to capture both fine-grained and global structural features.

The 3D CNN encoder extracts localized details crucial for small airway branches, while the 3D Swin Transformer encoder models long-range dependencies to maintain anatomical continuity.

The Multi-Cross Attention module enables bidirectional feature fusion, allowing local and global informa-

tion to complement each other and form semantically rich representations.

This architecture is specifically designed to address key challenges in airway segmentation, including low contrast, branch disconnection, and complex morphology. Its effectiveness is validated in the following experiments.

4. Experiments and Results

In this section, we will present the experimental details, encompassing comparative experiments and ablation Study for the automatic segmentation network, along with subjective experiments for the entire visualization interactive system. In particular, the evaluation of this system involved assessments and tests conducted by medical professionals.

4.1 Datasets

We have used two independent data sets for our experiments, one is a private data set provided by the Shandong University in China, and the other is a public data set based on the LIDC-IDRI project[10]. Firstly, we evaluate our method using a lung CT dataset from Shandong University in China, which consists of two parts: chest CT images from 50 patients and corresponding 3D masks manually annotated by medical professionals. The CT images have a size range of [512, 512, 600 800]. There are 35 training samples, 5 validation samples, and 10 testing samples in this private dataset. Considering the high resolution and hardware limitations, we adopted a patch-based strategy to pre-process the dataset, with overlapping patches to enrich the training set. The final input images fed into the network have a size of 128 × 128 × 128, and the segmentation stride is set to [48, 80, 80]. During the segmentation training process, we applied mask-based filtering to exclude irrelevant regions, resulting in a final dataset containing 4497 training samples, 2520 validation samples, and 5040 testing samples.

Secondly, we evaluate our method on lung CT dataset from LIDC-IDRI[10], which consists of two parts: chest CT images from 40 patients and corresponding 3D masks manually annotated by medical professionals. The CT images have a size range of [512, 512, 300-500]. There are 28 training samples, 4 validation samples, and 8 testing samples in this public dataset. Considering the high resolution and hardware limitations, we adopted a patch-based strategy to pre-process the dataset, with overlapping patches to enrich the training set. The final input images fed into the network have a size of 128 × 128 × 128, and the segmentation stride is set to [48, 80, 80]. During the segmentation training process, we applied mask-based filtering to exclude irrelevant regions, resulting in a final dataset containing 3272 training samples, 1368 validation samples, and 2736 testing samples.

4.2 Implementations details

For the both datasets, the Adam optimizer is adopted for model optimization with a batch size of 1. We train our model for 100 epochs. The initial learning rate is set to 0.003 and multiplied by 0.1 every 20 epochs. We set the $\alpha = 0.25$ in focal loss functions (7). All the data shuffle on Nvidia GeForce RTX 3090 24GB GPU.

4.3 Evaluation metrics

We employ four commonly used metrics in segmentation tasks to evaluate the effectiveness of our method, which are Sensitivity, Dice similarity coefficient (DSC), 95% Hausdorff distance (HD95), and Jaccard similarity coefficient. Sensitivity represents the percentage of correctly segmented regions in the Ground Truth. The Dice similarity coefficient (DSC) measures the degree of overlap or similarity between the model-generated segmentation result and the ground truth segmentation. A higher Dice coefficient indicates a higher degree of overlap, indicating better segmentation performance of the model. The 95% Hausdorff distance (HD95) represents the surface distance between the Prediction and Ground Truth, quantified as the maximum distance at the 95%. A smaller HD95 value indicates more accurate prediction of boundaries and closer proximity to the ground truth segmentation. Dice coefficient is sensitive to the internal filling of the mask, while Hausdorff distance is sensitive to the segmented boundaries. The Jaccard similarity coefficient is used to compare the similarity and dissimilarity between finite sets of samples. A higher Jaccard coefficient indicates higher similarity between samples. The primary evaluation parameter is DSC. We use FN for false negatives (positive cases predicted as negative), TN for true negatives (correctly predicted negative cases), TP for true positives (correctly predicted positive cases), and FP for false positives (negative cases predicted as positive). The specific formulas for these evaluation parameters are given below.

Sensitivity:

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

Dice Similarity Coefficient (DSC):

Table 1 Comparison result on SDU private dataset

Method	Sensitivity↑	DSC↑	HD95↓	Jaccard↑
3D U-Net[7]	0.9420	0.8698	70.5738	0.7717
Swin UNETR[27]	0.9274	0.8660	32.6641	0.7644
Ours	**0.9470**	**0.8959**	**9.1423**	**0.8125**

Table 2 Comparison result on LIDC-IDRI[10]public dataset

Method	Sensitivity↑	DSC↑	HD95↓	Jaccard↑
3D U-Net[7]	0.9018	0.8435	220.6875	0.7313
Swin UNETR[27]	0.8702	0.8531	167.9583	0.7443
Ours	**0.9087**	**0.8944**	**54.7887**	**0.8099**

$$DSC = \frac{(2 \times TP)}{((2 \times TP) + FP + FN)} \tag{9}$$

95% Hausdorff Distance (HD95):

$$\begin{aligned} HD95\langle X, Y\rangle &= 95\% \max\{d_{XY}, d_{YX}\} \\ &= 95\% \max\left\{\max_{x\in X}\min_{y\in Y} d\langle x, y\rangle, \right. \\ &\quad \left. \max_{y\in Y}\min_{x\in X} d\langle x, y\rangle\right\} \end{aligned} \tag{10}$$

Jaccard Similarity Coefficient (Jaccard):

$$Jaccard = \frac{TP}{TP + FN + FP} \tag{11}$$

4.4 Comparative experiments

Our experimental results on the private SDU dataset are showed in **Table 1**. The best results for each metric are indicated in bold. It can be observed that our proposed method achieves the best results without pre-training. Our results achieved a DSC score of 0.8938, which is 0.025 higher than that of the 3D U-Net. To compare with the state of art 3D medical segmentation method, we compared with Swin UNETR[27], which is an improve network using 3D Swin Transformer[28]. Furthermore, our method outperformed other methods significantly in terms of HD95 and Jaccard evaluation metrics.

Our experimental results on the public LIDC-IDRI dataset[10]are showed in **Table 2**. The best results for each metric are indicated in bold. It can be observed that our proposed method achieves the best results without pre-training. Our results achieved a DSC score of 0.8944, which is 0.051 higher than that of the 3D U-Net[7] . To compare with the network using 3D Swin Transformer[28], we compared with Swin UNETR[27]. Furthermore, our method outperformed other methods significantly in terms of HD95 and Jaccard evaluation metrics.

Figure 5 shows the 3D segmentation result examples. We present three examples of 3D segmentation results derived from the same CT image case for three methods in three orthogonal views. In the Fig. 5, the red region denotes the ground truth, and the yellow region represents the segmentation result from different method. We can see our proposed method is the most similar compare with the ground truth. Over-segmented regions are marked with red circles, clearly showing false positives in irrelevant lung areas.

In conventional methods (3D U-Net[7]and Swin UNETR[27]), numerous false positives appear outside the airway, as indicated by the red circles. These false positives are significantly reduced by the proposed method. Compared with that, our method generates smoother and more anatomically aligned segmentation outputs.

4.5 Ablation study

To assess the roles of individual encoders in our proposed W-Net and determine if performing two rounds of cross-attention is necessary, we conducted ablation experiments on the larger and higher-resolution Shandong University proprietary dataset. We experimented with networks composed of individual encoders, as well as with a simple concatenation of the two encoders (our previous research[11]). Additionally, we employed single-round cross-attention for feature enhancement and concatenated the results with different encoders, resulting in a total of four different combinations for experimentation. The experimental results are shown in **Table 3**.

Through the results of the ablation experiments, we found that the jointly trained W-Net, which combines two encoders, achieves a significant improvement compared to using either the 3D CNN alone or the 3D Swin Transformer alone with pre-trained weights loaded into

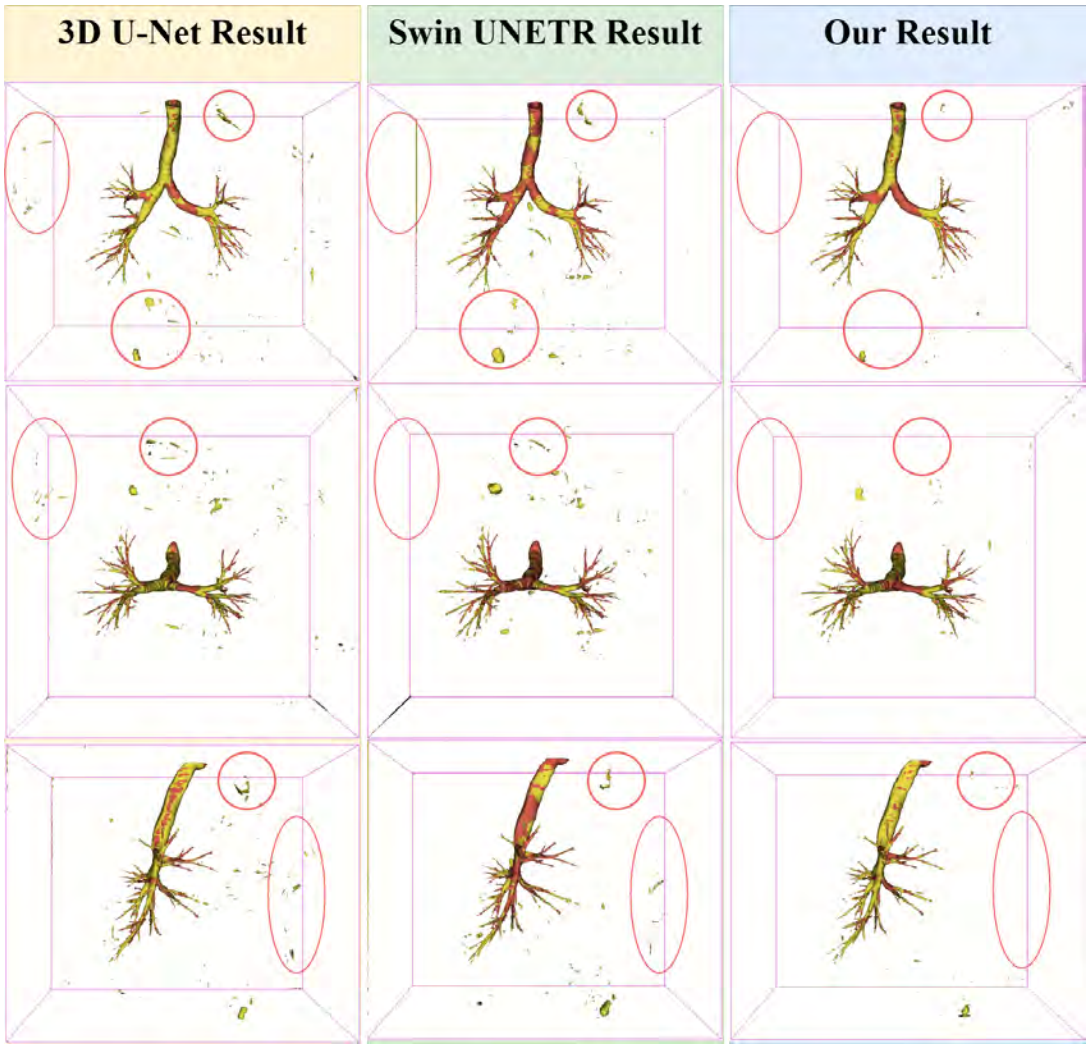


Fig. 5 Visualization comparisons on LIDC[10]dataset. In conventional methods (3D U-Net[7]and Swin UNETR[27]), numerous false positives appear outside the airway, as indicated by the red circles.

Table 3 Ablation study on SDU private dataset

Method	Sensitivity↑	DSC↑	HD95↓	Jaccard↑
3D U-Net[7]	0.9420	0.8698	70.5738	0.7717
3D Swin Transformer U-Net	0.9203	0.8496	130.3879	0.7400
W-Net[11]	0.9448	0.8938	9.7292	0.8090
CNN as Q concatenate Swin	0.9422	0.8847	44.5503	0.7948
CNN as Q concatenate CNN	0.9446	0.8939	21.7626	0.8094
Swin as Q concatenate CNN	0.9365	0.8824	25.6393	0.7906
Swin as Q concatenate Swin	0.9439	0.8912	13.0075	0.8048
Ours	**0.9470**	**0.8959**	**9.1423**	**0.8125**

the U-Net. Remarkably, our experiments demonstrate that our approach achieves good performance without the need for pre-training weights for the Swin Transformer.

In contrast to using single-round cross-attention, employing multi-cross attention allows for better interaction between the local information extracted by the CNN encoder and the global information captured by the Swin Transformer encoder. This enhancement leads to improved segmentation accuracy, particularly in the bronchial region, demonstrating the effectiveness of our proposed W-attention Net structure for lung airway segmentation.

5. Conclusion

In this article, we proposed a highly sensitive segmentation method for the pulmonary airways, particularly

the bronchi, to assist physicians in diagnosing airway-related diseases. The method employs a dual-encoder architecture and a multi-cross attention fusion strategy to combine local and global information. Experimental results demonstrate that this design improves segmentation accuracy while reducing structural inconsistencies. The method has achieved favorable performance on both public (LIDC-IDRI) and private datasets.

Due to the scarcity of high-quality publicly available data sets for airway segmentation, our results are based on a limited number of data sets, including a public and a private data set. Evaluating the generalizability of the proposed method on a broader range of datasets is left for future work.

Acknowlegment

This work is supported in part by the Japan Society of Promotion of Science (JSPS) under Grant Nos. 20KK0234 and 22H04736.

References

1) S. J. Howling, T. W. Evans and D. M. Hansell: "The significance of bronchial dilatation on CT in patients with adult respiratory distress syndrome", Clinical Radiology, Vol. 53, No. 2, pp. 105–109 (1998).
2) R. J. Shaw, R. Djukanovic, D. P. Tashkin, A. B. Millar, R. M. Du Bois and P. A. Corris: "The role of small airways in lung disease", Respiratory Medicine, Vol. 96, No. 2, pp. 67–80 (2002).
3) C. I. Fetita, F. Prêteux, C. Beigelman-Aubry and P. Grenier: "Pulmonary airways: 3-D reconstruction from multislice CT and clinical investigation", IEEE Transactions on Medical Imaging, Vol. 23, No. 11, pp. 1353–1364 (2004).
4) Y. Li, Y. Dai, X. Duan, W. Zhang, Y. Guo and J. Wang: "Application of automated bronchial 3-D CT measurement in pulmonary contusion complicated with acute respiratory distress syndrome", Journal of X-ray Science and Technology, Vol. 27, No. 4, pp. 641–654 (2019).
5) X. Wu, G. H. Kim, M. L. Salisbury *et al.*: "Computed tomographic biomarkers in idiopathic pulmonary fibrosis: the future of quantitative analysis", American Journal of Respiratory and Critical Care Medicine, Vol. 199, No. 1, pp. 12–21 (2019).
6) O. Ronneberger, P. Fischer and T. Brox: "U-Net: Convolutional networks for biomedical image segmentation", Proc. of MICCAI 2015, pp. 234–241 (2015).
7) Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger: "3-D U-Net: Learning dense volumetric segmentation from sparse annotation", Proc. of MICCAI 2016, pp. 424–432 (2016).
8) A. Vaswani, N. Shazeer, N. Parmar et al.: "Attention is all you need", Advances in Neural Information Processing Systems, Vol. 30 (2017).
9) H. Cao, Y. Wang, J. Chen et al.: "Swin-UNet: UNet-like pure transformer for medical image segmentation", Proc. of ECCV 2022, pp. 205–218 (2022).
10) Y. Qin, H. Zheng, Y. Gu et al.: "Learning tubule-sensitive CNNs for pulmonary airway and artery-vein segmentation in CT", IEEE Transactions on Medical Imaging, Vol. 40, No. 6, pp. 1603–1617 (2021).
11) L. Lyu, S. Chai, J. Liu, T. Tateyama, X. Qiao and Y.-W. Chen: "A 3-D Fusion U-Net with Dual CNN and Transformer Encoders for Lung Airway Segmentation", Proc. of IEEE Global Conf. on Consumer Electronics (GCCE), pp. 24–28 (2023).
12) E. M. van Rikxoort, W. Baggerman and B. van Ginneken: "Automatic segmentation of the airway tree from thoracic CT scans using a multi-threshold approach", Proc. of the 2nd International Workshop on Pulmonary Image Analysis, pp. 341–349 (2009).
13) N. Otsu: "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62–66 (1979).
14) P. Lo, B. van Ginneken, J. M. Reinhardt et al.: "Extraction of airways from CT (EXACT'09)", IEEE Transactions on Medical Imaging, Vol. 31, No. 11, pp. 2093–2107 (2012).
15) F. Wu, A. Souza, T. Zhang et al.: "Simplifying graph convolutional networks", Proc. of ICML 2019, pp. 6861–6871 (2019).
16) H. Gao, Z. Wang and S. Ji: "Large-scale learnable graph convolutional networks", Proc. of the 24th ACM SIGKDD, pp. 1416–1424 (2018).
17) M. Chen, Z. Wei, Z. Huang, B. Ding and Y. Li: "Simple and deep graph convolutional networks", Proc. of ICML 2020, pp. 1725–1735 (2020).
18) J. M. Wolterink, T. Leiner and I. Išgum: "Graph convolutional networks for coronary artery segmentation in cardiac CT angiography", Proc. of GLMI@MICCAI 2019, pp. 62–69 (2019).
19) J. Yun, J. Park, D. Yu et al.: "Improvement of fully automated airway segmentation on volumetric computed tomographic images using a 2.5-D convolutional neural net", Medical Image Analysis, Vol. 51, pp. 13–20 (2019).
20) J.-P. Charbonnier, E. M. van Rikxoort, A. A. Setio et al.: "Improving airway segmentation in computed tomography using leak detection with convolutional networks", Medical Image Analysis, Vol. 36, pp. 52–60 (2017).
21) Q. Meng, H. R. Roth, T. Kitasaka et al.: "Tracking and segmentation of the airways in chest CT using a fully convolutional network", Proc. of MICCAI 2017, pp. 198–207 (2017).
22) A. Dosovitskiy, L. Beyer, A. Kolesnikov et al.: "An image is worth 16 × 16 words: Transformers for image recognition at scale", arXiv:2010.11929 (2020).
23) A. Arnab, M. Dehghani, G. Heigold et al.: "ViViT: A video vision transformer", Proc. of ICCV 2021, pp. 6836–6846 (2021).
24) Z. Liu, Y. Lin, Y. Cao et al.: "Swin transformer: Hierarchical vision transformer using shifted windows", Proc. of ICCV 2021, pp. 10012–10022 (2021).
25) M. Ding, B. Xiao, N. Codella et al.: "DaViT: Dual attention vision transformers", Proc. of ECCV 2022, pp. 74–92 (2022).
26) H. Wang, P. Cao, J. Wang and O. R. Zaiane: "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer", Proc. of AAAI 2022, Vol. 36, No. 3, pp. 2441–2449 (2022).
27) A. Hatamizadeh, V. Nath, Y. Tang et al.: "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images", Proc. of BrainLes@MICCAI 2021, pp. 272–284 (2021).
28) Z. Liu, J. Ning, Y. Cao et al.: "Video Swin transformer", Proc. of CVPR 2022, pp. 3202–3211 (2022).

Appendix

APP 1. Interactive Visualization System

In this appendix, we introduce our entire interactive visualization system overview. Following that, we proceed to individually introduce the segmentation network and real 3D interactive visualization. After extracting the lung airway, we use volume rendering to obtain the 3D obj file. Then we use unity to develop the interactive visualization system.

To enable visualization and interaction with lung segmentation results, we used Unity and Microsoft's Mixed Reality Toolkit (MRTK).

Users can overlay the entire lung field, airway, or specific lung lobes. The system offers two interaction modes: close-range grabbing for direct manipulation, and long-range control using a hand-tracked beam and pinch gesture. As shown in **Fig. A1**, models can be rotated and zoomed via the bounding box. This intuitive interaction improves visibility of airways and lung regions, supporting diagnosis and bronchial surgery.

Two ways of interacting with the models were used to make the human-computer interaction work. In one case, when the user is close to the model, the user can interact with the model by grabbing and releasing it, simulating the action of grabbing real objects in the real world.

In the other case, when the user is further away from the model, the user can interact with the whole model through the beam released by the index finger part, and slightly pinch the index finger and thumb to complete the interaction operation.

Users can switch display areas using virtual buttons to intuitively observe infections in different lung regions. Buttons on both sides of the model control which areas are shown. The lung model is enclosed in a bounding box that supports gestures like rotation, zooming, and panning for enhanced visualization.

Our system uses a third-person perspective to let broader audiences observe HoloLens users interacting with virtual objects.

The shared scene feature allows multiple users to experience the same augmented reality (AR) environment. As shown in **Fig. A2**, the real-world view of the device is transmitted in real time to a computer using the Microsoft HoloLens application.

In order to better evaluate whether our system is user-friendly and beneficial in clinical medicine, a radiologist was invited to evaluate the usability of the system. The

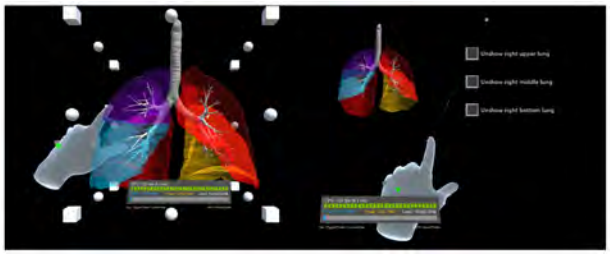


Fig. A1 Two ways of interacting with the models.

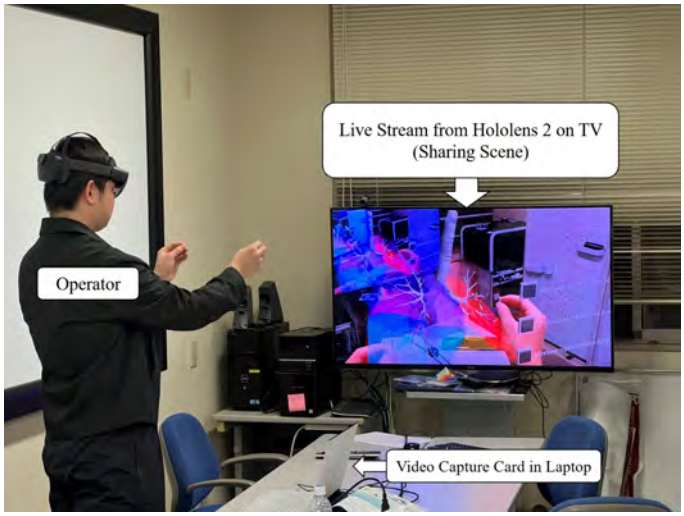


Fig. A2 Third-person sharing view on the TV

radiologist evaluation gives high evaluations for the system, suggesting that the proposed system is sufficiently satisfactory for use in clinical settings.

The system is shown to not only help in the clinical diagnosis of lung diseases, but also contribute to the innovation and improvement of medical education.

(Received November 6, 2024)
(Revised May 15, 2025)

Liang LYU
He received his B.E. from Dalian University of Technology in 2021 and his M.E. from Ritsumeikan University in 2023, where he is currently pursuing a Ph.D. in Information Science and Engineering. He is currently engaged in research on image processing and its applications in medical imaging.

Jiaqing LIU
He received the B.E. from Northeastern University in 2016, and the M.E. and D.E. from Ritsumeikan University in 2018 and 2021. He was a JSPS Research Fellow and Assistant Professor at Osaka University, and is currently with Ritsumeikan University. His interests include computer vision, medical engineering, and deep learning.

Shurong CHAI
He received his B.E. from Ritsumeikan University in 2021 and his M.E. from Ritsumeikan University in 2023, where he is currently pursuing a Ph.D. in Information Science and Engineering. He is currently engaged in research on image processing and its applications in medical imaging.

Fang WANG
She received her M.B. from Three Gorges University in 2019. She is currently pursuing a Ph.D. in medicine at Zhejiang University. Her research interests include liver imaging diagnosis and AI-based medical image analysis.

Tomoko TATEYAMA
She received her D.E. from the University of the Ryukyus in 2009. She is currently a Professor at the same university. Her research interests include medical image analysis and machine learning.

Xu QIAO
He received his B.S. and M.S. from Shandong University in 2004 and 2007, and Ph.D. from Ritsumeikan University in 2010. He conducted postdoctoral research at the University of Chicago in 2011. He is currently a Professor at Shandong University. His research interests include medical image pattern recognition and AI-based image processing.

Yen-Wei CHEN (*Member*)
He received his D.E. degree in Engineering from the Graduate School of Engineering, Osaka University, in March 1990. He served as a researcher at the Institute for Laser Technology, and later held positions as lecturer, associate professor, and professor at the Faculty of Engineering, University of the Ryukyus. In March 2004, he was appointed as a professor in the College of Information Science and Engineering, Ritsumeikan University, where he currently works. His research interests include computer vision, medical image analysis, and computational intelligence. He is a member of IEICE, the Japanese Society of Medical Imaging Technology, IEEE, and other professional societies.

