https://doi.org/10.2352/CIC.2025.33.1.33

3D CG HEVC Images Attention Mechanism Based on Vision Transformer

Norifumi Kawabata, Emerging Media Initiative, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa, Japan

Abstract

When creating media content, it is important to understand in advance where people focus their attention and what they are interested in when viewing images. Thus far, we have measured the area using saliency maps to determine the relationship between picture coding and gaze points. However, as an evaluation metric, it is not enough to consider factors other than area, and the accuracy was not enough. Therefore, we thought that we could achieve higher accuracy it estimating areas of interest by using the knowledge of Vision Transformer, which was developed in the field of computer vision based on natural language processing. In this study, we experimentally examined the evaluation of regions of interest in images by performing region of interest estimation using the attention mechanism of Vision Transformer on HEVC encoded 3D CG images and then analyzing the images.

1. Introduction

In everyday life after COVID-19, with the spread of telework and working from home, desk work has increased, and with it, the amount of time spent staring at displays and viewing images and videos has also increased. When viewing images and videos, humans routinely judge for themselves whether something is important or not, efficiently acquiring a wide variety of information, which they then process and ultimately understand. With the question of whether it is possible to reproduce such highly advanced human intelligence on a computer, research has been conducted in the 2010s research has been conducted using computer image processing to select image information based on importance. This includes studies based on subjective evaluations [2, 3] that consider image domain features and image encoding [1], as well as studies that research has been conducted using convolutional neural networks to construct architectures for super-resolution processing, based on objective evaluation [4, 5].

With the advent of high-definition, high-quality images and video, such as 4K and 8K broadcasting in the broadcasting field, it is extremely important for content creators to understand in advance what users are interested in and what patterns of interest they have before creating content. Actually, even if content creators create content with a specific intention, it does not necessarily match what users are looking for. Furthermore, subjective evaluations by humans differ from objective evaluations by computers. Therefore, from the perspective of visual attention, it is necessary to develop and verify an image information system that can obtain visual information in advance about where attention is focused. As a preliminary study, the author conducted research on the estimation of areas of interest in 3D CG images based on saliency maps and visual attention [6, 7]. As a result, there was little dependence on 3D CG content, and it was considered possible to identify saliency by utilizing both Spectral Residual (SR) and Fine Grained (FG).

Based on the findings of these prior studies, this study uses Vision Transformer to approach the problem. Figure 1 shows the evolution of Transformer. Since Transformer was proposed in the field of natural language processing in 2017, its range of applications has expanded rapidly. From around 2019, Transformer became popular in Vision and Language, which is a fusion problem of images and language. By adding images as input to the Transformer, which processes text data, domain-specific models such as document images and medical data, as well as various Vision and Language tasks, have been applied. Since around 2020, attempts have been made to input only images into Transformers, and research on Transformers in the field of computer vision has been conducted. This is known as Vision Transformer. Transformers take into account the attention mechanism, which plays an important role in selecting information necessary to solve the target task. Since 2017, Transformers have been used in the field of natural language processing, and since around 2019, they have been used in the fusion of images and language. In 2020, Transformers began to be used for problems involving images only. In this study, we considered whether it would be possible to efficiently estimate regions of interest using Vision Transformers for problems involving images only.

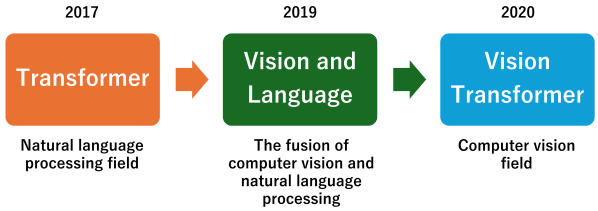

Figure 1. *The Evolution of Transformers*

In this study, we performed region of interest estimation by designing and using a Vision Transformer architecture that takes into account the attention mechanism for HEVC-encoded 3D CG images. We then analyzed image and visualized the results to experimentally examine the characteristics of regions of interest in images.

