February 2024

Computer Vision News

& Medical Imaging News

The Magazine of the Algorithm Community

BEST OF WACV 2024

30 Pages !!!

A publication by RSIP VISION

Improving the Effectiveness of Deep Generative Data

Ruyu Wang has just finished her PhD with the Bosch Center for Artificial Intelligence, where she is now starting a full-time position.

Her paper exploring the use of generative models to create synthetic defective data was presented at WACV last month. Fresh from the event, Ruyu is here to tell us more about her work.

Deep generative models continue to blur the boundaries between real and artificially generated content. **Through her work at Bosch**, Ruyu has identified a unique industrial application for these models to improve the overall efficiency of defect detection on production lines. However, a problem lies in the **lack of diverse defect images for training, meaning these models could easily be prone to errors**. Real-world production lines,

optimized against producing defects, pose a challenge for collecting sufficient data, as waiting for defects to occur naturally could be a perpetual task.

"*The way I'm tackling this problem is by using a generative model to generate defective data for training – it's synthetic data for data augmentation,*" Ruyu tells us, having previously published another paper in this domain. "*However, there is a gap between synthetic and real data. Models trained on synthetic data perform worse than those trained on real data.*"

This domain gap was first noticed in early works when **synthetic data from GAN engines simulating scenes and products** was used to train detection

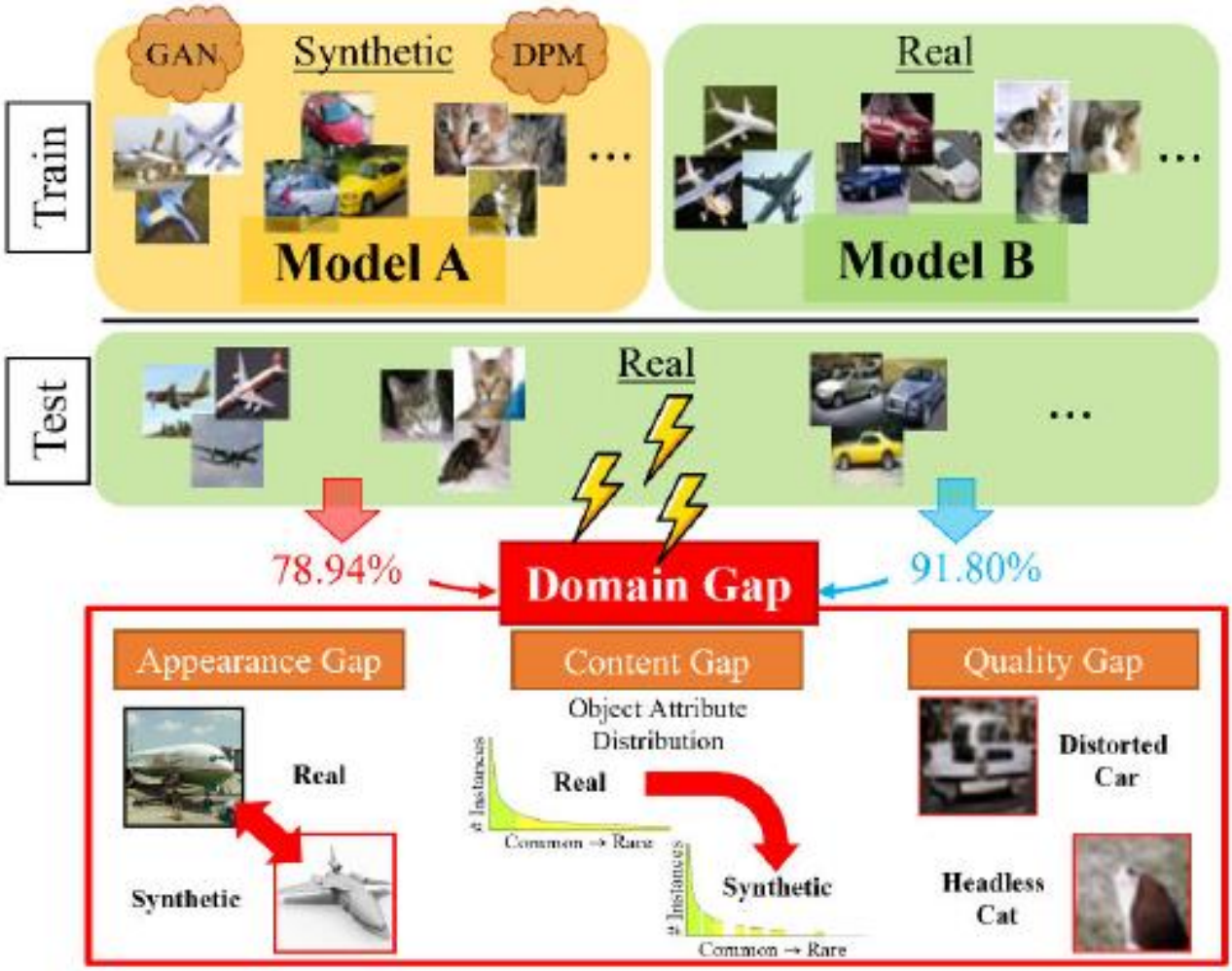


Figure 1. An illustration of the commonly observed performance degradation when training a model on synthetic data and testing it on real data. Such effects can be attributed to the *Domain Gap*, which we split into three aspects (appearance, content, quality).

and classification models. These models struggled because simulations oversimplified real-world cases with unrealistic textures and lighting conditions. Ruyu calls this the **Appearance Gap**. Modern deep generative models are optimized for realism, so it should not be an issue. However, there is still a noticeable performance drop when training.

She subsequently identified two further gaps: the **Content Gap** and the **Quality Gap**. The Content Gap is a gap in the attributes of the generated data compared to the real-world dataset. "*The most obvious gap is at the class level,*" Ruyu points out. "*Your model fails to generate a class of data in your dataset. Nowadays, powerful generative models don't make these kinds of stupid mistakes, but they can make a second mistake, which is more likely to be overlooked. You don't have a missing class, but for object-centered images,* ***some specific attributes might be missing because some combinations are rare***."

For example, a typical cat image will be classified as a cat, but if the cat is wearing a Christmas costume, it might be dropped or barely generated by the model, as it is much rarer. Even if the generative model learns it, straightforward sampling may struggle to capture it, leading to **a distribution mismatch between the sampled and training datasets**.

Lastly, the Quality Gap relates to anomalies in the deep generative data – for instance, **semantic artifacts like a two-headed cat**.

In this paper, Ruyu conducts several investigations to pinpoint the reasons behind the domain gap between synthetic-to-real data. Her first experiment

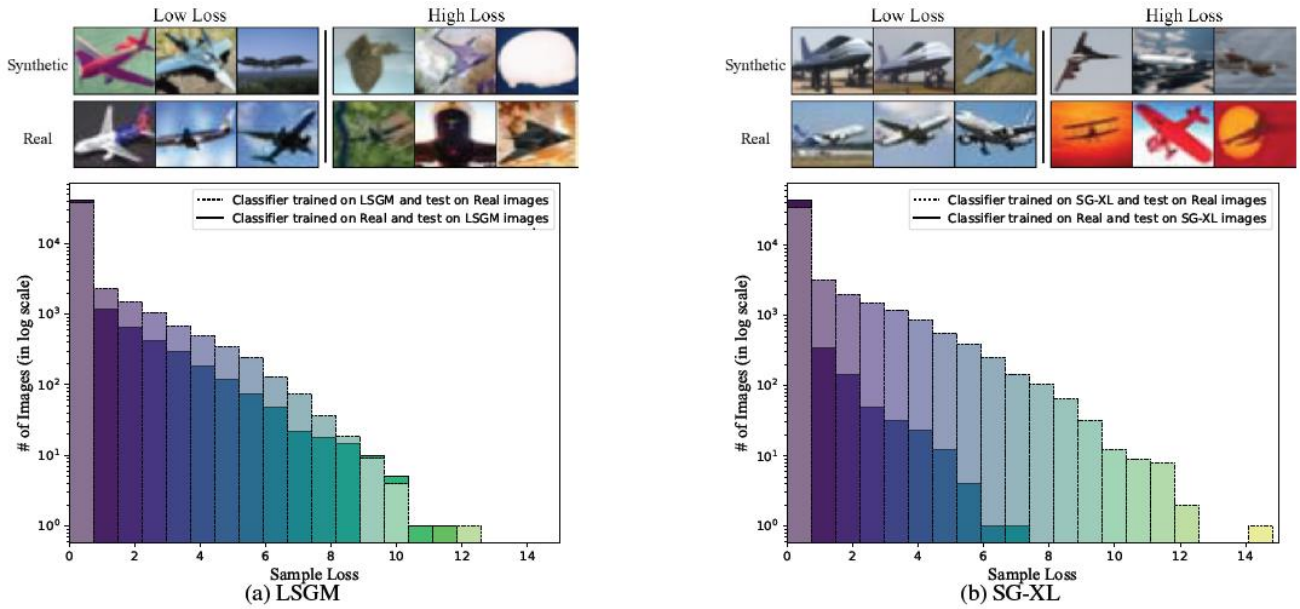


Figure 3. The loss distribution of the samples from different sources. We overlapped the evaluation of *Real* → *Synthetic* (Solid) and *Synthetic* → *Real* (Dotted) in each subgraph and visualized the associated low-loss and high-loss samples of class Airplane. Note that all models used for evaluation were initialized with ImageNet weights.

was to train a classifier on each dataset, including a real CIFAR-10 dataset and synthetic CIFAR-10 datasets from a diffusion-based and GAN-based model. She uses these classifiers to classify a real test set. *"If you train a classifier on real data and then test it on the real dataset, the performance is on par with your observation during training, so it's around 90% accurate,"* she confirms. *"But if you train on synthetic data, the performance drops. You test on the real dataset and get a performance of 85% or even 70%. This is a known issue. I said, let's use the classifier trained on the real dataset to classify the synthetic test set, and do you know what?* ***The real classifier can achieve more than 90% accuracy on the synthetic dataset!*** *The domain gap is not mutual. If you train on real data, then your classifier is super strong. The synthetic data seems to have some problem."*

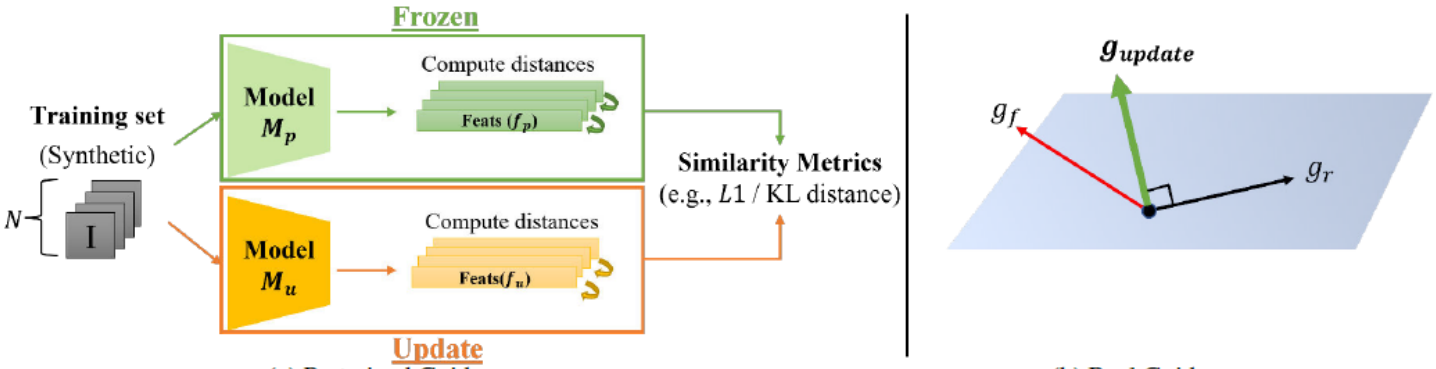
Ruyu's second experiment showed training on the synthetic dataset demonstrated **rapid accuracy convergence, reaching 99% in just a few epochs**, contrasting with slower progress observed on the real dataset. The classifier trained on synthetic data was solving the same task, yet it somehow picked up a signal to classify the images more easily.

Taking the hypothesis that the synthetic dataset was simpler than the real dataset, Ruyu's third experiment focused on **assessing the information content of the synthetic samples**. *"I trained a classifier on the real dataset and used it to examine the cross-entropy loss each sample brings to the classifier,"* she explains. *"The assumption is that if a synthetic sample contains some new information or it's different from what's already contained in the real dataset, it should stimulate a high loss for the classifier to update toward this direction."*

The majority of the 45,000 synthetic samples exhibited negligible loss, indicating a lack of substantial new information or divergence from the real dataset. The performance drop observed when training on synthetic data likely stems from it being a limited subset of the real dataset, missing crucial rare cases and unique attributes. Therefore, the classifier trained on top of it lacks the robustness to be effective when applied to a real-world dataset. *"If a generative model trained on a relatively large dataset, like a CIFAR-10 dataset, is having this problem,* ***what will happen if we train our data on limited defective samples?***" she poses. *"The problem will be more severe, our data diversity will be less, and data quality will probably drop to some extent."*

Despite the challenges, Ruyu is optimistic about the potential of integrating not-so-perfect synthetic data with real data to train a defect classifier, as even a modest improvement is valuable considering the scarcity of available

data for training. There may also be ways to improve its effectiveness. "*We propose a way to make the synthetic data we have at hand more effective in training the classifier by* ***introducing some prior information from pre-trained models****,*" she adds. "*Then, by using it to regularize the network, it improves the performance of the synthetic data without changing the generative model itself.*"

