November 2021

Computer Vision News

The Magazine of the Algorithm Community

Exclusive Interview with Raquel Urtasun: ***"We are actually going to change the world!"***

Best of ICCV 2021

with the new supplement

Medical Imaging News

page 46

Canadian Business Magazine, Studio Wyse, photo Miguel Arenillas.

DILBERT® BY SCOTT ADAMS

Computer Vision News

Editor:
Ralph Anzarouth

Engineering Editors:
Marica Muffoletto
Ioannis Valasakis

Head Designer:
Rotem Sahar

Publisher:
RSIP Vision
Contact us
Free subscription
Read previous magazines

Dear reader,

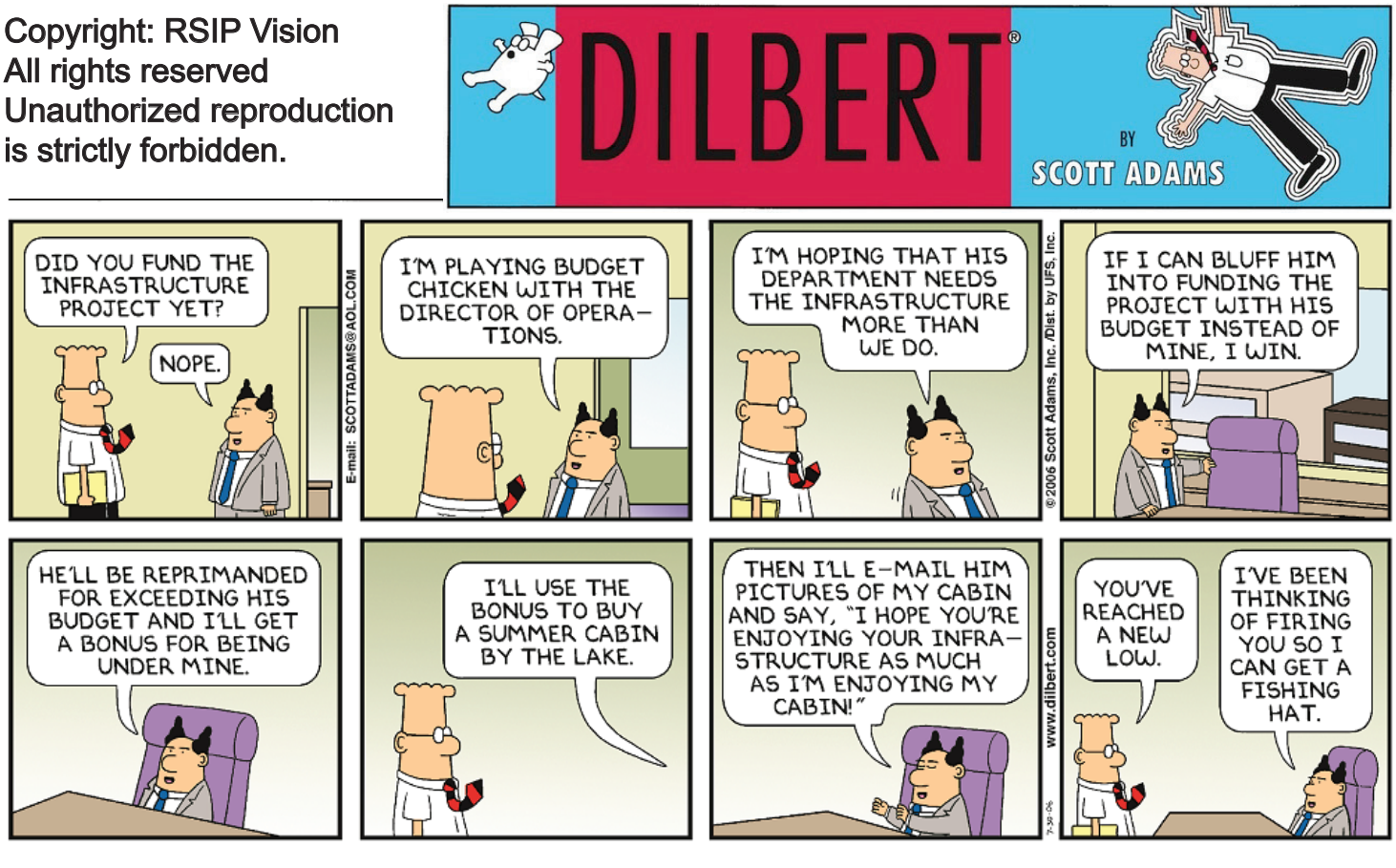
We're going to change the world! The cover story of this issue is our exclusive interview with **Raquel Urtasun**: this leading expert in the field of Autonomous Vehicles has ended her Chief Scientist stint at **Uber ATG** to launch a very promising startup! A very good friend of our magazine, Raquel accepted to tell the community what's behind this bold move and why she is confident that **Waabi** will change the world as we know it.

Last month's outstanding edition of **ICCV** was held online. Nonetheless, this virtual ICCV triumphed: over **1,500 accepted papers** were presented online, and some of them were reviewed in our **ICCV Daily** magazines. It is therefore natural to share with you the best of the best from ICCV 2021 in this November issue of **Computer Vision News**, published by **RSIP Vision**. Once again, our editorial choices were successful, and we featured the winner of the Best Paper Award as well as the Honorable Mention Award. We just liked their work! Our **BEST OF ICCV** section features only a tiny portion of the event, but you'll get an idea of some of the most meaningful scientific moments from this year's conference.

Finally, our new supplement **Medical Imaging News** was very favorable welcomed by the community. Keep sending us your feedback, subscribe to Computer Vision News, and share our link with your friends!

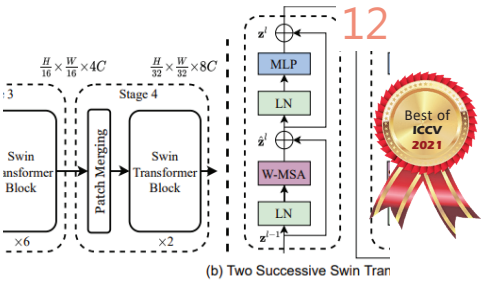
Enjoy the reading and **subscribe for free**!

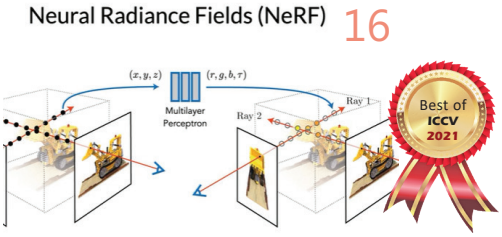
Ralph Anzarouth
Editor, **Computer Vision News**
Marketing Manager, **RSIP Vision**

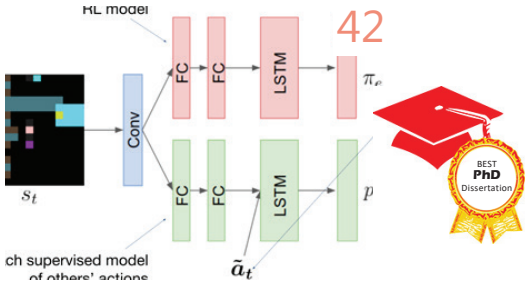


Summary 3

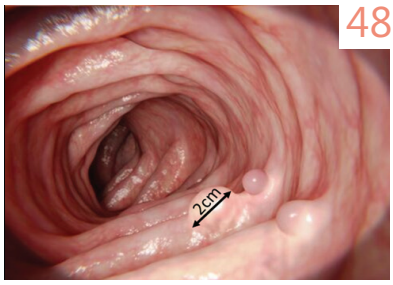
Computer Vision News

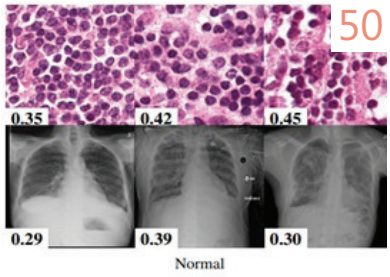




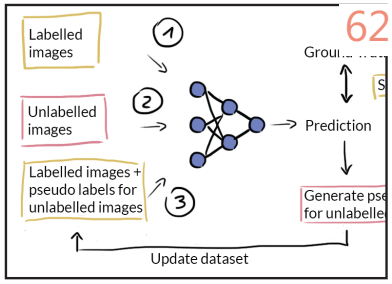


Medical Imaging News






```
'/case' + str(i) + '/Tran
itemMR = slicer.util.loadVolume(ite
itemSEG = slicer.util.loadVolume(it
itemCT = slicer.util.loadVolume(ite
itemTransform
= slicer.util.loadTransform(itemTransform_f
outputSEG_filename = filename_dir+'
+ str(i) + '.nii'
SEGresampledNode
= slicer.vtkSlicerVolumesLogic().CloneVolum
cene, itemSEG, 'out', True)
SEGparameters = {'inputVolume': ite
```

Exclusive Interview

"We are actually going to change the world!"