2. Related work

In this chapter, we will explain related work divided into (1) Vision Transformer architecture, (2) Vision Transformer in image processing and evaluation, and (3) Vision Transformer in medical images.

2.1. From the perspective of the Vision Transformer architecture

From the perspective of Vision Transformer architecture, Progressive Semantic Guidance for Vision Transformer for Zero-Shot Learning (ZSL) [8], Hierarchical Vision Transformer for Deformable Image Registration [9], Training Vision Transformer for Semi-Supervised Semantic Segmentation [10], multi-criteria

token fusion using one-step forward attention for efficient Vision Transformers [11], introducing ResNet and Vision Transformers into spiking neural networks [12], single-stage importance exploration and sparse exploration for Vision Transformer compression [13], High-Density Vision Transformer with Few Samples [14], Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation via Relational Descriptors [15], and other studies have been conducted.

2.2. From the perspective of Vision Transformer in image processing and evaluation

From the perspective of Vision Transformer in image processing and evaluation, we present a Transformer for image quality evaluation [16], dual-branch cross-modal calibration for RGB-D saliency instance segmentation [17], and vision and language consistency-guided multimodal prompt learning for blind AI-generated image quality evaluation [18]. Exploring Vision Transformers for 3D Human Action Language Models Using Motion Patches [19], A Top-Down Approach from Semantics to Distortion for Image Quality Evaluation [20], Improving Image Brightness Through the Interaction of Dual-Branch Structure and Texture Recognition Features [21], Self-supervised learning for image quality evaluation [22], Self-supervised pre-training model for image quality evaluation [23], and other studies have been conducted.

2.3. From the perspective of the Vision Transformer with medical imaging

From the perspective of Vision Transformer in medical imaging, triple supervised convolution transformation aggregation for robust monocular endoscope high-density depth estimation [24], dual input Transformer End-to-End model for preoperative evaluation of pathological complete response to neoadjuvant chemotherapy for breast cancer using ultrasound examination [25], a multi-task-based deep learning framework using Candmark Kenshuts for MRI segmentation [26], expert-guided knowledge distillation for semi-supervised vascular segmentation [27], a self-supervised network with convolutional and self-preserving mechanisms for unsupervised cell segmentation [28], Adaptive decomposition and shared weight volume transformation blocks for patch-free 3D medical image segmentation [29], Efficient medical image segmentation framework based on Transformer and U^2-Net [30], Dynamic hierarchical Transformer network for liver and tumor segmentation [31], and other studies have been conducted.

2.4. Conclusion

Overall, there has been no research on Vision Transformer that comprehensively considers computer graphics, image quality evaluation, regions of interest, and computer vision. In this study, we will conduct research on model construction and evaluation while taking these points into consideration.

3. Vision Transformer

3.1. Overview

In October 2020, a paper on ViT (Vision Transformer) was published on arXiv, and this paper was accepted at the International Conference on Learning Representations (ICLR) held in May 2021 [32]. Figure 2 shows an overview of ViT. In ViT, similar to DETR (DEtection TRansformer), which applied Transformers to object detection in May 2020 by the Facebook AI research team [33], images are divided into patches of equal size and treated as tokens. While DETR used ResNet to obtain feature vectors for

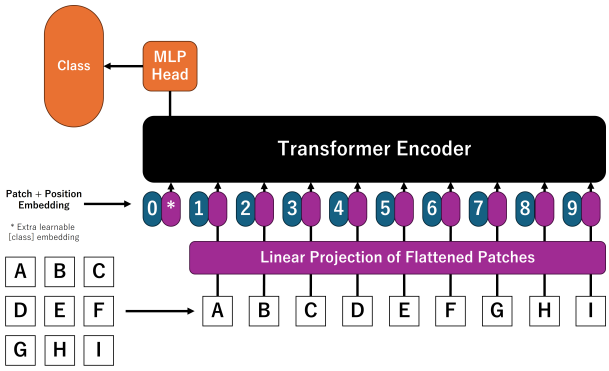

Figure 2. *Overview of Vision Transformer*

each region, ViT treats the pixel values within each patch as a vector and uses the linear transformation of this vector as the vector representation for each token. In addition to the conventional ImageNet dataset, the experiments used a proprietary dataset called JFT-300M, consisting of approximately 300 million images, for pre-training. As a result of the experiments, ViT pre-trained on ImageNet achieved recognition accuracy lower than conventional models, but ViT pre-trained on JFT-300M achieved recognition accuracy higher than conventional methods.