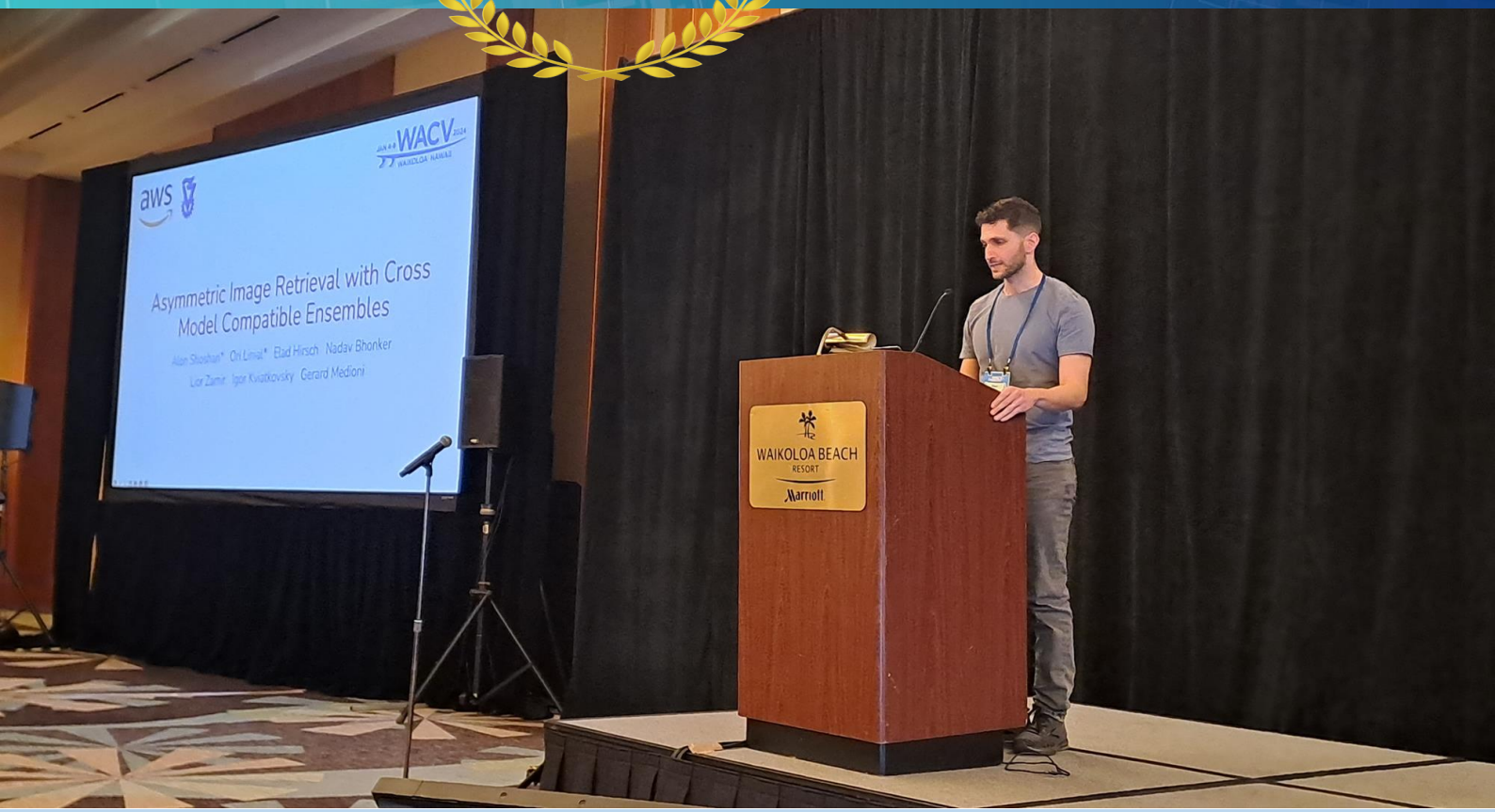


(a) Pretrained Guidance (b) Real Guidance

Figure 4. The proposed **Pretrained Guidance** and **Real Guidance**.

Reflecting on her journey, Ruyu tells us the initial spark that ignited her interest in computer science was her childhood love for computer games. However, realizing that game design was a lot about physics and mathematics led her to explore other directions. It was during her master's thesis in a lab focused on traffic-related applications that she first encountered anomaly detection, particularly in identifying irregularities on road surfaces. "*At that time, I really thought I was crazy, and my professor was crazy because he gave me the money to buy a drone!*" she laughs. "*I flew it to capture an aerial traffic video to compute traffic from the video. That's where I started to touch this deep learning and computer vision thing. It all started with* ***Faster R-CNN in 2015****. I used it to do the car detection. Then I used a* ***Kalman filter-based algorithm*** *to trace the cars and do the counting.*"

Ruyu came to Germany for her exchange semester, where, compared to Taiwan, which is more focused on semiconductor production, she recognized the potential for a future in computer vision and autonomous driving or different kinds of production. Wanting to stay, she sought opportunities for further exploration, leading to an internship at Bosch. "*That's where I got into generative models because my supervisor's idea at that time was to use* ***CycleGAN*** *to transfer an okay product to a defective product,*" she recalls. "*I just thought, these generative model things look pretty cool! My supervisor also said, 'Don't you think working on images is way cooler than other things because you see the images you produce?' I agree. It's very interesting. That's why I've kept doing it until now!*"



Elad Hirsch (top), presenting his oral work "Asymmetric Image Retrieval With Cross Model Compatible Ensembles".

Jie Zhang (bottom), presenting his oral work "Contextual Affinity Distillation for Image Anomaly Detection".

"I don't think we have solved Computer Vision..."

"I am not an ivory tower professor, where I invent a problem, solve it, and ask if anybody cares..."

"These people [at Amazon] are even crazier than I thought: they are willing to give me the keys to the kingdom..."

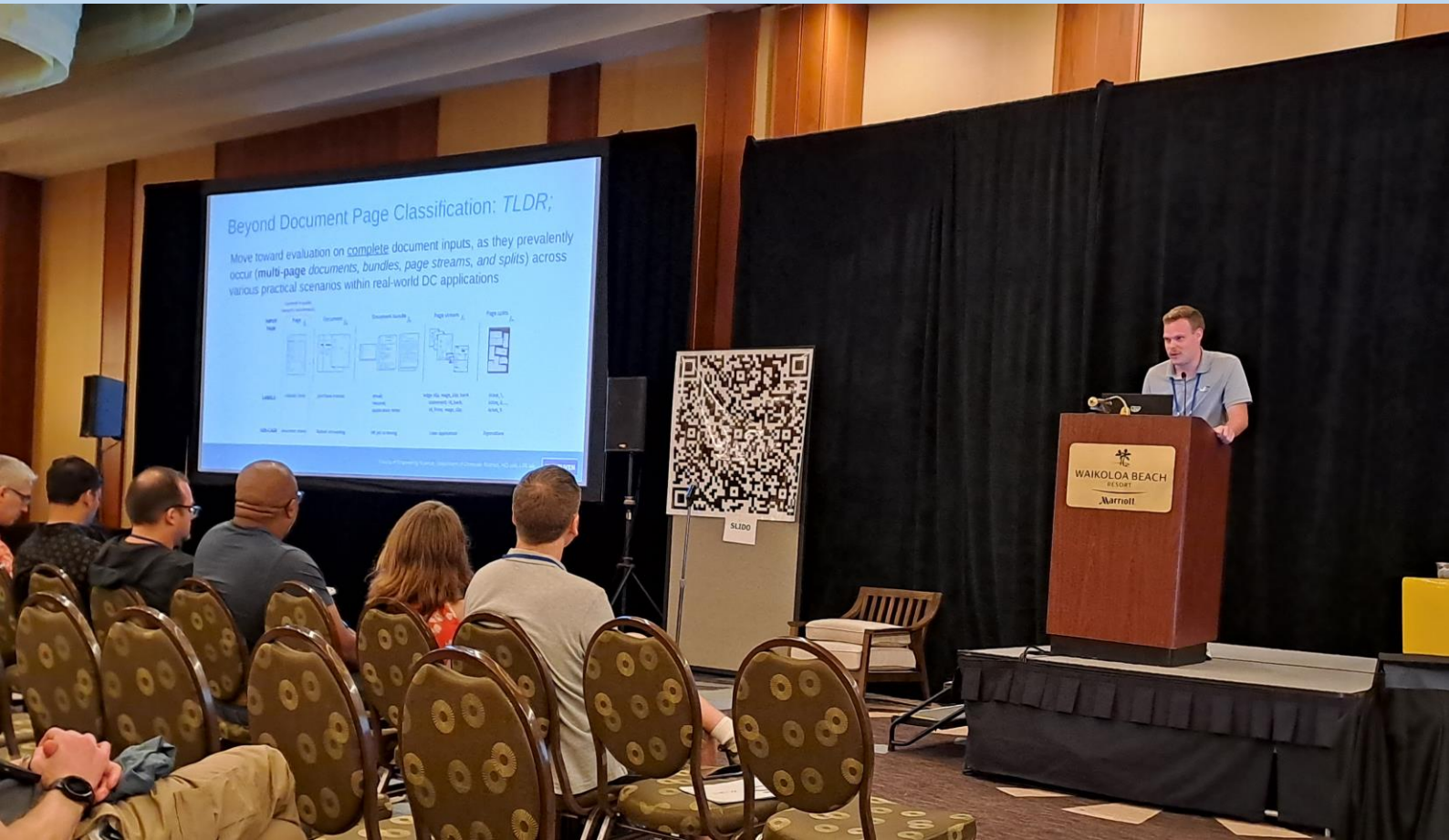
"I think that if you try to compete with labs that have tens or hundreds of people, working on [...] creating the next LLM that's going to require billions or trillions of data, is not a good idea for an academic..."

"As a grad student, it is always good to work at two things at a time. Because when you hit a wall, the wall is not gonna go away. But if you go away from the wall, for a little bit, you can get around it..."

Fireside chat with two giants, Gérard Médioni and Michael Black. Perhaps the most fascinating talk at WACV2024...

Paul Grimal (top), presenting his oral work “A Metric for Evaluating Alignment in Text-to-Image Generation”.

Jordy Van Landeghem (bottom), presenting his oral work “Beyond Document Page Classification: Design, Datasets, and Challenges”.



Controlling Rate, Distortion, and Realism: Towards a Single Comprehensive Neural Image Compression Model

Shoma Iwai is a second-year PhD student at Tohoku University in Miyagi Prefecture, Japan.

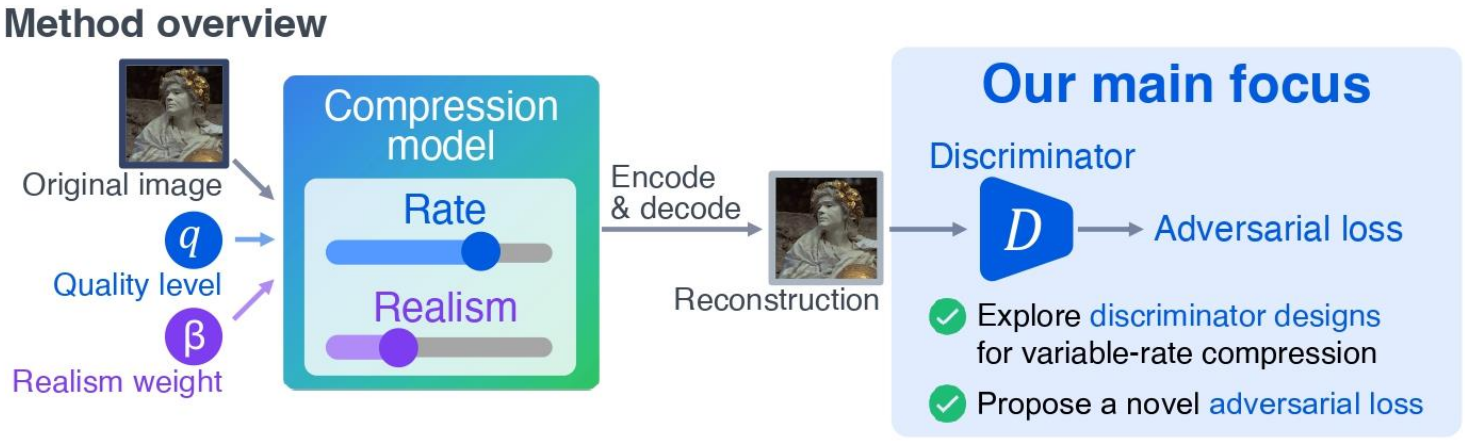
His paper proposes a single solution to two common challenges in image compression with deep learning. He spoke to us ahead of his oral presentation at AWCV 2024

In recent years, **image compression** has seen significant advancements, with neural network-driven techniques gaining more attention. Several works have employed deep generative methods like GANs and diffusion models to improve perceptual quality or realism. However, **optimizing models for different bit rates** remains a key challenge.

"*In image compression with deep learning, most models are optimized for* ***a single target bit rate****,*" Shoma begins. "*In other words, we need to train multiple models to compress images into different bit rates. Enhancing the perceptual quality of compressed images is another problem, especially when we compress images to a very small data size. In that case,* ***a lot of information is lost****.*"

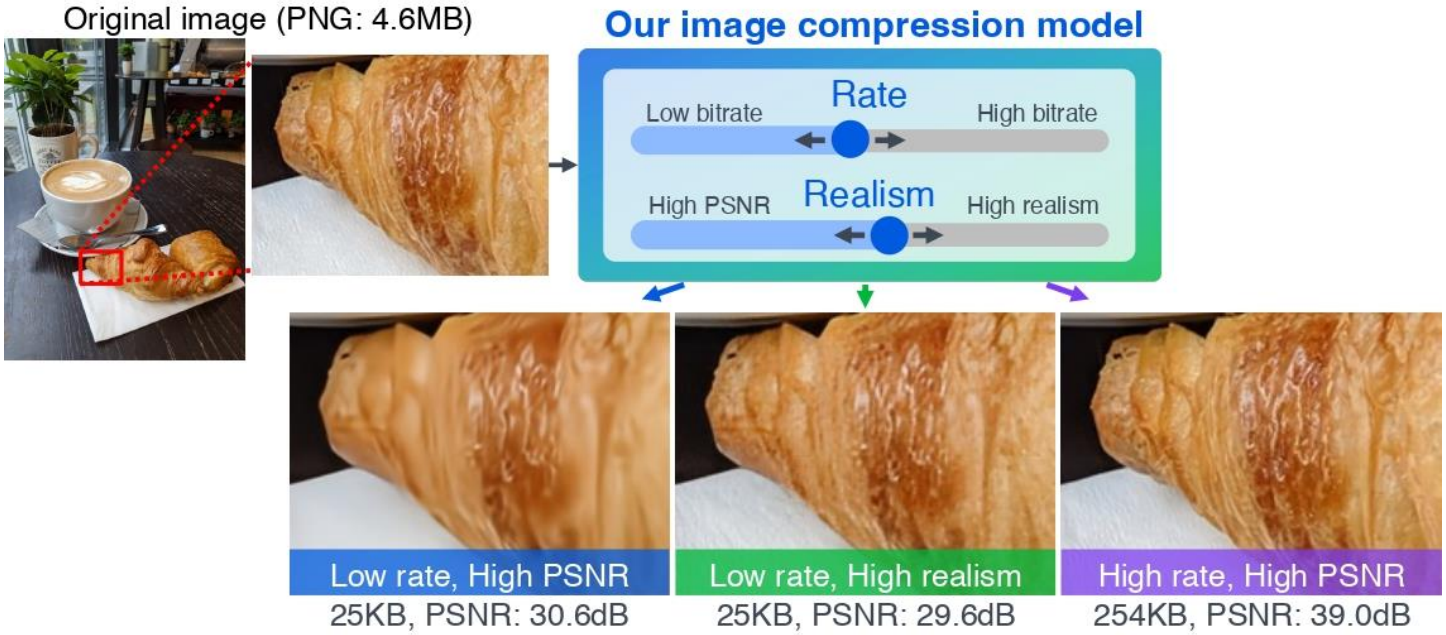
Although there are existing methods to tackle these issues individually, very few studies address both, which was the motivation behind this work. Its proposed variable-rate GAN-based approach places a key emphasis on the discriminator's role in training. Shoma explains he experimented with various discriminator designs to identify the one most suitable for the task and, additionally, introduced **a novel adversarial loss function**.