We last spoke to Raquel Urtasun four years ago when she had just started her first job in industry at Uber. Now, as the Founder and CEO of Waabi, she is at the forefront of the next generation of self-driving technology. The Entrepreneur just included Raquel among the 100 Women of Impact in 2021, and she has been starring for years now in our own inspiring group of 100 Women in Computer Vision. Waabi just raised a $100 million Series A initial funding. Raquel brings us up to date with what she's up to and tells us what the future holds for her exciting new venture.

Raquel, so much has changed since our last interview. Would you like to update our readers?

Sure! We've seen meaningful progress in self-driving over the past two decades since the DARPA Grand Challenge, but when it comes to commercial deployment, very little has changed. There are very small and very simple operational domains where we see self-driving vehicle testing, but it's far from the promise of self-driving at the scale that is going to transform our lives. Ifyou think about why that is the case, why we don't have this technology yet, in my expert opinion the main reason is

the technology itself that's employed in the industry. It's based on what I call the traditional engineering approach, which is an approach that doesn't scale nor generalize well across different domains. It's a technology that requires super expensive, time-consuming, and tedious manual tuning. You can think of the current software stacks as being like super complicated software stacks composed of many very tiny models, where there is a bit of AI in those models, but somehow they need to be combined into this final decision you're making. If you think about self-driving, it's a very, very hard task, and the technology doesn't scale, so there is a need for something else. This is where Waabi comes into play. If you really want to change technology and you know what technology you believe is really going to get us there, the best thing to do is to start a new company. That's basically why I decided to start Waabi this year.

I think you're hinting at things that you didn't know when you started at Uber, or that nobody knew at the time. What would you say is the main thing that you have learnt in the past few years?

I've been working in AI for 20 years and more than 10 in self-driving. At Uber, I learnt a lot about what it means to be a cutting-edge self-driving programme – one of the largest and most impressive out there. I also learnt a lot from Uber CEO, Dara Khosrowshahi. All these things became very, very useful when building a new company. I closed one chapter at the beginning of the year, and Waabi is a new chapter in my career. I built up my team over the last couple of years and it was called Waabi because it is very special: we are like a family, which is something that you don't necessarily see in other companies. We develop a lot of technology, and we learn a lot. It's not about Uber, it's not about the industry, this is something else! That's why I created Waabi in the first place.

"I believe I should utilize my skills to provide positive impact in this world!"

Tell us more about Waabi. Give us the elevator pitch!

[*laughs*] I think the best way to understand this is there has been meaningful progress in the industry, but this progress has kind of arrived at an impasse due to the type of thing that we employ in the industry. There is a need for a different solution. The key to this different solution is to really unleash the full power of AI. This is what Waabi is providing. It's a new generation of algorithms, very different than what you see in other companies, that really unleash that power of AI to build technology that

can be deployed safely and much faster than ever before. If you give me 30 more seconds, I can tell you about Waabi's second big differentiator.

Please go ahead!
The first one is the brain, our AI-first approach, and the second is our breakthrough simulator. If you look at the industry, to really understand whether the software works they need to drive many, many miles in the real world. This is why we see millions of miles driven by many, many teams. This is very expensive and not scalable. It's also not the safest thing to do. Instead, what Waabi has is this breakthrough simulator where, with an unprecedented level of fidelity, we can test the entire software stack and remove the reliance of driving all those millions of miles in the real world. So, we can develop much safer and much faster.

Is it a data augmentation thing?
No, it's much more than that. You can think of it as a digital world that is as big as the entire world and where everything looks real so that you can really test at scale – safety-critical scenarios as well as regular scenarios. You test close loop, which means that you test the reactions of the vehicle and how everything reacts to the vehicle as the vehicle does the manoeuvres. What is key to the simulator is that we can simulate how the sensors in real time will observe the scene in a way that looks the same as the real world. Mathematically, we can correlate if what happens in the digital world is the same thing that will happen in the real world.

Who will your clients be?
I guess the question is, what is the use case that Waabi is going to tackle? Because self-driving technology can be applied to many use cases. We believe that the first place this technology will be deployed is on self-driving trucks, so long-haul trucks. There are two reasons for this. One is that there is a chronic shortage of drivers. At the same time, it's one of the most dangerous professions in North America, so there is a need for improved safety. From the point of view of the technology, if you think of robotaxis in our cities versus

on our highways, while highways are still very difficult, it's much simpler than driving in cities. That's why we believe that long-haul driving is what makes the most sense in terms of our first use case. We're not an auto manufacturer, we're not going to build our own trucks, but we will be partnering with folks like OEMs in order to bring Waabi's software technology to driving those trucks.

How many are you there now?
The team right now is a little bit over 40, but we need to grow because we have very ambitious plans to deliver this technology so that society can really benefit from it. We are going to invest most of our investment money on growing the team.

So, you're hiring!
Yes! You can see a lot of hiring opportunities on our website at waabi.ai.

"If you really want to change technology and you know what technology you believe is really going to get us there, the best thing to do is to start a new company."

Okay, so now you have the money and the brain, what more do you need?
[*laughs*] Yeah, great question! Maybe it's worth talking a little bit about the money as you call it. We are backed up by best-in-class investors from deep technologies, such as Khosla Ventures, which is our lead investor, logistics like 8VC or Uber, as well as the thriving Canadian ecosystem that is growing very, very nicely in terms of entrepreneurship. We fundraised one of the largest Series A in Canadian history. Over CA$100 million, which is a little bit over US$83 million. We're very excited that people understand why there is a need for new technologies in self-driving, and that they understand this is the team, if any, that can deliver this new technology.

What might go wrong?
[*Raquel laughs*]

I'm sure that you have thought about it! I'm sure that sometimes in your mind you say, "What if…?"
[*Raquel pauses to think*] The solution to self-driving at this scale doesn't exist yet. As with everything that doesn't exist yet, there's always the risk of – will you get there? I believe that if there is a path to solve this, it's this path! But there is always the risk that because it's a very hard problem, can you solve it? I absolutely believe we can, which is why I've dedicated my life to self-driving for more than

10 years now. Also, to deploy at this scale, there are many things that have to come together, like being able to create these trucks, the sensors, the physical assets at the scale necessary. There are the regulators, and you need to build the trust of the public, etc. There are many factors. It's not just I build the technology and that's it. It's an entire ecosystem. As always, when you have many factors that need to come together, there is the risk that one of them is not ready, but I think everybody understands this technology can really change the world as we know it today for the better and we can create incredible opportunities. It is just a matter of working together towards creating a safe solution, right?

I know you well enough to say that you are not driven by money. You told me about your vision that this technology would save lives and give life back to the elderly and people with disabilities. Is it still so?
Yeah, it's definitely not the money. There are many ways to make money. So that's definitely not something that attracts me in any form! [*laughs*]


thestar.com

Life · Fashion & Style

U of T scientists create software to analyze outfits

New program, which they hope to turn into an app, determines whether an outfit is stylish and offers suggestions.

University of Toronto researchers Raquel Urtasun, left, and Sanja Fidler are creating an app that assesses clothing and recommends how to be more fashionable. (STEVE RUSSELL / TORONTO STAR) | ORDER THIS PHOTO

Archive photo from the Toronto Star with friend and colleague Sanja Fidler.

"I believe that if there is a path to solve this, it's this path!"