In this way, ViT, which appeared in October 2020, achieved many state-of-the-art (SoTA) results in image classification tasks by using self-attention instead of convolution. ViT is composed of the following three main components:

- Input Layer
- Encoder
- MLP Head

In the following sections, we will briefly explain these components.

3.2. Input layer

In the input layer, the following four processes are performed on the input image.

1. Divide into patches
2. Embedding
3. Class token
4. Position embedding

The input image is first divided into patches in the Input Layer. Figure 3 shows the structure for dividing the input image into patches. The input to ViT is a patch, which is a 3-dimensional tensor (height, width, number of channels). Therefore, it is necessary to convert the patch into a vector, similar to a word. This can be achieved by simply "flattening" the 3-dimensional tensor patches into vectors. This "flattening" process is called flattening. The structure for flattening patches into vectors is shown in Figure 4. By dividing the input image into patches and flattening them, each patch can be converted into a vector.

Next, we perform embedding. Embedding refers to representing data such as images, videos, and text as vectors. We apply the inner product of vectors to one-hot vectors to obtain better vectors that minimize the final loss (such as cross-entropy loss). To obtain better vectors, we use a linear layer called an embedding layer.

Then, perform class tokenization. The Class Token can be regarded as a vector that condenses the information of the entire image. The Class Token and all vectors for each patch