"*We show that these two methods improve performance and bridge the gap between the state-of-the-art and this high-controllability*



method," he tells us. *"Our performance matches the current state-of-the-art method proposed in **CVPR** last year. At the same time, our method can control the bit rate with a single model. It's a very good result."*

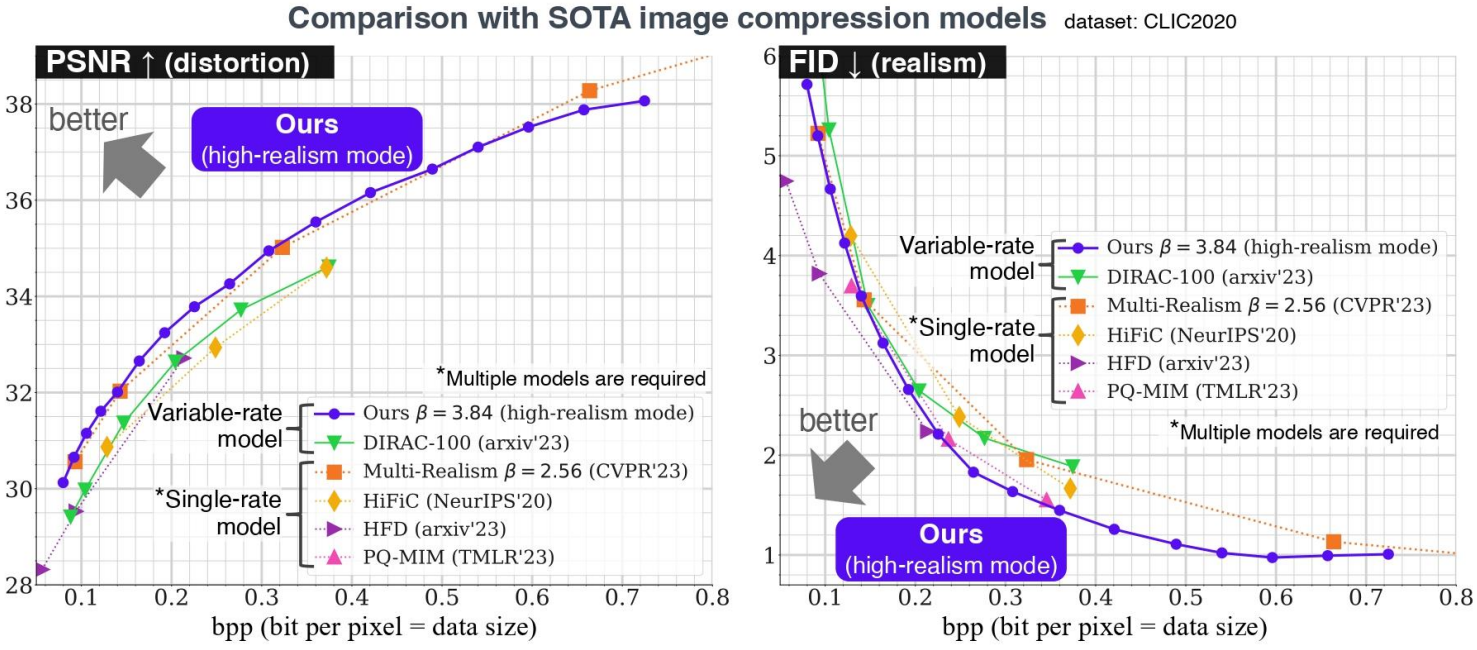
Thinking about next steps, Shoma would like to add even more controllability to the model by incorporating **Region of Interest (ROI) coding**. This feature enables pixel-level compression control, allowing users to prioritize the quality of specific regions of the image, a crucial advancement for practical applications. *"For example, if there are three people in a picture, in most cases, the quality of the people is important, but the quality of the background isn't as much,"* he says. *"We can maintain the high quality with the three people's regions and reduce the data size of the background."*



Shoma reveals his interest in image compression stemmed from an undergraduate class on image processing, where the details of JPEG compression were explained. JPEG is the most common image compression technique, with quality parameters to control the compressed data size. The statistical analysis and engineering work behind it and the implementation of standard techniques and widespread application fascinated him. So much so that he chose image compression as his research topic when he started work as a graduate student.

How does he feel now that he is about to present his work at this prestigious conference? No mean feat when you consider that not only was his paper chosen from thousands to be part of the event, but it was also picked to be one of a limited number of orals on the agenda.

"That's a good question, and frankly, I'm so surprised!" he laughs. *"Obviously, I was very happy when I heard it was accepted, and now it's chosen as an oral presentation. If we think about the applicability, this conference is about applications of computer vision, and our method can be applied to various use cases. Of course, there are a lot of challenges to implementation in the real world, but it has* ***high controllability and a lot of potential****. I think that's why it was chosen."*



Learning Robust Deep Visual Representations from EEG Brain Recordings

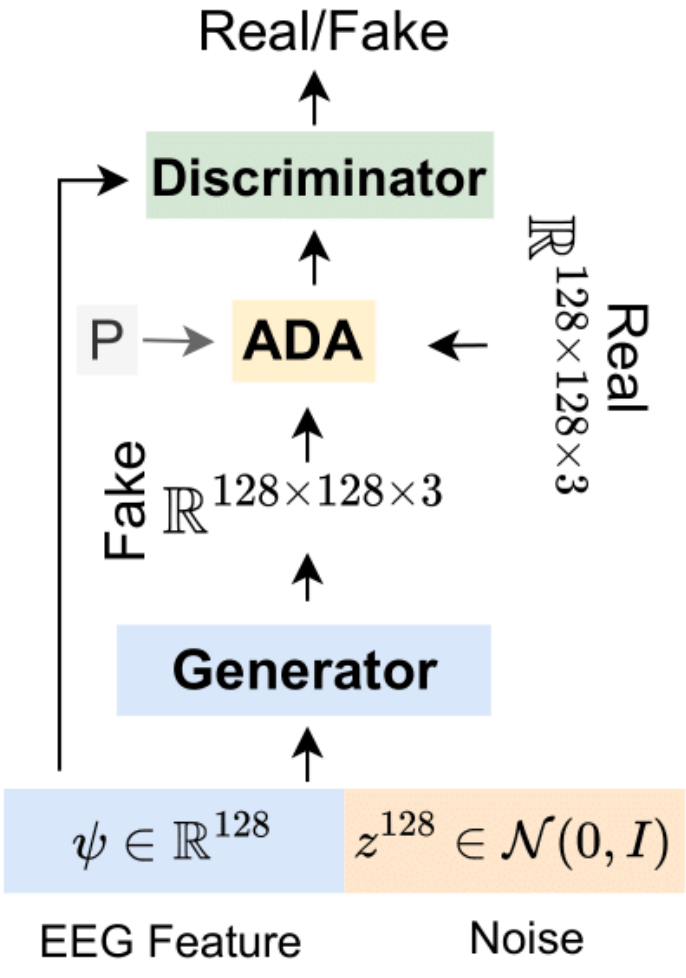
Prajwal Singh is a third-year Computer Science PhD student at IIT Gandhinagar, India, advised by Shanmuganathan Raman.

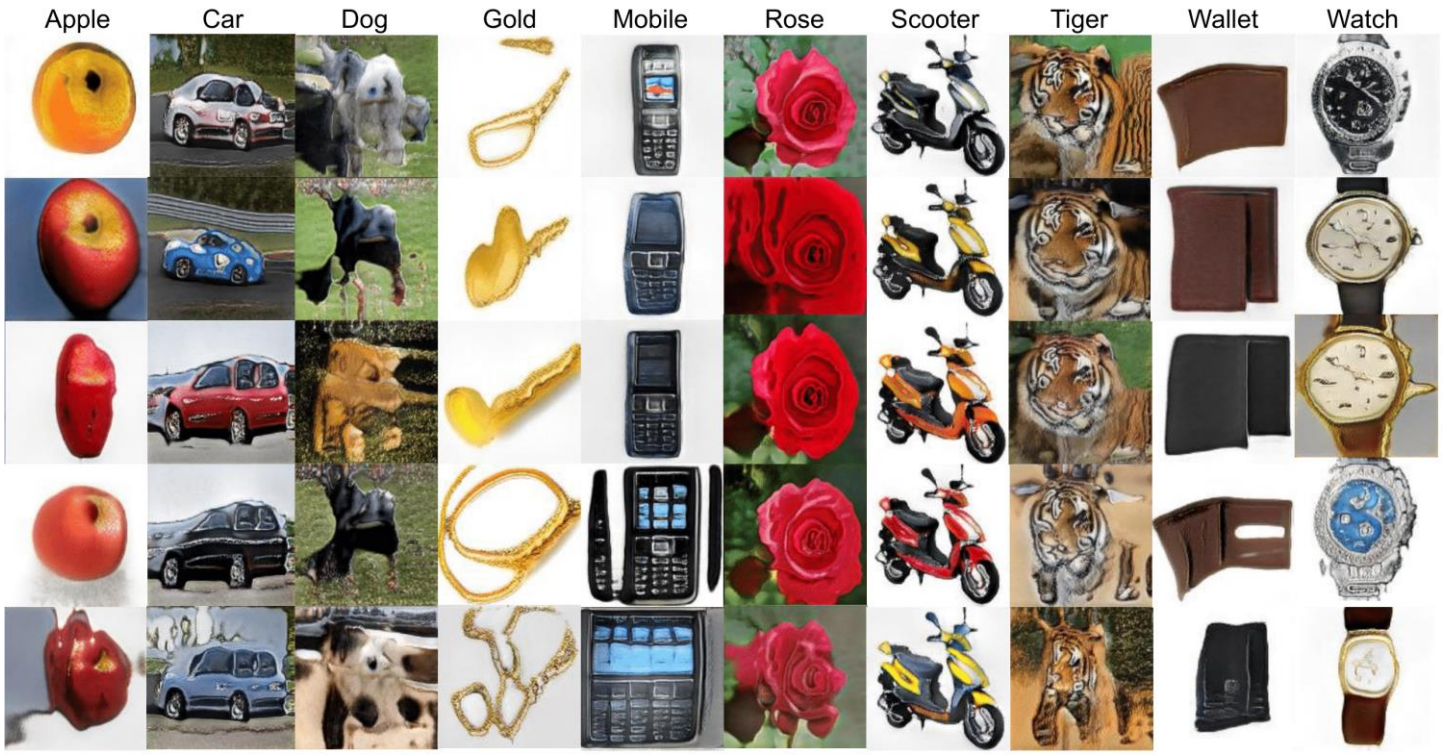
His paper attempts to answer a big question: How can computers reconstruct human thoughts? He spoke to us ahead of his virtual poster session.

In recent years, significant progress has been made in generative networks, particularly in **synthesizing images using text prompts**. The exploration of this concept has led to groundbreaking research in the field of **brain electroencephalography (EEG) signals** to visualize the images generated by the human mind.

The journey to reconstruct images from EEG data began in 2017, with pioneers like Spampinato and Palazzo laying the foundations for this work when they published their **Brain2image approach**. Their datasets have been a crucial resource for Prajwal, who is looking to advance their progress.

When humans view images, **their brains produce chemical responses and electrical impulses between neurons, which an EEG cap can**





Each image is generated with different EEG signals across different classes, Thoughtviz dataset.

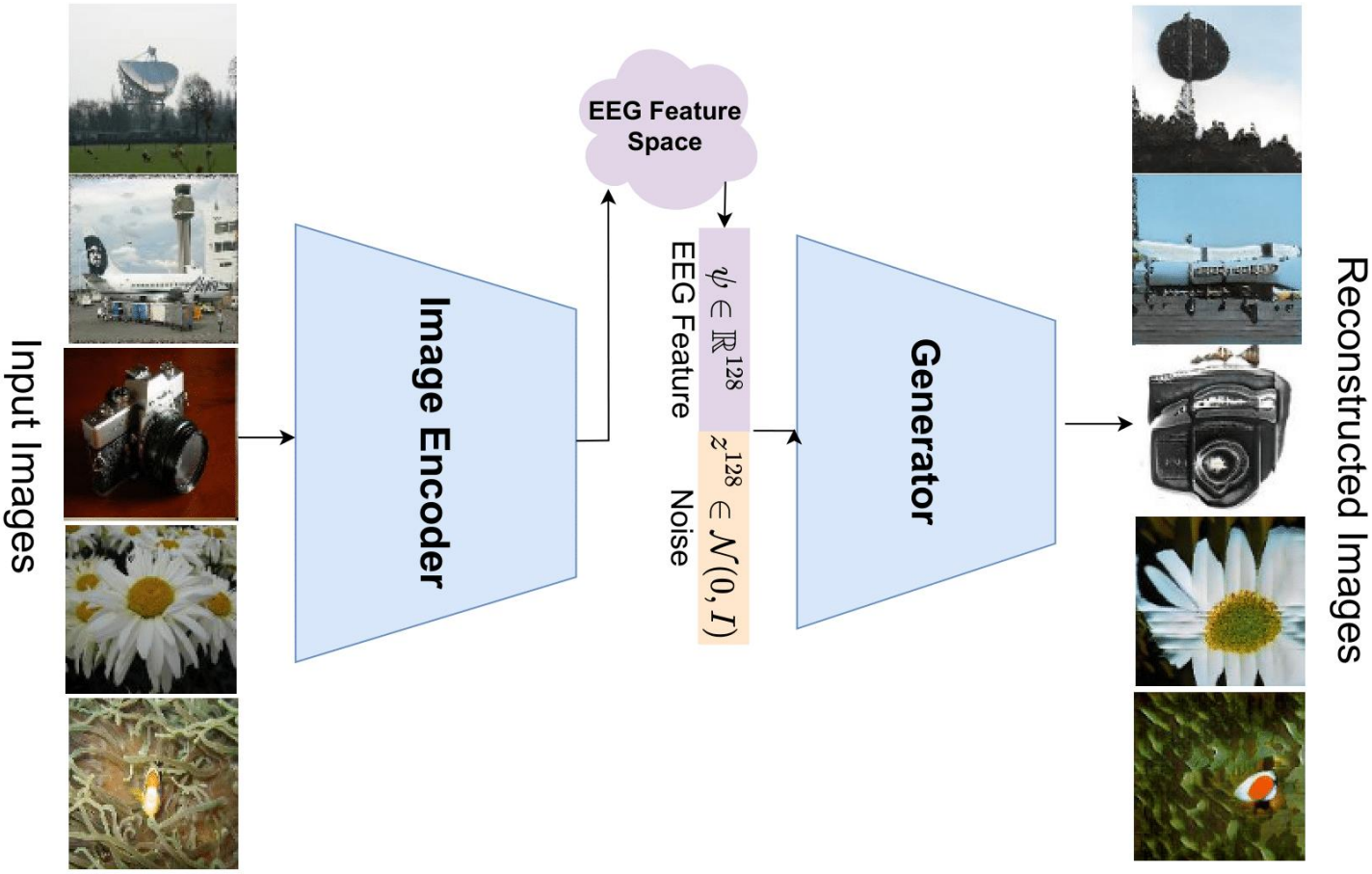
capture. This data can then be stored in a computer, but how do you extract the visual information from the EEG signal?