IIEEJ **Paper** -- *Special Issue on Image Electronics Technologies Related to VR/AR/MR/XR* --

Material Appearance Reproduction via GAN-Based Novel View Synthesis Considering the Naturalness of Intensity Changes in Viewpoint Transitions

Taishi IRIYAMA† (*Member*) , Chihiro HOSHIZAWA† , Takashi KOMURO†

† Department of Information and Computer Sciences, Saitama University

<**Summary**> In this paper, we propose a GAN-based novel view synthesis (NVS) method that focuses on generating naturally varying viewpoint-dependent material appearances during viewpoint transitions. While NVS has gained increasing attention for its ability to generate images from arbitrary viewpoints, reproducing viewpoint-dependent material appearances, particularly for specular and transparent objects, remains challenging. The proposed method introduces a 3D convolutional discriminator that evaluates the naturalness of generated image sequences across consecutive viewpoints to reproduce material appearance caused by intensity changes during viewpoint transitions. Experimental results using CG datasets with specular and transparent objects demonstrate that the proposed method effectively generates view-consistent images while preserving fine surface details and maintaining consistency across viewpoint transitions.

Keywords: material appearance reproduction, novel view synthesis, view interpolation, generative adversarial networks, 3D convolutional neural network

1. Introduction

Novel view synthesis (NVS) is a technique that renders a scene from unseen viewpoints based on images captured from some viewpoints. By rendering scenes from arbitrary viewpoints, there are significant potential applications such as immersive experiences and content creation for virtual reality (VR). Furthermore, the smooth synthesis of viewpoint-dependent intensity changes, such as those caused by specular reflection or refraction, enables the faithful reproduction of material appearance including surface roughness and glossiness, which holds promise for applications in material presentation within VR environments.

Recently, a variety of NVS methods using deep learning have been proposed, and these can mainly be classified into regression-based and generation-based methods. Regression-based methods directly predict the scenes from novel viewpoints without requiring explicit 3D scene reconstruction[1)–14)]. Among them, neural radiance fields (NeRF)[4)]and 3D Gaussian splatting (3DGS)[5)]are representative examples of regression-based methods. NeRF predicts color and density for any 3D coordinate and synthesizes novel views through ray-based volume rendering. 3DGS represents a scene as a collection of 3D Gaussian distributions and synthesizes novel views by directly projecting them onto the camera plane. However, both original NeRF and 3DGS assume that light travels in straight lines within the scene, limiting their ability to model effects such as refraction and subsurface scattering. Moreover, achieving high-quality geometry and reflectance properties requires numerous input images from various viewpoints, limiting its practical applicability.

On the other hand, generation-based methods use deep generative models, such as generative adversarial networks (GANs)[15)]and diffusion models[16)], to directly generate scenes from novel viewpoints[17)–23)]. These methods can photorealistically generate scenes from novel viewpoints by leveraging the learned data distribution, even when target viewpoint prediction involves significant uncertainty. However, these methods often struggle with maintaining geometric and texture consistency between viewpoints, which can result in artifacts such as unrealistic object deformations or flickering during viewpoint transitions.

In this paper, we propose a GAN-based NVS method that focuses on generating naturally varying viewpoint-dependent material appearances. The proposed method synthesizes image sequences that transition between viewpoints by leveraging information from multiple input viewpoints. To accurately reproduce material appearances caused by intensity changes during viewpoint

transitions, we introduce a 3D convolutional discriminator that evaluates the naturalness of generated image sequences across consecutive viewpoints. Experimental results using a CG dataset with specular and transparent objects demonstrate that the proposed method effectively generates view-consistent images while preserving fine surface details and maintaining visual coherence across viewpoint transitions.

2. Related Work

2.1 Geometry-based novel view synthesis

Geometry-based NVS methods use geometric information to synthesize images from novel viewpoints, either through explicit or implicit representations. Explicit geometry methods employ 3D data such as depth maps or point clouds. McMillan et al.[24]proposed a method that renders images from any nearby viewpoint by reprojecting pixels to their proper 3D locations in the novel view. Debevec et al.[25]proposed a method that recovers the basic geometry of the photographed scene using photogrammetric modeling and produces realistic renderings by compositing multiple views. Kang et al.[26]proposed a method that simultaneously estimates multiple depth maps with visibility modeling, enabling robust reconstruction from widely-spaced views.

Implicit geometry methods synthesize images by inferring pixel correspondences between viewpoints, without explicit 3D representations. Chen et al.[27]proposed a method that synthesizes intermediate views through linear interpolation of pixel correspondences computed using camera parameters and range data. Seitz et al.[28]proposed view morphing that achieves physically valid view interpolation by prewarping images to parallel views before interpolation. Stich et al.[29]proposed real-time view and time interpolation in image space using local homographies to handle occlusions without 3D reconstruction. Despite the flexibility and accuracy of these methods, their reliance on pixel warping and interpolation makes it difficult to reproduce viewpoint-dependent intensity changes, such as specular reflections.

2.2 Regression-based novel view synthesis

Regression-based NVS methods mainly use deep learning to predict images from novel viewpoints directly from input images without explicitly reconstructing the 3D scene structure[1)–14)]. Early methods, such as DeepStereo[1)], implicitly capture geometry by reprojecting images into a plane sweep volume for view synthesis. Zhou et al.[2)]proposed a multi-plane image (MPI) representation, where multiple front-parallel depth planes with assigned RGBA images allow view synthesis beyond a limited baseline. Tucker et al.[3)]extended the above method by predicting MPI representations directly from a single input image.

Recently, NeRF[4)]and 3DGS[5)]have attracted considerable attention. NeRF predicts color and density for any 3D coordinate and synthesizes novel views through ray-based volume rendering. 3DGS represents a scene as a collection of 3D Gaussian distributions and synthesizes novel views by directly projecting them onto the camera plane. However, both original NeRF and 3DGS assume that light travels in straight lines within the scene, limiting their ability to model effects such as refraction and subsurface scattering.

To address this limitation, various methods have been proposed that explicitly model complex light-material interactions such as reflections and refractions. Among NeRF-based approaches, a method has been proposed to handle specular reflection[6)], which estimates surface normals from scene features and models specular reflections by tracing reflection rays within the scene, as well as a method to handle transparent objects[7)], which reconstructs 3D geometry from a visual hull and models refraction based on Snell's law. Among 3DGS-based approaches, 3D Gaussian Ray Tracing[8)]integrates ray tracing into a particle-based scene representation and explicitly models reflections and refractions by tracing secondary rays. EnvGS[9)]introduces Gaussian-based environment representations and captures specular reflections by tracing rays in the reflection direction within the environment representation. However, these methods require a large number of densely sampled input images and per-scene optimization, which limits their practical applicability.

To address these limitations, several methods have been proposed to improve generalization and performance under limited input conditions. Among NeRF-based approaches, pixelNeRF[10)]conditions a pre-trained NeRF model on pixel-aligned local image features, enabling novel view synthesis without per-scene optimization. The image-based rendering method IBRNet[11)]achieves generalization by leveraging multi-view information with geometric consistency across views. Among recent 3DGS-based approaches for sparse view scenarios, DNGaussian[12)]uses random Gaussian initialization with depth

constraints to estimate 3D geometry without requiring precomputed point clouds. MCGS[13]improves the initialization of Gaussians from sparse-view inputs by incorporating robust sparse feature matching and progressive pruning FreeSplatter[14]synthesizes novel views from uncalibrated sparse-view images by jointly predicting per-pixel Gaussians while estimating their camera parameters.

While these methods independently address limitations such as transparent objects and a limited number of input images, it remains challenging to accurately predict material appearance with viewpoint-dependent intensity changes from only a few viewpoints.

2.3 Generation-based novel view synthesis

Generation-based NVS methods use deep generative models such as GANs and diffusion models[17)–19),22),23)]. These methods can photorealistically generate scenes from novel viewpoints by leveraging the learned data distribution, even when target viewpoint prediction involves significant uncertainty.

Several studies have explored the use of GANs to generate images from novel viewpoints. Zhou et al.[17)]proposed a method that generates multiple views from a single input image using variational inference and GANs. Shu et al.[18)]proposed a method that extracts viewpoint-independent features while integrating viewpoint information, enabling image generation from diverse perspectives. Hoshizawa et al.[19)]proposed a method focused on reproducing material appearance by capturing viewpoint-dependent intensity changes for specular objects. However, generating images for each viewpoint introduces flickering when attempting to generate continuous viewpoint transitions.

To address this issue, 3D-aware GANs and diffusion models have emerged to enhance 3D consistency. Nguyen et al.[20)]proposed a method that maintained internal 3D feature spaces to stabilize images across viewpoints, while Schwarz et al.[21)]proposed an approach that integrated neural radiance fields with GAN-based synthesis to improve consistency. Watson et al.[22)]proposed a diffusion model that achieves 3D consistency through stochastic conditioning of previous frames, while Chan et al.[23)]proposed a 3D-aware diffusion model that incorporates geometry priors via 3D feature volumes for consistent view synthesis. While these methods successfully achieve geometric consistency across viewpoints through their 3D-aware architectures, they do not explicitly address the consistency of view-dependent appearance properties, resulting in potential inconsistencies in specular reflections and refraction during continuous viewpoint transitions despite each view being individually photorealistic.

3. Proposed Method

The proposed method synthesizes image sequences that naturally vary in viewpoint-dependent material appearances between viewpoints by leveraging information from multiple input viewpoints. **Figure 1** illustrates the framework of the proposed method. First, each input RGBD image captured from some input viewpoints, consisting of an RGB image $x \in \mathbb{R}^{H\times W\times 3}$ and a depth image $d \in \mathbb{R}^{H\times W\times 1}$, is projected into 3D space using the intrinsic and extrinsic camera parameters, resulting in point clouds where each point includes spatial coordinates and RGB color information. For a target viewpoint with pose p_t, we project these 3D points back onto the image plane to obtain a warped image w_t:

$$w_t = \mathrm{warp}(x, d, p_t) \tag{1}$$

where t represents the frame number in the sequence, and p_t represents the camera pose of the output viewpoint at frame t.

Then, using these warped images as input, a generator G produces images from each output viewpoint. These generated images are sequentially concatenated to form an image sequence that corresponds to continuous viewpoint movement:

$$s^{\mathrm{gen}} = \{G(w_1), G(w_2), \ldots, G(w_T)\} \tag{2}$$

where $s^{\mathrm{gen}} \in \mathbb{R}^{T\times H\times W\times 3}$ is the generated image sequence, and T represents the total number of frames in the sequence. To ensure that the generated images exhibit natural appearance changes corresponding to viewpoint transitions, we employ adversarial learning utilizing a 3D convolutional discriminator. This discriminator considers both the temporal consistency and visual naturalness of the generated image sequences. During training, we randomly sample target viewpoint sequences instead of using fixed viewpoint sequences. This random sampling approach allows the model to encounter diverse viewpoint transitions and effectively learn complex view-dependent appearance changes, particularly in materials that exhibit specular reflection and refraction.

3.1 Network architecture

The generator is designed with a U-Net structure[30)]using 2D convolutions. **Figure 2** illustrates the network

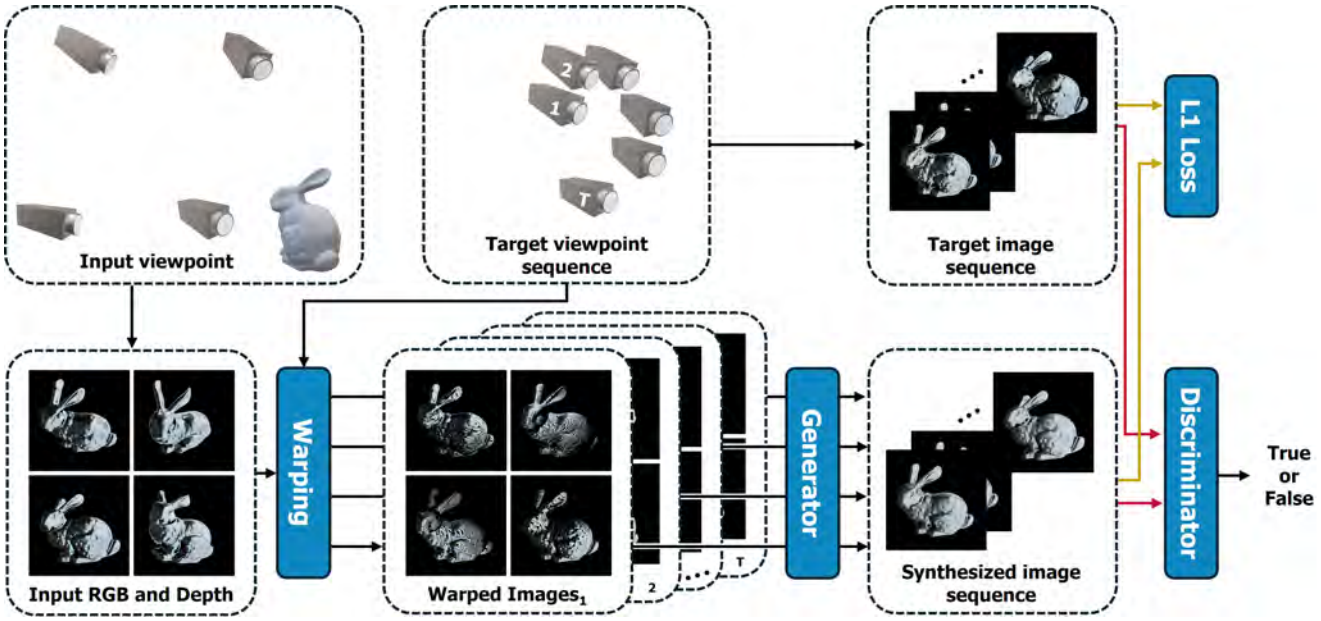
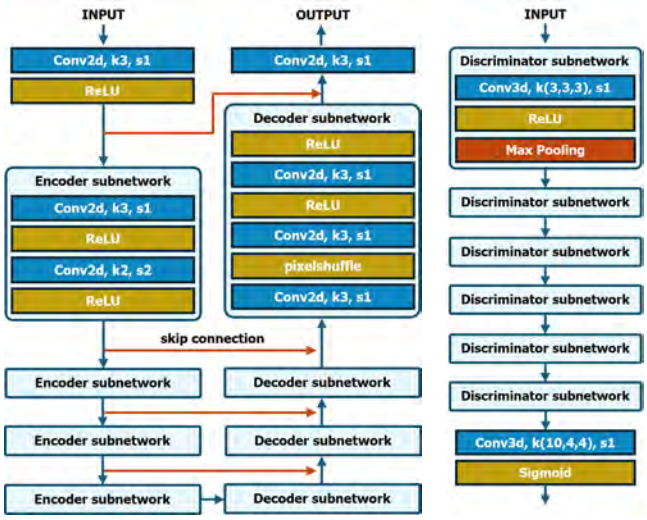


Fig. 1 Framework of the proposed method

architecture of generator and discriminator. As shown in Fig.2 (a), the encoder of generator comprises an initial convolution layer followed by four blocks. In each block in the encoder, a convolution uses stride 2 to reduce the spatial dimensions, halving both the height and width of the input while doubling the number of channels. The decoder of generator also consists of four blocks and an output layer. Each block in the decoder uses subpixel convolution to increase the spatial dimensions, doubling both the height and width of the input while halving the number of channels. Dropout layers are applied in the bottommost layer of the encoder and the topmost layer of the decoder to enhance generalization. As shown in Fig.2 (b), the discriminator is composed of six blocks utilizing 3D convolutions, followed by an output layer with a sigmoid activation function. Each block contains a 3D convolution layer, an activation function, and a 3D pooling layer.



(a) Generator (b) Discriminator

Fig. 2 Network architecture: (a) The generator with U-Net structure for image synthesis; (b) The 3D convolutional discriminator for evaluating the naturalness of viewpoint transitions

3.2 Loss function

The proposed model is optimized using a combination of $\mathcal{L}_1$ norm and adversarial losses. The $\mathcal{L}_1$ norm is applied between the generated image sequence s^{gen} and the ground truth image sequence s^{gt}, defined as follows:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} |s_i^{\text{gt}} - s_i^{\text{gen}}|_1 \tag{3}$$

where N represents the number of training samples, and s^{gt} denotes the sequence of ground truth images. The adversarial losses for the generator $\mathcal{L}_{G_{\text{adv}}}$ and discriminator $\mathcal{L}_{D_{\text{adv}}}$ are defined as:

$$\mathcal{L}_{G_{\text{adv}}} = \frac{1}{N}\sum_{i=1}^{N} ((1-y_i)\log(1-D(s_i^{\text{gen}}))) \tag{4}$$

$$\mathcal{L}_{D_{\text{adv}}} = -\frac{1}{N}\sum_{i=1}^{N} \big(y_i \log D(s_i^{\text{gt}}) + (1-y_i)\log(1-D(s_i^{\text{gen}}))\big) \tag{5}$$

where D represents the discriminator and y denotes the label. Combining these terms, we define the overall loss functions for the generator $\mathcal{L}_G$ and discriminator $\mathcal{L}_D$ as:

$$\mathcal{L}_G = \mathcal{L}_1 + \lambda \mathcal{L}_{G_{\text{adv}}} \tag{6}$$

$$\mathcal{L}_D = \mathcal{L}_{D_{\text{adv}}} \tag{7}$$

where λ is a weighting coefficient for the adversarial loss.

4. Experiments

We conducted an experiment to generate images from novel viewpoints using the proposed method and verified the reproduction of the material appearance for specular and transparent objects.

4.1 Dataset

We constructed a structured multi-view dataset using 3D models and environment maps for systematic evaluation. The images were captured using a virtual 5×5 camera array arranged on a spherical quadrilateral centered on the object. Each camera was positioned at a constant distance from the object and oriented towards its center. The vertices of the spherical quadrilateral were positioned with a $2\pi/9$ angular difference in Euler angles, with intermediate viewpoints distributed uniformly in a 5×5 grid between these vertices.

The training dataset consisted of 12,000 sets, generated using 6 different 3D objects and 500 environment maps. Each set contained 4 input RGBD images and 21 target RGB images, captured from a 5×5 camera array. All images were rendered at a resolution of 256×256. We divided this dataset into 95% for training and 5% for validation sets. The 3D objects were sourced from the Stanford 3D Scanning Repository[31] and included Stanford Bunny, Happy Buddha, Dragon, Lucy, Asian Dragon, and Thai Statue. The environment maps comprised 500 HDR images from Poly Haven[32]. For testing, we generated 100 image sets using the Armadillo 3D object and 100 environment maps from Poly Haven, ensuring no overlap with the training data.

For each object, we created both specular and transparent materials using Blender's physically-based renderer, Cycles. The material parameters and their ranges are detailed in **Table 1**. The object size, angle, position, and environment map orientation were randomized during dataset generation.

Table 1 Material parameters used for creating specular and transparent object datasets

Parameter	Specular		Transparent	
	Min	Max	Min	Max
Base Color (Red)	0	1	0.8	1
Base Color (Green)	0	1	0.8	1
Base Color (Blue)	0	1	0.8	1
Subsurface	Default		Default	
Specular	Default		0.5	
Roughness	0	0.1	0	
IOR	Default		1.5	
Metallic	1		Default	
Transmission	Default		1	
Transmission Roughness	Default		0	0.1

4.2 Training setup

We used the Adam optimizer[33] for model training, with hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rates were set to 20^{-5} for the generator and 10^{-5} for the discriminator. Other hyperparameters included a batch size of 1, a loss function weight λ of 10^{-4}, and a dropout rate of 0.5. The proposed method was trained on output view sequences consisting of 10 frames. The target viewpoint for each frame was determined by first randomly selecting an initial position from the 5×5 camera array excluding the vertices, then randomly selecting subsequent viewpoints from the eight adjacent positions in the array. We implemented our model using Python with PyTorch. All experiments were conducted on an NVIDIA GTX 1070 with 8 GB VRAM. The total training time was approximately 3 days for 10 epochs. During inference, the model processed each image in approximately 0.07 seconds.

4.3 Comparison with conventional methods

We evaluated the performance of the proposed method against three comparison methods, a method without GAN (w/o GAN), an existing GAN-based method (GAN), and IBRNet[12] with its pretrained weights.

The method without GAN uses only the generator with the same structure as the proposed method, optimized solely using the $\mathcal{L}_1$ norm without the GAN framework. The existing GAN-based method employs a generator identical to the proposed method, but with a discriminator modified to use 2D convolutions for a single output viewpoint. IBRNet is a generalizable neural rendering approach that aggregates features from multiple input views to synthesize novel viewpoints.

For subjective evaluation, **Figures 3** and **4** show the generated image sequences with specular and transparent objects using the proposed method and the comparison methods. Each row represents a different method, while each column represents a viewpoint position on the camera array. First, as shown in Fig.3, IBRNet suffers from geometric inaccuracies, particularly around object boundaries and occluded regions, leading to oversmoothed surfaces and loss of fine structural details. Additionally, the method struggles to reproduce view-dependent material properties, resulting in diminished specular reflections and a less realistic appearance. although the method without GAN preserves the basic geometry, it lacks the ability to reproduce specular re-

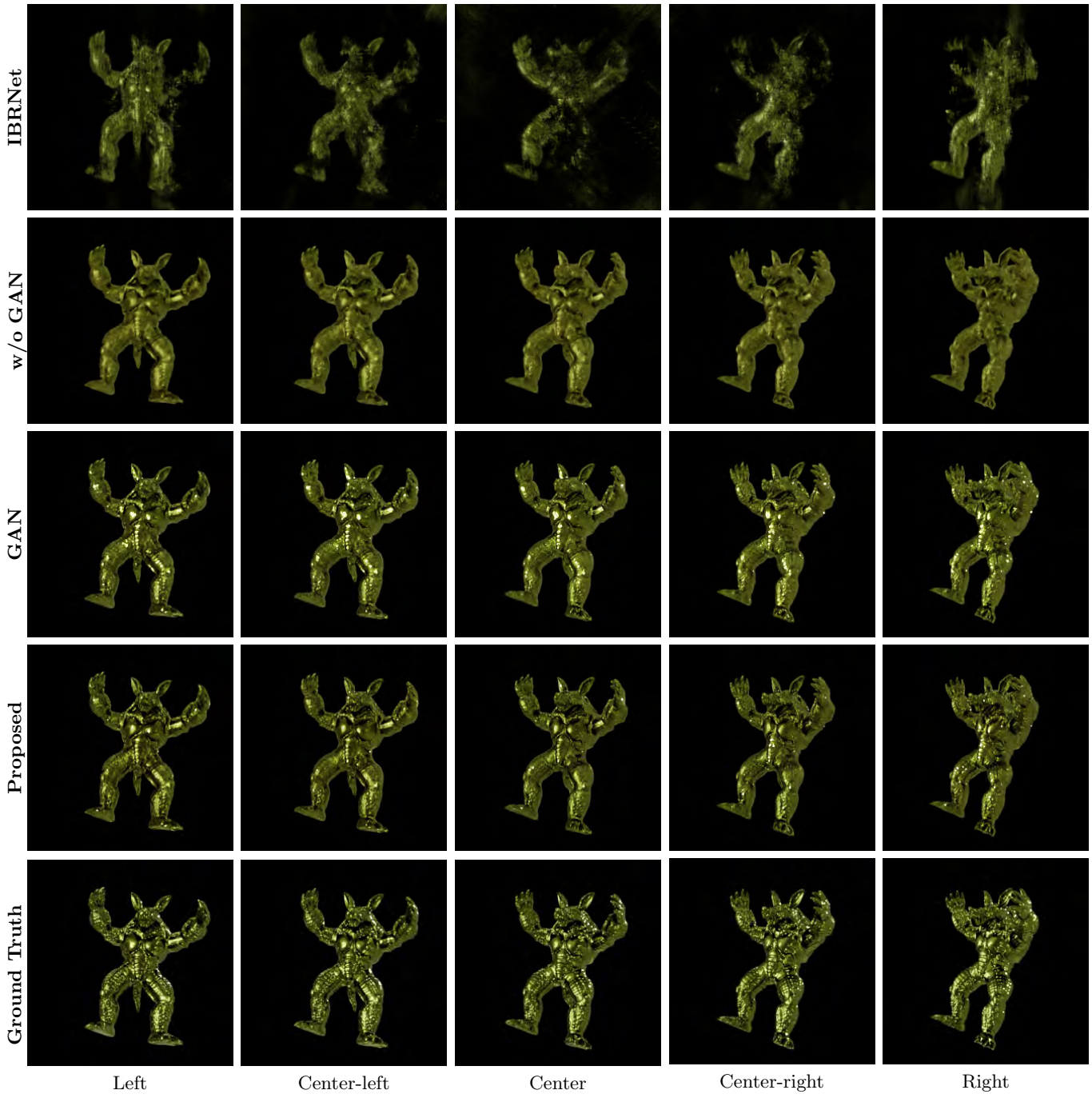


Fig. 3 Generated images with specular object showing horizontal viewpoint transitions at the middle row of the 5×5 camera array

flections, resulting in degraded material appearance with insufficient high-frequency details. In contrast, the existing GAN-based method better preserves specular reflections, resulting in more accurate material appearance. However, the method exhibits unnatural transitions of specular reflections between consecutive viewpoints, failing to reproduce the expected view-dependent changes in highlights. On the other hand, the proposed method successfully reproduces specular reflections while maintaining natural transitions of highlights between consecutive viewpoints. Next, as shown in Fig.4, IBRNet struggles with geometric inconsistencies in transparent regions, leading to over-smoothed surfaces and depth inconsistencies that reduce material clarity. The method without GAN reconstructs the basic shape but also produces blurry surfaces that fail to capture the characteristic transparency of the object. While both the GAN-based method and the proposed method successfully reproduce the basic transparent appearance, our approach better captures the fine highlights that appear and disappear depending on the viewpoint position, resulting in more convincing transparent material rendering. These