For me, I believe I should utilize my skills to provide positive impact in this world. There are many ways to do this. I'm very passionate about self-driving because, as you mentioned, there are many people that don't have access to transport. Quite often people tell me about how this technology will change their life or somebody else's. That's very motivating. At the same time, from a personal perspective, unfortunately I know many people who have had severe traffic accidents, so I'm very motivated to bring safety to our roads. Also, I think that self-driving vehicles will change the landscape of our cities. Less parking, more green spaces, less congestion, shared resources, reduced car ownership. One of the things we've seen in the pandemic is that the supply chain has been very effective. If we're able to transport goods so that everybody can have what they need to do what they want to do, I think that will be fantastic.

What should we expect to see next from Waabi?

As I said, we're focusing now on transportation of goods. That's the first thing that you will see from us. Again, it's about bringing this technology simply to the world and making a positive impact. But Waabi has a lot of ambitions meaning that the work doesn't stop there. We believe that we can do so much more, but we need to focus. That's why we are heads down on this first use case, but the technology that we are building generalizes across use cases, which is very important for the future after that.

In these last four years, do you think the panorama has become longer or shorter than you thought? Is the work you're doing with Waabi going to make this horizon even shorter?

That's a fantastic question! The more that I work on this domain, and in particular having industry experience and really understanding what everybody is doing, where they are going, their technology, etc., I can see that there has been a change of narrative. It's not just about what I think, but everybody has finally acknowledged that this is an extremely hard problem to solve and it's harder than people say it is. The narrative has changed from the idea that we will have this at scale imminently to a realization of the difficulty. From that point of view, for now there are much longer timelines. It's not that anything has changed, it's simply that there is more public recognition that that's the case. I believe with what we are doing at Waabi, we can commercialize this technology at scale much faster. So

yes, progressing the timelines but what is more important is that you should never compromise safety. We want a safer solution. The safest solution out there. Deploying it as soon as we can but never compromising safety.

"Everybody understands this technology can really change the world as we know it today!"

What do you want the general public to know about Waabi? Obviously, you have to direct your commercial enterprise towards specific goals, but what would you like everyone else to know?

[*Raquel pauses*] This is a very, very interesting question. I guess what I want the public to know is that self-driving is going to happen and that it's going to have a positive impact on all of us. That Waabi is going to be a very important part of the solution to self-driving, that we are working really hard on building that solution, and that they will see one way or another how it will positively impact them once it's ready.

The AI and computer vision community used to all meet up regularly at conferences in different places around the world, but because of the pandemic

we are all separated from each other for now. Do you have a message that will help the community feel a little more together again?

Let's see. [*pause*] I guess one of the most important things to know is that despite the pandemic, the world is full of opportunities for everybody that is reading this interview to shine. The world looks harder now in that you can't just go to a conference and directly see people. You can talk to people, all the same opportunities are still there, but you have to be more proactive. This community is fundamental for solving such an important problem as self-driving. When I was in academia I wrote

many papers, received many awards, was very successful, which was great, but I wasn't making a direct impact by bringing this technology to life. The moment I joined industry it was a game changer. I didn't leave research – we continue to publish and share with the community – but in the sense that not only are we solving difficult problems that we find challenging and are excited to solve, but we're actually going to change the world! This is such a unique thing. I hope more and more people will join us in our quest to solve self-driving.

"Despite the pandemic, the world is full of opportunities for everybody that is reading this interview to shine!"

Canadian Business Magazine, Studio Wyse, photo Miguel Arenillas.

Do you have any final thoughts to share?

I believe it's very important if you want to solve such a complex task that you have a diversity of perspectives because if we all think the same, we won't challenge each other and we won't build the best technology. It's by having different opinions and sharing those opinions and discussing and debating one another that we build better and better technology. We need lots of new ideas. This is why research is fantastic. We need a diversity of backgrounds to come up with different ideas. We need diverse teams to be able to provide that and we need to serve the diversity of our customers. That's the other bit that is very new and not really talked about. This is a very non-diverse field and that needs to change.

A young girl from the Asturias and a little guy from Northern Italy are talking about the future of the world: it means that opportunities are not missed!

[*both laugh*]

Thank you so much, Raquel, for a wonderful interview. You gave us diamonds!

Thanks Ralph. My pleasure as always to see you and to discuss with you!

Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows

Han Hu is a Principal Researcher in the Visual Computing Group at Microsoft Research Asia.

His work proposing a new general-purpose backbone for computer vision has just won the Marr Prize for Best Paper at this year's conference.

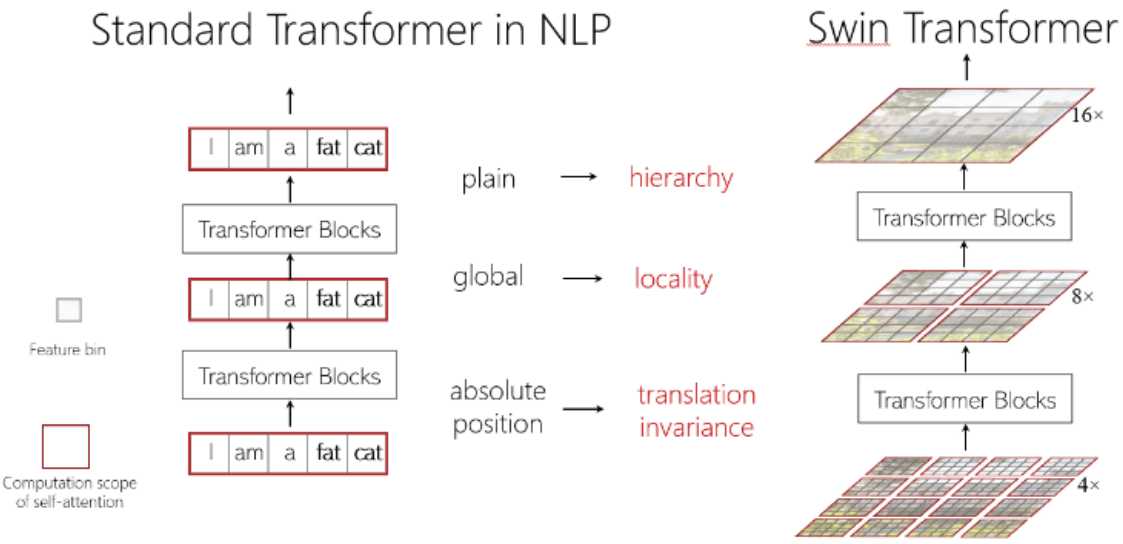
Huge congratulations to Han and his colleagues on taking home this prestigious award! He spoke to us ahead of his live Q&A session and before receiving the award.

The CNN backbone has dominated the field of computer vision for 30 years. Recent works have applied Transformers to a computer vision backbone for certain tasks, but an important question has not been answered – **can Transformers be a general-purpose backbone for computer vision**?

Transformers have been used in **natural language processing** (NLP) for years but are starting to show good results in computer vision. Their novelty is a self-attention layer rather than a convolutional neural layer.

This paper proposes a new architecture based on Transformer, **Swin Transformer**, and demonstrates that it performs much better than CNN backbones. It seeks to prove that this architecture can be applied to many different computer vision tasks.

To make Transformer work well in computer vision, Han and his colleagues looked carefully into the key differences between visual and text signals. They found three priors which are fundamental to vision signals and introduced them in this architecture: hierarchy, locality, and translation invariance. There is also a crucial design to make it practical in speed, named non-overlapping **S**hifted **win**dows (where the method name is from). It is much faster than the traditional sliding window approach due to its friendly memory access.