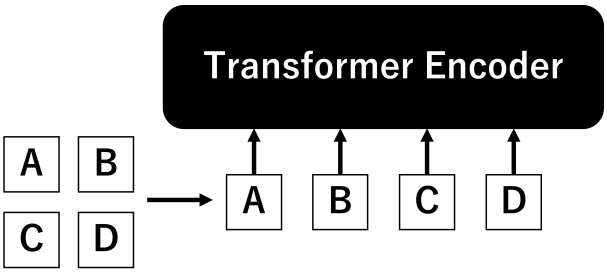

Figure 3. *Divide the input image into patches*

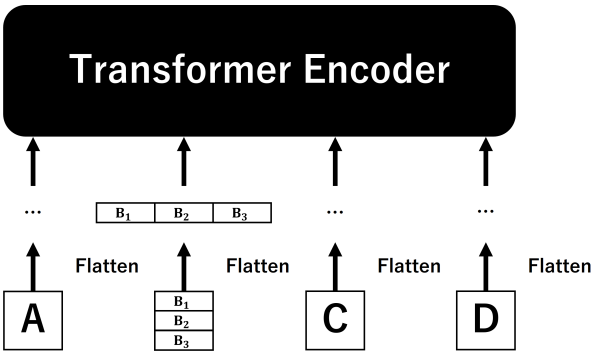

Figure 4. *Flatten patch to vector*

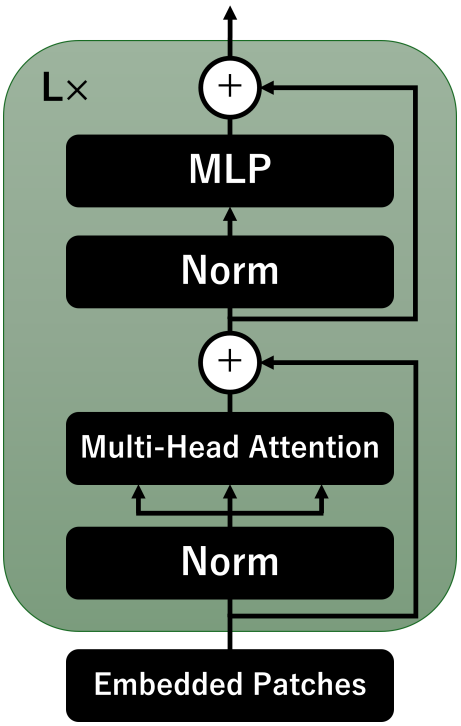

Figure 5. *Processing of Transformer Encoder*

are input to the Encoder, but the output from the Encoder is only the Class Token. The MLP (MultiLayer Perceptron) Head, which serves as a class classifier, is fed only this Class Token and performs classification tasks such as image classification. The class token condenses the information of the entire image through Self-Attention. The class token is a vector of the same length as the patch embeddings, and it itself is a learnable parameter. The initial value of the class token is set to a random number following a standard normal distribution. The class token is newly concatenated to the beginning of all patch embeddings and input into the Encoder.

Finally, positional embedding is performed. The reason why ViT requires this process is because self-attention alone has the weakness of being unable to learn patch position information. Like the class token, positional embedding is also a learnable parameter, and its initial value is a random number following a standard normal distribution.

The output from the input layer is a vector corresponding to the class token and each patch.

3.3. Encoder

The encoder takes the vector obtained from the input layer as input and outputs the class token. The structure of the Transformer encoder process is shown in Figure 5. The encoder uses a mechanism called self-attention. This self-attention plays an important role in ViT, such as in the encoder block. Self-Attention is completed by performing embedding, inner product, and weighted sum. On the other hand, the Encoder Block consists of three components: “LayerNorm”, “Multi-Head Self-Attention”, and “MLP”. LayerNorm refers to Layer Normalization, a normalization method with two skip connections. Normalization uses the mean and standard deviation of the data itself. The MLP in the Encoder Block is composed of two linear layers. GELU is used as the activation function after the first linear layer. On the other hand, no activation function is used for the second linear layer. Here, GELU (Gaussian Error Linear Unit) is an activation function often used in Transformer-type models, and

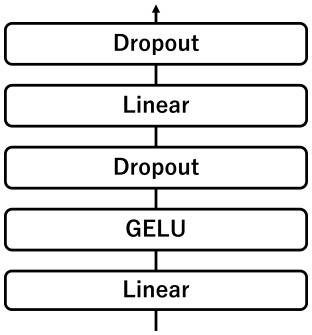

Figure 6. *MLP (MultiLayer Perceptron)*

is used in GPT and BERT instead of ReLU. Dropout randomly ignores some neurons during training.

3.4. MLP head

The final component of ViT, the MLP Head, is a classifier that performs class classification as shown in Figure 7. The classifier itself is simple, consisting of only two layers: LayerNorm and a linear layer. In image classification tasks, the length of the vector output by the MLP Head is the same as the number of classes.

4. Experiment

4.1. 3D CG image contents used in this study

The images used in this study are 3DCG content (Museum (M) and Wonder World (W)) such as the figure 8 provided free of charge by NICT [34]. To generate the images, we first constructed eight CG cameras, performed camera work and rendering [35], and generated eight still images. Although this content is originally multi-view 3D image content, in this study, we used single-view images from one of the 8 viewpoints.

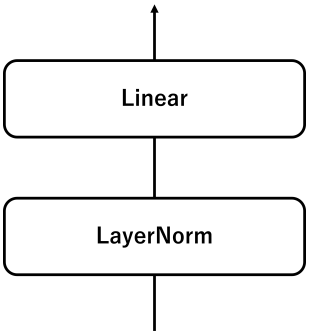

Figure 7. *Architecture of MLP Head*

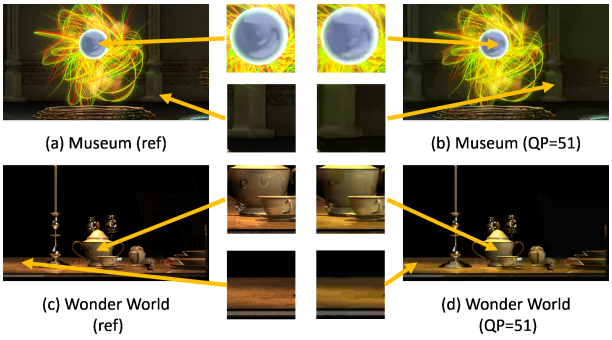

Figure 8. *3D CG image contents used in this study*

4.2. Experimental procedure

The experimental procedure is described below.