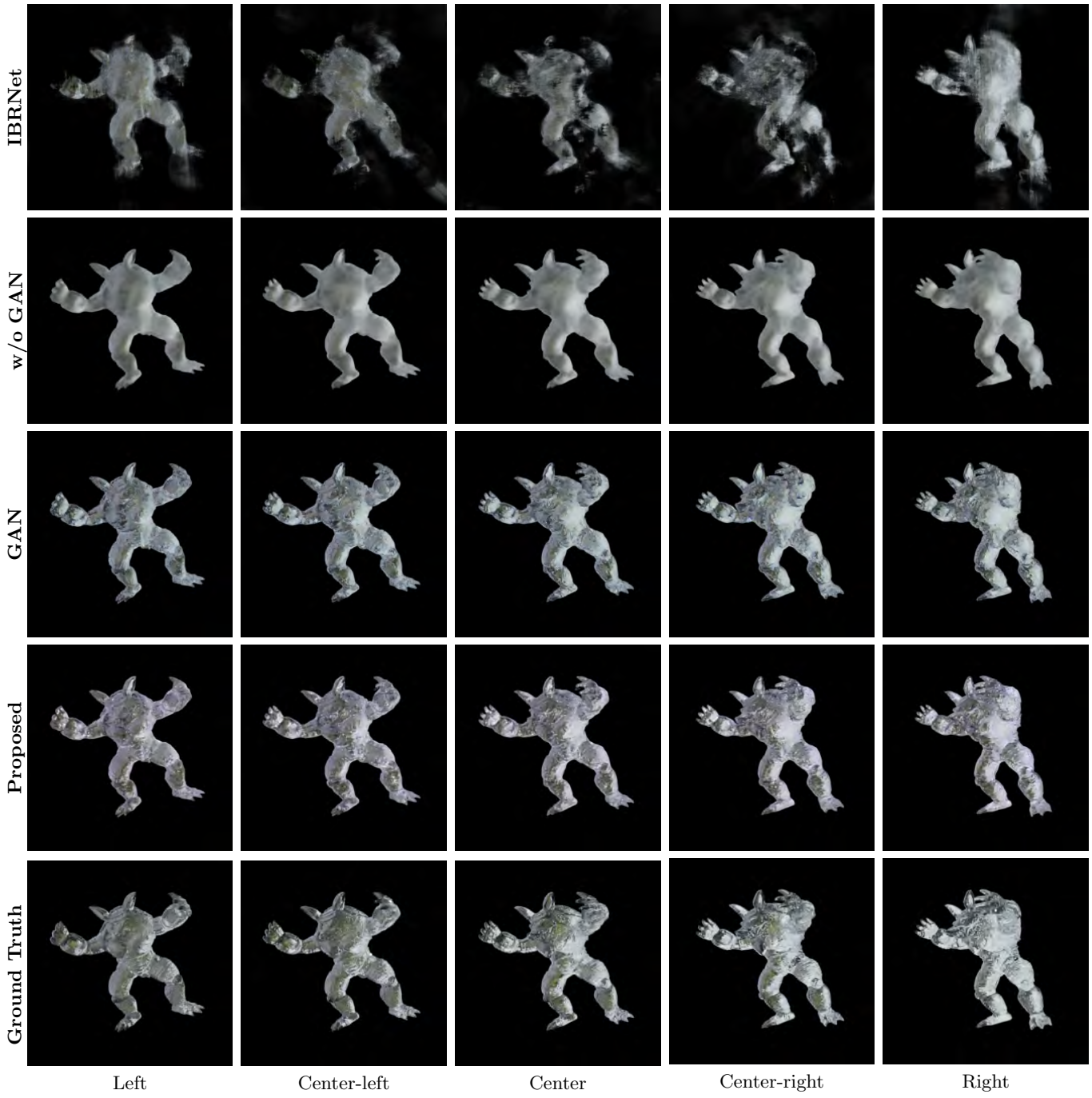


Fig. 4 Generated images with transparent object showing horizontal viewpoint transitions at the middle row of the 5×5 camera array

results demonstrate the ability of the proposed method to generate high-quality NVS that preserves both the fine details of specular and transparent surfaces while maintaining consistency of material appearance during viewpoint transitions.

For objective evaluation, we used peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS)[34]. In particular, LPIPS is an image quality metric closely aligned with human perception. **Table 2** shows the evaluation metrics for both specular and transparent materials. For specular objects, the proposed method achieves the best performance across all metrics. In contrast, for transparent objects, while the method without GAN shows better PSNR and SSIM, the proposed method achieves superior results in LPIPS. In particular, the LPIPS scores, which better correlate with human perception, show significant improvements with the proposed method for both materials.

Table 2 Quantitative evaluation results of proposed and comparison methods for specular and transparent materials

Specular	PSNR↑	SSIM↑	LPIPS↓
IBRNet	20.73	0.7423	0.1799
w/o GAN	25.85	0.9283	0.0795
GAN	25.72	0.9326	0.0474
Proposed	**27.25**	**0.9386**	**0.0414**
Transparent	PSNR↑	SSIM↑	LPIPS↓
IBRNet	21.07	0.8290	0.1758
w/o GAN	**27.73**	**0.9618**	0.1144
GAN	27.08	0.9579	0.0461
Proposed	27.02	0.9587	**0.0442**

Table 3 Ablation study with quantitative evaluation of different number of frames settings ($T = 5$, $T = 10$, $T = 15$) for specular and transparent materials

Specular	PSNR↑	SSIM↑	LPIPS↓
Proposed (T=5)	26.94	**0.9387**	0.0441
Proposed (T=10)	**27.25**	0.9386	**0.0414**
Proposed (T=15)	25.44	0.9286	0.0484
Transparent	PSNR↑	SSIM↑	LPIPS↓
Proposed (T=5)	**28.29**	**0.9663**	0.0486
Proposed (T=10)	27.02	0.9587	**0.0442**
Proposed (T=15)	27.27	0.9606	0.0540

4.4 Ablation study

To investigate the effect of temporal context given by the number of frames T input to our 3D convolutional discriminator, we conduct an ablation study with three different settings of $T = 5$, $T = 10$, and $T = 15$ frames.

Figures 5 and **6** show the generated image sequences for ablation study with specular and transparent objects, respectively. For specular objects, both $T = 5$ and $T = 10$ successfully reproduce specular reflections while maintaining natural transitions between consecutive viewpoints. In particular, $T = 10$ enhances the consistency of highlights across viewpoint transitions by using a longer sequence of frames for the discriminator. For transparent objects, $T = 5$ gives an impression closer to a matte texture rather than a transparent material, whereas $T = 10$ better reproduces the appearance of transparent objects. On the other hand, with $T = 15$ in both materials, the natural view-dependent transitions of appearance features are reduced. This result indicates that an excessive number of frames may strengthen the discriminator too much or cause overfitting.

Table 3 shows the quantitative results with PSNR, SSIM, and LPIPS evaluation metrics. For specular objects, $T = 10$ achieves the best performance in PSNR and LPIPS metrics, while $T = 5$ slightly outperforms SSIM. Performance deteriorates across all metrics when T increases to 15. For transparent objects, $T = 5$ shows superior PSNR and SSIM values, while $T = 10$ achieves the best LPIPS score. Increasing to $T = 15$ does not yield further improvements in any metric. Additionally, the computational cost of the 3D convolutional discriminator scales approximately linearly with sequence length. These results indicate that $T = 10$ provides a good balance between perceptual quality and computational efficiency in the proposed method.

5. Conclusion

In this paper, we proposed a GAN-based NVS method that focuses on generating naturally varying viewpoint-dependent material appearances during viewpoint transitions. Through adversarial training with a 3D convolutional discriminator, the proposed method reproduced natural material appearances derived from viewpoint-dependent intensity changes. Experimental results demonstrated the superiority of the proposed method in terms of both objective metrics and subjective evaluation, particularly in reproducing material appearances, and natural intensity changes during viewpoint transitions.

In our experiments, we focused on a structured camera array setup for systematic evaluation. Extending the experimental validation to more diverse camera configurations would further demonstrate the flexibility of the proposed method. Additionally, applying the proposed method to real objects captured by physical cameras presents challenges due to camera parameter uncertainties and complex lighting conditions. Furthermore, investigating the effectiveness of the proposed method for a broader range of materials, including anisotropic surfaces or materials with subsurface scattering, remains as future work. Addressing these directions could expand the potential applications of our approach in various fields, including VR. Moreover, the proposed method struggles to reproduce sharp specular reflections in objects with localized highlights due to dataset limitations and constraints on camera placement. To mitigate this limitation, it is necessary to introduce a more diverse set of 3D models that include shapes that clearly reveal specular reflections and relax constraints on camera placement.

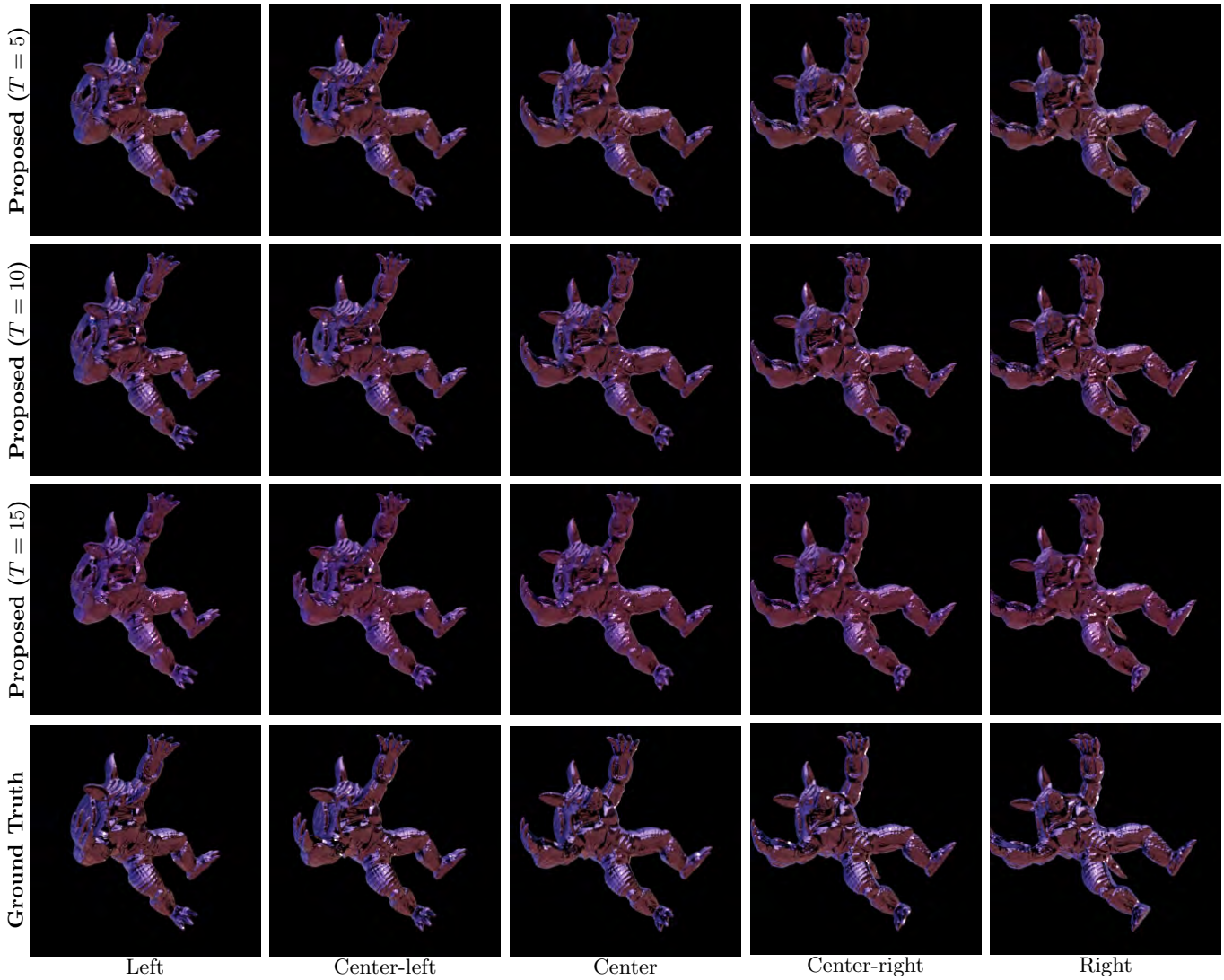


Fig. 5 Ablation study on the effect of temporal context given by the different number of frames settings ($T = 5$, $T = 10$, $T = 15$) in the 3D convolutional discriminator for specular objects

References

1) J. Flynn, I. Neulander, J. Philbin, N. Snavely: "Deep Stereo: Learning to Predict New Views from the World's Imagery", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5515–5524 (2016).
2) T. Zhou, R. Tucker, J. Flynn, G. Fyffe, N. Snavely: "Stereo Magnification: Learning View Synthesis Using Multiplane Images", ACM Transactions on Graphics, Vol. 37, No. 4, pp. 1–12 (2018).
3) R. Tucker, N. Snavely: "Single-View View Synthesis with Multiplane Images", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 551–560 (2020).
4) B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng: "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis", Communications of the ACM, Vol. 65, No. 1, pp. 99–106 (2021).
5) B. Kerbl, G. Kopanas, T. Leimkuehler, G. Drettakis: "3D Gaussian Splatting for Real-Time Radiance Field Rendering", ACM Transactions on Graphics, Vol. 42, No. 4, Article 139, pp. 1–14 (2023).
6) D. Verbin, P. P. Srinivasan, P. Hedman, B. Mildenhall, B. Attal, R. Szeliski, J. T. Barron: "NeRF-Casting: Improved View-Dependent Appearance with Consistent Reflections", SIGGRAPH Asia 2024 Conference Papers, Article 15, pp. 1–10 (2024).
7) H. Yoon, S. Lee: "Neural Radiance Fields for Transparent Object Using Visual Hull", The IEEE International Conference on Big Data and Smart Computing, pp. 375–376 (2024).
8) N. Moënne-Loccoz, A. Mirzaei, O. Perel, R. de Lutio, J. Martinez Esturo, G. State, S. Fidler, N. Sharp, Z. Gojcic: "3D Gaussian Ray Tracing: Fast Tracing of Particle Scenes", ACM Transactions on Graphics, Vol. 43, Article 232, pp. 1–19 (2024).
9) T. Xie, X. Chen, Z. Xu, Y. Xie, Y. Jin, Y. Shen, S. Peng, H. Bao, X. Zhou: "EnvGS: Modeling View-Dependent Appearance with Environment Gaussian", arXiv preprint arXiv:2412.15215 (2024).
10) A. Yu, V. Ye, M. Tancik, A. Kanazawa: "PixelNeRF: Neural Radiance Fields from One or Few Images", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4578–4587 (2021).
11) Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, T. Funkhouser: "IBRNet: Learning Multi-View Image-Based Rendering", Proceedings of the IEEE/CVF Conference on Computer Vision and

Fig. 6 Ablation study on the effect of temporal context given by the different number of frames settings ($T = 5$, $T = 10$, $T = 15$) in the 3D convolutional discriminator for transparent objects

Pattern Recognition (CVPR), pp. 4688–4697 (2021).

12) J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, L. Gu: "DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20775–20785 (2024).
13) Y. Xiao, D. Zhai, W. Zhao, K. Jiang, J. Jiang, X. Liu: "MCGS: Multiview Consistency Enhancement for Sparse-View 3D Gaussian Radiance Fields", arXiv preprint arXiv:2410.11394 (2024).
14) J. Xu, S. Gao, Y. Shan: "FreeSplatter: Pose-free Gaussian Splatting for Sparse-view 3D Reconstruction", arXiv preprint arXiv:2412.09573 (2024).
15) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio: "Generative Adversarial Nets", Advances in Neural Information Processing Systems, Vol. 27, pp. 2672–2680 (2014).
16) J. Ho, A. Jain, P. Abbeel: "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems, Vol. 33, pp. 6840–6851 (2020).
17) B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, J. Feng: "Multi-View Image Generation from a Single-View", Proceedings of the 26th ACM International Conference on Multimedia, pp. 383–391 (2018).
18) X. Xu, Y. Chen, J. Jia: "View Independent Generative Adversarial Network for Novel View Synthesis", Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7790–7799 (2019).
19) C. Hoshizawa, T. Komuro: "View Interpolation Networks for Reproducing the Material Appearance of Specular Objects", Virtual Reality & Intelligent Hardware, Vol. 5, No. 1, pp. 1–10 (2023).
20) T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, Y.-L. Yang: "HoloGAN: Unsupervised Learning of 3D Representations from Natural Images", Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7588–7597 (2019).
21) K. Schwarz, Y. Liao, M. Niemeyer, A. Geiger: "GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis", Advances in Neural Information Processing Systems, Vol. 33, pp. 20154–20166 (2020).
22) D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, M. Norouzi: "Novel View Synthesis with Diffusion Models", arXiv preprint arXiv:2210.04628 (2022).
23) E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J.

Park, A. Levy, M. Aittala, S. De Mello, T. Karras, G. Wetzstein: "Generative Novel View Synthesis with 3D-Aware Diffusion Models", Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4217–4229 (2023).

24) L. McMillan, G. Bishop: "Plenoptic Modeling: An Image-Based Rendering System", SIGGRAPH '95: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, pp. 39–46 (1995).

25) P. E. Debevec, C. J. Taylor, J. Malik: "Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach", SIGGRAPH '96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 11–20 (1996).

26) S. B. Kang, R. Szeliski: "Extracting View-Dependent Depth Maps from a Collection of Images", International Journal of Computer Vision, Vol. 58, pp. 139–163 (2004).

27) S. E. Chen, L. Williams: "View Interpolation for Image Synthesis", SIGGRAPH '93: Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques, pp. 279–288 (1993).

28) S. M. Seitz, C. R. Dyer: "View Morphing", SIGGRAPH '96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 21–30 (1996).

29) T. Stich, C. Linz, G. Albuquerque, M. A. Magnor: "View and Time Interpolation in Image Space", Computer Graphics Forum, Vol. 27 (2008).

30) O. Ronneberger, P. Fischer, T. Brox: "U-Net: Convolutional Networks for Biomedical Image Segmentation", Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015).

31) G. Turk, M. Levoy: "Zippered Polygon Meshes from Range Images", SIGGRAPH '94: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, pp. 311–318 (1994).

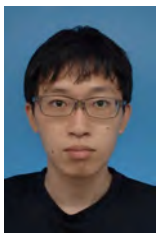
32) Poly Haven, `https://polyhaven.com/` (2024).

33) D. P. Kingma: "Adam: A Method for Stochastic Optimization", arXiv preprint arXiv:1412.6980 (2014).

34) R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang: "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018).

(Received October 31, 2024)
(Revised March 10, 2025)

Taishi IRIYAMA (*Member*)
He received the B.E., M.E. and Ph.D. degrees in electronic information engineering from Tamagawa University, Tokyo, Japan, in 2017, 2019 and 2022, respectively. Currently, he is assistant professor of mathematics, electronics and informatics at Saitama University.



Chihiro HOSHIZAWA
He received the B.E. and M.E. degrees in information and conmputer sciences from the Saitama University, Saitama, Japan, in 2022 and 2024, respectively. He is currently working as a software engineer at ibis inc.

Takashi KOMURO
He received the B.E., M.E., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo in 1996, 1998, and 2001, respectively. At present he is a professor of mathematics, electronics and informatics at Saitama University.

IIEEJ System Development Paper -- Special Issue on Image Electronics Technologies Related to VR/AR/MR/XR --

Crossing over Virtual and Real Experiences through Digital Twin Partner Robot

Ryota SUZUKI† (*Member*) , Yuto ISHIYAMA† , Yoshinori KOBAYASHI†

† Saitama University

<**Summary**> In this study, we propose a method of enhancing the user experience through collaboration with a digital twin partner robot in order to connect between VR experience and real experience. In recent years, as the performance of head-mounted displays has been improved and their prices have become lower, virtual reality (VR) technology has been used not only in the entertainment field, but also in a variety of other fields such as education and medical care. Against this backdrop, many studies have been conducted to explore the potential applications of VR technology. However, conventional VR technologies only provide a one-time and personal experience, which makes it difficult to reinforce the experience by reflecting on and sharing with family or friends, which is often done in real life. By using our system, an avatar agent with the same appearance and voice as a real-life personal robot goes to a VR space and experiences VR contents together with a user. After returning to reality, the user recalls the VR experience together with a real robot that has the same personality as the avatar agent in the VR space. We aim to enhance the VR experience by working with the digital twin partner robot in both real world and virtual world to connect their experience.

Keywords: human-robot interaction, virtual reality, digital twin

1. Introduction

In recent years, as the performance of Head-Mounted Displays (HMDs) has improved and become less expensive, Virtual Reality (VR) technology has been applied and studied in various fields, not only in the entertainment field such as games. VR technology can provide novel experiences that are difficult to realize in reality, and there is a movement to develop familiar services such as VR theme parks that provide VR content in shopping malls. In addition, VR can provide spatio-temporal augmentation of experiences in real space, such as immersive experiences of remote places and past events in the COVID-19 pandemic, such as VR museums, VR travel, and VR city walking experiences that guide you through a city while showing you virtual images of the past.

Immersion in VR provides new user experiences, but these experiences are often completed within VR. This means that virtual and real diverge, and the VR experience remains a personal, ad hoc experience. For example, after traveling in reality, it is common to look back on the experience with family and friends while viewing photos, which is thought to connect the past experience to current and simultaneously enhance the experience. On the other hand, it is difficult to look back on an experience in VR because of the discontinuity between the virtual experience and real. As a result, the effect of the VR experience such as excitement and impression like perceived presence should be limited.

Some studies attempting to connect between real and virtual space have been conducted. For example, there are those that use completely realistic images and sounds to stimulate the illusion of reality, those that blur the boundary between the VR world and real world through interactions such as opening an AR door then entering a virtual space, and those that provide mental support for returning from the VR game world to real world. However, most of these studies focus on spatial connectivity and do not maintain or enhance the persistence and connectivity of the experience.

Against the problem, we propose a method of enhancing the VR experience through collaboration with a digital twin partner robot. We aim to augment the effects of VR experiences, which are still limited to one-time, personal experiences, by connecting experiences mutually mediated by the experience itself, i.e., collaborative experiences in both spaces with social digital twin robots (see **Fig. 1**).

The purpose of this study is to verify the following two

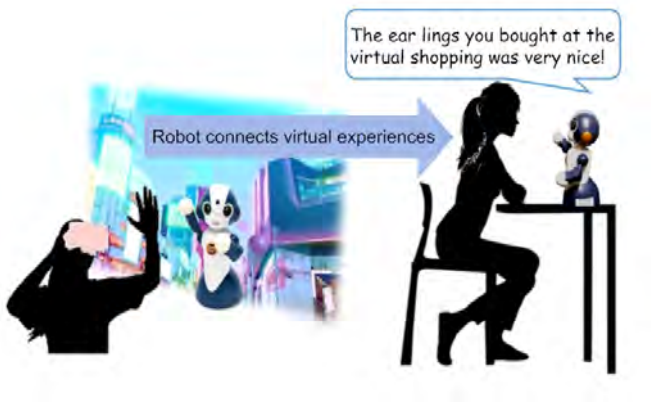


Fig. 1 A digital twin partner robot connects between experiences of virtual and real

points. (i) Can we give a sense of continuity between the virtual and real experiences? (ii) Does it augment the effects of the VR experience? Therefore, we develop a VR system that enables users to experience VR sightseeing contents with a personality robot that has the same appearance, voice, and behavior in the VR space. We then conduct experiments to verify whether the quality of the user experience improves with or without a partner robot.

2. Related Work

There has been vigorous debate about how to connect virtual and real, and how much social interaction affect to user experience in virtual reality. In this chapter, we introduce related works on the points, and discuss our position compared to the works.

2.1 Blurring spatial boundary between virtual and real

Steinicke et al. tackled with a primitive research about blurring spatial boundary between virtual and real by going to virtual reality via transitional environment, which is the virtual but has imitated appearance of real environment that a participant is in[1]. The work reported that usability score on questionnaire and behavioral movement got larger compared with the baseline of immediate transition to virtual reality.

George et al. also implemented a system that efficiently and mutually transit virtual and real world, on the point of usability of accessing resources in both worlds[2].