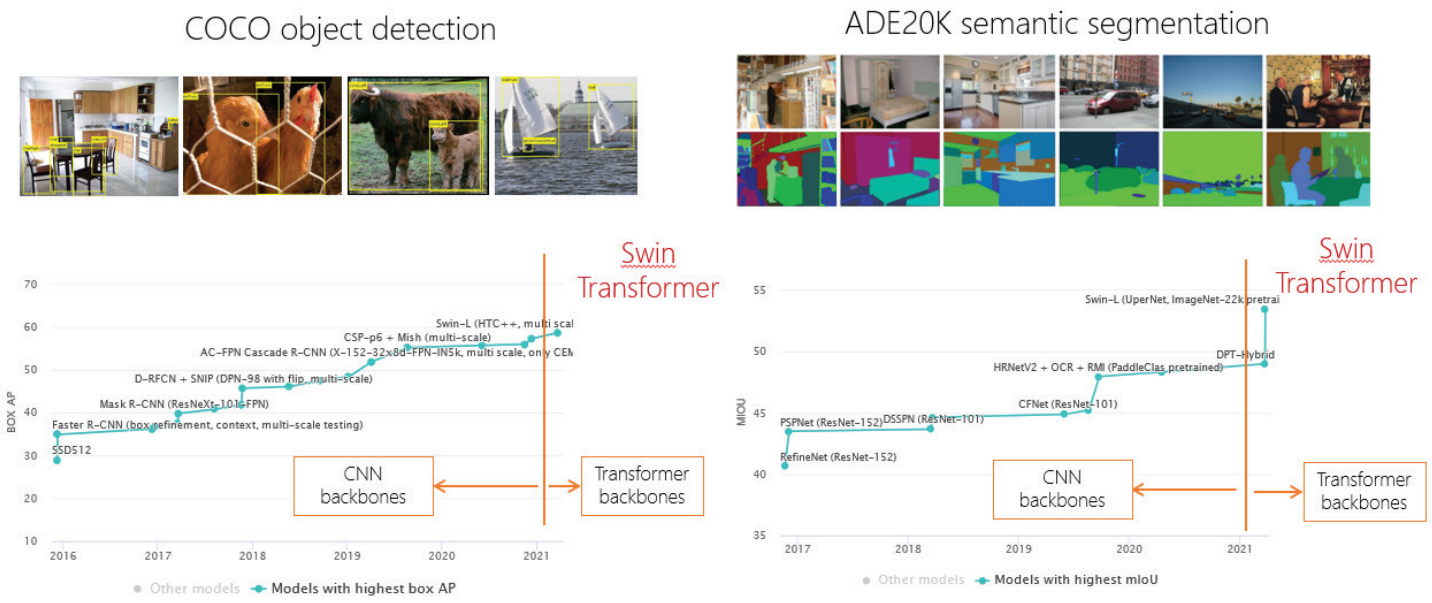
Top: from NLP Transformers to Swin Transformers (vision)

Bottom: a key design towards fast local Transformer computation by non-overlapping shifted window

"*We are at the start of an exciting new era for computer vision,*" he tells us.

"*Before this paper, I heard different voices about the future of Transformers in computer vision. Some said that Transformer performs well for some vision tasks but has a long way to go to be really popular in computer vision. Others said Transformer has a lot of potential and it should get more attention from the community. We are definitely in the second camp!*"

To prove that point, Han and his team showed Swin Transformer achieved strong performance on two important computer vision benchmarks: **COCO object detection** and **ADE20K semantic segmentation**, setting a new record on COCO and ADE20K. They demonstrated that the Transformer backbone can be around three points better than previous CNN backbones.



Han says one of the motivations for this paper is to unify the architectures of NLP and computer vision, and he and the team hope Swin Transformer's strong performance on various vision problems can drive this belief deeper in the community and encourage **modeling convergence** between the two. The next step is whether vision can have very big models, like GPT-3 in NLP, and whether NLP and computer vision signals can be better joined due to converged modeling.

"In previous years, NLP and computer vision progressed in parallel, but now because Transformers are used in both NLP and computer vision, they have the potential to be really joined together…"

"*In previous years, NLP and computer vision progressed in parallel, but now because Transformers are used in both NLP and computer vision, they have the potential to be really joined together,*" he tells us.

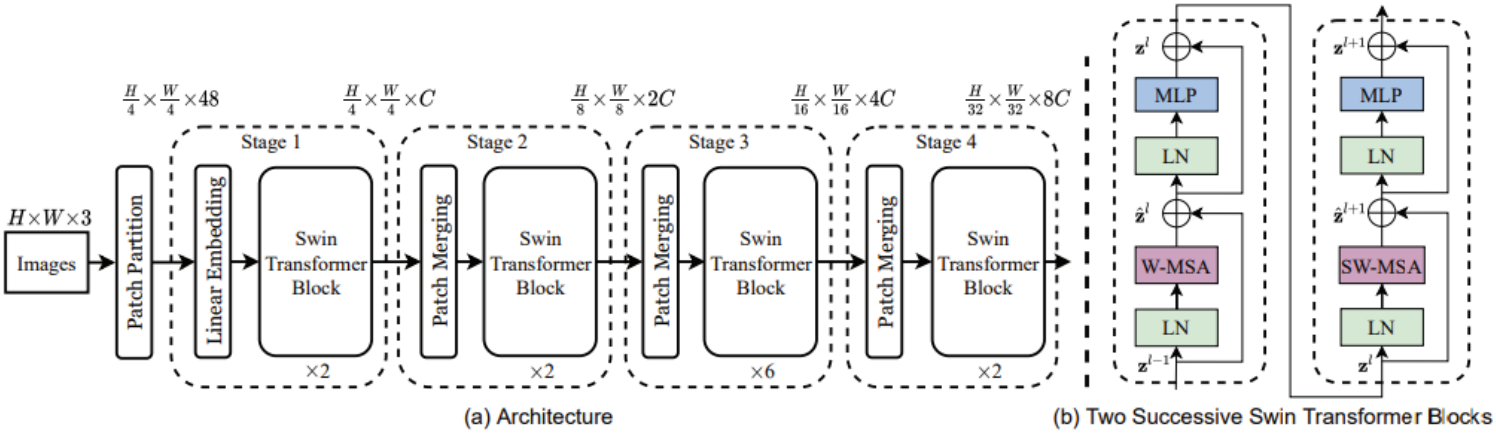
"*In previous years, computer vision tasks handled, for example, 100 categories, but now because NLP and computer vision can be better joined together, we can deal with almost all concepts for visual signals. I think this could be a big change and we want to study in this direction.*"

Han's work at **Microsoft Research Asia** involves research on how to build vision for **artificial general intelligence (AGI)**.

"Microsoft Research give us freedom. We're not limited to specific tasks – we can work on anything we're interested in," he reveals.

"As we're given so much freedom, we want to identify what the most important things to explore are and really push those. For now, we're mainly focused on two things. One is how to create a proper architecture for computer vision which is scalable – this is what this paper is about. The second is about the future way to do learning. There have been several promising ways, including self-supervised learning, classification-based supervised learning, and multi-modality learning, but their relationship and what the best practice for learning is unknown. We want to study this direction."

If Swin Transformer could help you in your own vision tasks, you can find the source code and pre-trained models on Github.



(a) Architecture (b) Two Successive Swin Transformer Blocks

"There have been several promising ways, including self-supervised learning, classification-based supervised learning, and multi-modality learning, but their relationship and what the best practice for learning is unknown. We want to study this direction."

Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields

Jon Barron is a staff research scientist in the Perception team at Google Research.

His work, which he describes as essentially a bug fix for NeRF, has been accepted as an oral presentation, and received an Honorable Mention in the Paper Awards. Big congratulations to Jon!

He spoke to us ahead of his live Q&A session and before receiving the award.

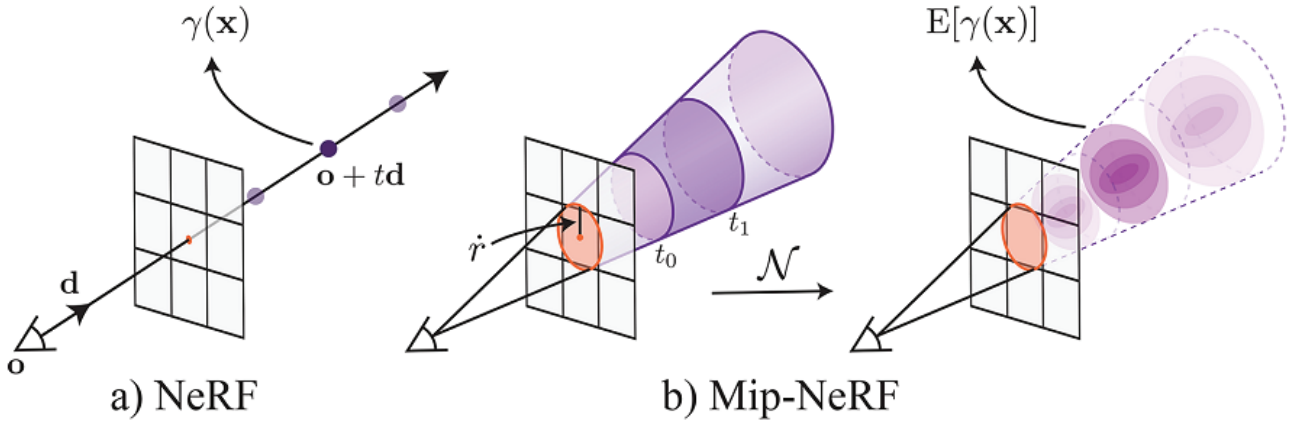
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis **was published at ECCV last year** and Jon tells us even as they were finishing that paper, they knew there were still several things on their list to solve. One of those was a critical problem NeRF had with aliasing.