"We use a contrastive learning approach to extract the features from the EEG data," Prajwal explains. *"Specifically, a triplet loss formulation. Once we have these features, a **StyleGAN** handles the synthesis part. Different generative networks have been used in the past, but for our work, we used StyleGAN, which generates photorealistic images and has been quite on trend recently."*

To address the scarcity of large datasets required to train deep learning architectures, researchers used a **StyleGAN-ADA network** that can generalize across different datasets. Previous methods all dealt with generating images from particular EEG datasets instead of working more broadly.

The potential applications of this work are significant, particularly in improving the quality of life of people with certain medical conditions, such as individuals who are mute or paralyzed, yet their brains remain active. In those scenarios, an EEG cap could be placed over that person's scalp to record their brain signals and, ultimately, reconstruct their thoughts.

Prajwal tells us the most challenging aspect of this research has been handling **the noisy nature of EEG**



data. The non-deterministic nature of brain signals poses difficulties in extracting useful information, making it a complex and intricate process.

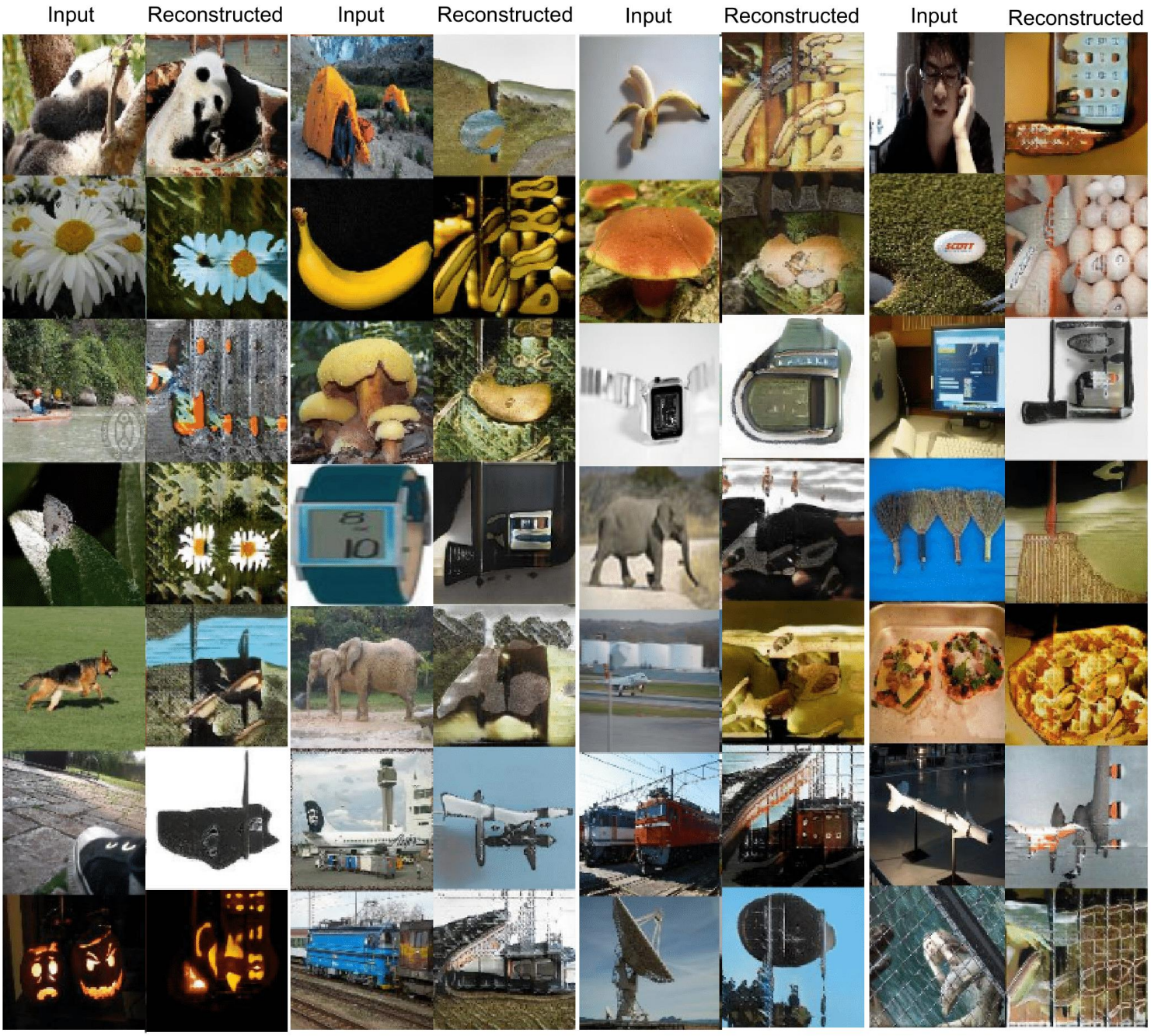
"*If I'm looking at a picture of a dog and record the EEG brain signal, when I repeat the experiment, the signal I get is going to be different because thoughts contain so many biases,*" he explains. "*While seeing one thing, we might think about something else or hear something from our surroundings, which makes it very difficult to extract useful information from the EEG data.*"

Synthesizing the information once extracted from the EEG was a further challenge. To solve this, he used **a self-supervised strategy** to train every approach, avoiding relying on supervised settings or ground truth.

EEG has been compared to a fingerprint for each and every person. Looking ahead, Prajwal hopes that future research explores the possibility of generalized EEG feature extraction methods. Currently, datasets are very controlled, but is it possible to move towards generalization and a better strategy for synthesizing images?

Recently, he has demonstrated the potential for deploying the model in a live setting by conducting an **image-to-image translation experiment**.

"We've got a collection of images that are never shown to the network, but similar classes have been shown to the StyleGAN for training," Prajwal tells us. *"We took a bunch of unseen images, transformed all those images into the EEG feature space, and tried to reconstruct the EEG features with the StyleGAN. Even though our network hadn't seen those images, it was able to reconstruct a close approximation of what this EEG might belong to. That shows we can deploy the model live if possible."*



Unseen images to EEG representation space and reconstructing back using EEGStyleGAN-ADA.

Luisa Verdoliva is a full professor at the University of Naples Federico II in Italy.

Luisa, you are also General Chair at WACV.

Yes, yes, it was hard work, so it's important to say that! [*laughs*]

What is your work in general?

My work is about media forensics. More specifically, I work on deepfake detection. I started around 12 years ago on this topic when I actually worked on detecting manipulated images. I started with a challenge. I joined the challenge. It was very fun at the time to detect if an image was manipulated or not. They were manipulations made using Photoshop, so very different from the ones that you can do right now.

The tools to detect it were not very good at the time.

Yes, they were not sophisticated, but in any case, it was challenging because if you spent time creating a good forged image, it was harder to detect.

Our readers remember you at a Media Forensics workshop with Cristian Canton a few years back. Is cooperating with these kinds of initiatives part of your work?

Yes, in that case, we joined that workshop with a paper. We won the Best Paper at the time.

I remember, you won!

Yes, but now I'm actually in the organisation of the CVPR Workshop on Media Forensics.

So, winning a workshop pays. You get another job!

[*Luisa laughs*] Yes, that's right!

You have seen remarkable progress in this field in the last 12 years. What is the most important thing that you have witnessed?

I think what was really a break was in 2018 with the generative adversarial networks that were used to create synthetic images, so the GAN images. That was really incredible to us. We couldn't believe that it was possible to obtain such high resolution in generated faces, and then not only faces but now you can generate whatever you want.

You can even describe what you want to generate, and you get it. You get it for images, and you get it for videos. This is really astonishing. There are things I couldn't believe that could ever happen.

Are you aware of the work done by Ilke Demir at Intel with her FakeCatcher?

Do you mean the deepfake detector? Yes, I know it. I think they actually were inspired by a paper, if I'm not wrong, on checking about the heartbeat. Yes, this is really interesting. We also worked on these biometric features in order to understand if a video of a person is the real person or not based on these biometrics. This is a very interesting direction.

What do we still need to solve in that area?

What I think is really challenging is the fact that often, all these videos and images can be of low quality, compressed, and resized. When you upload them over a social network, they can be strongly compressed, so the quality reduces, and also these tiny traces can be reduced…

Like artifacts?

Yes, these artifacts can be reduced, and it could be harder to detect them. Also, what is really important is to develop explainable detectors so that, as you say, you can look for some specific traces that you can explain. Otherwise, they're harder to interpret, and you don't know what's happening. If the detector says yes or no, why? Can I trust it? This is also very important.

It seems like a game of cat and mouse: how to create fakes that are so good that they cannot be detected and how to find them. Who is going to win in the end?

This is a really difficult question to answer, but note that even if a fake is perfect visually, this doesn't mean it doesn't embed some artifacts inside. It can be a perfect fake, but it can contain some artifacts that can be highlighted by some detectors. The main problem is if you have a very smart, malicious attacker that's able to hide the traces or even inject some specific traces if it knows the detector you're using. The problem is when this game is played with people who are also aware of the forensic detectors or have some knowledge so they can actually attack your detector.

It seems that, in some way, you believe in the power of your opponents, and they are making your task very difficult.

Yes, so you have to take this into consideration when you develop a detector, and you have to try to develop a method that is also robust to possible attacks.

You probably do not know the next technique they will develop, but you are confident that you will always find an answer. How can you be so confident?


BEST OF
WACV 2024

The point is that you can develop even different strategies that are based on different artifacts and this can help a lot because maybe you can attack a specific detector, but it's harder to attack several ones. It's really important never to rely on one single detector but to have different strategies. Each of them trying to detect a specific artifact.

Obviously, you are passionate about this subject. Is it strong enough to keep you interested for years to come?

For now, yeah. It depends on what will happen in the future. Also, in terms of protection of active methods. Methods for which you can maybe protect your data using some signatures or watermarks. Of course, maybe it can change, and it can evolve in the future, but I think there will be, in any case, some space for passive detectors and for a strategist that can integrate passive detectors with active ones.

What fascinates you about the subject?

It's like an investigation. You have some difficult traces you have to highlight, and this is quite challenging.

It is like, "*Elementary, Mr. Watson*"?

Yes, right. You have to find evidence. Sometimes, something looks perfect.

I remember CVPR 2016 in Las Vegas when Matthias Niessner showed his Face2Face.

Face2Face! Actually, I also worked with Matthias because we developed the FaceForensics++ dataset.

I was in the audience when he gave that live demo for the first time. It was very, very impressive.

Yeah, it was impressive.

It does not happen very much that we speak with researchers from your university in our magazine. I interviewed Fanny Ficuciello once, and that is pretty much it. Do you know why that is?

The main problem is, in general, probably the area of computer vision. For Fanny, it was robotics, I think.

Medical robotics.

Medical robotics, yes. It's probably expanding, so maybe you will have more interviews in the future! [*she laughs*]

How is it to work there?

It was the university where I studied, so I like it a lot. I think it's stimulating. I feel lucky to work with a lot of students who are very smart and who want to learn. Yes, I like the teaching aspect there a lot in terms of connection with the students, and I find a lot of stimulus with them.

What about the connection with the hometown, with Napoli?

Okay, I like Naples a lot. The food, the weather.

The people…

The people. What I like is the atmosphere you can feel there.

I think that everyone who has been to Napoli knows that. There is the famous saying: '*See Napoli and die!*'

[*laughs*] Yeah, yeah, you're right.

It is translated into most languages that I ever heard of. Why did they choose Napoli for that?

I think that Napoli is full of art, full of history, and probably not everybody knows that.

Let's say something about WACV, because you are General Chair. Tell us something about it. It is the biggest WACV ever.

Yes, we are very, very happy about that. It's a really great success with so many submissions compared to previous years. 2,000 submissions. It's great. There was a lot of interest from people who wanted to come, as well as workshops and tutorials. It was hard work, but it was worth it.

What was the most challenging part of this whole organization?

At least from my perspective, it was not a single part; it was the whole. [*laughs*] Everything needed supervision. The General Chair supervises everything, so it's not that you are doing all the work. There were great people in the organization who were doing everything, and you needed to make sure everything was done. It was really a lot. I didn't find one single, specific topic that was more challenging, but it was the whole.