Ito et al. are exploring a Substitutional Reality (SR) system that makes what we are seeing seem real to the subjective reality we perceive[3]. This system could automatically mix real time video with recorded video in time and space, using the eye of the user and image processing information. Specifically, they proposed a method that used a head mounted display equipped with an eye tracker to automatically switch images based on eye position information. The system was designed so that the user did not notice the video switching, utilizing the concepts of change blindness and inattentional blindness.

Soret et al. conducted a comparative study of ways to adapt transitions to the virtual environment to maximize the user experience[4]. This paper explores ways to ease the transition from the virtual environment to the real environment. In particular, they use a "door transition" method at the end of the virtual environment and compare its effectiveness to other basic transitions (direct and fading transitions). In the study, participants experienced stress management by watching images of their natural environment in virtual reality , after which they experienced different transition methods. It was found that the door transition provided a smoother transition to reality and gave participants a greater sense of control compared to the other methods. It was also found that this method caused less cybersickness (discomfort in virtual space, VR sickness) than other transition methods.

Billinghurst et al. demonstrated Macigbook, the book reader system using both AR and VR technology collaboratively. In the system, an AR character is shown on a book to play a scene written in the opened page, and then users can fly into immersive VR that the character plays the story[5].

Compared with these research, we intend to bring virtual experience itself to real through social interaction with a partner robot. We expect that connection of experiences should directly enhance experiences such as strength of satisfactory and memorability rather than slight connection of the spaces.

2.2 Effects of social interaction and context

Roth et al. researched about relationship between avatar realism and social interaction[6].

Imada et al. conducted an experiment to compare learning outcomes in online self paced learning with and without other avatars to overcome difficulty of persistence in online self paced learning[7]. This paper focused on ways to support the continuation of self paced learning (a method of learning at one's own pace) within a VR space. In particular, it examined how watching other avatars continue learning tasks in the same VR space affected the continuation and effectiveness of learning. They conducted a comparison experiment in which the presence of other avatars was a variable, and found that self paced

learning lasted twice as long and was 1.5 times more effective in a space where other avatars were present. The study also suggested that watching others' learning had a positive effect on learners' motivation and concentration. This study presented an effective means of increasing learner retention and effectiveness in online learning environments.

Some studies have investigated and reported the effects of VR co-experience with other people. Conventional studies have studied the effects of communication with others in media other than VR[8], and recently, researches have been conducted on its application to VR. For example, Bulu et al. investigated the influence of others' presence on satisfaction with VR educational content[9]. In addition, Cho et al. reported an empirical study investigating the presence of other avatars in educational situations and its impact on collaborative work[10]. Thus, it is expected that the presence of others can enhance the VR experience, but on the other hand, not all VR experiences can be conducted in collaboration with friends or other people. In contrast, our digital twin partner robot is expected to realize an effective cooperative experience without limitation of the availability of other people.

In our study, we aim to enhance VR experience by connecting the experience to real through interaction with a partner robot. This is the new challenge to apply robot interaction for enhancing VR experience, while the other works only consider for virtual avatars or interaction with the other human users.

3. VR Collaborative Experience System with a Partner Robot

In this study, we propose a method of enhancing the user experience through collaboration with partner robots in order to improve the connectivity between experiences of virtual reality and reality. In this system, an avatar agent with the same appearance and voice as a real life personal robot goes to a VR space and experiences VR sightseeing contents together with the avatar agent. After returning to reality, the user recalls the VR experience with a real robot that has the same personality as the avatar agent in the VR space. Our aim is to enhance the VR experience spatio-temporally, even after coming back to reality, through having shared experiences with a digital twin partner robot, that has the same appearance and memories between the virtual and real world.

(a) PLATEAU data around Shirakawa Station

(b) Reproduced data around Shirakawa Station

Fig. 2 The original and reproduced 3D model data around Shirakawa Station

3.1 VR sightseeing contents

We newly constructed a VR space for VR sightseeing experience. The destinations were Shirakawa Station and Shirakawa City Library in Shirakawa City, Fukushima Prefecture, Japan. The 3D model of VR space was created using PLATEAU provided by the Japan Ministry of Land, Infrastructure, Transport and Tourism[11] . PLATEAU is an open project that aims to develop 3D city models of all Japanese cities. The project aims to realize an urban digital twin, and is working on examples of solution development using 3D city models and solutions to social issues in various fields. It provides a data platform for 3D city models. We reproduced a downloaded VR city data of the area around Shirakawa Station in Shirakawa City, and bringing it into the 3D modelling application Blender. We then constructed more detailed 3D models of the destinations (Shirakawa Station and Shirakawa City Library) manually. PLATEAU data has shape of ground and its texture, and structures of buildings including height and site. However, it does not have detailed shapes and their texture of buildings. For our VR tourism content, the VR environment need have more detailed reconstruction of landmarks that enable a user to feel and discuss about the locations. The original and reproduced 3D model data is shown in **Fig. 2**.

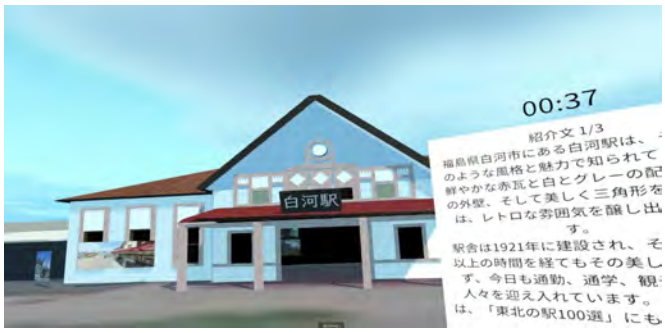


Fig. 3 Photos and a guidance board around the 3D model of Shirakawa Station

3.2 Implementation of VR sightseeing

The VR experience system was developed in Unity. The created 3D model of Shirakawa City was loaded in the Unity space. A user can experience immersive VR sightseeing through a VR hardware set VIVE pro (HTC Corporation). A user can walk around freely using a controller, and they also can look around by turning their neck.

Additionally, there were several real pictures of each destination around them. Moreover, guidance boards which showed explanation sentences stood around the destinations and their sentences were also read out by voice when a user come to the destinations (see **Fig. 3**).

The VR tour contents are designed to allow the participants to travel around each location in approximately 6 minutes.

3.3 Implementation of a digital twin partner robot

We use the communication robot Sota (Vstone co., ltd.) as a partner robot. Sota is programmable with the application Vstone Magic, and it interacts with a user using its arms, head, and voice. Moreover, we constructed a 3D model of Sota as its VR avatar. The avatar moves alongside a user in VR space, and so a user can travel the VR space with the partner robot. The avatar can interact to a user with its arms, head and voice as it does in real (see **Fig. 4**). As a detail, Sota in VR is always in front of a user, and looks towrad the destination while a user controls to move foward, and it turns toward a user when a user stops to see around.

4. Experiment

In order to verify whether this system can make users feel as if experiences between reality and virtual reality are connected, and to verify its effect, we conducted experiments. In the experiments, participants traveled the VR sightseeing contents with the partner robot. After the

(a) Sota (b) Moving with Sota in VR

(c) Talking with Sota in VR

Fig. 4 Behavior of the partner robot Sota in real and virtual

experience, we conducted a questionnaire and a memory test.

In this paper, we conducted two experiments. In the first experiment, the partner robot only play-backed the programmed scenario, and it did not react to the participants (Case I). On the other hand, in the second experiment, the partner robot reacted to the participants by Wizard-of-Oz (Case II). Through the two experiments, we observe how engagement between a user and the partner robot affects to user experience and connectivity between virtual and real.

4.1 Case I: Playback preset scenario

The subjects were seven students at university. Six were male and the other one was female. They were undergraduate and graduate students between the ages of 21 and 25.

Participants experienced the VR tour contents under both two conditions of with and without the digital twin partner robot. On the with-robot condition, at first, the robot made greeting to a participant in real world, and in the virtual world, it moved alongside the participant and spoke preset skit including explanation of buildings. We also prepared the without-robot condition as a baseline. For the condition, the staff instructed to wear HMD and move around the VR space, and the participants read sentences of the guidance board by themselves. The memory test and the questionnaires were conducted on sheets of paper. The order of the two conditions shuffled for each participant, and the content for each condition had been changed. In other words, a participant first experienced one of the locations under one of two robot conditions

and then experienced the other location under the other condition. The patterns were randomly selected for each participant.

Figure 5 shows samples of experimental scenes. During the experiment, the participants were totally free to walk around the environment, so the time of experience was varied for each participant and experience. Some participants moved directly to the location, and others were often stopped to look around the surroundings. When a participant wears an HMD, the partner robot in VR looks toward the participant. Then it talks to the participant to go forward to get to Shirakawa Station / Shirakawa City Library. In the VR environment, participants can move with a controller. While a participant is moving, the robot leads him/her. It takes almost 30 seconds to get to the target location if the participant keeps moving. While moving, the participants can see reconstructed virtual environments including other buildings such as an antenna tower. On reaching the target location, the guidance starts. The participant also looks and walks around as s/he likes during the guidance. After the guidance, the participants can walk around to watch the target location, among which several real photos are shown aside the location. The participants also can read the guidance board again. Then the participant feels enough, the participant takes off the HMD.

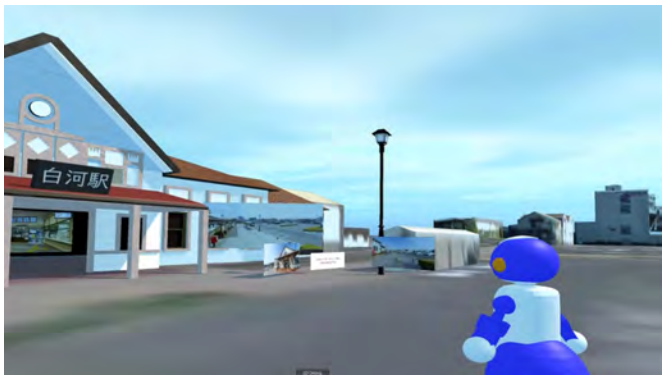
After traveling each location, they returned back to real world and had a memory test and a questionnaire. As the memory test, the robot asked the participant 10 questions about the contents. We marked the score as 2/1/0 points for a correct/related/wrong answer, respectively. **Tables 1** and **2** show questions and their correct answers of the tests. All correct answers were contained in the words of the guidance. And **Table 3** shows sentences of the questionnaire. The participants answered the questions by 7 level Likert scale (7 for strongly agree, 4 for neutral and 1 for strongly disagree).

Figure 6 shows the result of the questionnaire. As the summary of the result, it can be said that those who experienced with the partner robot felt motivation of experiencing VR and enjoyment more than without the partner robot. In Q.1-5, we asked peripheral questions to confirm whether the system has some barrier of usability or not on the experience. Among the questions, all the results about perceptual quality (Q.1), understandability (Q.2), ease of use (Q.3), concentration (Q.4), and stress (Q.5) were positive and there were little differences between the two conditions. Q.1-4 are the positive questions, and ev-

(a) Immediately after immersion

(b) 10 seconds after staring

(c) 20 seconds after staring



(d) Arrival

Fig. 5 Example scenes of VR sightseeing

Table 1 The questions and answers of memory test for trip to Shirakawa Station

Q.1	Which prefecture is Shirakawa Station in? (Fukushima)
Q.2	What color is the outer wall of Shirakawa Station? (white)
Q.3	In which year was Shirakawa Station built? (1887)
Q.4	What selection was Shirakawa Station selected? (Top 100 stations in Tohoku)
Q.5	What stained glass is in Shirakawa Station? (brightly colored stained glass)
Q.6	What kind of information is posted on the walls inside Shirakawa Station? (tourist information)
Q.7	What is the name of the restaurant in the waiting room of Shirakawa Station? (Eki-cafe SHIRAKAWA)
Q.8	What material are the window frames at Shirakawa Station made of? (old wood)
Q.9	What information does Shirakawa Station continue to convey to visitors? (its charm and history)
Q.10	What is the shape of the roof of Shirakawa Station? (triangle)

Table 2 The questions and answers of memory test for trip to Shirakawa City Library

Q.1	Which prefecture is Shirakawa City Library in? (Fukushima)
Q.2	How many minutes does it take to walk from Shirakawa Station to Shirakawa City Library? (about 5 minutes)
Q.3	Which historic site is Shirakawa City Library designed to preserve views to? (Komine castle)
Q.4	What is the distinctive element of Shirakawa City Library's design? (the large pitched roof)
Q.5	When did Shirakawa City Library open? (July 2011)
Q.6	What is the average number of visitors to Shirakawa City Library per day? (1000)
Q.7	How many square meters is the site area of Shirakawa City Library? (3900)
Q.8	Which company was responsible for the design of Shirakawa City Library? (Arup)
Q.9	What award has Shirakawa City Library received? (Lighting Promotion Award, Tohoku Architecture Award, Architecture Culture Award in Fukushima)
Q.10	How many meters is the maximum height of Shirakawa City Library? (12.7)

ery score was higher than the neutral score 4, and Q.5 is the negative question, and its score was lower than the neutral. So, it can be said that the usability of the system was not a barrier to the experiences. In Q.6-11, we asked user experiences. The largest difference was observed in the enjoyment of having a companion (Q.10). The next was the psychological distance and hurdles between the virtual and real worlds (Q.11). The third was enjoyment of the experience (Q.9). It indicates that the VR experience with a partner robot enhanced the enjoyment

Table 3 Questionnaire of Case I experiment

Q.1	Did you feel quality of the experience was high?
Q.2	Did you feel the system was easy to understand?
Q.3	Did you feel the system was easy to use?
Q.4	Could you concentrated on the experience?
Q.5	Did you feel the guidance of the sight stressful?
Q.6	Will you recommend to your friends?
Q.7	Do you think the system is usable for strong memory construction or enhancement?
Q.8	Did you feel positive through the experience?
Q.9	Did you feel enjoyment through the experiment?
Q.10	Did you feel enjoyment of a companion being there?
Q.11	Did you feel psychological distance or any hurdle while transiting to virtual and real?

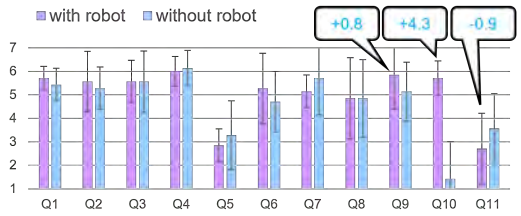


Fig. 6 Result of the questionnaire of Case I experiment

of the experience and degraded the psychological barrier between virtual reality and reality. The scores of motivations for recommendation (Q.6) and overall motivation (Q.8) were not so different between the two conditions. We also asked subjective presence of memorability in Q.7. There was not significant difference between the scores, but the average score of with-robot condition was larger than the score of without-robot condition.

In the memory test, the averages scores of with-robot and without-robot were 13.5 and 15.1 of 20, respectively. We executed T-test but significant difference was not shown. Although under the without-robot condition the participants were able to read the guidance board repeatedly without the guide of reading speed such as the spoken text for deep understanding, they got the compatible score against the with-robot condition. Therefore, the with-robot condition may have capability of enhancing memorability of experience.

4.2 Case II: Applying adaptive interaction by WoZ

The participants were seven university students. They were undergraduate and graduate students between the ages of 21 and 24. As same as Case I, each participant experienced two conditions of with- and without-robot, and the order of the experienced conditions were shuffled. Unlike the preliminary experiment, the participants chatted with Sota for 5 minutes before the VR experience.

Table 4 Questionnaire of Case II experiment

Q.1	I can remember the experience as I experience now.
Q.2	I can remember the experience as I see now.
Q.3	I can remember the sound and voice as I here now.
Q.4	I can clearly imagine the events of the experience.
Q.5	I can clearly imagine special layout (like which thing to where).
Q.6	I can remeber the words in the experience.
Q.7	I can tell the experience as the well-structured story.
Q.8	I can remember the detail and talk to others.
Q.9	Time and contents are not clear, and I only remember the brief.
Q.10	The same feeling coming up when I remember the events.
Q.11	The feeling is very negative.
Q.12	The feeling is very positive.
Q.13	I can use the system intuitively.
Q.14	It is easy to remember the VR experience.
Q.15	I feel the VR experience as it is in another world.
Q.16	I feel the VR experience as it is a part of my real experience.
Q.17	I feel me at VR is different from me at real.
Q.18	I feel me at VR is the same as me at real.

We had prepared about 100 patterns of conversation with behavior in advance, and the staff selected which pattern to play in realtime on the chat as Wizard-of-Oz. The participants chatted with Sota on two topics of "Dinner last night" and "Recent trips". The content is almost the same as that explained in Section 4.1. Moreover, Sota spoke to the participants several times during the VR experience. It uttered words that expressed empathy for the focusing target. The patterns of words and behaviors were also programmed in advance and the staff selected the patterns in realtime. The robot no more read out the guidance, and the guidance board read the skit instead, as the robot could be felt as a partner or a friend of the participants.

A memory test was also conducted, but the procedure was changed. Sota directly made unstructured interview by asking "Tell me the things you remember about the experience." and "Anything else?", and give some responses like "Yes". The questionnaire is conducted as same manner as the preliminary experiment. **Table 4** shows the sentences. The sentences are constructed with reference to Sekiguchi's psychological study about recall of episodic memory[12].

Figure 7 shows the result of questionnaire. Unlike in the preliminary experiment, significant differences were found in many of the questionnaire items in this experiment, with all items except Q.16 and Q.18 showing significant differences between with and without-robot condition. It can be said that the users obviously can have vividness of recalled experience (Q.1, Q.2, Q.3, Q.4, Q.5), the memory that could be structured by words (Q.6, Q.7, Q.8), clarity of experience in the memory (Q.9), synchronicity of feelings against the recalled experience (Q.10), motivation of the experience (Q.11, 12), usability of the system (Q.13), easiness of recall (Q.14), sense of continual presence of the past experience (Q.15, Q.17), and enough positive relative to neutral on sense of VR experience is the same as reality (Q.16, Q.18).

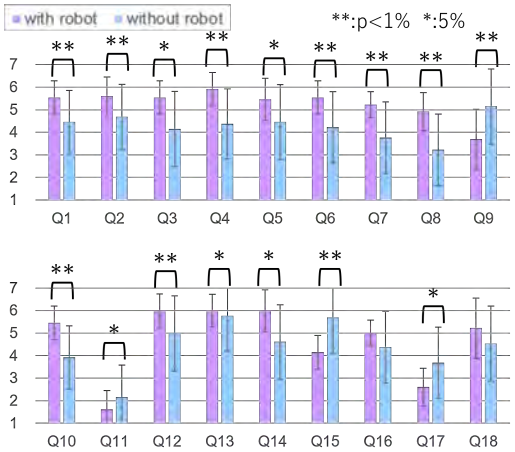


Fig. 7 Result of questionnaire of Case II experiment

As an evaluation of the memory test, we counted mora of pronounce in Japanese. Higher number of mora means the participant talked many and long time, and so the participant remembered many things. The averaged mora of with-robot condition is 1.37 times larger than without-robot condition. However, significant difference at 5% level was not observed. We then arbitrary categorized and counted a number of kinds of information in the words. For example, "The station had nameplate, and it was written in Kanji." scored as 2 ("station nameplate" and "written in Kanji"), and "I saw several houses and no cars, and the sky is blue." as 3 ("houses", "no cars" and "blue sky"). We adopted information even though it was wrong or not related to the target location, while our aim is to confirm that the system of experiencing and recalling together strengthens motivation for connecting virtual and real experiences and deepens the impression of the whole experience. The number of information of with-robot condition is 1.42 times larger than without-robot condition, and significant difference at 5% level was observed.

The experimental results showed that the proposed method produced positive effects on almost all items. It can be said that the proposed method significantly

enhanced the effects in terms of positive emotions, intention to use the method in the future, and memory enhancement. We found that the VR experience can predominantly increase satisfaction, awareness, and test scores. Therefore, we could summarise that this method enhanced the effect of the VR experience.

5. Conclusion

In this study, we measured the effects of a cooperative VR experience with a partner robot. After communicating with a real partner robot, the participants experienced a collaborative VR experience with an avatar that had the same appearance, voice, and behavior as the real partner robot. The results showed that there was a significant difference in the effect of the VR experience with and without the robot. It was confirmed that the experience gave a sense of continuity between the virtual and real worlds, triggered positive emotions, increased the awareness of reuse, and increased the amount of memory of the experience. Even though it does not aim to efficiently take bodily skills obtained in VR out to reality such as sports or surgical training, we expect to adapt the system not only to VR tourism but other situations that require continuity of experience including memory such as education and meetings.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 23K13286, 23KK0032, 24H00151.

References

1) F. Steinicke, G. Bruder, K. Hinrichs, A. Steed, A.L. Gerlach: "Does a Gradual Transition to the Virtual World increase Presence?," 2009 IEEE Virtual Reality Conference, pp.203–210 (2009).
2) C. George, A.N. Tien, H. Hussmann: "Seamless, Bi-directional Transitions along the Reality-Virtuality Continuum: A Conceptualization and Prototype Exploration," 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp.412–424 (2020).
3) D. Ito, K. Takahara, R. Sakamoto, A. Izumihara, S. Wakisaka, A. Hiyama, M. Inami: "EyeHacker: gaze-based automatic reality manipulation," ACM SIGGRAPH 2019 Emerging Technologies, pp.1–2 (2019).
4) R. Soret, A.-M. Montes-Solano, C. Manzini, V. Peysakhovich, E.F. Fabre: "Pushing open the door to reality: On facilitating the transitions from virtual to real environments," Applied Ergonomics, No.97, p.103535 (2021).
5) M. Billinghurst, H. Kato, I. Poupyrev: "MagicBook: transitioning between reality and virtuality," CHI '01 Extended Abstracts on Human Factors in Computing Systems, p.25–26, CHI EA '01, Association for Computing Machinery, New York, NY, USA (2001).
6) D. Roth, J.-L. Lugrin, D. Galakhov, A. Hofmann, G. Bente, M.E. Latoschik, A. Fuhrmann: "Avatar realism and social interaction quality in virtual reality," 2016 IEEE Virtual Reality (VR), pp.277–278 (2016).
7) S. Imada, N. Hayashida, H. Kuzuoka, K. Suzuki, M. Oki: "Making Others' Efforts Tangible," HCI International 2020 - Posters, eds. by C. Stephanidis, M. Antona, pp.239–247, Springer International Publishing, Cham (2020).
8) S. Morrison-Smith, J. Ruiz: "Challenges and barriers in virtual teams: a literature review," SN Applied Sciences, Vol.2, No.6, p.1096 (2020).
9) S.T. Bulu: "Place presence, social presence, co-presence, and satisfaction in virtual worlds," Computers & Education, Vol.58, No.1, pp.154–161 (2012).
10) Y.H. Cho, S.Y. Yim, S. Paik: "Physical and social presence in 3D virtual role-play for pre-service teachers," The Internet and Higher Education, Vol.25, pp.70–77 (2015).
11) The Japan Ministry of Land, Infrastructure, Transport and Tourism, PLATEAU, https://www.mlit.go.jp/plateau/ (2020).
12) R. Sekiguchi: "Phenomenological and Subjective Characteristics of Autobiographical Memory Recall: A Study Using the Subjective Characteristics Questionnaire for Autobiographical Episodic Memory (in Japanese)," Kansai University Psychological Research, Vol.2, pp.1–17 (2011).

(Received November 15, 2024)
(Revised March 31, 2025)

Ryota SUZUKI (*Member*)
After receiving the Ph.D. degree in 2016 from the Graduate School of Saitama University, he joined National Institute of Advanced Science and Technology until 2022. Since 2022, he currently working at Saitama University as Assistant Professor. His research interests include VR and AI technology and its application to human-robot interactoin.

Yuto ISHIYAMA
He had belonged to the master course of the Graduate School of Saitama University in 2022–2024, and he received M.Eng. degree in 2024.

Yoshinori KOBAYASHI
After completing the M.S. degree at the University of Electro-Communications in 2000, he joined Mitsubishi Electric Corporation. He received the Ph.D. degree in 2007 from the Graduate School of Information Science and Technology, the University of Tokyo. In 2007, he joined the Department of Information and Computer Sciences, Saitama University, as an Assistant Professor. Since 2020, he has been a Professor. His research interests include human sensing and its application to human-robot interaction.

IIEEJ Paper

Basic Design and its Performance of Multiplication-Free Multi-Alphabet Arithmetic Code for Markov-Model Sources

Fumitaka ONO[†,‡] ***(Honorary Member)***, Kazuto KAMIKURA[†] ***(Member)***, Yousun KANG[†] ***(Member)***

† Tokyo Polytechnic University, ‡ The University of Tokyo

<Summary> In order to apply the arithmetic coding for multi-level Markov-model sources, most conventional methods were using binary arithmetic codes by decomposing multi-level Markov-model sources into plural binary sources, which we here call B-B coding. The advantages of B-B coding are the possibility of utilizing the study results of binary arithmetic codes but it may need longer time by the increase of coding/decoding the expanded binary sources. The usage of multi-alphabet arithmetic code to multi-level Markov-model sources, which will be called as M-M coding here, and the comparison with B-B coding will be very interesting but it has been impossible since there are almost no report concerning the practical multi-alphabet arithmetic coding. In this paper, we will try to propose the basic design and the practical guideline of multiplication-free multi-alphabet arithmetic code (MFMAC) which will make the comparison of the total performance of B-B coding and M-M coding, possible. The basic design of MFMAC is composed of model source assumption, symbol area determination rule, designing of model parameter set for static coding, model parameter estimation for dynamic coding, and symbol ranking detection and update. We will also evaluate its performance of MFMAC, based on a design example. As there are many types of multi-level Markov-model sources, the general comparison of B-B coding and M-M coding will be left for individual cases, in which specific coding specification will be determined based on this designing guideline.