"*NeRF works by casting rays from the camera out into the world and then sampling discrete points along that ray and pushing them through a neural network,*" he explains.

"*This is generally a good idea, but it has a problem. You are shooting rays that are infinitely small. They are perfect rays, they do not have any thickness to them, and they fly through the air. Then you sample individual points along those rays. Those points are completely singular points. They have no size associated with them.*"

Like a pixel? "*A pixel isn't really a point;* ***a pixel is a box***," Jon responds.

"*A good way to think of a pixel is that the value of a pixel is the integral of everything that hit that pixel. This is a critical thing for people working in graphics and image processing to consider. You can't treat pixels like point estimates, you have to treat them like integrals over a little box to avoid aliasing and jaggies – bad looking images. In NeRF, we didn't do that, we treated the pixel like a little point. We shot this little ray out that had no thickness and didn't tell the ray it came from a small box, then when we would sample points along the ray, we would treat those points as if they were infinitely small. Maybe there's something inside the pixel, but the ray*



we shoot just happens to miss it, we don't see it at all, then the next pixel happens to hit it, and you get these bad aliasing artifacts."

Their solution is to rethink the image formation process. Instead of shooting rays, they start **shooting cones**. The idea being that a pixel is a little box, so you define a circle in that box and shoot a cone out that gets bigger as the cone gets further away. Instead of drawing points from a ray, they draw conical sub frustums. They take the cone and slice along its axis and then have these thick volumes of space. They want to featurize the entire volume, instead of just featurizing the center of it.

"*You can work through the geometry for how to cast cones, and how to cast rays, and it's not too hard,*" Jon tells us.

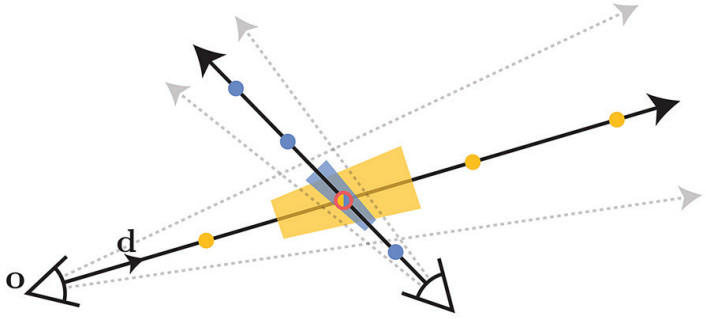
"*This is all stuff we understand well from geometry. But then there's this new problem, which is I now have this conical sub frustum in 3D space, and I need to featurize it as input to a neural network. We didn't know how to do that. We wanted to give it information about the size of this location, but we didn't want to do it in a naive way because what we want is to featurize this area of space so that the feature is invariant to small changes within that space. We want to reflect the fact that if the area is very large, then we don't know about the high frequencies of the location, and if the area is very small, then we confidently know where it is in terms of location.*"

Their solution here is to think about the positional encoding used in NeRF. In NeRF, these coordinates are encoded with what's called **positional encoding**. This is terminology from the transformer literature, but it is a straightforward idea. You take the coordinate, and you take its sine and cosine, and then sine and cosine twice that coordinate, and four times that coordinate. You scale it up and pass through some sinusoids, so you end up with a feature representation that has sinusoids at lower and higher frequencies. Those frequencies tell you where it is. You can think of it like an addressing scheme in a quad tree or an octree. The first sinusoids tell you if you are on the left or the right, and then the next sinusoid tells you if you're on the left side of the right part or the right side of the left part and you descend into a tree that way.

"For big regions, we didn't want to descend all the way down into the tree," Jon explains.

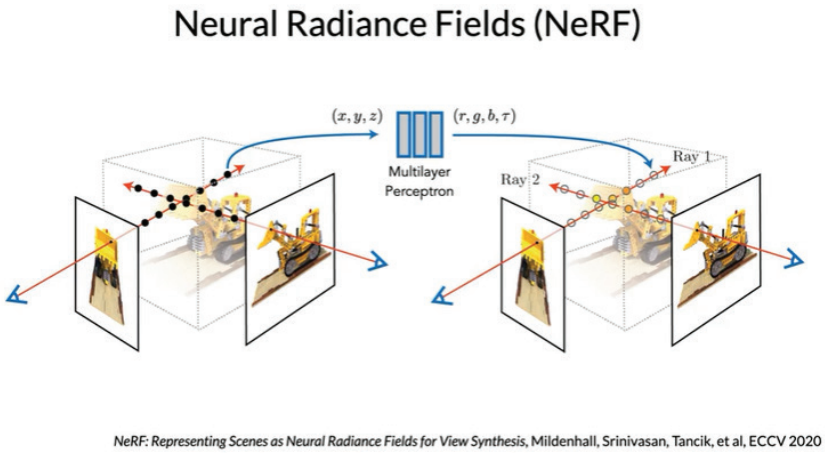
"The way we got the math working for this is we said we have this volume of space, and we don't want to just featurize one point in that space, we want to featurize the entire volume, and the way we're going to do that is we're going to marginalize over the volume. We're going to integrate out the entire volume of space. We compute what we call the integrated positional encoding. This is the expected positional encoding as you integrate over this volume. It is a hard thing to do, but if you approximate this conical frustum as just a standard multivariate Gaussian, this integration has a closed form."

The result is a nice map that simplifies out and you can say I'm going to featurize this conical sub frustrum using these positional encodings. The math simplifies greatly. If the volume is very big then the high frequencies get zeroed out, and if the volume is very small, the high frequencies stick around. The neural network that then takes these features has a way of saying, 'I know I'm exactly here,' or it can say, 'I know I'm somewhere in this whole space, but I don't know where.' The model can reason about scale naturally. It knows if something is big, it knows if something is small, and importantly, if something is big, it doesn't have false information about where it might be at these higher frequencies. They just don't exist. The frequencies get completely deleted.



"We were able to show an improvement on the kind of normal scenes people use NeRF for," Jon *reveals.*

"The main improvement is that NeRF would fail catastrophically in certain circumstances and this was whenever there was significant scale variation. If I take a bunch of pictures of an object, in NeRF we would always take these pictures from the same distance. That worked great because whenever we saw something it was at the same distance, so NeRF's issues with aliasing and scale didn't pose a problem for us. But then if we started moving the camera in or out, this became a big problem. If we zoomed the camera in too far, we were shooting these small rays that corresponded to small pixels but the model didn't know the pixels were small so we would see ghosting and weirdness. If the camera was far back, we'd shoot these rays that corresponded to big pixels, but it didn't know the pixels were big and we'd get bad aliasing. You can see this in the video results for the paper."



NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall, Srinivasan, Tancik, et al, ECCV 2020

Jon tells us he has seen evidence that NeRF is a kind of general-purpose tool for many things outside of computer vision, from computer graphics to robotics. The system is designed to take in images and to produce models from that, which is the essence of computer vision, but the techniques can be applied elsewhere. Everything this paper does is about the geometry and the optics of image formation, but then it is signal processing and linear algebra from that point onwards.