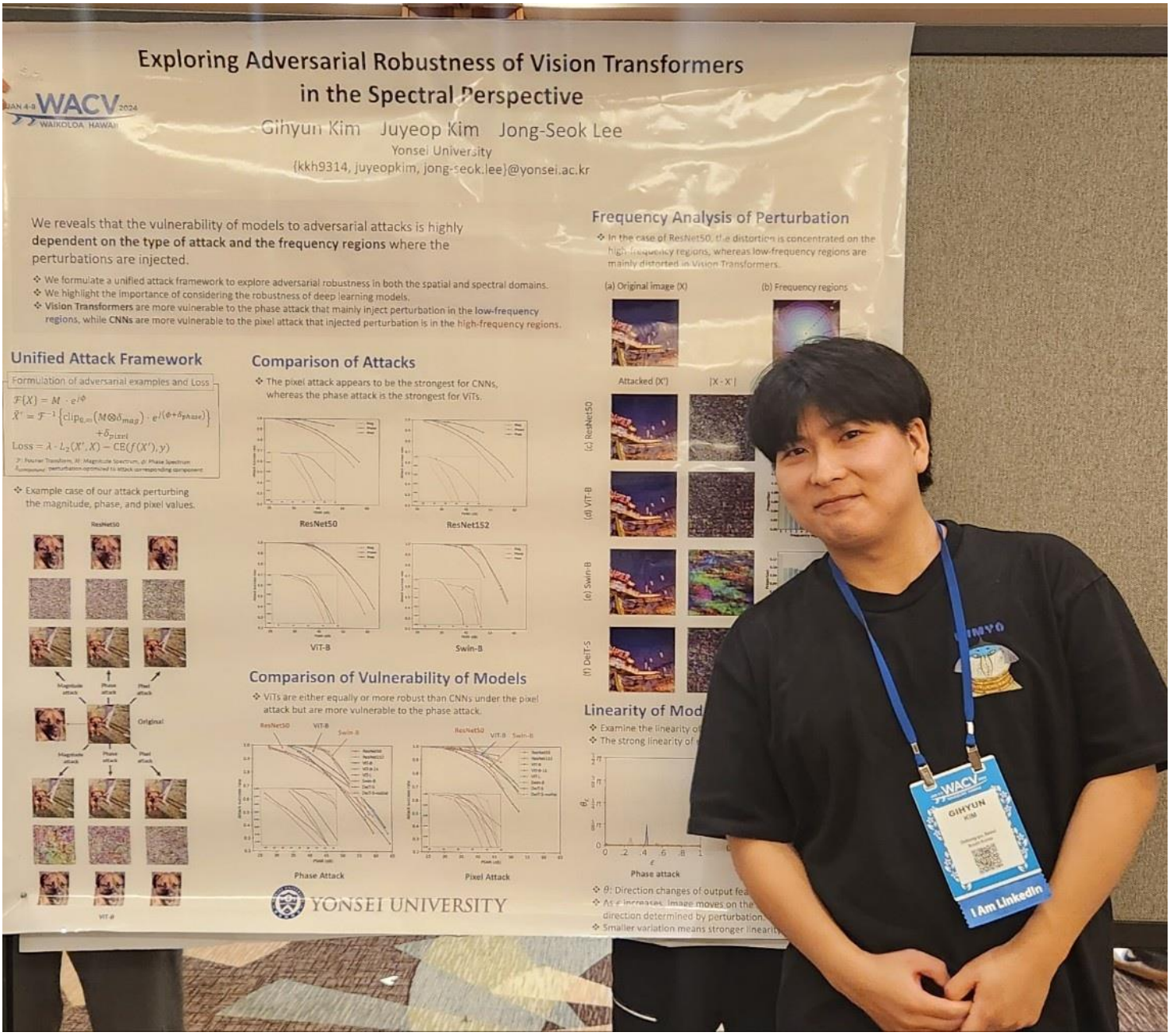
You started working in this community when you were one of very few women, right?

Now, I can tell that there are much more.

How have you found the progression?

Even at WACV, we have two keynote speakers who are women. I think this is really increasing a lot - slowly because probably this starts from when you're a child. I noticed even in my class at university that I have very few girls. This year, I had mostly males in my class, but I think it's something that should start from education – trying to stimulate girls to study maths and coding. I think they would be great!

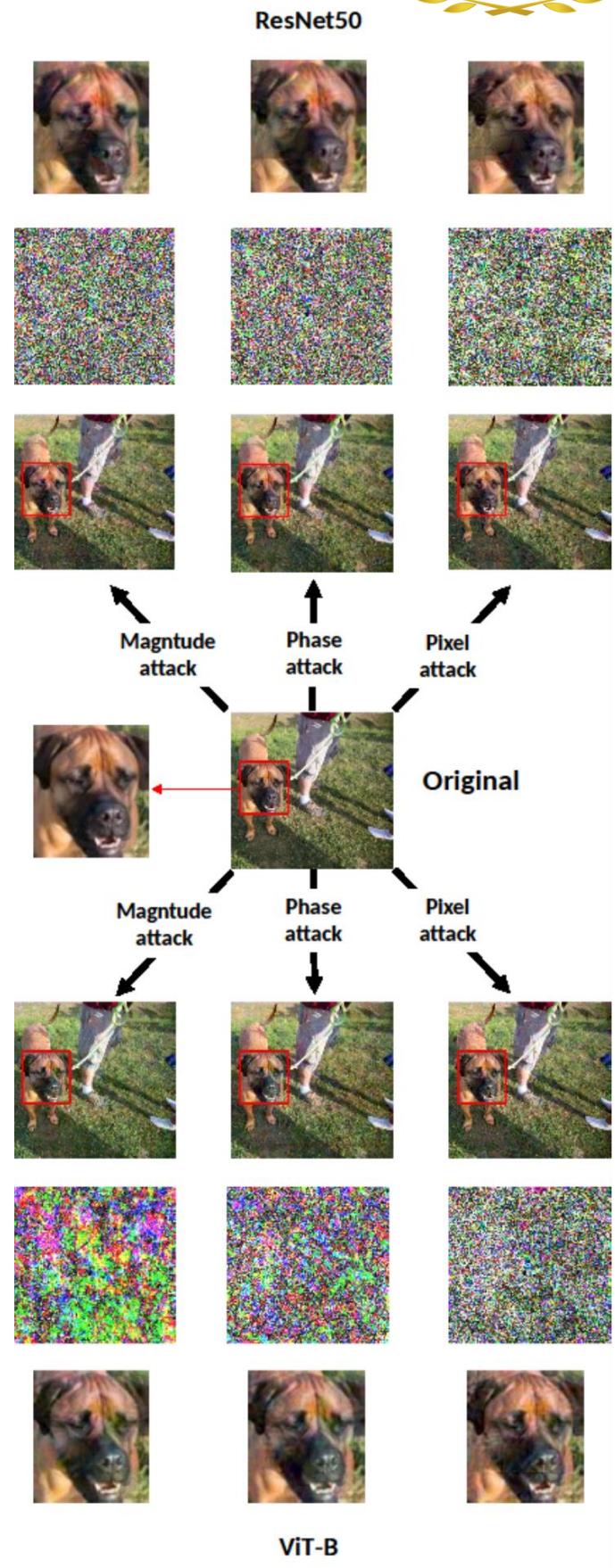
Exploring Adversarial Robustness of Vision Transformers in the Spectral Perspective



by Gihyun Kim

Gihyun Kim is a third-year Ph.D/M.S. integrated student at Yonsei University under the supervision of Jong-seok Lee.

In his research, he focuses on adversarial attacks, particularly exploring and comparing the adversarial robustness of CNNs and ViTs and analyzing their properties through the attacks.



The attacked image, the difference between the original and perturbed image, and an enlarged area of the perturbed image are shown in each case.

For an extended period, **Convolutional Neural Networks (CNNs)** have been the dominant architecture in computer vision. While CNNs show a powerful performance in the classification task, they are well known to be vulnerable to adversarial attacks which make CNNs misclassify by adding an extremely small (imperceptible) perturbation to an input image. The field of investigating adversarial attacks has emerged since the vulnerability issue is critical in applying CNNs to security-sensitive applications in real world and also in understanding of the operating mechanism of the model better.

With **Vision Transformers (ViTs)** emerging as new promising architectures, one question arises "**How vulnerable are Transformers compared to CNNs against adversarial attacks?**". While a recent series of research argued with this question, they do not reach consistent conclusions. In one group of studies, they claimed that ViTs are more robust against gradient-based attacks, and it is attributed to CNNs relying on high-frequency information, while ViTs are not. Another group of studies argues that ViTs are as vulnerable or more to attacks as CNNs in specific conditions, such as a training setup.

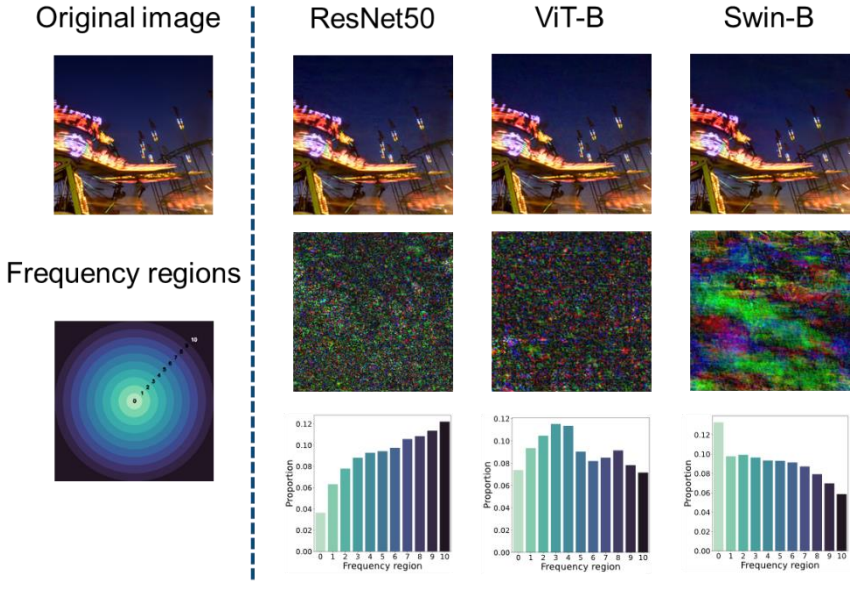
While the two conflicting groups compare adversarial robustness

CNNs and ViTs with various experiments, they only used gradient-based attack methods. In previous research, it was observed that ViTs focus more on low-frequency features while CNNs rely more on high-frequency features. From this point of view, gradient-based attacks, which tend to perturb high-frequency features in images through spatial domain perturbations, might cause CNNs to be fooled more easily than ViTs.

In order to mitigate such bias, we formulate an attack framework that can directly attack the pixel values, magnitude spectrum, and phase spectrum of an image to allow flexible perturbations in both spatial and spectral domains. It can be observed that attacking different components induces different distortion patterns in the image (Fig 1 in the previous page). The distortion pattern also varies depending on the target model.

Our research reveals that Vision Transformers exhibit similar or more vulnerability to the phase attack, which primarily injects perturbations in the low-frequency regions while Convolutional Neural Networks are more vulnerable to the pixel attack that injects perturbations mainly in the high-frequency regions.

To examine the effect of the phase attack on the spectral characteristics of images, we employ the Fourier transform on the difference between the original and attacked images, analyzing the magnitudes in different frequency regions (Fig 2 below). For ResNet50, the high-frequency regions are mainly distorted whereas the distortion is concentrated on the low-frequency region in ViTs. This aligns with the information that CNNs and ViTs rely more on high and low-frequency information, respectively.



Example of the perturbed image resulting from the phase attack, distortion in the pixel domain, and distribution of the distortion over different frequency region.

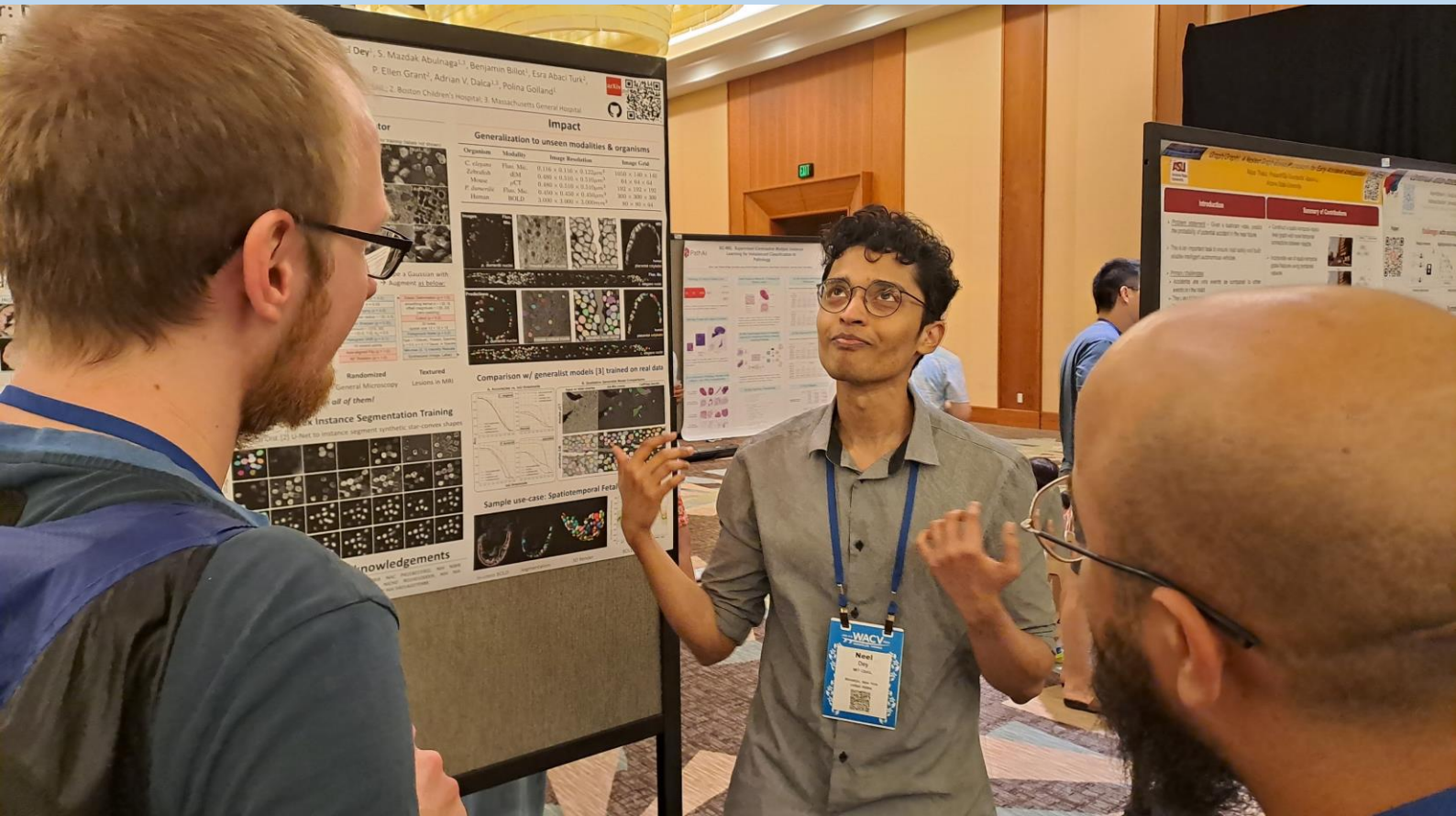
Waikoloa Sunset

BEST OF WACV 2024

One day after Dima Damen, another fascinating keynote speech by Lihi Zelnik (Technion): “Digitizing Touch”.