1. Set camera work and rendering from 3D CG content to generate 3D CG images.
2. Perform H.265/HEVC encoding processing on the generated 3DCG images. In this study, we set seven types of quantization parameters $Q = ref, 20, 25, 30, 35, 40, 51$ and generated evaluation content.
3. We constructed the Vision Transformer architecture and adjusted the parameters. In this study, we used the ailia SDK to construct and use the ViT architecture. All parameters were set to their default (initial) values.
4. Using the ViT architecture and framework, we performed region of interest estimation and prediction on the evaluation content, generating input images, region of interest color maps, prediction labels, and gaze point maps as result images.
5. Based on the results of Experiments 1 and 2, we discuss 3D CG content, regions of interest, and label prediction.

4.3. Experimental method and assessment method

As an experimental method, we conducted the following two types of experiments in this study.

- Experiment 1 (Exp. 1): Region of interest estimation and prediction using Vision Transformer in 3D CG content "Museum"
- Experiment 2 (Exp. 2): Region of interest estimation and prediction using Vision Transformer in 3D CG content "WonderWorld"

As an evaluation method, 3D CG images were input into the Vision Transformer architecture to generate color maps of regions of interest, prediction labels, and attention maps, and evaluations were performed on the differences between 3D CG content and changes in quantization parameters.

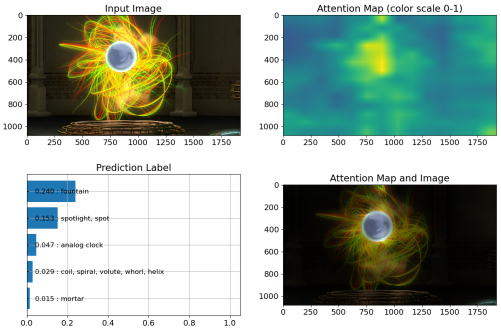

Figure 9. *Result of Exp. 1 ($Q = ref$)*

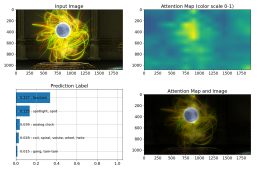

Figure 10. *Result of Exp. 1 ($Q = 20$)*

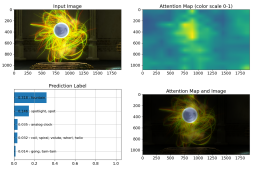

Figure 11. *Result of Exp. 1 ($Q = 25$)*

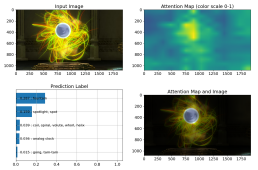

Figure 12. *Result of Exp. 1 ($Q = 30$)*

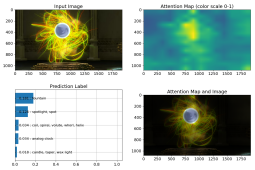

Figure 13. *Result of Exp. 1 ($Q = 35$)*

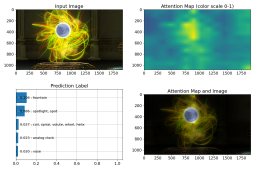

Figure 14. *Result of Exp. 1 ($Q = 40$)*

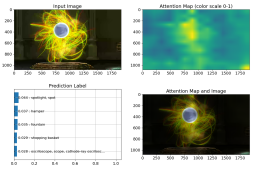

Figure 15. *Result of Exp. 1 ($Q = 51$)*

5. Experimental results and discussion

5.1. Results and discussion of Experiment 1 ("Museum")

The results of Experiment 1 are shown in Figures 9 to 15.