Thinking about next steps, Jon tells us there are still limitations to NeRF. The models are slow and making them fast is an ongoing problem to solve. Also, NeRF bakes in all the lighting effects, so you can move the camera, but you cannot change the light or the textures. In the future, Jon would like NeRF to be a **general-purpose inverse rendering tool** that can do relighting, ray tracing, material editing, and all these things you would expect from sophisticated graphics engines.

"*This I feel is step one of at least a 10-step process of modernising NeRF,*" he adds. "*NeRF is simple in many ways. It uses primitive technology, ray tracing, and fully connected neural networks. Everything we've been doing is from the '80s and we're slowly bringing NeRF into the '90s and the '00s by adding more sophisticated things that are well understood in graphics but haven't ported over to computer vision and deep learning. We're still on aliasing, which is a simple concept, but making these things neural is challenging.*"

Jon tells us his tastes have always leaned towards older papers from the '80s and '90s.

"*Things were simpler back then,*" he adds. "*People were focused on pixels and low-level things, which allowed for simple mathematical treatments and elegant algorithms. I'm excited to use my knowledge and enthusiasm for old computer vision and image processing papers and apply them in a modern context.*"

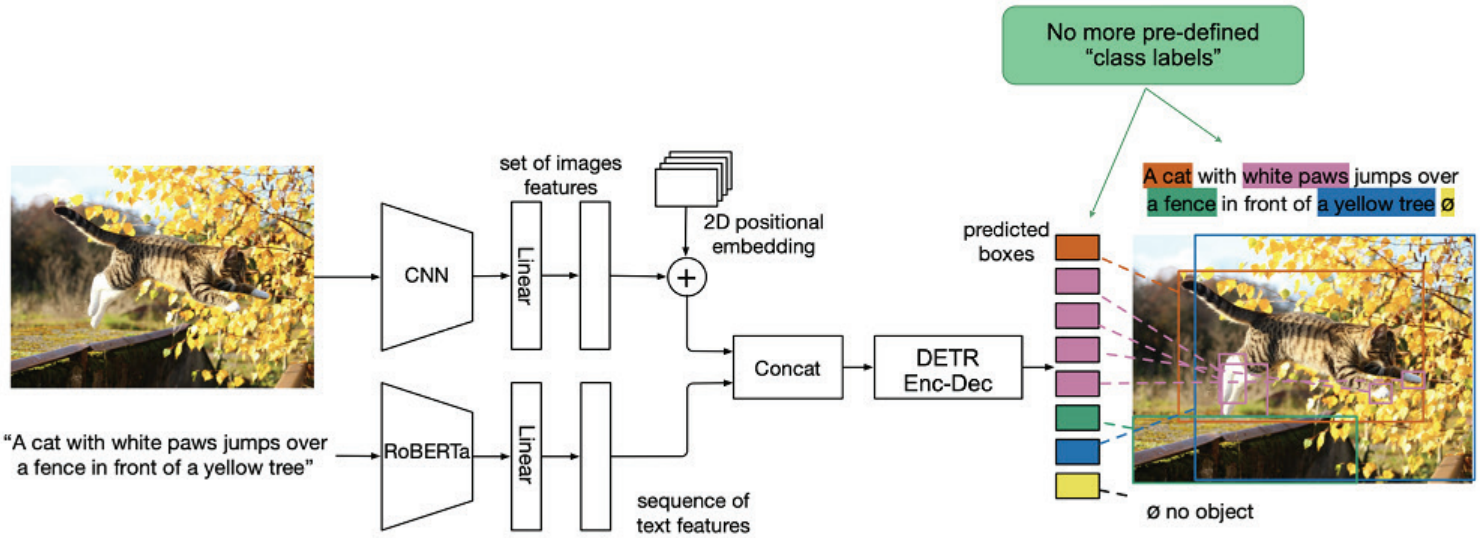
To learn more about Jon's work see the project page.

MDETR – Modulated Detection for End-to-End Multi-Modal Understanding

Aishwarya Kamath is a PhD student at New York University, under the supervision of Yann LeCun and Kyunghyun Cho. Her work, which has been accepted for an oral presentation, proposes a novel multi-modal approach to understanding images and text. She spoke to us ahead of her live Q&A session.

The main idea of this work is that when we want to have some understanding of what is in an image, we should not be bound by what an object detector can find.

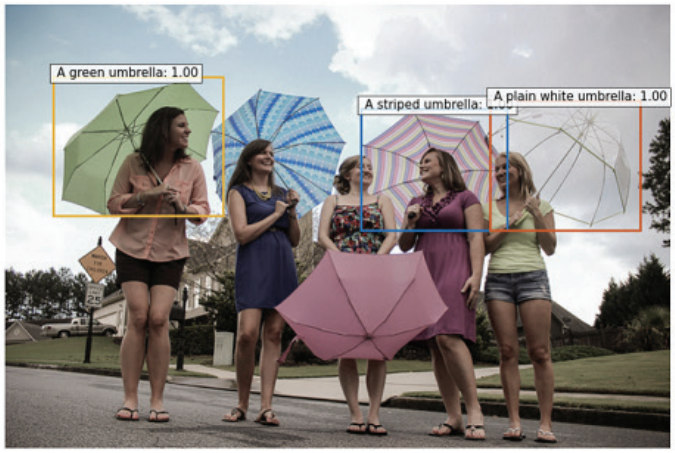
In previous work, most object detectors are trained using **a fixed label set**. That means if there are 1,600 classes the object detector can find, once it finds those classes, the downstream multi-modal understanding uses one of those detected objects to reason over open-ended questions like, what is on the table? If the object detector does not have this class in its fixed label set, it will no longer be able to find it and the downstream reasoning will fail.



*"To solve this, we propose a new method called **modulated detection**,"* Aishwarya tells us. *"The idea is you can find anything in an image if you can describe it in the text. In the paper, we have this example of a **pink elephant**, which is not something that you ever see in the real world. The model can leverage compositionality of language and is trained to detect whatever you speak about in the text in an end-to-end manner. It can put together the fact that it knows what pink is and it knows what an elephant is to find things like a pink elephant. It can find many more things than would be possible with a fixed classification based on a fixed label set."*

The end-to-end part is important because having the object detector in the loop means detection is a main part of the model. It trains the features in conjunction with the downstream tasks. Without this, you could have many features that are not relevant to the task which the model has to learn to ignore. This paper proposes giving the model exactly what it needs. By only detecting objects that are relevant to the query, it performs much better.

*"The biggest takeaway for the computer vision community is that we push for no more fixed class labels and instead just use **free-form language**,"* Aishwarya advises. *"This is applicable for any pure computer vision task."*



"A green umbrella. A pink striped umbrella. A plain white umbrella"

"A car. An electricity box"

One challenge for Aishwarya and her colleagues stems from the fact that these models are so large and so they do not have many iteration cycles because it takes such a long time to train them.

*"I think that was the most stressful part because the results in the paper all depend on **one large run of the model which takes more than a week!**"* she reveals. *"Everything has to be perfect before that week because **you only have one shot at it**. It takes a lot of compute, and a lot of money goes into it,*

so I think that's the most difficult part to make sure that one run is perfect for these kinds of large models."

The idea for the work was a joint one between Aishwarya and **Nicolas Carion**, a postdoc in her lab and the last author on the paper. His paper, **End-to-End Object Detection with Transformers** – which was an oral at **ECCV 2020** – presented a new object detection framework using transformers called **DEtection TRansformer**. It proposed object detection in an end-to-end way that no longer required non-differentiable components, like non-maximum suppression. In previous works, it was not possible to train end to end because the gradients do not flow through these non-differentiable components.

"*This work started in a slightly different way,*" Aishwarya discloses. "*We started out working on extending DETR to take in multi-modal input. Then after a few months of trying simpler versions, we came up with this novel approach of doing it through modulated detection.*"

There have been a number of **multi-modal understanding papers** in the last couple of years since transformers began to be used everywhere, but with this paper taking a different approach to the others, it might just be what caught the **ICCV reviewers'** attention.

"*You can see on many benchmarks that we show a huge improvement, even compared to papers that came out one or two months before us,*" Aishwarya points out. "*On one of the benchmarks, we made the error rate half of what it was before! We had more than five points increase on the referring expressions dataset. I think that will have convinced them because we offer a new approach to the same problem, and it clearly works very well.*"

(a) Query: "street lamp" (b) Query: "major league logo" (c) Query: "zebras on savanna"

Aishwarya reveals she is already working on a follow-up paper for CVPR next year. Can she reveal any details at this early stage?