Jialin Yuan (top), a PhD Student from Oregon State University, presenting her work about detection of harmful multimodal content from the asymmetric angle in vision content and language content.

Neel Dey (bottom), a Postdoctoral Researcher @ MIT CSAIL in Polina Golland's Medical Vision Group, presenting his work that presents a synthetically-trained segmentor for 3D irregular blob-like shapes in *any* new and unseen radiology and microscopy dataset without needing any retraining, interaction, or adaptation.



Mathias Unberath receives Johns Hopkins Career Impact Award

The award, supported by PHutures, celebrates individuals who provide outstanding mentorship or professional development for students and trainees.

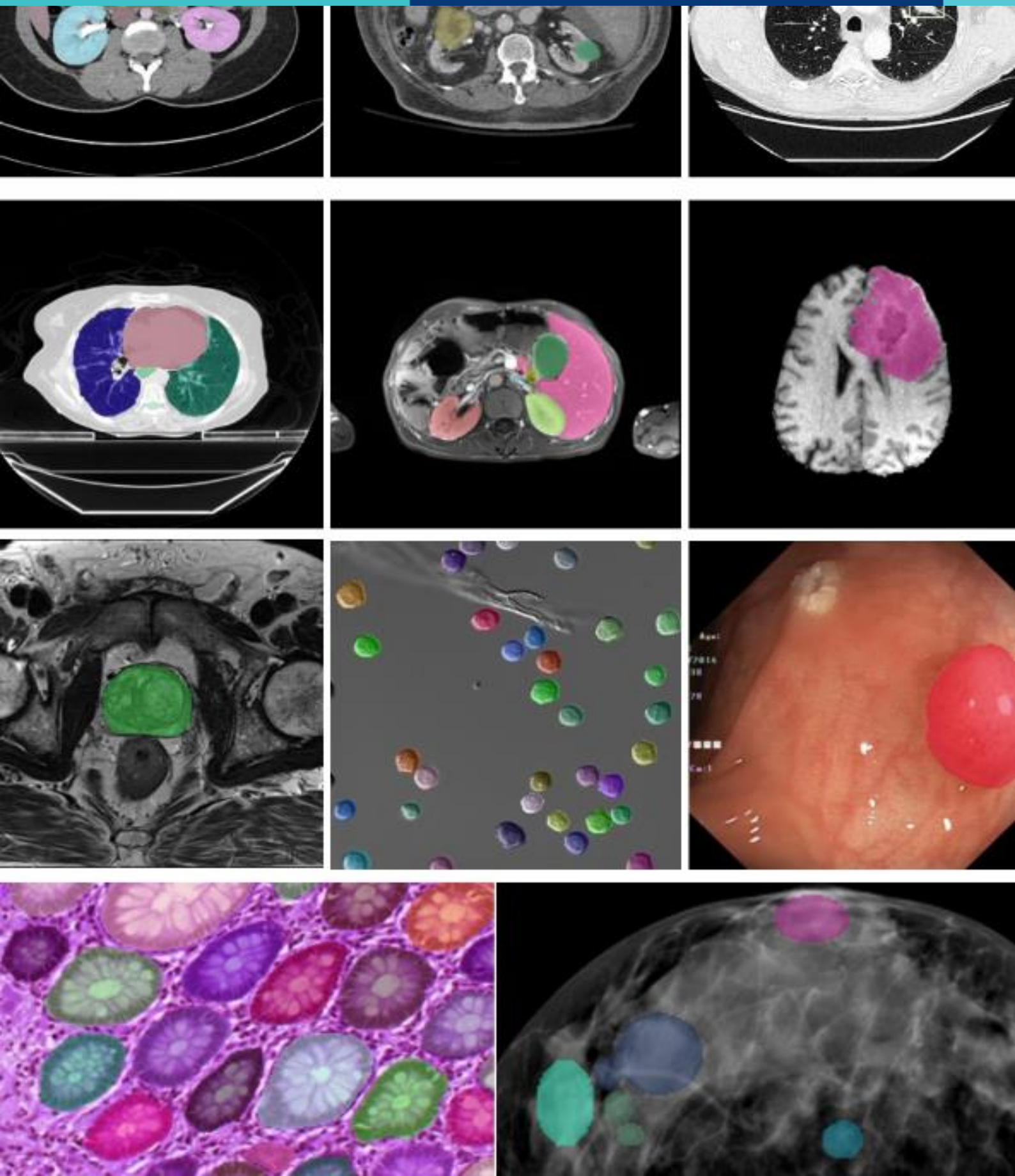
Mathias Unberath, an assistant professor of computer science at the Whiting School of Engineering at Johns Hopkins University - with secondary appointments in the School of Medicine - is one of 10 members of the JHU community who were awarded Career Impact Awards in November.

The awards recognize individuals who have provided outstanding contributions to the professional development of their students and trainees.

Mathias had also a poster at WACV 2024 in the session of Friday evening - RobustCLEVR: A Benchmark and Framework for Evaluating Robustness in Object-Centric Learning.

MEDICAL IMAGING NEWS

FEBRUARY 2024



Segment Anything in Medical Images - MedSAM

Bo Wang (left) is a tenure-track Assistant Professor in the Departments of Computer Science and Laboratory Medicine & Pathobiology at the University of Toronto. He is the inaugural Temerty Professor in AI Research and Education in Medicine and the Chief AI scientist at the University Health Network, the largest research hospital in Canada. He also holds a CIFAR AI Chair at Vector Institute.

His research focuses on machine learning, computational biology, and computer vision, with a particular emphasis on their applications in biomedicine.

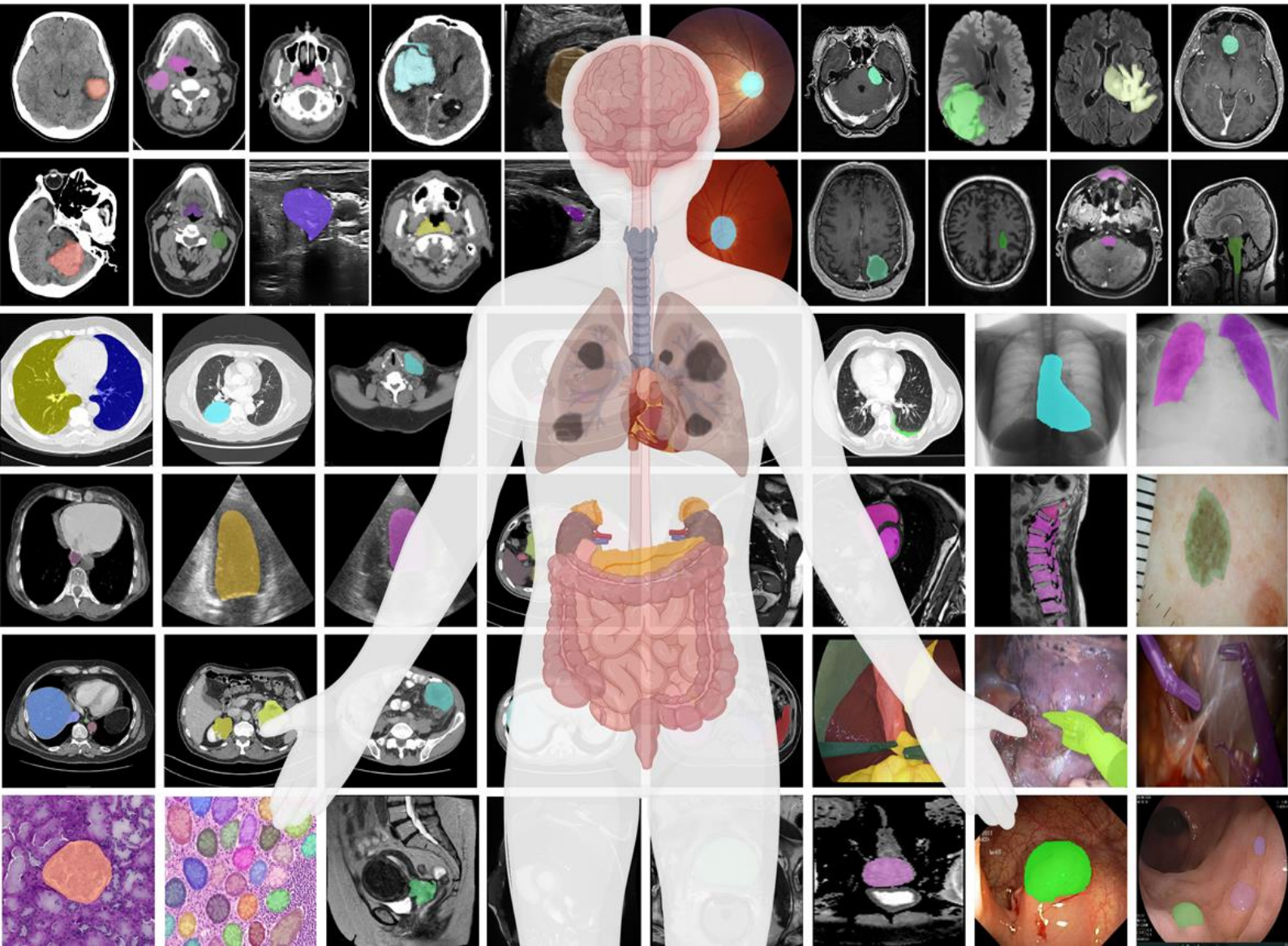
Jun Ma (right) is a Postdoctoral Fellow at the University of Toronto, Vector Institute, and University Health Network (UHN). He will be joining UHN AI Hub as a Machine Learning Lead. His research focuses on using advanced AI technologies to provide accurate and automatic cancer quantification, speed up diagnoses and personalize patient care.

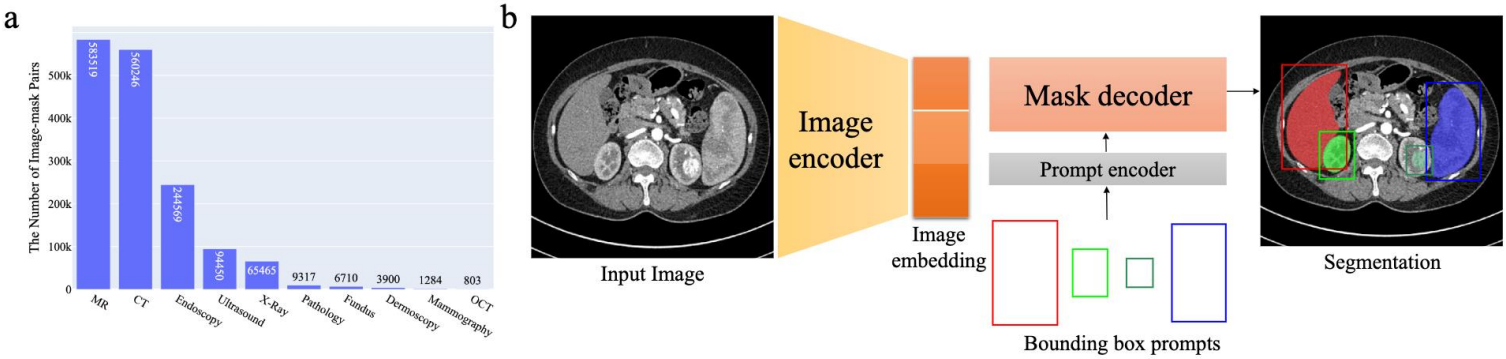
Fresh from being published in the prestigious Nature Communications journal, they are both here to discuss their groundbreaking medical image segmentation method.

Medical image segmentation has long been a specialized field, requiring tailored models for different imaging modalities and datasets. However, a novel approach said to be the first foundation model for promptable medical image segmentation promises to change the landscape.

The model in question, **Segment Anything in Medical Images – MedSAM for short –** is unique in its ability to handle diverse medical imaging modalities accurately and efficiently, providing a one-size-fits-all solution that can significantly improve a physician's workflow efficiency.

"*Before MedSAM, most medical image segmentation was a specialist approach where, for different modalities and datasets, you had to train different models,*" Bo tells us. "*This was for two reasons. One, we didn't have much data. Now, we're at a tipping point where we can access millions of medical images. Two was the model architecture. MedSAM inherited from the Segment Anything Model (SAM) published by Meta Research. We took that base model and continuously fine-tuned SAM with millions of medical image masks.*"





MedSAM is developed on a large-scale medical image dataset with over 1.5M+ image-mask pairs.

While initially unsure of its applicability to medical images, Bo and Jun immediately recognized the potential impact of adapting SAM for this task. Foundation models were still in their infancy, so there were many uncertainties, but deciding they were onto something, they got to work. In the end, through meticulous hyperparameter tuning and engineering efforts, they discovered it could effectively handle diverse medical image modalities.

"*Nowadays, it's very common for us to use deep learning for medical image segmentation, but before the deep learning era, mathematical models were the most popular method for this task,*" Jun recalls. "*These models are inherently transparent but have* ***limited usability because they require much hyperparameter tuning when segmenting a new image****.*"

There has been a paradigm shift from the original models to the current popular deep learning models. In 2015, with the emergence of fully convolutional neural networks, such as **U-Net, FCN, and W-Net**, features could be learned end-to-end, significantly improving the automaticity compared to previous mathematical models. The inference process can be fully automatic, too, without additional parameter tuning. However, these models are usually customized for specific images or modalities with limited generalizability and adaptability.

