公開

研究終了報告書

「解釈可能なインタラクティブ深層学習」

研究期間：2019年10月～2022年3月
研究者：谷 林
加速フェーズ期間：2022年4月～2023年3月

1．研究のねらい

This project aims to propose a novel framework to automatically generate annotation, contextual labelling and semantically reasoning via invasively collecting user' gaze and texture input. In the research plan, there are three goals: (1) Automatically generate accurate pixel-wise label from user' gaze and texture input. (2) Interpret deep learning of its reasoning. (3) A large scale system to generate the pixel-wise label for medical image.

In the acceleration phase, this research extends it to the real 3D world based on the human gaze. Combining the human gaze would significantly enhance 3D vision tasks such as 3D point cloud segmentation.

2．研究成果

（1）概要

Thanks to the support from ACT-X, this project has resulted in the fruitful outputs. These results are from joint international collaborations across the world. In the original plan, there are three goals: (1) Automatically generate accurate pixel-wise label from user' gaze and texture input. (2) Interpret deep learning of its reasoning. (3) A large scale system to generate the pixel-wise label for medical image.
I have completed first two goals. However, the goal 3 is not completed due to the COVID-19.

In the acceleration phase, all of goals including the new goals proposed in acceleration phase are completed.

The whole project including acceleration phase comprises four major research themes: (1) Gaze integrated attention mechanism. (2) Medical vision language (VL) model. (3) Low-quality vision. (4) Human-Centric AI.

The first research theme explores integrating the human gaze into artificial intelligence (AI) to enhance performance under limited training data. Under this theme, I have developed several algorithms that mimic a human's eye-gazing pattern to guide the neuron network to focus on the task-relevant region.

The second research direction extends the VL model, a rapidly developing AI technique, to the medical domain. In 2021, our team was awarded the third prize in the medical visual question answering (VQA) challenge. I have worked with NIH to identify the limitations in existing medical VL benchmarks to collect the world's most significant, high-quality medical image difference-aware VQA benchmark and state-of-the-art solution.

During the research, I found that one of the major bottlenecks is the gap between ideal training data and actual testing data. The quality and scale differences seriously deteriorate the performance of algorithms. Therefore, I have also made several breakthroughs in low-quality vision. The representative research utilises self-supervised learning to construct a quality-invariant manifold. This latent representation significantly enhances the existing detection methods in low-quality conditions with minimal modification.

Throughout the project, I strictly obeyed the Social Principles of Human-Centric AI to ensure my research ensures global sustainability outlined in the Sustainable Development Goals (SDGs) through qualitative changes in social conditions and genuine innovation. For example, I have proposed a practical and systematic solution to protect face information while enhancing AI fairness in the full-process pipeline from camera to final users.

（2）詳細

Thanks to the support from ACT-X, this project has resulted in the fruitful outputs including 6 international journals (IEEE JHBI, IEEE TMM, PR, etc.) and 8 international conference proceeding papers (CVPR, ICCV, ECCV, BMVC, etc.). Now there are still 4 AAAI submissions under stage 2 review, 3 journals submissions under review and 6 CVPR papers to be submitted. These results are from joint international collaborations involving Yale University, Princeton University, University of Pennsylvania, National Institutes of Health (NIH), University of York (UK), University of Amsterdam, The University of Melbourne, Shanghai Jiaotong University, etc. Now, multiple academic and industrial institutes are following my proposed work and investing on collecting large scale gaze to enhance the attention mechanism of Artificial Intelligence (AI).

This project aims to propose a novel framework to automatically generate annotation, contextual labelling and semantically reasoning via invasively collecting user' gaze and texture input.

Goal 1: Research on generate accurate pixel-wise label from user' gaze and texture input.
This goal is completed. Our CVPR 2021 work achieves discriminative attribute localization guided by the actual human gaze and the attribute descriptions on general image. To further extend it to medical image, our BMVC 2021 work (1) generates the pixel-wise segmentation on both 2D and 3D medical data based on the actual human gaze. However, due to the physical contact restriction of Covid-19, I have only collected limited number of gaze data. Therefore, I have proposed a novel self-supervised method that allows the model training under limited and low quality data. This work (2) is published in a top AI conference: ICCV 2021.