Keywords: arithmetic code, Markov-model, entropy coding, state transition model, ranking conversion

1. Introduction

The arithmetic code has come to be used for image coding about 40 years ago[1)]. It was first adopted in bi-level image coding standard[2)] about 30 years ago for JBIG. And the target has expanded from bi-level image standard to multi-level image and video standards.

Figure 1 shows the block diagram of image coding/decoding process. The image coding technology is composed of the image modeling and the entropy coding. The image modeling is the source conversion into the source having less entropy than that of the original. The quantization of the parameter of the converted source can be also included.

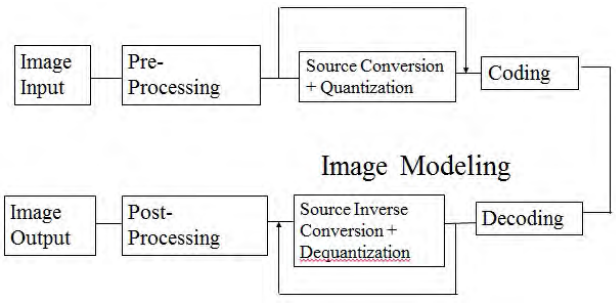


Fig.1 Block diagram of image coder and decoder

The entropy coding is the process of converting the observed event to the codeword having the corresponding information amount (entropy).

The entropy coding can be classified into block coding (table base) and non-block coding (procedure base) in information theory. The representative of the first type is Huffman coding and that of the second is the arithmetic coding[3)].

The advantage of arithmetic coding to Huffman coding will be the separation of code and the source, since arithmetic coding can be applied more generally than Huffman code. The recognition of the context procedure can be introduced in both source conversion and entropy coding stages and Markov-model source can be decomposed into memoryless sources by this procedure.

In order to apply the arithmetic coding for multi-level Markov-model sources, most of known methods were using binary arithmetic codes by decomposing multi-level Markov-model sources into plural binary sources, which we here call B-B coding. The advantages of B-B coding are the possibility of utilizing the study results of binary arithmetic codes but it may take longer time by the increase of coding/decoding procedures for the expanded binary sources. The usage of multi-alphabet arithmetic code

to multi-level Markov-model sources, which will be called as M-M coding here, and the comparison with B-B coding will be very interesting but it has not been tried because there are almost no report concerning the practical multi-alphabet arithmetic coding. Witten et al[4] have proposed multi-alphabet arithmetic coding, but it is quite basic one and the multiplication process is required in order to cope with the fluctuation of the valid area in the numerical line, which will need complicate processing and considerable time.

As we believe that multiplication-free process will be inevitable in assigning the area to each symbol against the valid fluctuating numerical line area, we will try to newly propose the multiplication-free multi-alphabet arithmetic code (MFMAC) in this paper. We assume the occurrence ranking conversion as pre-processing, and the fluctuation of the valid area will be absorbed by the area of the most probable symbol. On the other hand, the areas of all the symbols other than the most probable symbol, are fixed based on the source model parameter and the ordering values, regardless of the size of the current area. The basic design will be composed of source model assumption, the designing of representative value set for static coding, the parameter estimation for dynamic coding, and the ranking data estimation and update. Its performance will be also evaluated but the comparison with the case of using binary arithmetic code is left for further study, since such comparison will become meaningful when applied on real image sources.

2. Basic Principle of M-M Coding

In arithmetic coding, each symbol is assigned a corresponding area in the numerical line between 0.0 and 1.0, according to its probability. If the occurred symbol sequence is observed, the area in the numerical line is limited to the corresponding area of which size is given by the probability of the sequence. As the coding procedure progresses, the current valid area (VA), will decrease. If VA decreases less than half of the full size of the coding register, VA is doubled and the accuracy of the area handled in the coding register will be also doubled. If the size is still less than half of the full size, such procedure will be repeated, and the accuracy will be finally recovered. This procedure is called "renormalization". Therefore, VA will take the value between 0.5 and 1.0 if the full size of coding register bits is expressed as 1.0. In M-M coding, VA is divided into plural area, of which number is same as the number of source symbols.

In the source conversion stage shown in Fig.1, we assume the context recognition and the ranking conversion to convert the original symbol into the ordering value of the occurrence probability.

At the decoder, the possible numerical line area is detected from the transmitted code, and based on the addresses of the bounds of neighboring symbols, corresponding ordering value will be retrieved. Then the ordering value will be converted into the alphabet symbol. The reason of the ranking conversion is to unify the ordering of the numerical line sizes of the symbols in any context, which will be convenient in detecting the corresponding area from the transmitted code. The most probable symbol can be located either at the lowest part or highest part of the numeric line, by the way.

Based on this idea, we will propose the specification of MFMAC, which is composed of model source assumption, symbol area determination, model parameter estimation, and symbol ranking update. We will also investigate the basic performance of MFMAC being proposed.

3. Source Model Assumption

3.1 Definition of s-s model

In binary arithmetic coding, the number of parameters is only one and the parameter is given as the probability of MPS (More Probable Symbol) or of LPS (Less Probable Symbol). In considering the multi-level source model, it will be simple to consider a model having just one parameter, and the further variation of the sources will be handled with different source models. Here we will consider a single side probability decreasing model (s-s model), in which the probability of ($\boldsymbol{n}$+1)th symbol in occurrence ranking will be given by the probability of $\boldsymbol{n}$th symbol multiplied by $\boldsymbol{R}$ ($0<\boldsymbol{R}\leqq 1$). We call the parameter $\boldsymbol{R}$ as "declining parameter (DP)". The distribution of the probability for s-s model is usually called as geometric distribution.

3.2 Nature of s-s model

In a s-s model having the declining parameter of $\boldsymbol{R}$ ($\boldsymbol{R} \neq 0$), and the symbol number of N, the probability of top probability symbol (TS: Most frequent symbol) which is called "PTS" will be given

as $(1-R)/(1-R^N)$ when $R \neq 1$, or $1/N$ when $R=1$. Also, if N is enough large, PTS can be approximated by $(1-R)$.

4. Area Assignment Rule

4.1 Multiplication-free procedure

In the ideal arithmetic coding, the symbol area in the numerical line is calculated by the multiplication of the symbol probability and the size of the current valid area. However, for that purpose, the symbol area calculation will be needed not only for the occurred symbol but for all the symbols located under the occurred symbol, on the numerical line. Therefore, multiplication-free procedure will be inevitable for the practical multi-alphabet arithmetic coding. As the multiplication-free procedures adopted in various binary arithmetic coding methods was found to be effective[2),5),6)], its introduction into multi-alphabet arithmetic coding will be also promising[7)].

The area assignment rule we propose here is to fix the total area to symbols other than most probable symbol, based on the model parameter regardless of the current valid area. As we defined "TS" as "top probability symbol" in **3.2**, we define "OS" as "other than top probability symbol". The total area assigned to the symbols other than TS is called as "OSA", which means "other than TS area", and each symbol in OS is assigned an area proportional to its probability in OS. The rest of the OSA in current valid area will be assigned to TS, and it is called as "TSA". Therefore, TSA is given by (VA− OSA), and it will be varying according to the size of VA.

This idea is called as multiplication-free assignment (MFA). The probability of TS, which is called as "PTS", is calculated by (VA− OSA)/VA. The probability of i-th OS is also calculated by (OSA/VA)*x_i, where x_i is the probability of i-th OS in all OSs, and (OSA/VA) is varying according to the size of current VA. The efficiency caused by MFA compared to multiplication-based or ideal area assignment is called as "MFA efficiency".

4.2 General design of OSA

As VA will take the value between 0.5 and 1.0, and as TSA, given by (VA− OSA), must be positive value, OSA must be less than 0.5. The nominal value of PTS expressed as (VA− OSA)/VA will be minimum when VA=0.5, and will be maximum when VA=1.0. Therefore, (VA− OSA)/VA will be equal to the original PTS at specific VA value between 0.5 and 1.0. We call such VA value as D1. If VA=D1, PTS is same as original one, and as the probabilities of all OSs are also equal to the original ones, optimum coding efficiency of 100% will be achieved.

The coding efficiency will become minimum value when VA is either 0.5 or 1.0. Here, we will design OSA to satisfy the condition that the efficiency at VA=0.5 will become same as that at VA=1.0.

4.3 Calculation of OSA for s-s model

For s-s model, let a=1/PTS. Then following two relations will be introduced, if the number of symbols is large enough. These relations will be helpful to know the values from the parameter of s-s model.

OSA=(2^(a-1)-1)/(2^a-1)

D1=OSA*a/(a-1)

The range of D1 is found as following.

0.5<D1<ln(2)≒0.69317

As R increases, D1 will come closer to 0.5, and as R decreases, D1 will come closer to ln(2).

4.4 Coding efficiency versus DP value

At this stage, the code length of i-th symbol m_i is calculated by $\log_2(1/q_i)$, where q_i is the ratio of the area of i-th symbol to the current VA. The coding efficiency is calculated by the ratio of the source entropy and the average code length. When the current VA is equal to D1, the efficiency will be 1, which is the best case.

When the current VA is 0.5 or 1, the efficiency will become worst, and its value will be noted as WF, as they are set to be same value as described in Section **4.2**. Concerning the average efficiency AF for the distribution of the current VA, we will use the medium value of WF and 1, which is moderate since the efficiency curve is convex upwards.

Figure 2 shows the relation of R and the MFA efficiency when the number of symbols N is set to 256, and this curve is named as F(R).

From Fig. 2, the average efficiency will become lower as R increases. It can be said that the case, when R equals to 1, which means equal probability source, will provide the worst case. In **Table 1**, we show the various values of N, the number of symbols, when R equals to 1. It was found that the worst case is when N equals to 17, and though we limited the value of N less than 256, the efficiency will become better if we set N to be larger than 256.

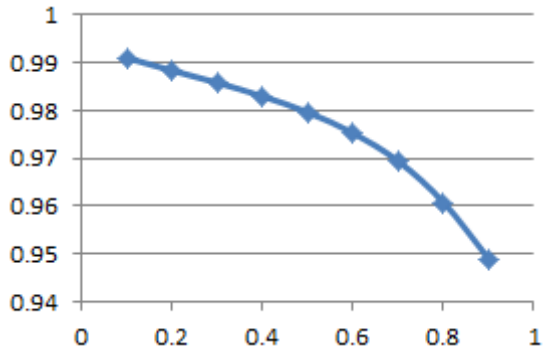


Fig.2 MFA efficiency versus R value: Vertical axis is the efficiency, and horizontal axis is the parameter value ***R***

Table 1 Equal probability sources and their efficiency

Symbol	Average Eff.
2	0.960845
4	0.946409
8	0.933440
16	0.928933
17	0.928930
32	0.931081
64	0.935800
128	0.941135
256	0.946273

5. Representative Values of DP

5.1 Designing of representative DP values

In order to code the target sources having variety of DP values, we need to define the set of representative values of DP, and select the most efficient one, since it is not realistic to use the value of DP strictly tuned for the target source. Such case is called as static coding case. From the nature of arithmetic coding, an area assignment method is considered to be derived from a source parameter.

In designing the set of representative values, the efficiency at the border values of neighboring representative values of ***R***, which will provide the lowest efficiency, will become a parameter. The border value will be the switching point of two neighboring DP values. We will call the worst efficiency occurred at the switching point as “RR efficiency”, and as MFA efficiency defined in **4.1** is not flat, RR efficiency is defined as a relative decreasing factor in addition to the MFA efficiency.

We will show the designing example by setting the number of symbols to be 256, and the RR efficiency to be 99%, as following.

Let us start the procedure of designing the set of representative values of ***R***, by assuming that switching point to be 0.5, and let the representative value of ***R*** which is smallest but larger than 0.5 to be $\boldsymbol{R}_1$. The efficiency of the source having ***R***=0.5 when coded by the area assignment method of ***R***=0.5, the efficiency will be 0.9796 from Fig. 2. Then, $\boldsymbol{R}_1$ will satisfy that the efficiency of coding the source having ***R***=0.5 by the area assignment based on $\boldsymbol{R}_1$ will be 0.9796×0.99= 0.9628. From this condition $\boldsymbol{R}_1$ will be found to be 0.5564. Let’s call the switching point between $\boldsymbol{R}_1$ and the next representative value $\boldsymbol{R}_2$ as $\boldsymbol{W}_1$. Then $\boldsymbol{W}_1$ will be introduced by the relation that the efficiency to encode the source of which DP parameter is $\boldsymbol{W}_1$ by area assignment method based on $\boldsymbol{R}_1$ will become 99% of its MFA efficiency. By such procedure, next peak $\boldsymbol{R}_2$, next switching point $\boldsymbol{W}_2$ and so on for larger ***R*** and ***W*** values will be introduced.

The representative values of ***R*** under 0.5 will be also introduced by similar procedure. Let us call the representative value of DP which is largest but smaller than 0.5 as $\boldsymbol{R}_0$. Then $\boldsymbol{R}_0$ will be introduced by the relation that the efficiency of coding the source, of which DP is 0.5, using the area assignment method based on $\boldsymbol{R}_0$, will become 99% of MFA efficiency at ***R***=0.5. Then $\boldsymbol{R}_0$ will be introduced to be 0.4280. Let the representative value of DP next to $\boldsymbol{R}_0$ as $\boldsymbol{R}_{-1}$. The switching point between $\boldsymbol{R}_0$ and $\boldsymbol{R}_{-1}$, which is called $\boldsymbol{W}_{-1}$ will be also introduced by coding the source of which DP is $\boldsymbol{W}_{-1}$ by the coding method based on $\boldsymbol{R}_0$, will become 99% of MFA efficiency at $\boldsymbol{W}_{-1}$. Then $\boldsymbol{W}_{-1}$ will be introduced to be 0.3699. By such procedure, smaller ***R*** and ***W*** values will be also introduced.

5.2 Set of representative DP values

Table 2 shows the representative values and switching values of DP. It can be said that the source of which ***R*** is between 0.04 and 0.96 will be coded by using the set of twelve representative values of DP, with the cost of maximum 1% from the MFA efficiency given in Figure 2.

Figure 3 shows the MFA efficiency (reddish brown), which is same as the curve shown in Fig.2, and the composite efficiency caused by RR efficiency (blue). The minimum of RR efficiency is set to be 99%.

Table 2 Representative values and switching values

R_{-5}	0.0460	0.0354	W_{-6}
R_{-4}	0.0795	0.0623	W_{-5}
R_{-3}	0.1314	0.1048	W_{-4}
R_{-2}	0.2064	0.1683	W_{-3}
R_{-1}	0.3069	0.2567	W_{-2}
R_0	0.4280	0.3699	W_{-1}
R_1	0.5564	0.5000	W_0
R_2	0.6761	0.6311	W_1
R_3	0.7762	0.7462	W_2
R_4	0.8532	0.8351	W_3
R_5	0.9075	0.8965	W_4
R_6	0.9431	0.9361	W_5
		0.9612	W_6

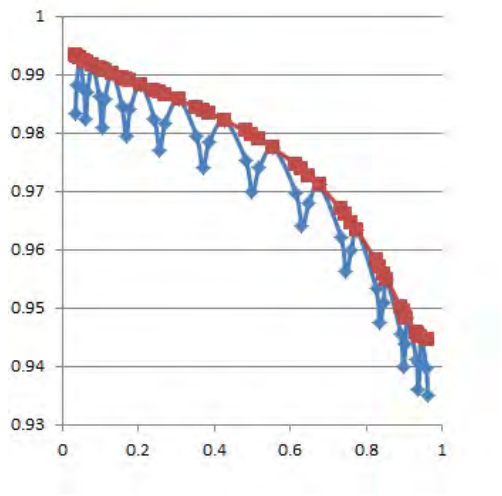


Fig.3 MFA efficiency (reddish brown) and its decrease by RR efficiency (blue) : Vertical axis is the efficiency and horizontal axis is the parameter value ***R***

6. Dynamic Coding

6.1 State transition diagram

The study described in Chapter 5 is assuming that DP of the source is known beforehand. However, if the DP value is not known or time-varying, dynamic coding by estimating the statistics of DP value will be effective. For dynamic coding, state transition diagram method is known as a quite popular one in binary arithmetic coding. As for multiplication-free multi-alphabet arithmetic coding, similar procedure will be considered to be effective. Therefore we will utilize the outcome introduced in dynamic coding of STT-coder[8)-10)].

Figure 4 shows the state transition diagram of MFMAC based on the study result of STT-coder. The state transition of binary arithmetic coding is defined

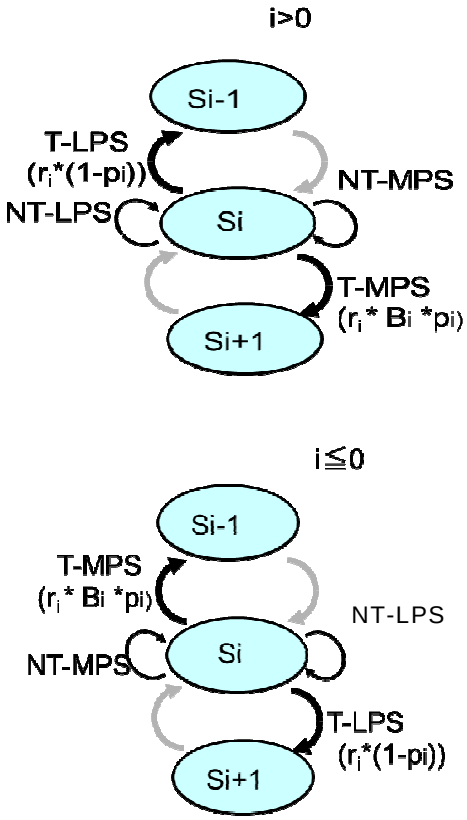


Fig.4 State transition in dynamic MFMAC

whether observed symbol is MPS (More Probable symbol) or LPS (Less Probable Symbol), and the state transition of multi-alphabet arithmetic coding is defined whether the observed symbol is TS or OS, and which one will become MPS or LPS will depend on ***R*** value. In Fig.4, T-MPS shows the MPS of state transition case, and NT-MPS shows the MPS of non-transition case, that means the transition to the same state. The meaning of T-LPS and NT-LPS are likewise. The black arrows show the transition from the state of Si, and the grey arrows show the movement into the state of Si from other states.

If ***R*** is equal or larger than 0.5, OS will be MPS, and TS will be LPS. If R is less than 0.5, TS will be MPS, and OS will be LPS. It is shown in Fig.4 and also in **Table 3**. The range of Table 3 is based on the representative ***R*** set which is given in Table 2. Using this concept, the idea similar to dynamic coding of STT-coder can be applied to the dynamic coding of MFMAC, but as the total number line area for OS is limited to less than 0.5 in any ***R***, the design and the performance is not symmetric reflecting the value of 0.5, which means the difference from the case of binary arithmetic coder.

Table 3 Parameters of states

i	R	PTS	T-LPS	MPS	Bi	T-MPS
−5	0.0460	0.9540	1.0000	TS	0.0482	0.0482
−4	0.0795	0.9205	0.8000	TS	0.0864	0.0691
−3	0.1314	0.8686	0.6400	TS	0.1512	0.0968
−2	0.2064	0.7936	0.5120	TS	0.2601	0.1332
−1	0.3069	0.6931	0.4096	TS	0.4428	0.1814
0	0.4280	0.5721	0.3277	TS	0.7481	0.2451
1	0.5564	0.4436	0.3277	OS	0.7972	0.2612
2	0.6761	0.3239	0.4096	OS	0.4791	0.1962
3	0.7762	0.2238	0.5120	OS	0.2883	0.1476
4	0.8532	0.1468	0.6400	OS	0.1720	0.1101
5	0.9075	0.0925	0.8000	OS	0.1019	0.0815
6	0.9431	0.0569	1.0000	OS	0.0603	0.0603

6.2 Transition ratio

The occurrence of TS can be the trigger to the adjacent state having smaller DP value, and the occurrence of OS can be the trigger to the adjacent state having larger DP value.

The transition ratio of MPS and LPS should be based on the occurrence ratio of both symbols but the absolute transition rate of LPS will better be controlled based on the absolute probability of LPS.

Let p_i equal to the MPS probability at state S_i. Then $(1-p_i)$ will be the LPS probability at state S_i. Then B_i, which is the ratio of the probability of T-MPS (transition of MPS) and probability of T-LPS (transition of LPS) at state i will satisfy following equality to keep staying the current state, when current state i is optimum for the source

$$1-p_i = B_i \cdot p_i$$

Let the probability of T-LPS in the state S_i as r_i. Then the probability of T-MPS will be $B_i \times r_i$. In the study of STT-coder, it was introduced that following condition should be satisfied:

$$\frac{B_{i+1}}{B_i} < \frac{r_i}{r_{i+1}} < 1$$

Based on the result of setting ratio of probability $g = r_i/r_{i+1}$ to be 0.7, 0.8, 0.9, it was found that the performance of g=0.8 is the best. The r_i is better to set as 1 for quite large MPS probability.

In the state transition of MFMAC, the state of which R is around 0.5 will have the MPS probability around 0.5. Therefore, from the center of the state transition diagram, as R increases, the probability of MPS will increases, and as R decreases, the probability of MPS will also increases. Therefore, r_i will be set as shown in **Table 3**.

According to the study result of STT-coder, the dynamic coding efficiency is about 92% of static coding efficiency, when the representative value set is designed by assuming the efficiency of switching point to be 99%, and will be increased to 96% if the number of states will be doubled.

Comparing the case of STT-coder and MFMAC, the MPS probability of each state will take quite similar values. Therefore we can assume similar performance will be achieved in MFMAC too. So, we can estimate the dynamic coding efficiency will be given as 92% of static coding efficiency shown as blue in Fig.3, and 96% of that will be possible if we increase the number of states to be doubled. By the way, the decision of T-MPS or T-LPS based on the probability shown in Table 3, will be done using pseudo random value such as the size of VA.

7. Ranking of Symbols

7.1 Symbol alphabet and ordering value

In MFMAC, the occurred symbol alphabet must be converted into ordering value, at the encoder, as described in Chapter 2. At the decoder, the decoded ordering value should be converted to the symbol alphabet. According to the occurrence histogram, the ranking data should be updated in both encoding side and decoding side. Therefore the conversion from ordering value to symbol alphabet is also needed at encoder to speed up the procedure.

7.2 Occurrence counter of each symbol

Let ***i-th*** symbol u_i has an occurrence counter CTR showing its accumulated occurrence number of v_i. Let us assume the CTR having 8bit. At the initial stage, any v_i will be set to zero. If any of the symbol counter may reach to 256, all the symbol counter values, will be halved, by one bit shifting. The ranking data will be produced based on CTR value. At the initial stage, the ranking data will be given by the alphabetical order though every symbol counter value is the same.

7.3 Ranking data update

To update the ranking data of the symbols based on the strict order of CTR value will provide strictly correct ranking information, but to check the strict order of CTR whenever new event is observed will take too long time. As the theoretical code length of the symbol will be given by the logarithm of the

reciprocal of the probability, to recognize the value of the probability by the number of valid bits of CTR will save the time while keeping the approximate ranking data. If the symbol counter is set to 8bit, let us classify the counter value CTR into 9 groups shown in **Table 4**.

If a symbol occurs and of which CTR is kept in the same group, its ranking data will not be modified. If its CTR is moved to lower group, the ranking data of the symbol is checked and if its ranking data is the highest one in the current group, the ranking data is kept same. But if it is not, the symbol having the highest ranking data in the current group is looked up and the ranking data of the occurred symbol will be swapped with it.

Let us define the number of symbols in #i group to be t_i. Then, if the youngest number group having non-zero set is #j, the ranking data of symbols in #j group will be between one and t_j. For group #(j+1), the ranking data will be between t_j+1 and t_j+t_{j+1}, if t_{j+1} is not zero. Such procedure will be followed to increase the number of group to t_9. As the number of symbols for i-th group t_i will take the value between 0 and 256, each group needs 8bit counter plus 1bit indicator to distinguish 0 and 256.

The update of the ranking data will be done if any symbol will move to another group, and the update of t_i will be also executed accordingly.

Based on this grouping, the number of the symbols in each category, and the total number of the symbols included in the groups having less CTR values are kept, and whether ranking data should be changed or not will be judged.

In the s-s model, if R is smaller than 0.5, the ratio of the occurrence number of two successive ranks will be more than two times. Therefore, this grouping will not usually affect the coding efficiency. For larger R, the ranking data of this method may not be strictly correct but as R increases, the ratio of area size between two successive rankings will become closer and the approximation of the ranking will not harm the coding efficiency so much.