"*The SAM paper was released in early April last year, and I looked at its demo on natural images and realized it was a great breakthrough,*" Jun tells us. "*I experienced the traditional mathematical segmentation methods, which were kind of painful because they required a lot of hyperparameter tuning. We tried SAM on our medical images and found that performance wasn't very good since its training set mainly contained natural images.*"

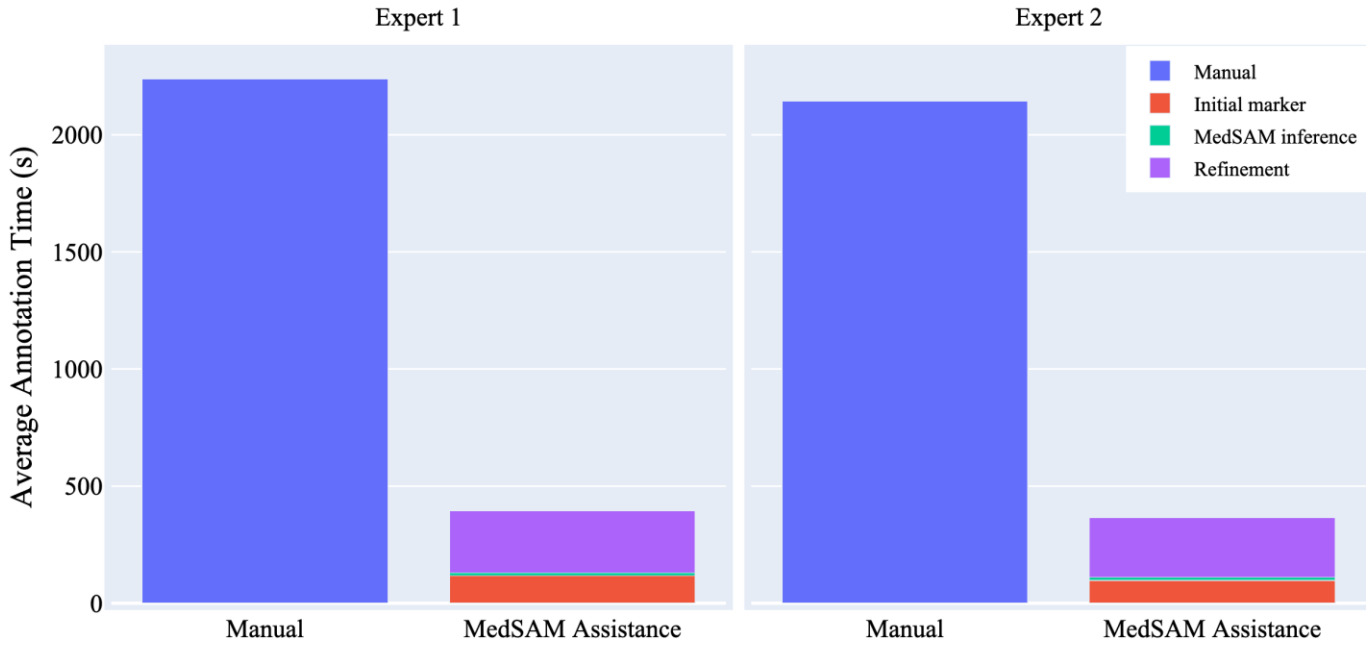
MedSAM represents another paradigm shift from specialist models to generalist or foundation models. Jun identifies three key components to develop a generalized model. "*First, a great network architecture; second, a*

massive dataset; and third, lots of computing," he explains. "*SAM already demonstrated that the recent transformer is a great network. In the beginning, our two main bottlenecks were the data and the computing. We spent nearly two months creating data. These datasets are scattered across the network and stored in different formats. We invested lots of time in standardizing these images to make them in the same format. Then, since training from scratch is almost impossible, we decided to use transfer learning.*"

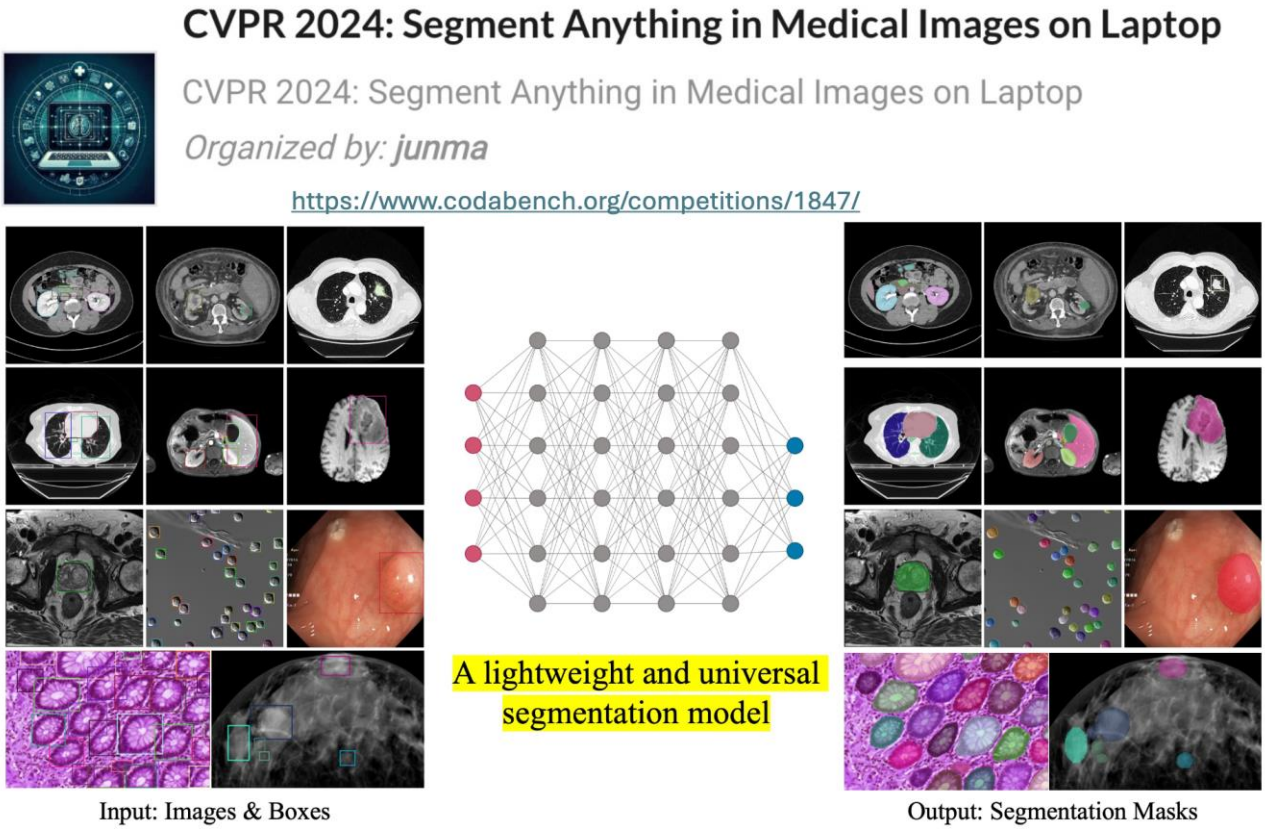
The team created **a large-scale medical image dataset with more than 1.5m image-mask pairs**. They gathered all the publicly available datasets they could find. Their original plan was to train the model from scratch, but they realized the computing effort required was too huge to be affordable. **Meta used more than 200 GPUs for their original paper** – far beyond the means of most academic groups. Ultimately, they employed transfer learning to enhance the model's ability to handle medical images.

"*This method is **a promptable segmentation method**, so it contains an **image encoder**, which is a vision transformer to extract the features from the image and obtain the image embedding, and a **prompt encoder**, which can encode the bounding box prompt to obtain the prompt embedding,*" Jun reveals. "*Then, we also have a transformer-based mask decoder to merge the image embedding and the prompt embedding to generate the final masks.*"

For Bo, the clinical potential of this approach was the most exciting part. A trial with two groups of doctors showcased **a 80%+ increase in workflow efficiency for those doctors assisted by MedSAM over those following a**



MedSAM can be used to reduce the annotation time cost by 80+%.



The CVPR 2024 challenge seeks universal medical image segmentation models that are deployable on laptops or other edge devices without reliance on GPUs.

standard manual workflow. *"That's a huge boost of efficiency,"* he points out. *"You can imagine how we could use MedSAM to help doctors do a much faster job in segmentations. This is the biggest pain point for physicians, radiologists, and CT scan readers because they spend significant time doing manual contouring."*

Looking at broader applications of MedSAM in clinical practice, the team points to its potential for **revolutionizing tumor measurement - the RECIST criteria**. RECIST is a standard approved by the FDA to assess tumor progression. They choose two points representing the longest diameter of tumors in MRI or CT scans and then use this length as the metric to indicate tumor progression before and after treatment.

Outside of this paper, Jun says his research is focused on employing advanced AI models in clinical practice for more personalized treatment and to improve patient outcomes. Plans are underway to organize **challenges at CVPR** and **MICCAI** this year, hoping to **optimize MedSAM for use on a laptop** and further develop it for **generalized cancer segmentation in whole-body CT scans**.

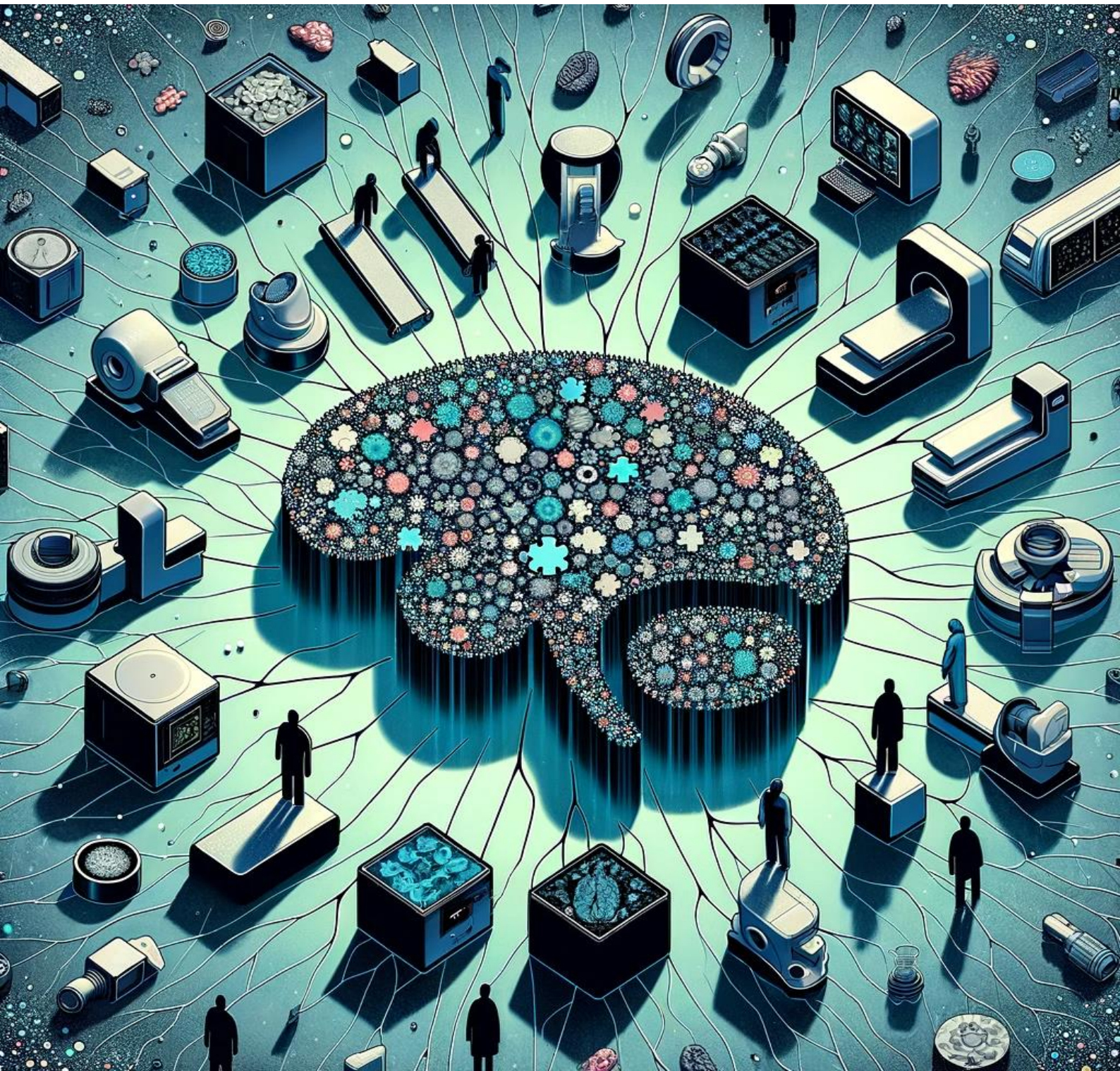
"I really appreciated the support of Professor Bo Wang, my collaborators and our institutions for their computing and resources for me to do this work," he adds. *"Now, we want to gather the efforts from the whole community to make the model deployable in clinical practice!"*

Data Availability and Quantity in Medical AI Projects (part 2)

Data is key in all AI projects. Training and optimization of neural networks rely on large quantities of data. This is all the more critical for medical AI.

It is also true that medical AI projects often lack a sufficient amount of data. The deficiency might be in quantity and/or in quality. When developing a next-generation medical device, data may even be totally absent.

We started last month to suggest some mitigation tips by RSIP Vision: part 1 is here. Here are more tips.



Working with the expertise of RSIP Vision enables us to mitigate this challenge in many ways. Our research and development team has implemented data augmentation tasks that are specific to each modality and application.

Balancing datasets is an effective technique to optimize the training process. Let's consider a scenario where we obtain data from different sources. In such cases, we may have one source that has provided significantly more data than the others. If we train our network with this unbalanced data as a single block, the performance will be better on the larger dataset and not as good, or even very poor, on the smaller ones. Therefore, we need to balance the datasets. The data engineer must do this with great care, with the help of medical staff, to determine **the best weight for each dataset** while giving enough consideration to the smallest one.

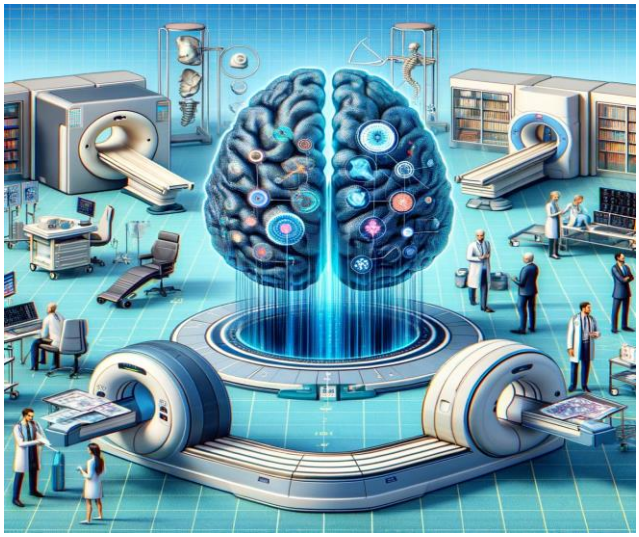
We have a technique for generating synthetic data, which involves creating new data by **using available**

data collected with different modalities. One practical application of this technique is to **generate synthetic X-ray images from CT images (DRRs)**, which significantly enlarges the smaller dataset. This technique is different from data augmentation, which adds noise and other changes to existing real data to make it different in a clinically coherent way. Synthetic data generation is done using **Generative Adversarial Networks (GANs)** or other simulation tools and can produce completely generated data.