Goal 2: Research on deep learning interpretability.
To achieve the goal 2, several strategies are proposed to explain the reasoning of deep learning. The representative work (3) is published in a top medical journal IEEE JHBI that explains the deep neural network from the perspective of Koch's Postulates, the foundation in evidence-based medicine. The proposed method could automatically find the location and types of symptoms that the DR detector identifies as evidence to make prediction. To incorporate more external knowledge, a knowledge graph based method has been proposed to automatically reason the semantic dependencies among objects. This method is published in ICIP and could be used to explain the high-level semantic contexts.

Goal 3: Research on a large scale system to generate the pixel-wise label for medical image.
However, the goal 3 of a system for large-scale medical data is not completed due to the COVID-19. My research is particularly disturbed by the restricted physical contact for two reasons: 1. The medical relevant data is highly sensitive that could only be assessed locally. 2. The collection of gaze requires recording the doctors' action with a special device, gaze tracker.

In acceleration phase, this research has yielded additional 7 international journals and 10 international conference proceedings papers (AAAI, CVPR, ECCV, BMVC, etc.). These results are from joint international collaborations involving Yale University, Princeton University, the University of Pennsylvania, the National Institutes of Health (NIH), the University of York (UK), the University of Amsterdam, The University of Melbourne, Shanghai Jiaotong University, etc.

Research Theme A [Foundation of Gaze Analysis] [Acceleration phase]
The project aims to integrate the human's gaze and language to automatically generate pixel-level labels while endowing AI the capacity to explain how it makes decisions semantically. Therefore, gaze analysis is the foundation of the whole project. I have devoted myself to exploring how to integrate the human gaze into artificial intelligence (AI) to enhance performance under limited training data.

In 2022 European Conference on Computer Vision (ECCV) work, we further advance this field by collecting the world's largest human real-time gaze data for fine-grained classification on different hierarchies. This dataset allows the community to learn different granularity-wise attention regions with multi-grained classification tasks. Under this research theme, I have effectively explored and promoted the field of gaze analysis which receives increasing attention in the artificial intelligence community.

Research Theme B [Medical Vision Language Field] [Acceleration phase]
To fully utilise the text input of doctors, I have made several attempts to apply the most advanced language vision model for medical applications.

In 2021, I led a team to win the third prize for the world's highest medical VQA challenge. During the research, I find that the existing benchmark is limited in the data size and label quality. For example, prior VQA datasets in medical images have significant errors in their labels due to issues with their text mining strategies (for example, issues with rule-based systems). Moreover, existing models tend to adopt the learned data distributions of the dataset (e.g., the connections between the question type and its corresponding frequent answers) to make predictions instead of understanding the medical input and medical knowledge.

Therefore, I am working with NIH to collect and release the world's largest Chest-Xray Different VQA dataset. Unlike the standard medical VQA dataset that directly follows the common VQA task that answers questions on a single image, the proposed dataset is consistent with the radiologist's diagnosis practice that compares the current image with the reference before concluding the report. We have also proposed a novel expert knowledge-aware graph representation learning model to leverage expert knowledge such as anatomical structure prior, semantic and spatial understanding. Thanks to the support of ACT-X, I can fill a well-known gap in the medical VL research community.

Research Theme C [Low-quality vision] [Acceleration phase]
During the research, I found that one of the major bottlenecks in artificial medical intelligence is the gap between ideal training data and actual testing data. The quality and scale differences seriously deteriorate the performance of algorithms. Therefore, I have also made several breakthroughs in addressing this gap. At International Conference on Computer Vision (ICCV) 2021, I proposed a novel self-supervised method that allows model training under limited and low-quality data. In 2022, I continue tackling the challenging scale gap problem. For the first time in the world, I propose to use the resolution as a self-supervised signal[2]. The novel self-supervised framework can detect objects in degraded low-resolution images. Specifically, the downsampling degradation is used to transform self-supervised signals to explore the equivariant representation against various resolutions and other degradation conditions. The proposed generic framework could be implemented on various mainstream AI architectures.