Yet, as the judgement of TS may affect the efficiency severely, it will be done strictly, as shown in the next section.

7.4 Top probability symbol

Concerning the detection and updating of the TS (top probability symbol), we will adopt strict processing. We will memorize the counter value of current TS(s) as MAX, and the number of symbols of which CTR value is equal to MAX, as NT.

Table 4 Group# and its CTR range

group #	CTR range	group #	CTR range
1	CTR≧128	6	7≧CTR≧4
2	127≧CTR≧64	7	3≧CTR≧2
3	63≧CTR≧32	8	CTR=1
4	31≧CTR≧16	9	CTR=0
5	15≧CTR≧8		

a) Increment of MAX and checking NT

The increment of MAX will happen if the newly occurred symbol's current CTR is MAX. We set its CTR to be MAX+1, and increment the maximum occurrence value MAX by one. Concerning NT, if NT is currently one, the ranking data of the occurred symbol is already one and not need to be modified. If NT is not one, then the occurring symbol's ranking data is checked and if it is not one, we look up the symbol of which ranking data is one, and swap the ranking data of two symbols. NT is then set to one.

b) Increment of NT

The increment of NT will happen if the occurring symbol's CTR value is MAX− 1. We increment NT by one. If the occurring symbol's ranking data is equal to new NT, the ranking data of the occurred symbol is not modified. If the occurring symbol's ranking data is larger than new NT, we look for the symbol of which ranking data is equal to new NT, and we swap the ranking data of the two symbols.

c) Procedure when CTR overflows

The counter overflow will happen when MAX is 255, if the CTR is 8bit, and a symbol of which counter is 255 occurs. At this time NT may take any value, but only the symbol of which counter overflow will be given the ranking data of one (which means TS) after this procedure and NT will become one. After the overflow, the counter value of all symbols will be halved by shifting, and MAX is set to 128. New t_1 will be 1, and new t_2 will be (old t_1) − 1. New t_i (for $3 \leqq i \leqq 8$) will be old t_{i-1}. New t_9 will be (old t_9)+(old t_8).

8. Necessary Memories

Based on the above study result, the total memory needed for sources having N(=2^n) symbols

Table 5 Coding efficiency compared with Huffman

R	Huffman	Arithmetic
	14bits	14bits
0.841	0.988	0.946
0.707	0.985	0.958
0.5	0.986	0.964
0.382	0.955	0.964
0.293	0.871	0.961
0.245	0.803	0.959

and 8bit occurrence counter will be as follows.

1) Symbol counter $N\times$8bits to express CTR
2) Conversion from symbol alphabet to ordering value: $N\times n$ bit
3) Conversion from ordering value to symbol alphabet: $N\times n$ bit
4) Number of symbols categorized by valid bits of CTR: $(8+1)\times\log(N+1)$bit

This set of memories is required for each context.

Let us set the number of transition states in dynamic coding to be M, coding register size to be CR bit, then fixed area data is composed of $(N-1)\times$ CR bits ROM. In dynamic coding, to show the current estimated state, log M bit is needed for each context.

Concerning the coding register size, which is Kbit, the longer will provide the higher efficiency, since the minimum numerical line area for symbols is 2^(− K). So, even if the theoretical numerical line area to be assigned is less than minimum unit area, we will have to assign minimum unit area to the symbol.

So, if the number of symbols is N(=2^n) and the coding register size is K bit, the rough estimation of the loss will be 2^(− K+n). For example, if n=8 and K=14, the rough estimation of the loss will be about 2^(− 6) which will be about 1.6%. **Table 5** shows the comparison of MFMAC with Huffman codes for some R values[11], being convenient to construct Huffman codes, when K is 14. This result is assuming static coding, and will provide the convenient performance.

9. Conclusion

In this paper, we have proposed multiplication-free multi-alphabet arithmetic coding method (MFMAC), to show the practical possibility of applying multi-alphabet arithmetic codes for multi-level Markov-model sources. The basic design of MFMAC is composed of model source assumption, symbol area determination rule, designing representative model parameter set, model parameter estimation, and symbol ranking update.

We have assumed single side decreasing probability model, in which one parameter R is needed to express the decreasing ratio of the probability in its ranking. The proposed area assignment rule is to fix the total area of symbols other than most probable symbol, which is called as “OSA”, regardless of the current valid area based on the model parameter value. In “OSA” each symbol is assigned a predetermined area which is proportional to its probability in symbols other than TS. The rest of the current valid area will be assigned to TS.

For static coding, we showed how to design a representative parameter set, as each parameter value can be considered to correspond to specific range of R. For dynamic coding, we refer the study result of STT-coder, one of the binary arithmetic codes, which is expected to show the similar performance. Concerning the ranking data, we introduced the speedy and efficient detection and updating method. We also considered the coding register size and its effect to the coding efficiency. Based on these studies, we have checked the performance and we judged that the proposed system is very promising.

As a conclusion, we proposed multiplication-free multi-alphabet arithmetic code and showed the basic design method and its performance. Though, we assumed the single side probability decreasing model, and the number of symbols to be 256, to evaluate different source models and different numbers of symbols will be possible based on our basic design and the shown design guideline.

The comparison with the case in which binary arithmetic code will be applied to bi-level sources derived from the multi-level source will be left for individual cases, in which specific coding specification will be determined based on this designing guideline. We hope this study will contribute to the future investment of practical arithmetic coding.

Acknowledgements

A part of this research has been supported by Canon Inc. and authors are deeply grateful for it. The authors also would like to thank Mr. Shigetaka Ogawa (ICT-Link, Ltd) for joining the discussion in the early stage of this research.

References

1) G. Langdon, J. Rissanen: “Compression of Black-White Images with Arithmetic Coding”, IEEE Trans. on Commun. COM-29, No.6, pp. 858–867 (June 1981).
2) F. Ono et al. “Definition of the QM-coder”, ISO/IEC JTC 1/SC2/WG8/JBIG N224 R1(1990).
3) P. Elias: Note 1 in Chapter 3 (pp.61-62) of “Information Theory and Coding” by N. Abramson, McGraw-Hill Inc. (1963).
4) I. H. Witten, R. M. Neal, J. G. Cleary: “Arithmetic coding for data compression”, Communication of ACM, Vol.30, No.6, pp. 520–540 (June 1987).
5) W. B. Pennebaker, J. L. Mitchell, G. G. Langdon Jr., R. B. Arps: “An Overview of the Basic Principles of Q-Coder”, IBM Journal of Research and Development, Vol.32, No.6, pp.717–726 (1988).
6) F.Ono, S.Kino. M.Yoshida, T.Kimura : “Bi-level Image coding with MELCODE– Comparison of Block Type Code and Arithmetic type code– ”, 7.6.1–7.6.6. Globecom ’89 (1989).
7) F.Ono: “Usage of Multi-Alphabet Arithmetic Codes for Multi-Symbol Source Coding”, IIEEJ Annual Conference (2021)
8) I. Ueno, F. Ono: “Designing Method of State Transition Table-Driven Arithmetic Coder STT-coder”, Journal of IIEEJ, Vol.43, No.1, pp. 62–70 (Jan. 2014).(In Japanese)
9) I. Ueno, F. Ono: “Probability Estimation and Dynamic Coding Method for State Transition Table-Driven Arithmetic Coder STT-coder”, Journal of IIEEJ, Vol.43, No.1, pp. 71–78 (Jan. 2014). (In Japanese)
10) F. Ono, I. Ueno: “Dynamic Coding Design for State-Transition-Table-Driven Arithmetic Coder: STT-Coder” IEVC2017, 5A-3(2017).
11) F. Ono: “Arithmetic Codes for Image Coding: Their Advantages and the Efforts to Make Them Practical”, Trans on IEVC, IIEEJ, Vol.11, No.1, pp. 1–12 (June 2023).

(Received Nov. 6, 2024)
(Revised Feb.18, 2025)

Fumitaka ONO (***Honorary Member)***

He has received his BE, ME and Ph.D from The University of Tokyo respectively. He has been with Mitsubishi Electric Corp., and then with Tokyo Polytechnic University. His interested area covers image coding, entropy coding, and image processing. He is IEEE Life Fellow, IEICE Fellow and IIEEJ Fellow. He has been engaged in the international standardization work since 1985, and had been ISO/IEC JTC 1/SC29/WG1 JBIG Rapporteur. He has received Award from Ministry of Education and Science, Award from Ministry of Industry and Trade, SCAT Chairman’s Grand Award, Contribution Award from IPSJ/ITSCJ, and so on. He is currently Professor Emeritus of TPU, and Visiting Researcher of The Univ. of Tokyo.

Kazuto KAMIKURA (***Member)***

He received his BE and ME from Tokyo University of Science, and his Ph.D in Engineering from The University of Tokyo. He worked for NTT Corporation from 1986 to 2014, and has been with Tokyo Polytechnic University since 2014. His research interests include video coding and image processing. Several of the technologies he developed have been adopted by international standards such as MPEG.

Yousun KANG (***Member)***

She received the Ph.D. degree in engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2010, where she conducted research under the supervision of Prof. Hiroshi Nagahashi. From 2007 to 2010, she was with Toyota Central R&D Labs., Inc., Aichi, Japan. She was then a researcher at the National Institute of Informatics, Tokyo. Since 2011, she has been with Tokyo Polytechnic University, where she is currently a Professor. Her research interests include deep learning for object detection, image segmentation, generative models, multimodal analysis for video understanding and Japanese Sign Language recognition, and medical AI for pathology report generation using gigapixel whole slide images.

IIEEJ **Paper**

Online Hand Drawing Pattern Classification Using Sketch-RNN

Shione ISHIDA† , Kyoko SUDO† (*Member*)

† Toho University

<**Summary**> One recent approach to unsupervised anomaly detection is training an Autoencoder and using the reconstruction error as an index. In this work, we adopt Sketch-RNN, a sequential Variational Autoencoder (VAE) model, to classify the online hand drawing pattern of healthy persons or Parkinson's disease patients. We train the Sketch-RNN with no labeled data and evaluate the accuracy of the attribute recognition using the reconstruction error. The proposed method, online drawing data and Sketch-RNN model, outperforms the conventional method, recognition with still image data and CNN model, showing the pipeline for online data classification using VAE.

Keywords: VAE, Pattern classification, Hand drawing pattern

1. Introduction

Detecting the signs of change or abnormality is a fundamental issue in all fields, including marketing, finance, medical diagnostics, and detecting defects in factory products. An "anomaly" in this context is data that behaves differently from most data defined as normal. In anomaly detection in real-world data, labeled anomaly data is often absent or scarce. Even when labels are available, they are insufficient to fully characterize all concepts of anomaly. Therefore, unsupervised or semi-supervised learning is commonly used to learn a model of normality from normal data and detect anomalies by deviation from the model. One of the recent approaches is the unsupervised learning of the Variational Autoencoder (VAE)[1]model and judging standard or anomaly by reconstruction loss. In this work, we handle the classification of online hand-drawing patterns. There are conventional approaches for detecting anomalies in online data using LSTMs or VAE. However, they primarily take one-dimensional sensor signals. We enlarge this approach to multidimensional online patterns. We focus on a sequence-to-sequence Variational Autoencoder, Sketch-RNN, to learn online drawing patterns and classify data using the reconstruction loss. We apply our method to the online hand drawing patterns for the test of Parkinson's disease and estimate the accuracy of classifying the data of patients and the data of healthy persons.

2. Related Works

2.1 Anomaly detection method using unsupervised learning

Unsupervised anomaly detection is realized as outlier detection in the feature space. There are several methods for finding outliers, including a) methods based on clustering, b) methods based on probabilistic models, and c) methods based on reconstruction models. In the classification of anomaly detection methods by Ruff et al[2], methods such as the One-Class Support Vector Machine(OC-SVM)[3]belong to a), methods using Mahalanobis distance[4]or GMM model approximation[5]belong to b), and methods using principal component analysis(PCA)[6]belong to c). Methods using deep generative models such as VAE or GAN fall between b) and c).

2.2 Anomaly detection using VAE

The VAE consists of an encoder that compresses the input into low-dimensional latent variables and a decoder that decodes the input from the latent variables to the original number of dimensions and outputs the reconstructed signal. The latent variable is assumed to be controlled by a probability distribution, typically a Gaussian distribution. We learn the model to minimize the sum of the reconstruction error to bring the outputs and inputs closer together and the Kullback-Leibler Divergence error to bring the latent variable closer to a Gaussian distribution. When we train the VAE model with specific patterns, similar patterns will be projected to close in distribution on the latent space, and the reconstruction

error will be low. On the other hand, those far from the trained data have high reconstruction errors and cannot be reconstructed well. Based on this property, the VAE is trained to learn only normal patterns and discriminate between abnormal and normal based on whether it can reconstruct them. For anomaly detection using VAE, An et al.[7]showed that outliers in MNIST data can be detected using the Reconstruction Loss as an indicator. Recent applications to various actual data include anomaly detection of industrial parts images[8], anomaly detection of 3D point clouds[9], and anomaly detection of time-series patterns of sensor signals[10].

2.3 Model for Multidimensional online patterns

Zhang et al. used an RNN to model the pen movements when writing Kanji characters and attempted to generate vectorized characters[11]. Compared to traditional convolutional neural network(CNN) based approaches that require image-like representations, their RNN-based approach is a complete end-to-end system that deals directly with raw sequence data. At that time, their discriminative RNN model of Chinese characters achieved state-of-the-art performance in the ICDAR-2013 competition[12] database due to its straightforward use of spatial and temporal information.

The Sketch-RNN[13]is a model for simple hand-drawn sketches created by online pen input. A recurrent neural network (RNN) based on the VAE structure enables the generation of sketches with information on pen movements, including displacement and stroke order. This model has the same structure as VAE and allows the input of line drawing data containing time series information. Therefore, we assume this model is suitable for recognizing handwritten exam data from healthy subjects and patients, including similar time-series information as simple hand-drawn sketches.

3. Methods

3.1 Model

Sketch-RNN model based on Sequence-to-Sequence Variational AutoEncoder. **Figure 1** shows an overview of the model used in the experiment. The input sketch sequence S to the Sketch-RNN model is a list of five vectors $(\Delta x, \Delta y, p1, p2, p3)$. $(\Delta x, \Delta y)$ denotes the pen position. Δx is the offset distance in x-direction from the previous point and Δy is the offset distance in y-direction from the previous point. $(p1, p2, p3)$ is a one-hot-vector representing the state of the pen. $p1$ indicates that the pen is touching the paper and the next line will be drawn. $p2$ indicates that the pen has left the paper and the next line is not drawn. $p3$ indicates the end of drawing by the end condition. The Encoder is a bi-directional RNN[14]which takes as input the sketch sequence S and S reverse(inverse order of S) and outputs a latent vector z of size Nz. Specifically, the hidden state $h(h_{\rightarrow}$ and $h_{\leftarrow}$ connected) acquired from the bi-directional RNN is projected onto two vectors μ and $\hat{\sigma}$ of size Nz, respectively, using a fully connected layer. $\hat{\sigma}$ is converted to a non-negative standard deviation parameter σ using exponential operations. μ and σ are used together with an IID Gaussian variable vector $N(0, I)$ of size Nz to construct the latent vector z. The Decoder is an autoregressive RNN that generates and outputs a sketch sequence S' based on the latent vector z from the Encoder. At each step i of the decoder RNN, the input is the concatenation of the previous point S_{i-1} and the latent vector z. The start of the sketch sequence, S_0, is defined as $(0, 0, 1, 0, 0)$. The output of each time step is the parameter of the probability distribution of the next data point S_i. In Sketch-RNN, $(\Delta x, \Delta y)$ is modeled by a Gaussian mixture model (GMM) consisting of M normal distributions. And $(p1, p2, p3)$ is modeled by the categorical distribution $(q1, q2, q3)(q1 + q2 + q3 = 1)$. The sampling process generates parameters for both the GMM and categorical distributions at each time step and samples the result S' of that time step. The sample's randomness level during the sampling process can be controlled by introducing a temperature parameter τ. τ is typically set between 0 and 1. In the limit of $\tau \to 0$, the model becomes deterministic and the sample consists of the most likely points of the probability density function. Equation (1) is the loss function. As with VAE, it consists of the reconstruction error L_R, which brings the input and output closer together, and the KL divergence L_{KL}, which brings the distribution of the latent vector z closer to a normal distribution. The reconstruction error L_R represents the log-likelihood of the probability distribution explaining the training data S and is learned to maximize. The L_R is composed of the term L_s for the coordinates and the term L_p for the pen states. It is represented by the Equation (2). The L_s can be represented by the Equation (3) and L_p by the Equation (4). In this case, N_{max} is the length of the longest sketch in the training data set. (In principle, it can be thought of as a hyperparameter.) The Π is a vector of length M, corresponding to the weights of each normal distribution in the Gaussian mixture model. The KL divergence loss

represents how far the latent vector z is from the standard normal distribution and is learned to minimize. It is represented by the Equation (5). We use the reconstruction error as the index of classification.

$$Loss = L_R + w_{KL} L_{KL} \tag{1}$$

$$L_R = L_s + L_p \tag{2}$$

$$L_s = -\frac{1}{N_{max}} \sum_{i=1}^{N_s} \log\left(\sum_{j=1}^{M} \Pi_{j,i} N(\Delta x_i, \Delta y_i)\right) \tag{3}$$

$$L_p = -\frac{1}{N_{max}} \sum_{i=1}^{N_{max}} \sum_{k=1}^{3} p_{k,i} \log(q_{k,i}) \tag{4}$$

$$L_{KL} = -\frac{1}{2N_z}(1 + \hat{\sigma} - \mu^2 - \exp(\hat{\sigma})) \tag{5}$$

3.2 Datasets

(a) QuickDraw dataset

The QuickDraw dataset is a large hand drawing dataset from the online game "Quick Draw!"[15], where the players are asked to draw objects belonging to a particular object class in less than 20 seconds. The QuickDraw dataset consists of hundreds of general object classes like "Donut", "Book", "Circle", and "Square". We show the example of the Quickdraw dataset "Donut", "Book", "Circle" and "Square" in **Fig. 2**. This dataset is used to train Sketch-RNN. Each class of QuickDraw is a dataset of 70K training samples, in addition to 2.5K validation and 2.5K test samples. The representation of the dataset is the vector image. The vector is $(\Delta x, \Delta y, \Delta z, p1, p2, p3)$. $(\Delta x, \Delta y)$ is the difference from the previous point (x, y), and $(p1, p2, p3)$ is the one-hot vector that shows the state of the pen. $p1$ indicates that the pen is currently touching the paper and that a line will be drawn connecting the next point to the current point. $p2$ indicates that the pen is lifted off the paper after the current point, and no line is drawn next. $p3$ indicates that the sketch has been drawn.

(b) Original Artificial Hand Drawing Dataset

We generate the Original Artificial Hand Drawing Dataset for training Sketch-RNN. First, we make an application interface using Python to collect hand-drawing patterns. The subjects draw patterns using a pen device on a tablet as shown. We collect two patterns, spiral, and meander, mimicking the patterns of the NewHandPD dataset. **Figure 3** shows the application we made to collect the drawing patterns. We obtained 900 samples. The number of the subjects is 9, all healthy people and right-handed, which we use as the training data. We need some "anomaly" data to pre-test the model after training, so we collected an extra set of data by intentionally written with the left hand. This dataset is created in the same vector format $(\Delta x, \Delta y, \Delta z, p1, p2, p3)$ as the QuickDraw dataset. We show the example of the Original Artificial Hand Drawing Dataset in **Fig. 4**.

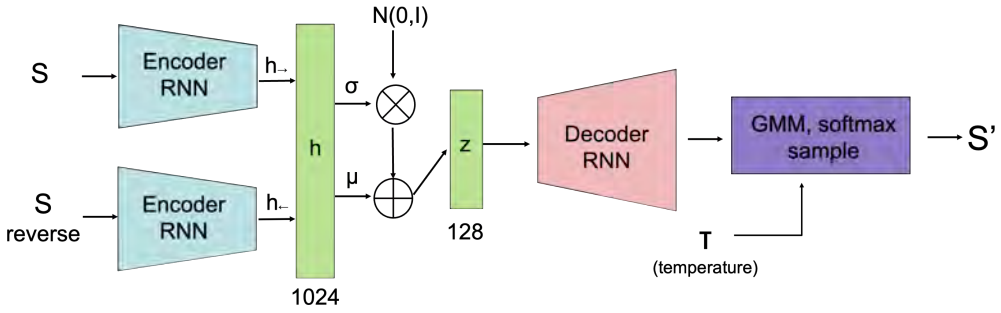
(c) NewHandPD dataset

The NewHandPD dataset[16]contains images and signals acquired during handwriting exams of 31 Parkinson's disease (PD) patients and 35 healthy. The exams consists of a task to draw spirals, meanders and circle shapes. Each individual holds the smart-pen with a straight arm and draws the shapes by moving their wrist. The signal data in the NewHandPD dataset consists of six different signals obtained from the smart pen during the exam: CH1: Microphone, CH2: Fingergrip, CH3: Axial Pressure of Ink Refill, CH4: Tilt and Acceleration in "X direction", CH5: Tilt and Acceleration in "Y direction" and CH6: Tilt and Acceleration in "Z direction". In this study, the spiral and meander of the NewHandPD dataset are used to recognize the attributes of healthy person/Parkinson's disease patients. The coordinates obtained by the pen are in three dimensions, but the test is to draw a figure with the arm extended, assuming a piece of paper. Therefore, We use CH5 in the y-axis direction and CH6 in the z-axis direction as inputs to the sketch-RNN model. It is known that patients with Parkinson's disease have symptoms of tremors. Jeon[17]measured tremors with an accelerometer. It is also known that there is a symptom that "writing velocity, writing fluency and drawing may deteriorate[18]". For these reasons, we incorporate both acceleration and velocity as the feature. We use the acceleration vectors of CH5 and CH6 converted to velocity vectors as new meander and spiral.

4. Experimental Formulation

4.1 Overall view of the experiment

The experimental process's overall picture is shown in **Fig. 5**. First, the Sketch-RNN model is trained with handwriting patterns. Since the NewHandPD dataset of domain data is insufficient in number, we will use two types of large volume data for training the Sketch-RNN model. One is the QuickDraw dataset, and the other is the Original Artificial Hand Drawing Dataset. We made a pen tablet application to create our own training data. Next, test the data in the NewHandPD dataset is used to test the Sketch-RNN model. Currently, the NewHandPD



$$Loss = L_R + w_{KL}L_{KL}$$

Fig. 1 An overview of the model used in the experiment

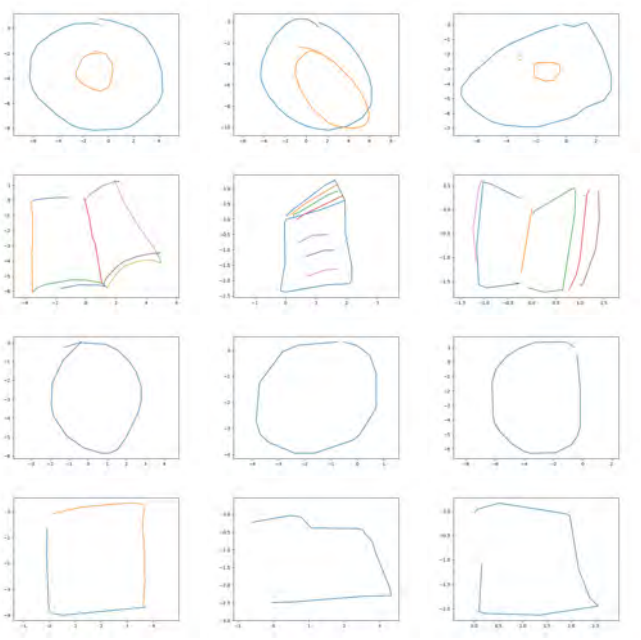


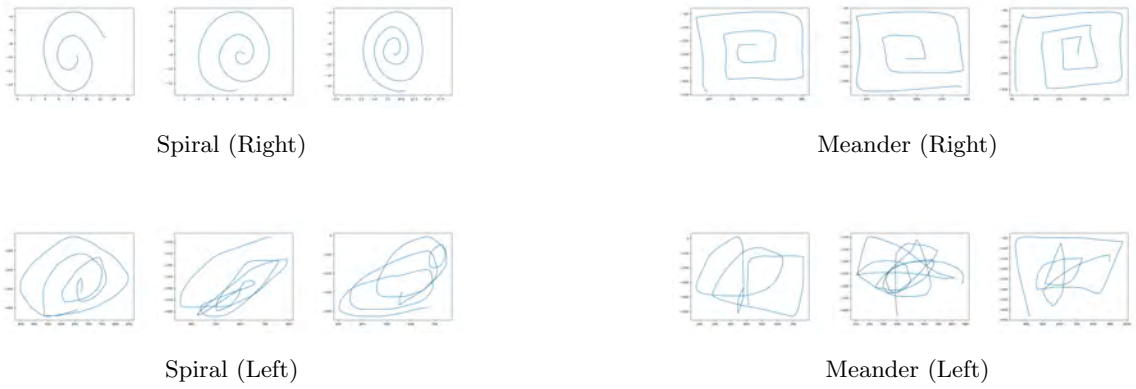
Fig. 2 The example of the Quickdraw dataset "Donut", "Book", "Circle" and "Square" with the line colored corresponds to the sequential stroke ordering

Fig. 3 The way drawing a pattern using a pen device on a tablet to create the Original Artificial Hand Drawing Dataset

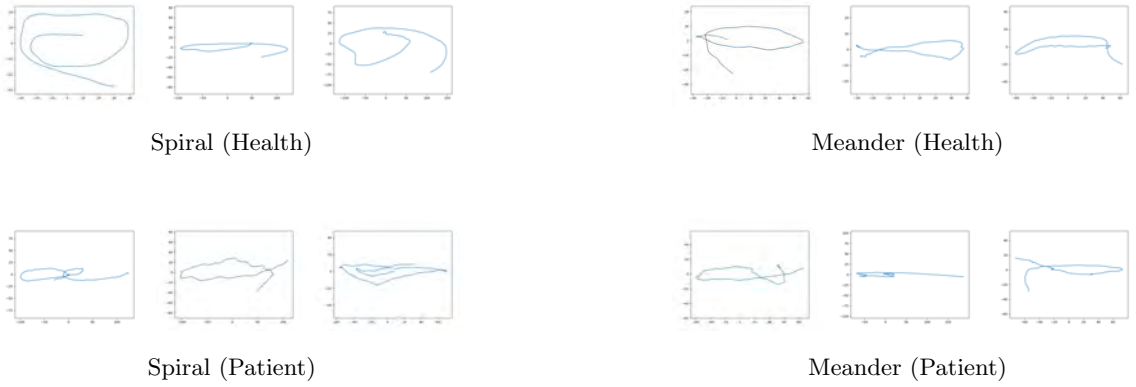
dataset undergoes preprocessing and conversion from acceleration data to sketch sequence S. The transformed sketch sequence S is used as input to the Sketch-RNN model, which classifies whether the input data is healthy or patient. The Sketch-RNN encodes vector sequences and reconstructs vector sequences, not encodes and decodes from image to image directly. The appearance of the reconstructed image pattern that is obtained by rendering from the reconstructed vector sequence generated by Sketch-RNN does not necessarily match intuitively the appearance of the input image pattern.