Artifacts and effects generation is a technique that is similar to data augmentation and is used to address the problem of variety in data. Sometimes, the data contains artifacts that are due to the type of device used rather than the true condition of the patient. For instance, CT images scanned in regions that include metals can produce significant artifacts that can obscure important areas or even alter key clinical values. Patient movement is a common cause of artifacts in medical images. Ultrasound images have specific artifacts like **reverberation, shadowing, mirroring, anisotropy, and more**.

RSIP Vision employs a technique called **the support task** to improve the accuracy of its models. This involves adding a secondary task, such as segmenting a nearby structure, to support the main task of segmenting an organ or disease. By doing so, the network is less likely to become confused and produce inaccurate results. This technique is particularly useful when dealing with limited data, as it improves the network's accuracy in identifying the region of interest. For example, when segmenting a specific bone using CT, it is helpful to segment all the surrounding bones as well, so that the network can distinguish between them accurately. This technique is also useful for more complex tasks such as **tumor segmentation**, where the network can learn to differentiate the tumor from surrounding tissues. Additionally, this technique can be used to add contrast to images with less defined contours, such as when **segmenting a bone in an MRI scan or analyzing a surgical video**.

These are just some of the options that can greatly enhance the accuracy, generalization, and robustness of AI modules when working with limited datasets. Contact RSIP Vision to ensure your project is executed in a highly professional manner.



The Summer School on Deep Learning for Medical Imaging, now in its fifth year, brings together medical imaging beginners and experts from diverse backgrounds keen to delve into the fundamentals of deep learning and its applications in medical imaging. Organiser Jose Dolz, an associate professor at ÉTS Montréal, is here to tell us more about July's event.

The collaborative effort behind the summer school involves a dedicated team from ÉTS Montréal, the University of Sherbrooke, and CREATIS – INSA Lyon in France, where it was established in 2019. Hosting duties alternate between Lyon and Montreal each year, enabling participation from different parts of the world and creating a genuinely global community. "*One year, it's in France, which attracts people from Europe,*" Jose tells us. "*The next year, it comes to Montreal so that we can attract more people from the US and North America.*"

Another distinctive feature is the school's inclusive nature. Across five days, the curriculum spans a wide spectrum, accommodating students and professionals with varying levels of expertise. "*It's not limited to only AI or medical imaging experts,*" Jose points out. "*We cover many different topics, from the basics of machine learning and deep learning to more complex, advanced topics. This year, we have talks about foundation models, generative models, normalizing flows, diffusion tensor models, weakly supervised learning, and few-shot learning, so it actually covers many different aspects.*"

The event makes time for plenty of hands-on practice, reinforcing the theoretical knowledge shared during the talks. This interactive element ensures a deeper understanding for participants and facilitates engagement with tools for those on the clinical side of medical imaging, including doctors and physicians. Breaking down barriers between disciplines, the school enables a fruitful exchange of ideas and collaboration among professionals with different backgrounds.

Although less interactive, the talks and hands-on sessions are recorded, allowing the content to be shared with a larger virtual audience.

Notably, the **Summer School on Deep Learning for Medical Imaging** has received recognition from the **MICCAI Society**, further solidifying its reputation in the community. Reflecting on previous editions, Jose reveals that students often join to broaden their horizons after working in labs, as it is a great way to expand their knowledge and access crucial networking and mentoring opportunities.

Aside from academic pursuits, the week-long event embraces a lighter side, fostering community building and camaraderie between participants. "*During the day, we have the school, and it's more serious, but after that, we all go out together for dinner and drinks and have some fun!*" he adds.

Jose is keen not to take full credit for the event, emphasizing its success is a collective effort alongside his fellow organizers, Pierre-Marc Jodoin, Christian Desrosiers, Thomas Grenier, and Michaël Sdika.

The 2024 Summer School on Deep Learning for Medical Imaging takes place at ÉTS Montréal from 8-12 July. Registration is open now.

Anne-Marie Rickmann recently completed her doctoral degree at the Ludwig-Maximilians University Munich (LMU) under the supervision of Prof. Christian Wachinger.

Her research focused on developing deep learning techniques for medical image segmentation of 3D MRI and CT scans.

Anne will continue her research as a postdoctoral researcher at Yale School of Medicine in the Radiology and Biomedical department, where she will work with James Duncan and Albert Sinusas on developing deep learning methods for multimodal cardiac image analysis. Congrats Doctor Anne-Marie!

Medical Image Segmentation is an important task in image analysis pipelines, serving as a prerequisite for many downstream applications. While some may consider segmentation a solved task, I believe there are still challenges in segmentation that are interesting to work on.

Here I highlight some methods we worked on in the last years.

Organ hallucinations:

We introduced HALOS at IPMI 2023, addressing a unique challenge – organ hallucinations.

When investigating the performance of usually well-performing models like nnU-Net on large-scale and diverse datasets like the UK-Biobank, we found that these models struggle when they are confronted with anatomical changes post-organ resection surgery, such as cholecystectomy or nephrectomy. The models, interestingly, tend to segment organs that no longer exist. We termed this phenomenon organ hallucination.

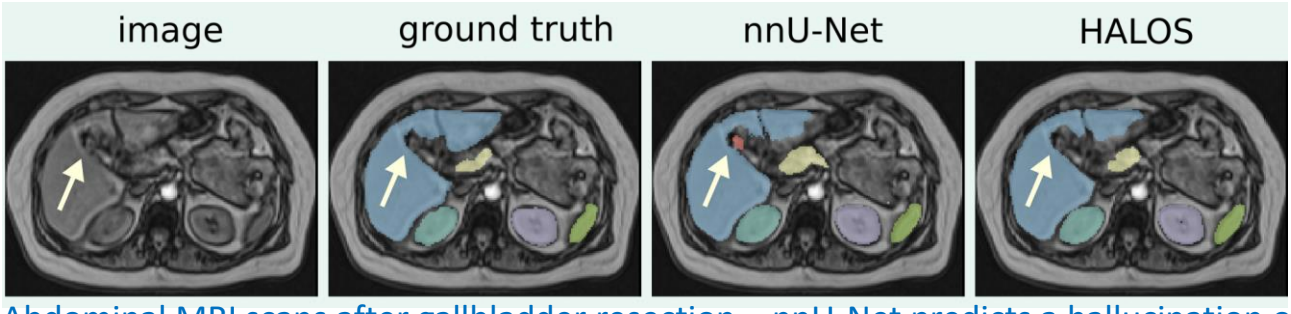
HALOS tackles this by integrating an organ existence classifier and using feature fusion modules to enhance the segmentation process with information about organ existence.

A standout feature of HALOS is its flexibility: it can utilize ground truth labels for organ existence when available, or alternatively, rely on the classifier's output.

This approach shows promise for extension to other organ resection cases and might be useful in other scenarios where anatomy significantly differs from the majority of training data scans.

Cortical Surface Reconstruction:

A large part of my research focused on cortical surface reconstruction.



Abdominal MRI scans after gallbladder resection – nnU-Net predicts a hallucination of the gallbladder (white arrow) and HALOS doesn't.

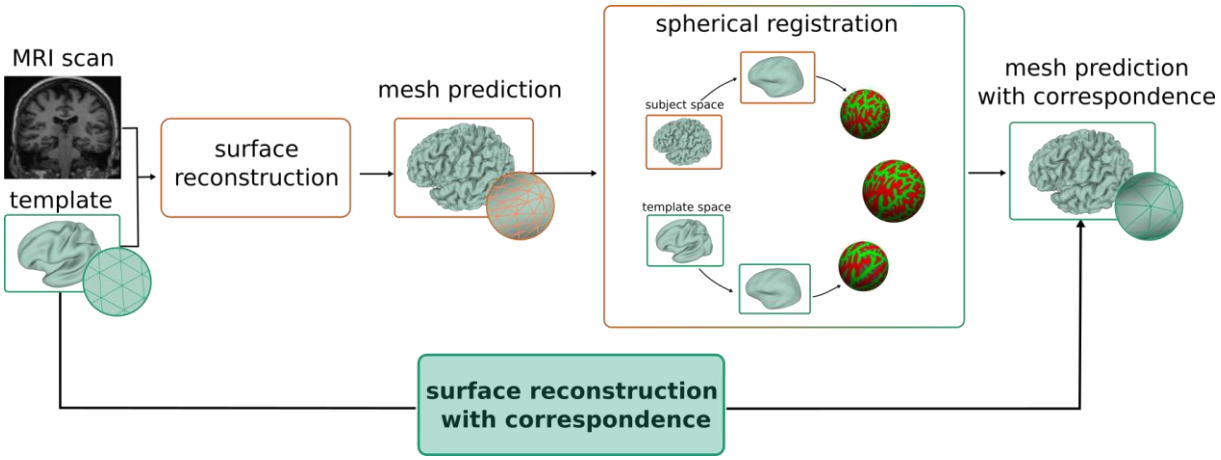
This is the task of extracting surface representations of the cerebral cortex from brain MRI scans.

Our work Vox2Cortex, which we presented at CVPR 2022, is a combination of a 3D U-Net which learns a voxel-wise segmentation of the brain – and a graph convolutional network that takes a template mesh as input and learns vertex vise displacement vectors. Information is passed from the U-net to the graph network to guide the deformation process.

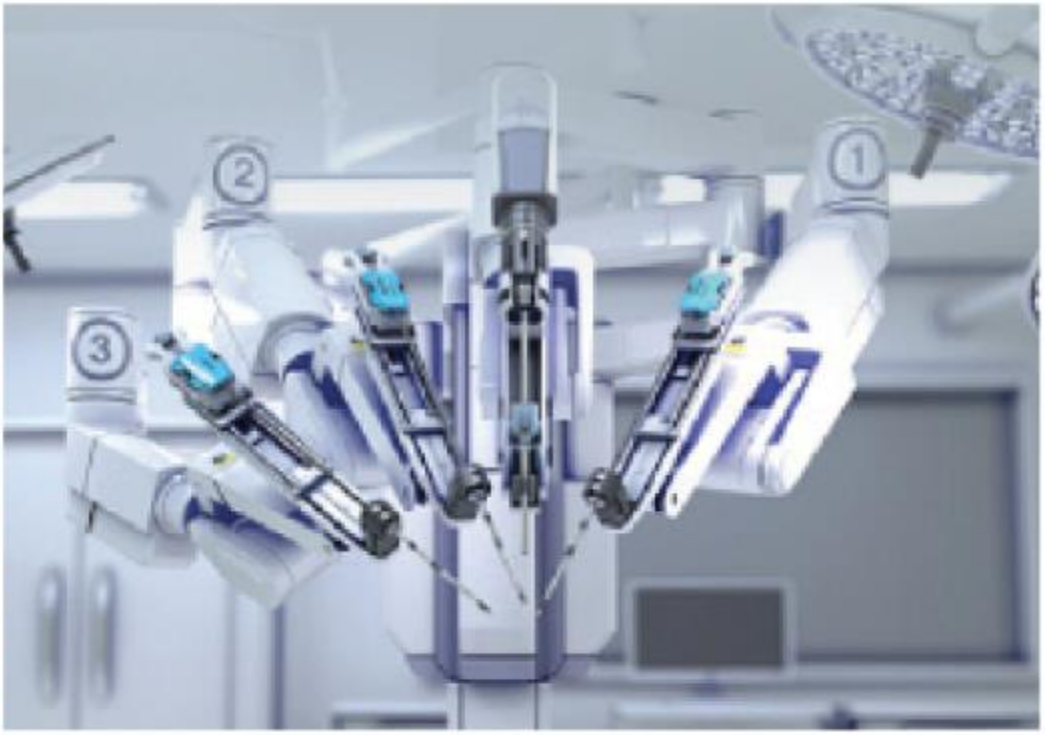
The strength of this approach lies in its use of template deformation, maintaining consistent topology and mesh connectivity, thus simplifying group comparisons.

Later we presented an extension of this work, V2CC, at MICCAI 2023. V2CC enhances the process by establishing vertex correspondence with the template. This development eliminates the need for cumbersome post-processing registration steps, allowing for straightforward comparisons between subjects or with reference groups. We achieved this by pre-registering training data to the template and optimizing an L1 loss. This seemingly simple strategy has proven highly effective. It also eases applications like comparing cortical thickness in Alzheimer's patients with healthy controls, which is now much faster and does not require registration.

Big thank you to my supervisor Christian Wachinger, for his support throughout the last years and my colleagues at AI-Med!



Typical template deformation approach – which does not guarantee vertex correspondence and relies on a post-processing step of registration. Our V2CC approach directly predicts a mesh with correspondence to the template



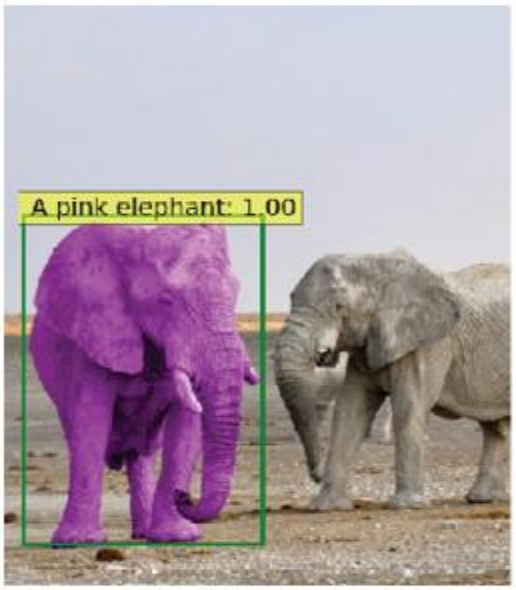
IMPROVE YOUR VISION WITH

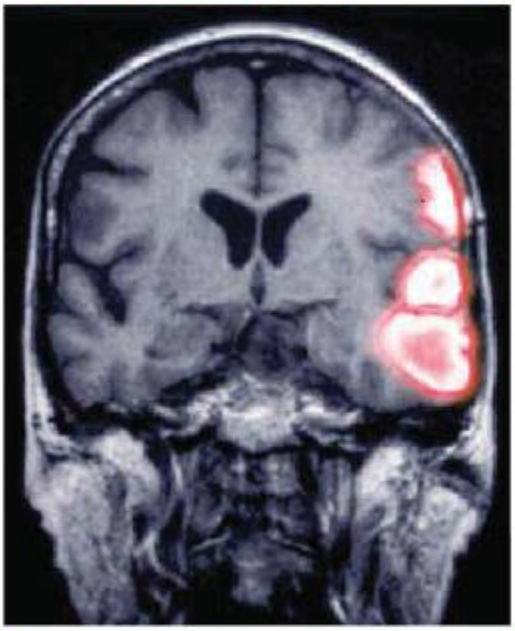
Computer Vision News

SUBSCRIBE

to the magazine of the algorithm community

and get also the new supplement Medical Imaging News!





RSIP VISION