I have also contributed to several works that enhance the low-quality image. One representative work addresses challenging illumination conditions (low light, under-exposure and over-exposure) in the real world. In this fiscal year, a lightweight fast Illumination Adaptive Transformer (IAT) is proposed to restore the standard lit sRGB image from either low-light or under/overexposure conditions. Specifically, IAT uses attention queries to represent and adjust the image processing pipeline (ISP) related parameters such as colour correction and gamma correction. With only ~90k parameters and ~0.004s processing speed, our IAT consistently achieves superior performance over State-of-The-Art (SOTA) on the benchmark low-light enhancement and exposure correction datasets. The extremely lightweight and fast speed makes the proposed algorithm suitable for cheap embedding devices, thus allowing the universal application, especially in resource-poor countries. Our method would, therefore "leave no one behind", as called by SDGs.

Research Theme D [Human-Centric AI] [Acceleration phase]
When applying AI for actual social implementation, it is essential to use it for the public good of humanity and to ensure global sustainability outlined in the SDGs through qualitative changes in social conditions and true innovation.

Privacy protection and fairness are essential in human-centric AI. I propose a practical and systematic solution[3] to protect face information and ensure fairness in the full-process pipeline from camera to final users. Specifically, a novel lightweight Flow-based Face Encryption Method (FFEM) is implemented on the local embedded system privately connected to the camera, minimising the risk of eavesdropping during data transmission. FFEM uses a flow-based face encoder to encode each face to a Gaussian distribution and encrypts the encoded face feature by randomly rotating the Gaussian distribution with the rotation matrix as the password. While encrypted latent-variable face images are sent to users through public but less reliable channels, the password will be protected through more secure channels through asymmetric encryption, blockchain, or other sophisticated security schemes. The user could select to decode an image with fake faces from the encrypted image on the public channel. Only trusted users can recover the original face using the encrypted matrix transmitted in a secure channel. More interestingly, by tuning the Gaussian ball in latent space, the fairness of the replaced face on attributes such as gender and race could be controlled. This work is accepted in AAAI Conference on Artificial Intelligence (AAAI) 2023.

３． 今後の展開

Now this research has attracted global attention from both academic and industrial fields. As far as I know, several groups across the world are starting collecting and utlising the gaze data following my proposed methodology.

The proposed system has potential to become the standard system for clinics to conduct routine medical imaging examine. This would also be used to for the annotation of general 2D image and even 3D data. With this system, large scale data with labelling could be collected at almost zero cost.

Thanks to the support in acceleration phase, I have following research plan.

Future Direction A [Social Implementation of Gaze-Assisted Annotation System]
The system is developed for 2D and 3D medical images in this project. I will extend this system to ultrasound videos. I will promote the social implementation of the current system as the standard system for clinics to conduct routine medical imaging examinations. It would allow the community to collect large-scale data with labelling at almost zero cost. The collected labels would significantly improve medical imaging and general AI research hungry for data.

Future Direction B [Gaze Analysis for Mental Disorder Diagnose]
Based on my research achievements on gaze analysis supported by ACT-X, I have been granted RIKEN-MOST Collaborative Research Project for using human gaze and other multi-modal signals to subtype schizophrenia. I will fully utilise the gaze-relevant techniques developed in this project to diagnose and assess mental disorders such as schizophrenia, depression, and Alzheimer's disease (AD). This research will promote Goal 3 of SDGs for Good health and well-being.

Future Direction C [Computational Terahertz Imaging Device]
Terahertz imaging is an emerging and significant nondestructive evaluation technique for various applications. Various materials are transparent to terahertz radiation, allowing the measurement of the thickness, density, and structural properties of materials that are otherwise difficult to detect. Since terahertz is not ionising radiation, terahertz imaging does not cause damage to living tissue, making it a safe, non-invasive biomedical imaging technique. I will utilise the low-quality vision techniques supported by ACT-X to develop a new generation of the computational terahertz imaging device. This imaging has the potential for fundamental scientific research and applications in various fields, such as biomedical, communication, and security inspection.