"*We're trying to learn how to train our model with less supervision than we have in this paper because even though it worked really well as a method, it still requires quite strong supervision, so one direction would be to try to*

"He was happy about that and recognized that there's a lot to be explored in this area..."

reduce that. The other direction would be to give it generative capabilities so that's basically the direction I'm working on now."

We have to ask Aishwarya what it is like to be advised by Yann LeCun. Did she feel a lot of pressure when she first found out?

"*Of course!*" she laughs. "*When I went for my NYU interview, I had an open mind about all the professors that I could end up working with, but then I had an interview with Yann, and it went very well. He told me he was planning to take a PhD student and asked, 'Do you want to be my PhD student?' I was like, 'Whaaat?' Yeah, definitely I was a bit stressed in the beginning! But once I joined and we started working together it went away.*"

What is he like to work with?

"*He's very chilled and very easy to talk to. He's patient and doesn't push me to work on things that he's interested in, but more supports me to explore the things that I want to explore. In general, text is not something that he's used to working with, he's more of a computer vision person, but we discussed that my interests in the past were mostly text-based, so I would work at the intersection of language and vision. He was happy about that and recognized that there's a lot to be explored in this area.*"

On that subject, Aishwarya tells us she will be attending the 4th Workshop on Closing the Loop Between Vision and Language on Sunday.

Recently, Aishwarya has been working at **Google AI** as an intern on different projects, but still in the same space of **image-to-text understanding**.

"*For my PhD, multi-modal understanding is the kind of space that I'm interested in,*" she says. "*I'm trying to get experienced with many kinds of applications for this work and trying to figure out how I can use the technology to help people at some point. That's my end goal.*"

Long-Term Temporally Consistent Unpaired Video Translation from Simulated Surgical 3D Data

Dominik Rivoir is a PhD student at the National Centre for Tumor Diseases, under the supervision of Stefanie Speidel.

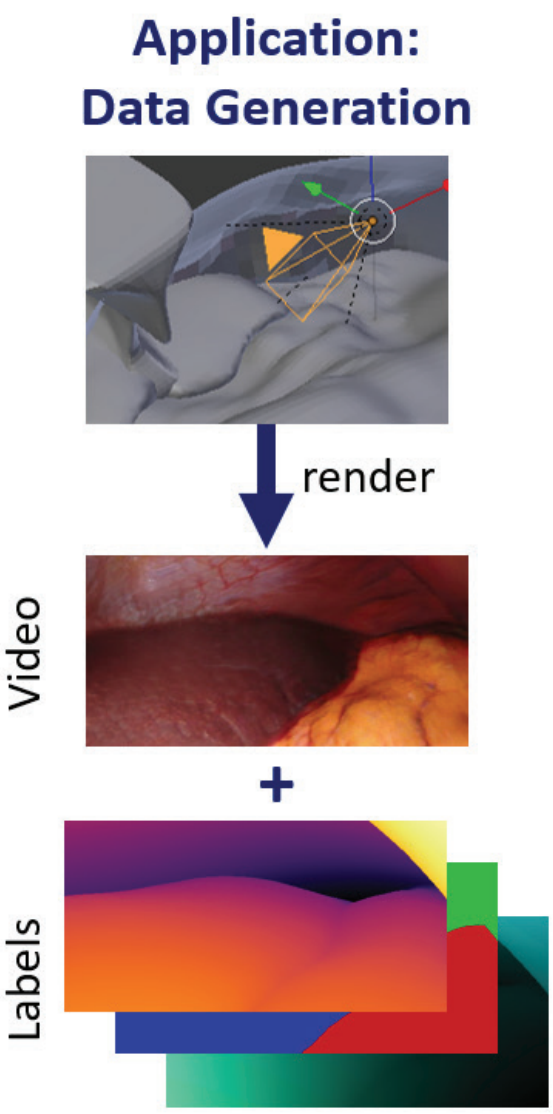
His work proposes a new approach to providing simulated 3D data in a surgical setting.

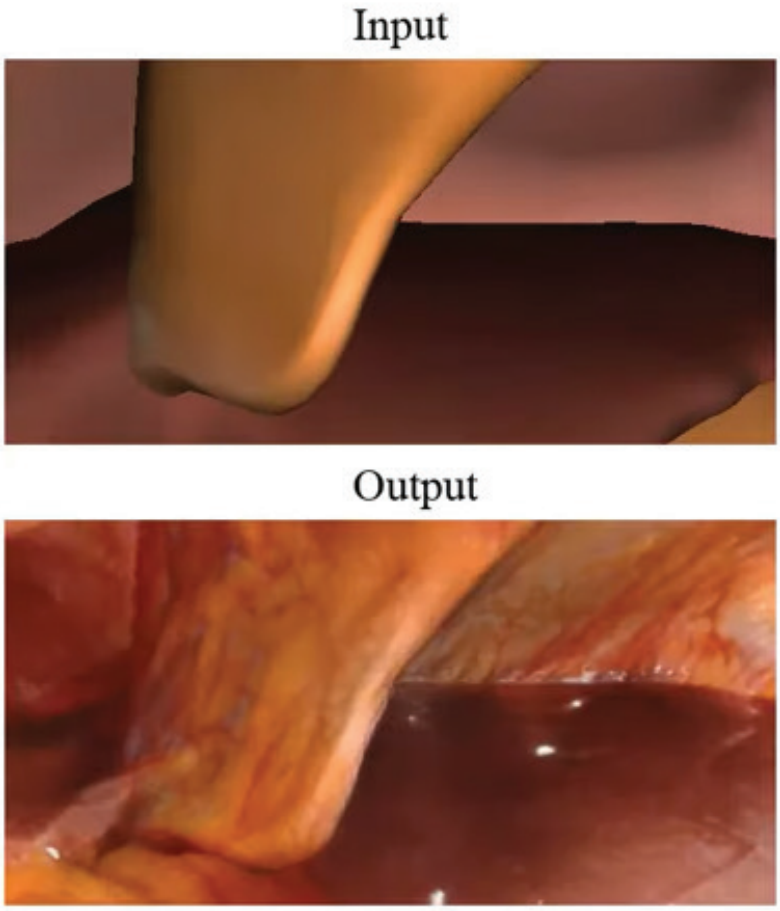
It has been accepted as a poster this year and Dominik spoke to us ahead of his live Q&A session.

Computer-assisted surgery involves **analyzing video feed from the surgery to provide feedback**. This may be to **identify structures**, like organs or unsafe regions the surgeon should stay away from, or to **recognize instruments** to understand what the surgeon is doing.

One of the biggest challenges in the surgical setting is the lack of labeled data. Obtaining labeled video data in the surgical setting, such as point correspondences over time, or training 3D information like depth or 3D positions, is especially difficult. This work aims to solve that by providing synthetically rendered, but realistic looking data, with the goal of **achieving view-consistency, or long-term temporal consistency, in the generated videos**.

"*This work is a mix between **GAN-based translation**, which has been a big focus for a few years now, and the more recent field of **neural rendering**,*" Dominik explains.