4.2 Preprocessing and conversion from acceleration signal to velocity signal

The raw acceleration signal has noise due to the body movement and has bias due to gravity. We preprocess the acceleration signal by applying a high pass filter to remove the low-frequency component, then using the moving average for noise reduction, and then subtracting the value of mode to remove the influence of acceleration due to gravity. Then, we subtract the mean value so that the mean of the signal is 0. We integrate this pre-processed acceleration data to obtain the velocity data. The original data includes the signal of drawing a specific pattern and an offsets that capture the movement before and after drawing a specific pattern. Some original data has a duration of over 30,000 frames, while some short data is within 5,000 frames. So, we manually cut each frame to remove offsets and extract the signal around 5,000 frames to at least contain a part of one drawing duration for a specific pattern. We thus obtain the three-dimensional velocity data. However, we need to decrease the dimension of the data to 2 since Sketch-RNN inputs two-dimensional data. So, we use the y-axis direction and the z-axis direction. The x-axis mainly represents the rise and lower movement of the pen, and y-axis and z-axis represent the shape of the drawing pattern. The signal waveforms for each process are shown in the **Fig. 6**. The upper left is the original signal, the upper right is the original data cut out according to the time information, the lower right is



(a) Example of Original Artificial Hand Drawing Dataset



(b) Example of NewHandPD Dataset

Fig. 4 Example of Datasets

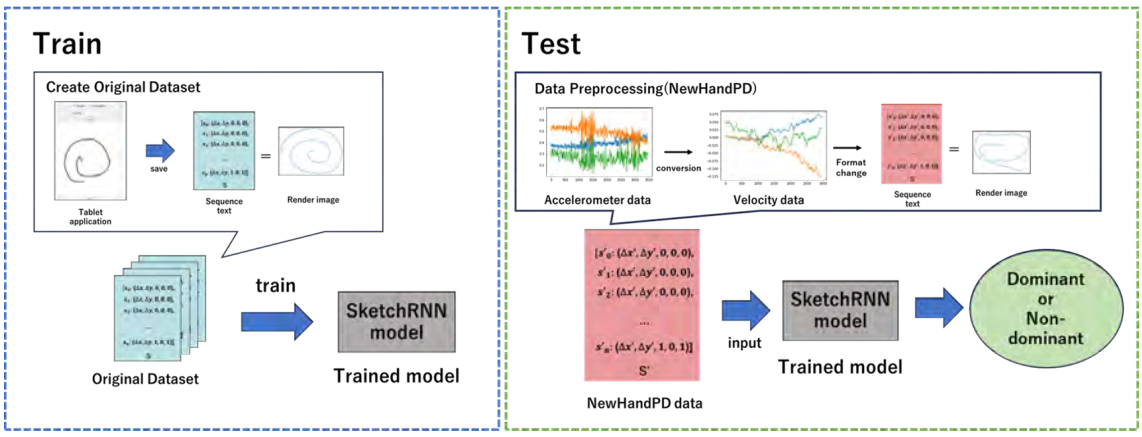


Fig. 5 Overall view of the experiment

the signal from the upper right with a high pass applied, and the lower left or final signal.

4.3 Experimental conditions

(a) Training the model

The classification accuracy depends on whether the test data's standard patterns fit the encoder's embedding space. So, the sketch category for pretraining has to be close to the expected test patterns. For these reasons, we used artificial data to train the Sketch-RNN model. The number of epochs for training is 200, and the batch size is 10. The parameters of the Sketch-RNN model were set to 0.0001 for τ, N_{max} were set to 2000 and the other parameters were set to the default values of the model. The Original Artificial Hand Drawing Dataset was divided into 840 train data, 40 validation data, and 40 test data (including 20 left hand data) for both spiral and meander. The NewHandPD dataset meander was set to 40 test data (26 for health subjects and 10 for patients) and spiral was set to 40 test data (33 for health subjects and

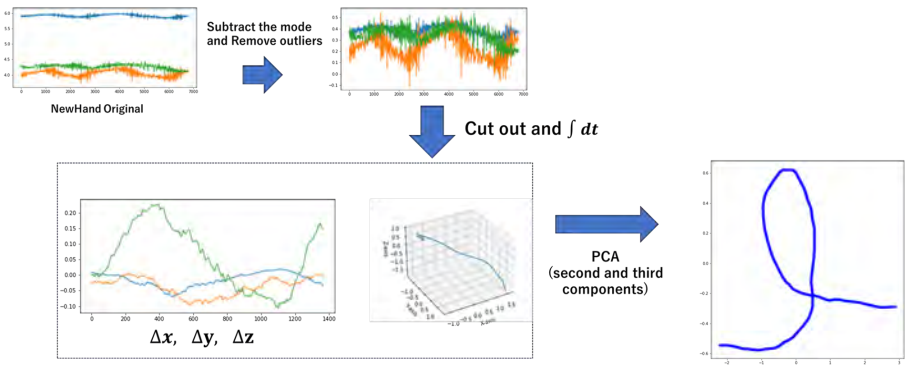


Fig. 6 The process of converting acceleration signals to velocity signals

7 for patients). When converting NewHandPD to sketch sequence, the pen states $p1$, $p2$, and $p3$ are set to $(1, 0, 1)$ for the last point only, and $(0, 0, 0)$ for the rest, based on the assumption that the pen is not lifted due to the characteristics of the NewHandPD dataset. The high-pass parameters are Passband end frequency [Hz]: 3000, Frequency at the end of the stopband [Hz]: 150, Maximum loss [dB] at the passband end: 3, minimum loss at the end of the stopband [dB]: 40, sample rate: 30.0.

(b) Pre-test of the Model

As the model's pre-test, we test the accuracy of the recognition rate for distinguishing the standard pattern of a healthy right-handed person from the pattern drawn by the left hand on purpose. For this pre-test, we prepared only the spiral category. We didn't use the meander category because the meander pattern is hard to classify as normal or other.

(c) Test of the Model

The NewHandPD dataset consisted of 100 train data (healthy subjects only), 40 validation data, and 40 test data. In this work, we used only the test data. We selected two patterns, spiral, and meander, from the categories contained in the NewHandPD dataset; spiral, meander, circle, and diag, as the test data. We did not use the patterns of circle and diag in this work since the circle is too simple and the diag is too complex, even though it seems complicated to distinguish between a healthy person and a patient. We used the NewHandPD dataset for test data.

4.4 Result

(a) Input data and the reconstruction output

Figure 7 shows the example of the test data of the spiral and meander categories of the NewHandPD dataset and the reconstruction outputs of the Sketch-RNN models. The first row shows the input spiral patterns and the output reconstruction patterns by the model trained with the Original Artificial Hand Drawing Dataset, corresponding to the input spiral patterns. The second row shows the input meander patterns and the output reconstruction patterns by the model trained with Original Artificial Hand Drawing Dataset, corresponding to the input meander patterns. The models are those trained with the Original Artificial Hand Drawing Dataset.

Table 1 ROC-AUC values and recognition rates(%) for each training model and each test data

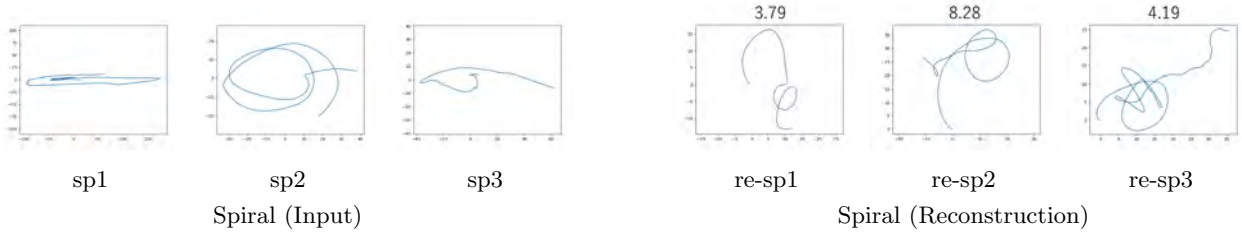
model	test data meander	test data spiral
Donut	0.60 62.50%	0.67 67.50%
Book	0.42 62.50%	0.36 55.00%
Circle	0.60 62.50%	0.61 60.00%
Square	0.63 67.50%	0.67 62.50%

(b) Recognition of the Attribute (Healthy / Patient)

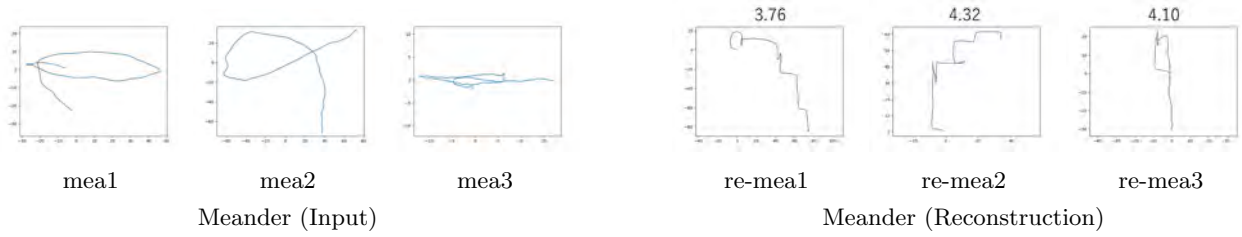
We use ROC-AUC (The area below the ROC curve) and the recognition rate (1 - mean of errors). **Table 1** and **Table 2** shows the result of the recognition rate based on the reconstruction error by the models trained with the QuickDraw dasaset and the Original Artificial Hand Drawing Dataset, respectively. **Figure 8** shows the ROC curves. The former work of the NewHandPD dataset[16] transfers the online vector patterns to gray-scale images and obtains the recognition rate as the image classification task by CNN. We cannot directly compare the recognition rate; however, we show the number as the reference.

4.5 Discussion

Table 1 shows the ROC-AUC values and recognition rates(%) for the test data (meander and spiral). The overall recognition rate is not high. When the model is



(a) Example of Input/Output of VAE Model for Spiral Pattern



(b) Example of Input/Output of VAE Model for Meander Pattern

Fig. 7 Examples of the input spiral and meander data of the NewHandPD and each corresponding output reconstruction pattern

Table 2 ROC-AUC values and recognition rates(%) for the pre-test data (artificially made Right/Left hand data), and the test data (health/patient data of the NewHandPD dataset)

Category	Pre-test (Right/Left hand)		Test (Health/Patient)	
	ROC	Accuracy	ROC	Accuracy
spiral	0.89	80.0%	0.91	95.0%
meander	-	-	0.57	82.5%

Table 3 Comparison of recognition rates using the ImageNet pre-trained caffe model with the Our method, according to the original literature (Pereira[16])

		meander	spiral
Pereira[16]	64 × 64	85.00%	80.19%
	128 × 128	87.14%	77.53%
Our Method		82.5.00%	95.00%

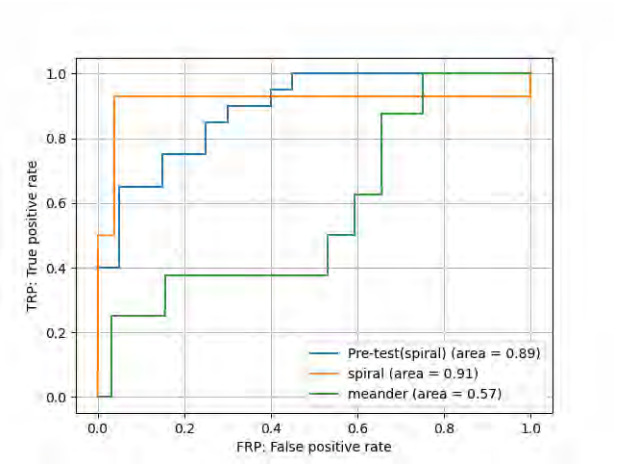


Fig. 8 The ROC curve of the recognition of health or patient of NewHandPD dataset

trained with the data of Book category, and the test category is spiral, the recognition rate is lower than when the model is trained with the data of "Donut", "Circle" and "Square" categories. This result might reflect that the data distribution of book category is more different from the data distribution of spiral, than the data distribution of "Donut", "Circle" and "Square" categories are. Table 2 shows the ROC-AUC values and recognition rates(%) for the pre-test data (artificially made Right/Left-hand data) and the test data (health/patient data of the NewHandPD dataset). The pre-test recognizes artificially made regular and anomaly data. This time, we regard the data written with the left hand by the right-handed person as an anomaly. The accuracy is 80%, which suggests a lot of average data well trains the Model. The result of the test of recognizing health or patient data of the NewHandPD dataset is 95% and 82% for Spiral and Meander patterns each. We obtained higher accuracy for spiral than meander, while the reference value of Pereira's work in **Table 3** shows the reverse order. We consider it because Pereira's CNN model for 2D rendering images tends to capture the edge characteristics and is good at meandering. On the other hand, our Model based on RNN will learn smooth and isotropic movements like spirals rather than discontinuous movements like meanders.

5. Conclusion

We proposed a method of classifying the attributes of hand-drawing signal sequences using Sketch-RNN, an Autoencoder that inputs and outputs vector sequences by a recurrent process. We trained the model with a large hand drawing dataset QuickDraw, and a certain amount of artificial data that we originally made by hand drawing an interface and tested the model with the small public data of a rehabilitation test, ensuring the thoroughness and validity of our experiment. The experiment shows that the model trained standard patterns and can recognize healthy patients. It worked better for the artificial data than QuickDraw, which suggests that Sketch-RNN as an anomaly detection model requires the training data whose generation process is close to that of the test data. It worked better for the spiral pattern than the meander pattern, which suggests that the Sketch-RNN model learns smooth and isotropic movements better than discontinuous movements. This work was limited to 2D trajectories, but future work will include improving the model for 3D trajectories.

References

1) D. P. Kingma, M. Welling: "Auto-Encoding Variational Bayes," Proc. of 2nd International Conference on Learning Representations (2014).
2) L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich and K. Muller: "A Unifying Review of Deep and Shallow Anomaly Detection," arXiv:2009.11732v3 (2021).
3) B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson: "Estimating the support of a high-dimensional distribution," Neural Computation, Vol. 13, No. 7, pp. 1443-1471 (2001).
4) J. Laurikkala, M. Juhola, E. Kentala, N. Lavrac, S. Miksch, B. Kavsek: "Informal identification of outliers in medical data," Proc. of 5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, Vol. 1, pp. 20-24 (2000).
5) C. M. Bishop: "Novelty detection and neural network validation," IEE Proceedings - Vision, Image and Signal Processing, Vol. 141, No. 4, pp. 217-222 (1994).
6) K. Pearson: "On lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol. 2, No. 11, pp. 559-572 (1901).
7) J. An, S. Cho: "Variational Autoencoder based Anomaly Detection using Reconstruction Probability," Special Lecture on IE, Vol.2, pp. 1-18 (2015).
8) T. Matsubara, K. Sato, K. Hama, R.Tachibana, K. Uehara: "Deep Generative Model Using Unregularized Score for Anomaly Detection With Heterogeneous Complexity," Proc. of IEEE Transactions on Cybernetics, Vol. 52, No. 6, pp. 5161-5173 (2022).
9) M. Masuda, R. Hachiuma, R. Fujii, H. Saito, Y. Sekikawa: "Toward Unsupevised 3d Point Cloud Anomaly Detection Using Variational Autoencoder," Proc. of IEEE International Conference on Image Processing, pp. 3118-3122 (2021).
10) S. Lin, R. Clark, R. Birke, S. Schönborn, N. Trigoni, S. Roberts: "Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model," Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4322-4326 (2020).
11) X. -Y. Zhang, F. Yin, Y. -M. Zhang, C. -L. Liu, Y. Bengio: "Drawing and Recognizing Chinese Characters with Recurrent Neural Network," Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, No. 4, pp. 849-862 (2018).
12) F. Yin, Q.-F. Wang, X.-Y. Zhang, C.-L. Liu: "ICDAR 2013 Chinese Handwriting Recognition Competition," Proc. of International Conference on Document Analysis and Recognition, pp. 1464-1470 (2013).
13) D. Ha, D. Eck: "A Neural Representation of Sketch Drawings," Proc. of International Conference on Learning Representations (2018).
14) M. Schuster, K. K. Paliwal: "Bidirectional Recurrent Neural Networks," IEEE Trans. on Signal Processing, Vol. 45, No. 11, pp. 2673-2681 (1997).
15) J. Jongejan, H. Rowley, T. Kawashima, J. Kim, N. Fox-Gieg, Google Creative Lab / Data Arts Team: The Quick, Draw!, https://quickdraw.withgoogle.com (2016).
16) C. R. Pereira, S. A. T. Weber, C. Hook, G. H. Rosa and J. Papa: "Deep Learning-aided Parkinson's Disease Diagnosis from Handwritten Dynamics," 29th The Conference on Graphics, Patterns and Images (SIBGRAPI 2016).
17) H. Jeon, S. K. Kim, B. Jeon, K. S. Park: "Distance estimation from acceleration for quantitative evaluation of Parkinson tremor," Proc. of Annual International Conference of the IEEE Engineering in Medicene and Biology Society, pp. 393-396 (2011).
18) M. Eklund, S. Nuuttila, J. Joutsa, E. Jaakkola, E. Mäkinen, E. A. Honkanen, K. Lindholm, T. Vahlberg, T. Noponen, T. Ihalainen, K. Murtomäki, T. Nojone, R. Levo, T. Mertsalmi, F. Scheperjans, V. Kaasinen: "Diagnostic value of micrographia in Parkinson's disease: a study with FP-CIT SPECT," Journal of Neural Transmission (Vienna), 129(7), pp. 895-904 (2022).

(Received December 29, 2023)
(Revised March 12, 2025)

Shione ISHIDA
She received B.S. in Information Science from Toho University in 2022, and M.S. in Information Science from Toho University in 2024. She joined Hitachi Solutions, Ltd. in 2024.

Kyoko SUDO (*Member*)
She received the B.E. and M.E. degrees in mathematical engineering and information physics and a Ph.D. in information physics and computing from the University of Tokyo in 1991, 1993, and 2006. She was with NTT Corporation from 1993 to 2016. She is currently a professor at Toho University, Japan. Her research interests include computer vision, machine learning, and applications. She is a member of ITE, IPSJ, and IEICE.

IIEEJ **Short Paper**

Interactive Bayesian Optimization of Level of Abstraction for Stylized Image Composition

Ryoma HASHIMOTO† , Yoshinori DOBASHI† (*Member*)

† Hokkaido University

<**Summary**> We introduce an interactive system for stylization of a photograph to match the level of abstraction of a given foreground illustration. The stylized photograph can then be used as the background. We employ an existing automatic stylization method but it often has parameters to control the degree of abstraction. Our system allows the user to interactively find the optimal parameters that produce the desired appearance. We conduct a simple user-study to investigate the usefulness of our system.

Keywords: interactive system, image stylization, Bayesian optimization

1. Introduction

In recent years, artists have used image processing techniques to create stylized illustrations, often converting photographs into stylized backgrounds for foreground objects. However, adjusting the color and atmosphere of photographs to match the foreground can be time-consuming, especially for those without the necessary skills.

Automatic style transfer methods, such as those based on machine learning, can transfer the style of a foreground object to a background photograph. The method proposed by Huang and Belongi[1)]abstracts and stylizes the photograph to match the foreground image. However, the method often produces unnatural results and requires time-consuming trial and error to find the optimal parameters for the stylization. Additionally, existing methods apply uniform abstraction across the entire image, which may not be ideal; the level of abstraction should vary across an image depending on its contents.

To address this, we propose an interactive system that helps users find the optimal parameters. Our system subdivides the photograph into segments, transferring the foreground style to each segment. Using Bayesian optimization, it estimates the best parameters based on user preferences. We demonstrate the effectiveness of our system through a user study.

2. Related work

Style transfer has been a key research topic for over 20 years, beginning with procedural methods like Image Analogies[2)]and Image Quilting[3)]. Recent methods employ neural networks for more flexible and high-quality results. Gatys et al. were the first to achieve style transfer using neural networks[4)], but their approach requires retraining the network for each new image. Xun et al. introduced real-time style transfer using Adaptive Instance Normalization (AdaIN)[1)], which aligns feature map statistics between content and style images. However, this method applies a uniform style transfer to the entire content image. So, there may be insufficient processing for detailed regions or excessive transformation for regions with fewer details

For interactive parameter tuning, Koyama et al. proposed Sequential Gallery[5)], where users select from a set of images created with different parameter settings, and the system refines the parameters using Bayesian optimization. Our method extends this idea to stylized image composition.

3. Proposed method

An overview of our system is illustrated in **Fig.1**. The user provides a stylized image of a foreground object and a photograph to serve as the background. We assume the photograph has been subdivided into segments, such as sky and ground, each with similar detail levels. This segmentation is user-provided via an interactive tool[6)].

The style of the foreground image is transferred to each segment, and the final image is created by combining the results. For style transfer, we use the AdaIN model[1)], which has a single parameter, α, controlling the level of

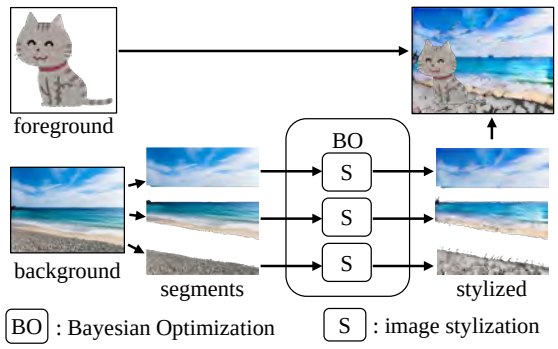


Fig. 1 Overview of the proposed method

abstraction.

To match the generated image to the user's preference, we must optimize the parameter for each segment. However, since the user's preference is difficult to define mathematically, this becomes a black-box optimization problem. To address this, we develop an interactive system based on Bayesian optimization[7]. The system presents multiple images with different parameters, and the user scores each one. The system then estimates the user's preferences and updates the parameters accordingly.

3.1 Overview of Bayesian optimization

Bayesian optimization is designed for black-box functions that facilitates finding the global optimum[7]. We employ standard Bayesian optimization. It consists of three main steps. First, based on the results obtained from past experiments, a function called the acquisition function is determined to identify the next point to explore in the objective function. A Gaussian process is used to construct the acquisition function. The Gaussian process predicts the output for unknown data, providing the mean and standard deviation of the predictions. In our case, the objective function reflects the user's preference, and the data points are the parameters for stylizing the photograph segments. We use the Upper Confidence Bound (UCB) acquisition function to balance exploration and exploitation[7]. Next, the point that maximizes is determined. Finally, the newly selected exploration point and its corresponding value of the objective function are added to the existing experimental results, and the acquisition function is updated.