4．自己評価

Thanks to the support from ACT-X, I have the unique opportunity to collaborate with global researchers on the breakthrough research on utlising human gaze to enhance the artificial intelligence. During the research, I have published several papers on top AI conferences and journals.

In the original plan, I have proposed three goals. Though the first two goals are completed, the third goal is postponed due to Covid-19. My research is particularly disturbed by the restricted physical contact for two reasons: 1. The medical relevant data is highly sensitive that could only be assessed locally. 2. The collection of gaze requires recording the doctors' action with a special device, gaze tracker. As the result, there is 1,000,000 yen for travelling and the personnel expenses that I could not spend.

Apart from the grant itself, the support from ACT-X advisors and other peer ACT-X researchers is equally important for my research. Without these advice and support during 2 years, it is impossible to achieve the existing progress. ACT-X also provides a valuable chance to communicate and collaborate with other peer ACT-X researchers.

The proposed research supported by ACT-X has attracted global interesting. Now, multiple academic and industrial institutes are following this work and investing on collecting large scale gaze to enhance the attention mechanism of AI. It would a keystone technique to relieve the data-hungry bottleneck of AI by providing large-scale pixel-wise label at almost zero-cost.

My evaluation after acceleration phase.

Thanks to the support from ACT-X, I have the unique opportunity to collaborate with global researchers on breakthrough research on utilising the human gaze to enhance AI. Under this project, I can propose the world's first system to integrate human gaze and text input to generate pixel-level label automatically. After 2019 when this project started, human gaze research attracted global attention, with more and more academic and industrial institutes following this work and investing in collecting large-scale gaze to enhance the attention mechanism of AI.

This endeavour would hot have been possible without the financial support of ACT-X. The project allows me to purchase special devices like a gaze tracker to undertake this research. Besides, ACT-X also allows me to attend conferences and discuss with leading experts in different fields. Remarkably, the supported visit to Oxford University helped me pave the way for commercial implementation of the proposed porotype and developing a new generation imaging device.

Apart from essential financial support, the patient guidance, encouragement and advice from my ACT-X advisor, Prof Uchida Seiichi, who cared so much about my work, and who responded to my questions and queries so promptly. The immense knowledge, plentiful experience and magical humour have been invaluable for me to grow my academic abilities and catapult my progress.

Throughout seven times field meetings, I have received warm and detailed feedback from Program Officer Prof. Kawarabayashi Ken-ichi, fields advisors. The insights and suggestions were constructive and improved my work significantly. I also benefit much from ACT-X colleagues from different fields. Their insights on how to apply AI in multiple domains were genuinely fascinating. From the presentation and discussion, I can figure out how to extend the research in the current projects to other fields, such as designing terahertz imaging devices.

Apart from financial and academic support, I also attribute the achievements of this project to the administration support from Japan Science and Technology Agency (JST). Only in the hospitable environment that JST staff have helped create for foreign researchers like myself, I could finally navigate the difficult situation caused by Covid and achieve the research goal in ACT-X.

With ACT-X's nearly four years of support, I've acquired and practised many new skills, especially mentoring and supervising students. During the research in this project, I have trained several students, including the undergraduate ones, to conduct high-quality research and publish their first top conference or journal paper. Some PhD students who participated in the project have graduated and enrolled in faculty positions at leading universities like Oxford University, The University of British Columbia, etc.

There is still plenty of achievable lined up, and I will try my best to keep up with excellence.

５． 主な研究成果リスト

（１）代表的な論文（原著論文）発表

研究期間累積件数：3件

研究期間累積件数：6件（加速フェーズ実施後更新）

Yifei Huang, Xiaoxiao Li, Lijin Yang, **Lin Gu***, Yingying Zhu, Hirofumi Seo, Qiuming Meng, Tatsuya Harada, Yoichi Sato. Leveraging Human Selective Attention for Medical Image Analysis with Limited Training Data. Proceeding of The British Machine Vision Conference (BMVC). 2021. *Lin Gu is the corresponding author.