From the results of Experiment 1, the prediction labels "fountain" and "spotlight, spot" ranked high, and as Q increased, the accuracy decreased. This is thought to be due to the variability of HEVC and the complexity and color tones of the "Museum" content affecting the accuracy of the predicted labels.

Although the Attention Map showed no visual effects from HEVC changes, as Q increased, the depth of the color map became harder to distinguish, and the accuracy of the predicted labels also decreased. This suggests that HEVC visually affects the region of interest.

5.2. Results and discussion of Experiment 2 ("Wonder World")

On the other hand, the results of Experiment 2 are shown in Figures 16 to 22.

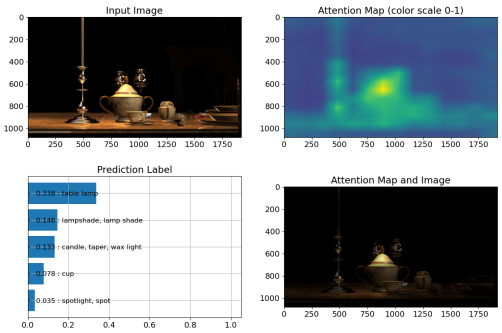

Figure 16. *Result of Exp. 2 ($Q = ref$)*

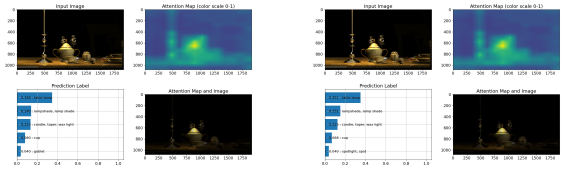

Figure 17. *Result of Exp. 2 ($Q = 20$)* ***Figure 18.*** *Result of Exp. 2 ($Q = 25$)*

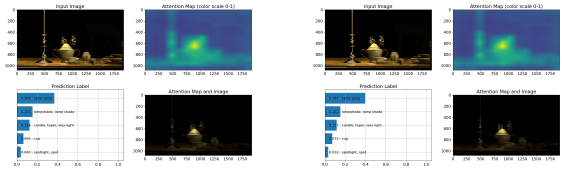

Figure 19. *Result of Exp. 2 ($Q = 30$)* ***Figure 20.*** *Result of Exp. 2 ($Q = 35$)*

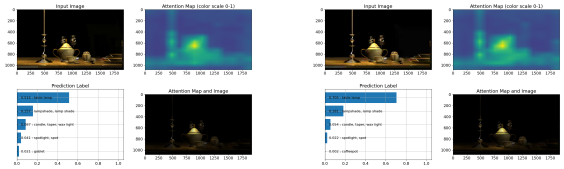

Figure 21. *Result of Exp. 2 ($Q = 40$)* ***Figure 22.*** *Result of Exp. 2 ($Q = 51$)*

The results of Experiment 2 show that the predicted labels "table lamp" and "lamp shade" ranked high, while "spotlight, spot" ranked low. Compared to Experiment 1, the hotspots for color labels were limited, and there was little change even when Q varied. Regarding the predicted labels, unlike Experiment 1, the accuracy decreased when Q was around 30, and the accuracy was high at both ends. This is thought to be due to the object area being narrower than in Experiment 1 and the depth of the color map being clear, which reduced the influence of HEVC and affected the results.

6. Conclusion

In this study, we estimated regions of interest by considering the attention mechanism using Vision Transformer on 3D CG images encoded with HEVC.

The results of this study show that using Vision Transformer enables more accurate and detailed estimation of regions of interest compared to conventional methods, and that it can understand 3D CG image patterns and HEVC encoding characteristics, as clarified by the experimental results.

Future prospects include adjustment of Vision Transformer internal parameters, improvement of AI architecture, enhancement of content patterns, application to videos, support for cross reality (XR), and application to medical images.