3.2 Parameter estimation using Bayesian optimization

Let the input photograph be subdivided into n segments. For each segment, we need to determine a corresponding parameter for style transfer for each segment. Applying style transfer to each segment individually results in high computation time, making the optimization process slow.

To address this, we sample m parameters at regular intervals and precompute a set of stylized images for each segment. We use $m = 20$ for the examples in this paper. The problem then becomes finding the optimal combination of n parameters from the discrete set $\alpha_{ij}, (i = 0, \ldots, n-1; j = 0, \ldots, m-1)$. Our Bayesian optimization process consist of two phases, initial evaluations and Bayesian exploration, as described in the following subsections.

3.2.1 Initial evaluations

The first phase is the initial evaluation. Bayesian optimization estimates promising parameters sequentially based on accumulated results, but initially, there are insufficient evaluations. Therefore, in this phase, we heuristically select multiple parameter sets and ask the user to evaluate them.

To guide the initial selection, we leverage Bayesian optimization's tendency to explore unexplored areas. Thus, we use extreme parameter sets (zeros and ones) for the evaluations, which helps make the subsequent estimation more efficient. Additionally, we include several randomly chosen sets of parameters to increase the number of initial evaluations. The number of random sets is proportional to the number of the segments.

The system first presents images with zero parameters, followed by those with random sets, and finally with parameters set to one. The user scores each image.

3.2.2 Bayesian exploration

Based on the evaluation results, our system estimates promising parameters using the acquisition function and presents the resulting stylized image to the user. The user then provides a score for the image. This process repeats until the user is satisfied. The goal is to achieve an optimal image style that aligns with the user's preferences. By scoring each image, Bayesian optimization refines the parameter estimation, progressively adjusting the image style to match the user's needs.

We developed a simple interface for interactive optimization, as shown in **Fig. 2**. The stylized image, rendered with the suggested parameters, is displayed in the top left (1). To the right, segmentation results are shown to help the user understand how each region is stylized (2). Several buttons at the bottom allow the user to score the current image (3). A "satisfaction button" lets the user indicate their overall satisfaction with the result. Additionally, a button on the right enables toggling the

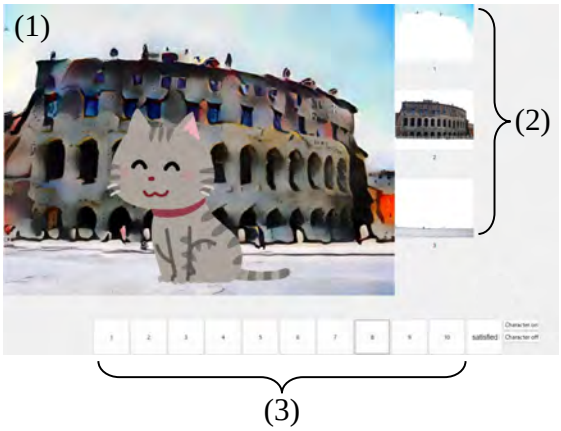


Fig. 2 Our user interface for Bayesian optimization


(a) photograph for background (b) foreground illustration

Fig. 3 Input images used for the experiment

display of the foreground object (e.g., the stylized cat in this example). This interface allows users to intuitively participate in the optimization process and create images that best suit their preferences.

4. Result

In this chapter, we present experimental results to evaluate the effectiveness of the proposed method. The experiments were conducted on a desktop PC with an Intel Core i7 13700 CPU, 32GB of memory, and an NVIDIA GeForce RTX 4070 Ti GPU.

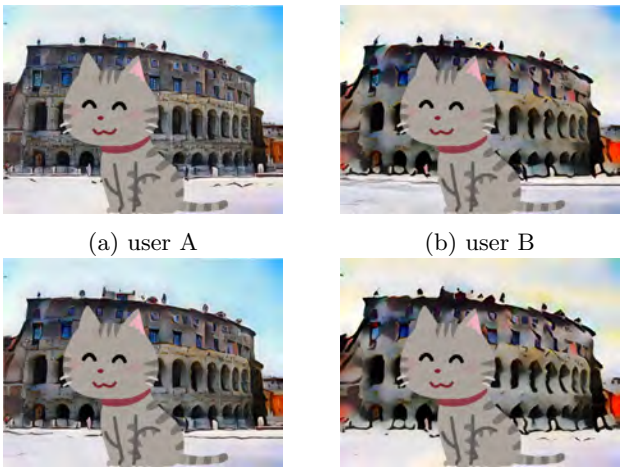
For the following examples, we use the images shown in **Fig. 3**, where the foreground is an illustration of a cat and the background is a photograph of the Colosseum in Rome. The background photograph is segmented into three regions: the sky, the building, and the road.

Figure 4 shows examples generated by our method. Figures 4(a) through (c) are produced by three different users. For comparison, Fig. 4(d) is generated without segmentation.

As these examples illustrate, each user prefers a different level of abstraction depending on the content of each segment. This demonstrates the ability of our method to generate stylized images tailored to individual user preferences.

4.1 User study

To verify the effectiveness of the proposed method, we



(a) user A (b) user B

(c) user C (d) without segmentation

Fig. 4 Example of results generated by using our system

conducted a comparative experiment against manual parameter adjustment. The study had two main objectives:

1. To assess the usefulness of our interface in helping users, even those without knowledge of style transfer, find a satisfactory parameter set.
2. To evaluate the efficiency of our system in generating the desired image compared to manual parameter exploration.

Nine university students in their twenties participated in this study.

4.2 Method

Participants were first briefed on the procedures for manual adjustment and Bayesian optimization. Each participant then tried both methods twice. To minimize the order effect, the sequence of methods was alternated, with the experiment proceeding as follows: Bayesian optimization, manual adjustment, manual adjustment, and Bayesian optimization.

The user interface for manual adjustments is shown in **Fig. 5**. It is similar to the one used for Bayesian optimization (Fig. 2), but instead of evaluation buttons, slider bars are provided to adjust the parameters for each segment. After selecting parameters with the sliders, the user clicks a button to generate the stylized image. The satisfaction button is also included below the sliders.

Each task was allocated three minutes, and the operation was forcibly terminated after that time, regardless of user satisfaction. Four pairs of foreground and background images were used. The background images had already been segmented into multiple regions. During the briefing, participants worked with a background image segmented into four regions and a foreground image

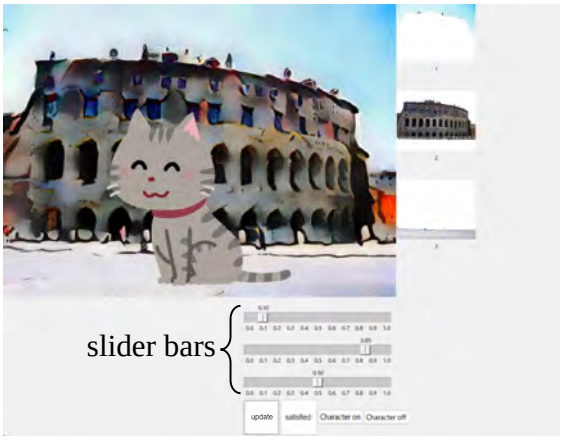


Fig. 5 User interface for manual adjustment

with moderate details. For the first round, a background image with three regions and a foreground image with low details were used, and for the second round, a background image with five regions and a foreground image with high details were used.

Participants were asked to assume they were using the system to post the resulting image on social media. They clicked the satisfaction button when they felt the stylized image was of sufficient quality. The time taken to reach the satisfaction button was recorded.

After completing all tasks, participants answered the following two questions to assess the system's effectiveness:

Q1 Which method do you prefer for generating images: Bayesian optimization or manual adjustment?

Q2 Which type of image do you prefer: those generated without background segmentation or those with segmentation?

The responses were rated on a 7-point Likert scale, with a score above 3.5 indicating a preference for our method.

4.3 Results

Table 1 summarizes the results of the user study. The second column compares the average time to satisfaction between Bayesian optimization (BO) and manual adjustment (MA). With Bayesian optimization, the time was about half as long as with manual adjustment.

The third column compares the average number of image generation attempts within three minutes. While there was no significant difference in the number of attempts, the number was slightly lower with Bayesian optimization as the number of segments increased. This suggests that our system enables users to more efficiently explore the parameter space to find a satisfactory result.

Finally, the questionnaire results, summarized in Table 2, further support the usefulness of our system.

Table 1 Results of the user study. N_{seg} is the number of segments in the input photograph. BO and MA indicate Bayesian optimization and manual adjustment, respectively.

N_{seg}	time to satisfy		# attempts	
	BO	MA	BO	MA
3	62.1s	121.1s	21.8	20.2
5	71.5s	151.6s	14.4	18.4

Table 2 Results of the questionnaires. The average scores (Avg) and the standard deviations (SD) are shown.

	Q1	Q2
Avg	4.67	5.22
SD	1.87	1.79

5. Conclusion

This paper introduces a method for efficiently generating background images suitable for character illustrations using style transfer. We demonstrated that the method can produce user-satisfying results by optimizing parameters based on user evaluations.

A direction for future work is reducing the computational cost of generating stylized backgrounds, especially as the number of regions increases.

References

1) X.Huang, S.Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization" Proc. of 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1510-1519 (2017).
2) A.Hertzmann, C.E.Jacobs, N.Oliver, B.Curless, D.H.Salesin, "Image Analogies," Proc. of ACM SIGGRAPH 2001, pp.327-340 (2001).
3) A.A.Efros, W.T.Freeman, "Image quilting for texture synthesis and transfer," Proc. of ACM SIGGRAPH 2001, pp.341-346 (2001).
4) L. A. Gatys, A. S. Ecker, M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414-2423 (2016).
5) Y.Koyama, I.Sato, M.Goto, "Sequential Gallery for Interactive Visual Design Optimization" ACM Transactions on Graphics, Volume 39, Issue 4, Article No.:88, pp.88:1-88:12 (2020).
6) S.Beucher, F.Meyer, "The Morphological Approach to Segmentation: The Watershed Transformation" Proc. of Mathematical Morphology in Image Processing, pp.433-481 (1993).
7) B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," Proc. of the IEEE, Vol. 104, No. 1, pp. 148-175 (Jan. 2016).

(Received Jan. 8, 2025)
(Revised Mar. 25, 2025)

Ryoma HASHIMOTO
He is a master course student at Hokkaido University, Japan. His research interests include image stylization by deep neural network.

Yoshinori DOBASHI (*Member*)
He is a Professor at Hokkaido University, Japan. His research interests center on computer graphics, including realistic image synthesis, efficient rendering, and sound modeling for virtual reality applications. He received his BE, ME and PhD in Engineering in 1992, 1994, and 1997, respectively, from Hiroshima University. He worked at Hiroshima City University from 1997 to 2000 as a research associate.

Call for Papers

Special Issue on
Image Electronics Technologies Related to AI

IIEEJ Editorial Committee

The rapid advancements in artificial intelligence (AI) technologies in recent years have profoundly accelerated and enhanced various image-electronics-related fields, including image and video processing, recognition, and generation. These technologies have vast potential applications, spanning autonomous driving, medical image diagnostics, facial recognition systems, anomaly detection in industrial settings, surveillance cameras, drones, and beyond. At the same time, addressing societal challenges arising from the misuse of these technologies—such as the generation of fake images—is expected to become a pressing issue. Nonetheless, it is evident that AI technologies will continue to grow in importance, playing an increasingly pivotal role in the field of image electronics technologies.

This special issue invites a broad range of submissions focusing on research advancements in AI and their impact on image-electronics-related technologies, as well as evaluations of their practical applications. Accepted contributions may include research papers, system development papers, practice-oriented papers, and survey papers.

1. Topics covered include but are not limited to
- Application of AI in image processing (recognition, classification, generation)
- Image recognition technologies using machine learning and deep learning
- Improvements in video compression and transmission technologies using AI
- Computer vision technologies using AI
- Image generation and editing using generative AI
- Fusion technologies between natural language processing and image processing
- Application of AI in medical image processing
- AI in automated driving
- Application of AI in video analysis using surveillance cameras and drones

2. Treatment of papers

   The submission paper style format and double-blind peer review process are the same as the regular paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as the regular contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:

   IIEEJ Transactions on Image Electronics and Visual Computing Vol.14, No.1 (June 2026)

4. Submission Deadline:

   Friday, October 31, 2025

5. Contact details for Inquiries:

   IIEEJ Office E-mail: hensyu@iieej.org

6. Online Submission URL: http://www.editorialmanager.com/iieej/

First Call for Papers

The 9th IIEEJ International Conference on Image Electronics and Visual Computing 2026 (IEVC2026)

Hiroshima Univ., Hiroshima, Japan / March 16-19, 2026

https://www.iieej.org/en/ievc2026/ (in preparation)

Purpose:

The International Conference on Image Electronics and Visual Computing 2026 (IEVC2026) will be held in Hiroshima City, Japan, on March 16-19, 2026, as the 9th international academic event of the Institute of Image Electronics Engineers of Japan (IIEEJ). Past IEVCs were held in Cairns, Australia (2007); Nice, France (2010); Kuching, Malaysia (2012); Koh Samui, Thailand (2014); Danang, Vietnam (2017); Bali, Indonesia (2019); online (2021); and Tainan, Taiwan (2024). The conference aims to bring together researchers, engineers, developers, and students from various fields in both academia and industry for discussing the latest researches, standards, developments, implementations and application systems in all areas of image electronics and visual computing.

Topics:

The conference will cover a broad set of research topics including, but not limited to, the following:

- 3D image processing
- Bioinformatics and authentication
- Computer vision
- Data hiding
- Image analysis and recognition
- Image and video coding
- Image and video retrieval
- Image assessment
- Image restoration
- Mobile image communication
- Motion analysis
- Object detection
- Printing and display technologies
- Segmentation and classification
- Smart display
- Versatile media appliance
- Animation
- Content production
- Extended Reality
- Metaverse
- Modeling
- Non - photorealistic rendering
- Rendering
- Visual computing
- Visualization
- Architectural industry mondiale
- Artificial intelligence and deep learning
- Big data and cloud computing
- Content delivery network
- Digital museum, digital archiving
- Generative AI
- Hardware and software implementation
- Interaction
- International standards
- Security and privacy
- Social secured cybertechnology
- Unmanned Aerial Vehicle
- Visual and hearing impaired support
- Visual communication

Paper submission:

The official language is English, and authors should submit their papers as PDF through the online submission system, which will be available around June 2025 at the following IEVC2026 official website:

https://www.iieej.org/en/ievc2026/

The paper submission guide and IEVC formats (TeX format / MS Word format) will be also provided at this site. The organizing committee particularly encourages graduate students to present their works in the special sessions that are now planned by the committee of the conference.

General Papers:

The general papers category is divided into two types: journal track and conference track.

✧ Journal Track:

Journal track aims to publish the papers on the journal in addition to the publishing in the conference, with a quick review process. This type of paper will appear in a special issue on "Journal Track Papers in IEVC2026" in the IIEEJ Transactions on Image Electronics and Visual Computing, Vol. 14, No. 1 (June 2026), if accepted through the journal review process. The authors have to prepare two types of papers different in the amount, the paper for the conference and the paper for the journal. The latter one is the extended version of the former one. **Note that the paper for the journal should follow the "guidance for paper submission" available from the website of IIEEJ, to be finally published in the IIEEJ Transactions.**

Important Dates

- Pre-Entry Submission (title, authors, 100 words abstract):	Sept. 12, Friday, 2025
- Paper Submission (2-4 pages, for the conference):	Sept. 26, Friday, 2025
- Paper Submission (6-8 pages, for the journal):	Oct. 24, Friday, 2025
- Notification of Conference Acceptance:	Nov. 21, Friday, 2025
- Camera-Ready Paper (2-4 pages, for the conference):	Dec. 12, Friday, 2025

✧ Conference Track:

Conference track aims to present the papers about recent results and preliminary work at IEVC2026. The authors are required to submit a paper of which length is 2-4 pages. Accepted papers will be published both in the online proceedings of IEVC2026 (indexed by J-stage) and in the USB proceedings. Rejected papers in the conference track can be resubmitted as late breaking papers.

Important Dates

- Pre-Entry Submission (title, authors, 100 words abstract):	Sept. 12, Friday, 2025
- Paper Submission (2-4 pages):	Sept. 26, Friday, 2025
- Notification of Acceptance:	Nov. 21, Friday, 2025
- Camera-Ready Paper (2-4 pages):	Dec. 12, Friday, 2025

Late Breaking Papers:

All suitably submitted papers for this category will be accepted for the conference. The authors must submit an abstract of which length is 1-2 pages, and select one from the following two types: 1) Technical papers or 2) Art/Demo papers. All the registered papers as late breaking papers will be published only in the USB proceedings of IEVC2026.

Important Dates

- Pre-Entry Submission (title, authors):	Nov. 24, Monday, 2025
- Abstract Submission (1-2 pages):	Nov. 28, Friday, 2025
- Notification of Acceptance:	Dec. 5, Friday, 2025
- Camera-Ready Paper (1-2 pages):	Dec. 12, Friday, 2025

Further information:

After the conference, the Trans. on IEVC of IIEEJ is planning a forthcoming special issue on "Extended Papers Presented in IEVC2026", which will be published in Dec. 2026. More detailed information will be notified on the IEVC2026 website and the Journal of IIEEJ.

Revised: January 6, 2017

Revised: July 6, 2018

Revised: Dec. 10, 2024

Guidance for Paper Submission

1. Submission of Papers

(1) Preparation before submission

- The authors should download "Guidance for Paper Submission" and "Style Format" from the "Academic Journals", "English Journals" section of the Society website and prepare the paper for submission.
- Two versions of "Style Format" are available, TeX and MS Word. To reduce publishing costs and effort, use of TeX version is recommended.
- There are four categories of manuscripts as follows:
  - Ordinary paper: It should be a scholarly thesis on a unique study, development or investigation concerning image electronics engineering. This is an ordinary paper to propose new ideas and will be evaluated for novelty, utility, reliability and comprehensibility. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Short paper: It is not yet a completed full paper, but instead a quick report of the partial result obtained at the preliminary stage as well as the knowledge obtained from the said result. As a general rule, the authors are requested to summarize a paper within four pages.
  - System development paper: It is a paper that is a combination of existing technology or it has its own novelty in addition to the novelty and utility of an ordinary paper, and the development results are superior to conventional methods or can be applied to other systems and demonstrates new knowledge. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Data Paper: A summary of data obtained in the process of a survey, product development, test, application, and so on, which are the beneficial information for readers even though its novelty is not high. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Survey Paper: A summary of existing Research and Developments, organized under some viewpoint, compared for the sake of positioning purpose, observed as the changes in generations. Comprehensive references, overall perspective, objective evaluation, are needed without advertising specific organizations. It is also appreciated that the status and problems of the field, and the effect of them to the researchers and concerned people are undersood by the author, and the resulant paper encourages the new entry into the field, accelerates further development of related technologies, and prompts the development in even other fields or brand new researches. As a general rule, the authors are requested to summarize a paper within eight pages.

- To submit the manuscript for ordinary paper, short paper, system development paper, or data paper, at least one of the authors must be a member or a student member of the society.
- We prohibit the duplicate submission of a paper. If a full paper, short paper, system development paper, or data paper with the same content has been published or submitted to other open publishing forums by the same author, or at least one of the co-authors, it shall not be accepted as a rule. Open publishing forum implies internal or external books, magazines, bulletins and newsletters from government offices, schools, company organizations, etc. This regulation does not apply to a preliminary draft to be used at an

annual meeting, seminar, symposium, conference, and lecture meeting of our society or other societies (including overseas societies). A paper that was once approved as a short paper and being submitted again as the full paper after completion is not regarded as a duplicate submission.

(2) Submission stage of a paper

- Delete all author information at the time of submission. However, deletion of reference information is the author's discretion.
- At first, please register your name on the paper submission page of the following URL, and then log in again and fill in the necessary information. Use the "Style Format" to upload your manuscript. An applicant should use PDF format (converted from dvi of TeX or MS Word format) for the manuscript. As a rule, charts (figures and tables) shall be inserted into the manuscript to use the "Style Format". (a different type of data file, such as audio and video, can be uploaded at the same time for reference.)

  http://www.editorialmanager.com/iieej/
- If you have any questions regarding the submission, please consult the editor at our office.

Contact:
Person in charge of editing
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-Ku, Tokyo 116-0002, Japan
E-mail: hensyu@iieej.org
Tel: +81-3-5615-2893, Fax: +81-3-5615-2894

2. Review of Papers and Procedures

(1) Review of a paper

- A manuscript is reviewed by professional reviewers of the relevant field. The reviewer will deem the paper "acceptance", "conditionally acceptance" or "returned". The applicant is notified of the result of the review by E-mail.
- Evaluation method

  Ordinary papers are usually evaluated on the following criteria:
  - ✓ Novelty: The contents of the paper are novel.
  - ✓ Utility: The contents are useful for academic and industrial development.
  - ✓ Reliability: The contents are considered trustworthy by the reviewer.
  - ✓ Comprehensibility: The contents of the paper are clearly described and understood by the reviewer without misunderstanding.

A short paper can be evaluated by having a quickness on the research content and evaluated to have new knowledge with results even if that is partial or for specific use, apart from the novelty and utility of an ordinary paper.

A system development paper is evaluated based on the following criteria, apart from the novelty and utility of an ordinary paper.

- ✓ Novelty of system development: Even when integrated with existing technologies, the novelty of the combination, novelty of the system, novelty of knowledge obtained from the developed system, etc. are recognized as the novelty of the system.
- ✓ Utility of system development: It is comprehensively or partially superior compared to similar systems. Demonstrates a pioneering new application concept as a system. The combination has appropriate optimality for practical use. Demonstrates performance

limitations and examples of performance of the system when put to practical use.

A data paper is considered novel if new deliverables of test, application and manufacturing, the introduction of new technology and proposals in the worksite have any priority, even though they are not necessarily original, apart from the novelty and utility of an ordinary paper. Also, if the new deliverables are superior compared to the existing technology and are useful for academic and industrial development, they should be evaluated.

A survey paper is evaluated by comprehensiveness, overviewing point, and objectiveness apart from the novelty of an ordinary paper. Reliability, comprehensibility, completeness of reference papers are common to those in an ordinary paper. Utility is evaluated how the paper will enlighten the readers in the target fields.

(2) Procedure after a review

- In case of acceptance, the author prepares a final manuscript (as mentioned in 3.).
- In the case of acceptance with comments by the reviewer, the author may revise the paper in consideration of the reviewer's opinion and proceed to prepare the final manuscript (as mentioned in 3.).
- In case of conditional acceptance, the author shall modify a paper based on the reviewer's requirements by a specified date (within 60 days), and submit the modified paper for approval. The corrected parts must be colored or underlined. A reply letter must be attached that carefully explains the corrections, assertions and future issues, etc., for all of the acceptance conditions.
- In case a paper is returned, the author cannot proceed to the next step. Please look at the reasons the reviewer lists for the return. We expect an applicant to try again after reviewing the content of the paper.

(3) Review request for a revised manuscript

- If you want to submit your paper after conditional acceptance, please submit the reply letter to the comments of the reviewers, and the revised manuscript with revision history to the submission site. Please note the designated date for submission. Revised manuscripts delayed more than the designated date be treated as new applications.
- In principle, a revised manuscript willl be reviewed by the same reviewer. It is judged either acceptance or returned.
- After the judgment, please follow the same procedure as (2).

3. Submission of final manuscript for publication

(1) Submission of a final manuscript

- An author, who has received the notice of "Acceptance", will receive an email regarding the creation of the final manuscript. The author shall prepare a complete set of the final manuscript (electronic data) following the instructions given and send it to the office by the designated date.
- The final manuscript shall contain a source file (TeX edition or MS Word version) and a PDF file, eps files for all drawings (including bmp, jpg, png), an eps file for author's photograph (eps or jpg file of more than 300 dpi with length and breadth ratio 3:2, upper part of the body) for authors' introduction. Please submit these in a compressed format, such as a zip file.
- In the final manuscript, write the name of the authors, name of an organizations, introduction of authors, and if necessary, an appreciation acknowledgment. (cancel macros in the Style file)

- An author whose paper is accepted shall pay a page charge before publishing. It is the author's decision to purchase offprints. (ref. page charge and offprint price information)

(2) Galley print proof

- The author is requested to check the galley (hard copy) a couple of weeks before the paper is published in the journal. Please check the galley by the designated date (within one week). After making any corrections, scan the data and prepare a PDF file, and send it to our office by email. At that time, fill in the Offprint Purchase Slip and Copyright Form and return the scanned data to our office in PDF file form.
- In principle, the copyrights of all articles published in our journal, including electronic form, belong to our society.
- You can download the Offprint Purchase Slip and the Copyright Form from the journal on our homepage. (ref. Attachment 2: Offprint Purchase Slip, Attachment 3: Copyright Form)

(3) Publication

- After final proofreading, a paper is published in the Academic journal or English transaction (both in electronic format) and will also be posted on our homepage.

IIEEJ Trans. on Image Electronics and Visual Computing

Vol. 13, No. 1 2025

Editor in Chief: Osamu Uchida
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan

Print: ISSN 2188-1898
Online: ISSN 2188-1901
CD-ROM: ISSN 2188-191x