Abstract: Human gaze is a cost-efficient physiological data that reveals human underlying attentional patterns. The selective attention mechanism helps the cognition system focus on task-relevant visual clues by ignoring the presence of distractors. Thanks to this ability, human beings can efficiently learn from a very limited number of training samples. Inspired by this mechanism, we aim to leverage gaze for medical image analysis tasks with small training data. Our proposed framework includes a backbone encoder and a Selective Attention Network (SAN) that simulates the underlying attention. The SAN implicitly encodes information such as suspicious regions that is relevant to the medical diagnose tasks by estimating the actual human gaze. Then we design a novel Auxiliary Attention Block (AAB) to allow information from SAN to be utilized by the backbone encoder to focus on selective areas. Specifically, this block uses a modified version of a multi-head attention layer to simulate the human visual search procedure. Note that the SAN and AAB can be plugged into different backbones, and the framework can be used for multiple medical image analysis tasks when equipped with task-specific heads. Our method is demonstrated to achieve superior performance on both 3D tumor segmentation and 2D chest X-ray classification tasks. We also show that the estimated gaze probability map of the SAN is consistent with an actual gaze fixation map obtained by board-certified doctors.

2. Multitask AET with Orthogonal Tangent Regularity for Dark Object Detection. Ziteng Cui, Guo-Jun Qi, **Lin Gu***, Shaodi You, Zenghui Zhang, Tatsuya Harada. Proceeding of International Conference on Computer Vision (ICCV). 2021. *Lin Gu is the corresponding author

Abstract: Dark environment becomes a challenge for computer vision algorithms owing to insufficient photons and undesirable noises. Most of the existing studies tackle this by either targeting human vision for better visual perception or improving the machine vision for specific high-level tasks. In addition, these methods rely on data argumentation and directly train their models based on real-world or over-simplified synthetic datasets without exploring the intrinsic pattern behind illumination translation. Here, we propose a novel multitask auto encoding transformation (MAET) model that combines human vision and machine vision tasks to enhance object detection in a dark environment. With a self-supervision learning, the MAET learns an intrinsic visual structure by encoding and decoding the realistic illumination-degrading transformation considering the physical noise model and image signal processing (ISP). Based on this representation, we achieve object detection task by decoding the bounding box coordinates and classes. To avoid the over-entanglement of two tasks, our MAET disentangles the object and degrading features by imposing an orthogonal tangent regularity. This forms a parametric manifold along which multitask predictions can be geometrically formulated by maximizing the orthogonality between the tangents along the outputs of respective tasks. Our framework can be implemented based on the mainstream object detection architecture and directly trained end-to-end using the normal target detection datasets, such as COCO and VOC. We have achieved the state-of-the-art performance using synthetic and real-world datasets.

3. Yuhao Niu, **Lin Gu**, Yitian Zhao, Feng Lu. Explainable Diabetic Retinopathy Detection and Retinal Image Generation. IEEE Journal of Biomedical and Health Informatics. 2021.

Though deep learning has shown successful performance in classifying the label and severity stage of certain diseases, most of them give few explanations on how to make predictions. Inspired by Koch's Postulates, the foundation in evidence-based medicine (EBM) to identify the pathogen, we propose to exploit the interpretability of deep learning application in medical diagnosis. By isolating neuron activation patterns from a diabetic retinopathy (DR) detector and visualizing them, we can determine the symptoms that the DR detector identifies as evidence to make prediction. To be specific, we first define novel pathological descriptors using activated neurons of the DR detector to encode both spatial and appearance information of lesions. Then, to visualize the symptom encoded in the descriptor, we propose Patho-GAN, a new network to synthesize medically plausible retinal images. By manipulating these descriptors, we could even arbitrarily control the position, quantity, and categories of generated lesions. We also show that our synthesized images carry the symptoms directly related to diabetic retinopathy diagnosis. Our generated images are both qualitatively and quantitatively superior to the ones by previous methods. Besides, compared to existing methods that take hours to generate an image, **our** second level speed endows the potential to be an effective solution for data augmentation.