References

[1] F. Bossen, D. Flynn, Karsten Sühring, "HM Software Manual," `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/`, accessed September 15, 2025.

[2] N. Kawabata and M. Miyao: "3D CG Image Quality Metrics by Regions with 8 Viewpoints Parallax Barrier Method," *IEICE Trans. on Fundamentals*, Vol.E98-A, No.08, pp.1696–1708, August 2015. DOI: `10.1587/transfun.E98.A.1696`

[3] N. Kawabata and M. Miyao: "Multi-view 3D CG Image Quality Assessment for Contrast Enhancement Based on S-CIELAB Color Space," *IEICE Trans. on Information and Systems*, Vol.E100-D, No.07, pp.1448–1462, July 2017. DOI: `10.1587/transinf.2016EDP7367`

[4] N. Kawabata: "HEVC Image Quality Assessment of the Multi-view and Super-resolution Images Based on CNN," *Proc. of GCCE 2018*, POS1A-3, pp.38–39, October 2018. DOI: `10.1109/GCCE.2018.8574768`

[5] N. Kawabata and T. Nakaguchi: "Optimal Design of Color Laparoscopic Super-Resolution Image Quality Based on Generative Adversarial Networks," *Proc. of the 2023 International Conference on Computer Graphics and Image Processing* (CGIP 2023), 2023, S2-6, pp.1–7, January 2023. DOI: `10.1109/CGIP58526.2023.00009`

[6] N. Kawabata: "3D CG Image Region of Interest Estimation and Visual Attention Based on Saliency Map," Proc. of the 29th International Display Workshops (IDW'22), 29th, VHFp1-6L, pp.709–710, December 2022. DOI: `10.36463/idw.2022.0709`

[7] N. Kawabata: "ROI Estimation of Laparoscopic Image Including Contrast Enhancement Based on Saliency Map," *Proc. of the 27th International Workshop on Advanced Image Technology* (IWAIT2024), 2024, 7A-2, pp.1–6, January 2024.

[8] S. Chen et al., "Progressive Semantic-Guided Vision Transformer for Zero-Shot Learning," *Proc. of CVPR 2024*, pp.23964–23974, 2024.

[9] M. Ghahremani et al., "H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration," *Proc. of CVPR 2024*, pp.11513–11523, 2024.

[10] X. Hu et al., "Training Vision Transformer for Semi-Supervised Semantic Segmentation," *Proc. of CVPR 2024*, pp.4007–4017, 2024.

[11] S. Lee et al., "Multi-criteria Token Fusion with One-step-ahead Attention for Efficient Vision Transformers," *Proc. of CVPR 2024*, pp.15741–15750, 2024.

[12] X. Shi et al., "SpikingResformer: Bringing ResNet and Vision Transformer in Spiking Neural Networks," *Proc. of CVPR 2024*, pp.5610–5619, 2024.

[13] H. Ye et al., "Once for Both: Single Stage of Importance and Sparsity Search for Vision Transformer Compression," *Proc. of CVPR 2024*, pp.5578–5588, 2024.

[14] H. Zhang et al., "Dense Vision Transformer Compression with Few Samples," *Proc. of CVPR 2024*, pp.15825–15834, 2024.

[15] Z. Zhou et al., "Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation through Relationship Descriptors," *Proc. of CVPR 2024*, pp.3817–3827, 2024.

[16] J. You et al., "Transformer for Image Quality Assessment," *Proc. of ICIP 2021*, 2021. DOI: 10.1109/ICIP42928.2021.9506075

[17] J. Pei et al., "CalibNet: Dual-Branch Cross-Modal Calibration for RGB-D Salient Instance Segmentation," *IEEE Trans. Image Process.*, Vol.33, pp.4348–4362, 2024. DOI: 10.1109/TIP.2024.3432328

[18] J. Fu et al., "Vision-Language Consistency Guided Multi-Modal Prompt Learning for Blind AI Generated Image Quality Assessment," *IEEE Signal Process. Lett.*, Vol.31, pp.1820–1824, 2024. DOI: 10.1109/LSP.2024.3420083

[19] Q. Yu et al., "Exploring Vision Transformer for 3D Human Motion-Language Models with Motion Patches," *Proc. of CVPR 2024*, pp.937–946, 2024.