4. Zhenqiang Li, **Lin Gu**, Weimin Wang, Ryosuke Nakamura, Yoichi Sato. Surgical Skill Assessment via Video Semantic Aggregation. Proceeding of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

Abstract: Automated video-based assessment of surgical skills is a promising task in assisting young surgical trainees, especially in poor-resource areas. Existing works often resort to a CNN-LSTM joint framework that models long-term relationships by LSTMs on spatially pooled short-term CNN features. However, this practice would inevitably neglect the difference among semantic concepts such as tools, tissues, and background in the spatial dimension, impeding the subsequent temporal relationship modeling. In this paper, we propose a novel skill assessment framework, Video Semantic Aggregation (ViSA), which discovers different semantic parts and aggregates them across spatiotemporal dimensions. The explicit discovery of semantic parts provides an explanatory visualization that helps understand the neural network's decisions. It also enables us to further incorporate auxiliary information such as the kinematic data to improve representation learning and performance. The experiments on two datasets show the competitiveness of ViSA compared to state-of-the-art methods.

5. Ziteng Cui, Yingying Zhu, **Lin Gu***, Guo-Jun Qi, Xiaoxiao Li, Renrui Zhang, Zenghui Zhang, Tatsuya Harada. Proceeding of Exploring Resolution and Degradation Clues as Self-supervised Signal for Low Quality Object Detection. Proceeding of European Conference on Computer Vision (ECCV) 2022 *Lin Gu is the corresponding author.

Abstract: Image restoration algorithms such as super resolution (SR) are indispensable pre-processing modules for object detection in low quality images. Most of these algorithms assume the degradation is fixed and known a priori. However, in practical, either the real degradation or optimal up-sampling ratio rate is unknown or differs from assumption, leading to a deteriorating performance for both the pre-processing module and the consequent high-level task such as object detection. Here, we propose a novel self-supervised framework to detect objects in degraded low resolution images. We utilises the downsampling degradation as a kind of transformation for self-supervised signals to explore the equivariant representation against various resolutions and other degradation conditions. The Auto Encoding Resolution in Self-supervision (AERIS) framework could further take the advantage of advanced SR architectures with an arbitrary resolution restoring decoder to reconstruct the original correspondence from the degraded input image. Both the representation learning and object detection are optimized jointly in an end-to-end training fashion. The generic AERIS framework could be implemented on various mainstream object detection architectures with different backbones. The extensive experiments show that our methods has achieved superior performance compared with existing methods when facing variant degradation situations

6. Junjie Zhu, **Lin Gu**, Xiaoxiao Wu, zheng li, Tatsuya Harada, Yingying Zhu. People taking photos that faces never share: Privacy Protection and Fairness Enhancement

from Camera to User. Proceeding of AAAI Conference on Artificial Intelligence (AAAI) 2023

The soaring number of personal mobile devices and public cameras poses a threat to fundamental human rights and ethical principles. For example, the stolen of private information such as face image by malicious third parties will lead to catastrophic consequences. By manipulating appearance of face in the image, most of existing protection algorithms are effective but irreversible. Here, we propose a practical and systematic solution to invertibly protect face information in the full-process pipeline from camera to final users. Specifically, We design a novel lightweight Flow-based Face Encryption Method (FFEM) on the local embedded system privately connected to the camera, minimizing the risk of eavesdropping during data transmission. FFEM uses a flow-based face encoder to encode each face to a Gaussian distribution and encrypts the encoded face feature by random rotating the Gaussian distribution with the rotation matrix is as the password. While encrypted latent-variable face images are sent to users through public but less reliable channels, password will be protected through more secure channels through technologies such as asymmetric encryption, blockchain, or other sophisticated security schemes. User could select to decode an image with fake faces from the encrypted image on the public channel. Only trusted users are able to recover the original face using the encrypted matrix transmitted in secure channel. More interestingly, by tuning Gaussian ball in latent space, we could control the fairness of the replaced face on attributes such as gender and race. Extensive experiments demonstrate that our solution could protect privacy and enhance fairness with minimal effect on high-level downstream task.

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

Cardiology and Cardiac Surgery chapter in the clinical textbook Artificial Intelligence in Clinical Medicine