[20] C. Chen et al., "TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment," *IEEE Trans. Image Process.*, Vol.33, pp.2404–2418, 2024. DOI: 10.1109/TIP.2024.3378466

[21] Y. Huang et al., "Brighten up Images via Dual-Branch Structure-Texture Awareness Feature Interaction," *IEEE Signal Process. Lett.*, Vol.31, pp.46–50, 2024. DOI: 10.1109/LSP.2023.3340999

[22] P. C. Madhusudana et al., "Image Quality Assessment Using Contrastive Learning," *IEEE Trans. Image Process.*, Vol.31, pp.4149–4161, 2022. DOI: 10.1109/TIP.2022.3181496

[23] P. Chen et al., "SPIQ: A Self-Supervised Pre-Trained Model for Image Quality Assessment," *IEEE Signal Process. Lett.*, Vol.29, pp.513–517, 2022. DOI: 10.1109/LSP.2022.3145326

[24] W. Fan et al., "Triple-Supervised Convolutional Transformer Aggregation for Robust Monocular Endoscopic Dense Depth Estimation," *IEEE Trans. Med. Robot. Bionics*, Vol.6, No.3, pp.1017–1029, August 2024. DOI: 10.1109/TMRB.2024.3407384

[25] T. Tong et al., "Dual-Input Transformer: An End-to-End Model for Preoperative Assessment of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Ultrasonography," *IEEE J. Biomed. Health Inform.*, Vol.27, No.1, pp.251–262, January 2023. DOI: 10.1109/JBHI.2022.3216031

[26] D. Miao et al., "A Multi-Task Based Deep Learning Framework With Kandmark Detection for MRI Couinaud Segmentation," *IEEE J. Transl. Eng. Health Med.*, Vol.12, pp.697–710, 2024. DOI: 10.1109/JTEHM.2024.3491612

[27] N. Shen at al., "Expert-Guided Knowledge Distillation for Semi-Supervised Vessel Segmentation," *IEEE J. Biomed. Health Inform.*, Vol.27, No.11, pp.5542–5553, November 2023. DOI: 10.1109/JBHI.2023.3312338

[28] Y. Zhao et al., "The Contrastive Network With Convolution and Self-Attention Mechanism for Unsupervised Cell Segmentation," *IEEE J. Biomed. Health Inform.*, Vol.27, No.12, pp.5837–5847, December 2023. DOI: 10.1109/JBHI.2023.3310507

[29] H. Wang et al., "Adaptive Decomposition and Shared Weight Volumetric Transformer Blocks for Efficient Patch-Free 3D Medical Image Segmentation," *IEEE J. Biomed. Health Inform.*, Vol.27, No.10, pp.4854–4865, October 2023. DOI: 10.1109/JBHI.2023.3305830

[30] X. Li et al., "TransU^2-Net: An Effective Medical Image Segmentation Framework Based on Transformer and U^2-Net," *IEEE J. Transl. Eng. Health Med.*, Vol.11, pp.441–450, 2023. DOI: 10.1109/JTEHM.2023.3289990

[31] R. Li et al., "DHT-Net: Dynamic Hierarchical Transformer Network for Liver and Tumor Segmentation," *IEEE J. Biomed. Health Inform.*, Vol.27, No.7, pp.3443–3554, July 2023. DOI: 10.1109/JBHI.2023.3268218

[32] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proc. of ICLR 2021*, 2021.

[33] N. Carion et al., "End-to-End Object Detection with Transformers," *Proc. of ECCV 2020*, pp.213–229, 2020.

[34] NICT, http://www.nict.go.jp/, accessed September 15, 2025.

[35] Autodesk, http://www.autodesk.co.jp/, accessed September 15, 2025.