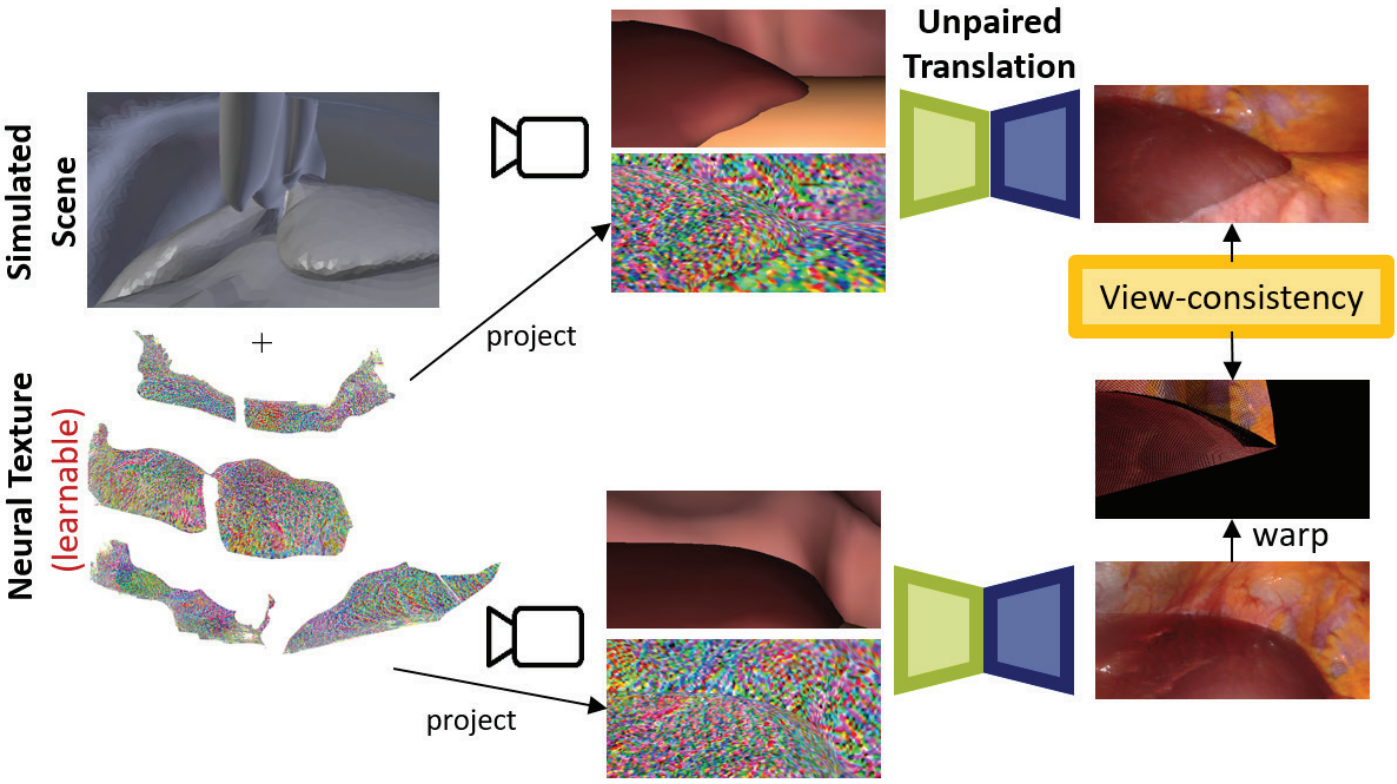
"Neural rendering is where we take traditional rendering pipelines from computer graphics and try to make them differentiable and incorporate neural networks…"

"***Neural rendering*** *is where we take traditional rendering pipelines from computer graphics and try to make them differentiable and incorporate neural networks. We take the 3D information that we have from our simulated scene and try to incorporate that into the learning process.*"

Obtaining the data needed for this was, as ever, a challenge in the surgical domain, as was finding the corresponding **real data and simulated data**. The team had to design a simulated scene that somehow resembled the realistic data. For that, they had to build the scenes first and then extract all the 3D information from that and design the model on top.

"*We use something called a neural texture,*" Dominik explains.

"*We have a 3D representation of the entire scene's texture. In traditional computer vision, at each texture location in the scene you would store something like the color of the object that you want to render, then how it reflects, something like that. But we want to learn these features and not explicitly define them, so at each spatial location on the texture we have a* ***learnable feature vector****, and we project those feature vectors from 3D into 2D space when we want to capture an image. Then we use a neural network to translate these abstract features into an image. Since this whole pipeline including the projection into 3D space is differentiable, we can just learn the network combined with the neural texture together.*"



The neural texture enables the model to **store global information about the scene**, such as the location of vessels, in 3D space and thus these details can be rendered consistently across different viewpoints.

One of the work's main achievements was **combining the two areas of neural rendering and image translation**. In image translation, there has been a lot of work done on unsupervised learning, but in neural rendering, that is not really the case. In neural rendering, people have been doing amazing research in how to make this pipeline differentiable, but they have been looking at the supervised setting. Dominik and his colleagues have been trying to combine these two areas and **make the 3D neural rendering unsupervised also**.

He thinks the work could **have an impact** in the real world sooner rather than later.

"*We're not directly designing an assistance system, but rather we're providing the data to evaluate or to train other systems,*" he tells us.

"*We're providing **an evaluation platform for other methods**. We hope that this way we can help other researchers show their methods are useful. In turn, that will support better translation.*"

In terms of next steps for this work, one issue with the current method the team would like to solve is that **it always has to make some assumptions**, particularly in the unsupervised setting, to make sure that what it's learning

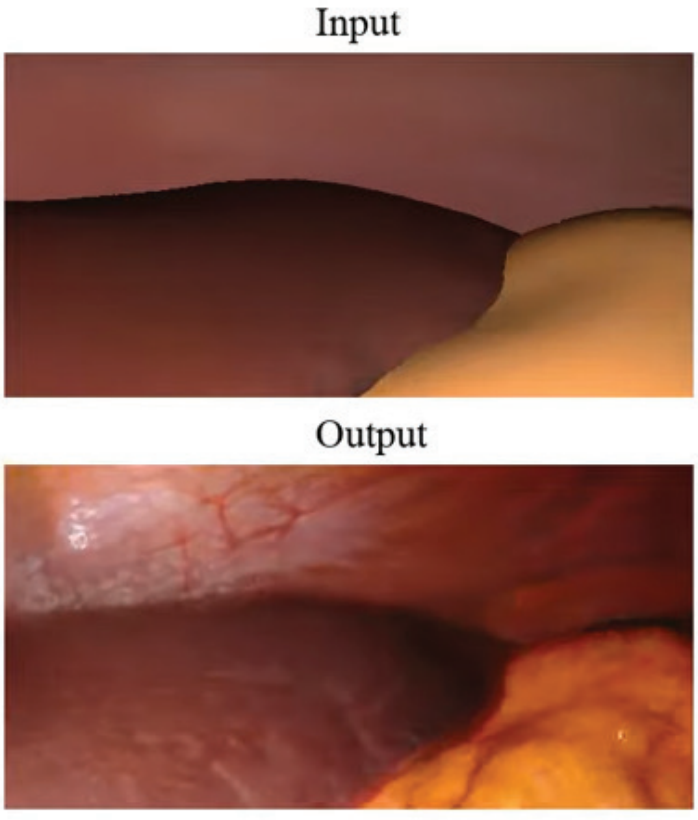
or what the model is translating is useful. The 3D shapes have to be modeled and the model fails as soon as the 3D shape is not realistic. A future direction for the work could be to **look at lessening those restrictions and trying to incorporate learning the 3D shapes** or other assumptions that are being made.

We cannot let Dominik leave without asking him about working with Stefanie Speidel, who regular readers will know is a dear friend of our magazine.

"*It's really amazing to work with Stefanie!*" he smiles.

"*She's one of the pioneers in the surgical assistance field and has so much experience. On the one hand, she knows what's important and can lead you in which direction to go, but on the other, she provides a lot of freedom in terms of which way you want to go. She's always there to guide us and make sure we're going in the right direction. It's a great mix of having the freedom to work on what you're passionate about but also having somebody who has all this experience to supervise it. I'm very happy here!*"

"We're not directly designing an assistance system, but rather we are providing the data to evaluate or to train other systems!



Probabilistic Modeling for Human Mesh Recovery

Nikos Kolotouros is a fifth-year PhD student at the University of Pennsylvania, advised by Kostas Daniilidis, and is currently a research intern at Facebook Reality Labs.

Georgios Pavlakos is a postdoc at UC Berkeley, working with Angjoo Kanazawa and Jitendra Malik.

Their work introduces a new paradigm for human pose estimation and has been accepted as a poster at ICCV this year. Nikos and Georgios spoke to us ahead of their live Q&A session.

Most previous works treat **human pose estimation** from a single image as a typical regression problem, where given an image as input, they try to predict a single pose result.

In this paper, Nikos and Georgios argue that this is suboptimal, since the problem of predicting a 3D pose from a single image is inherently ambiguous. Other works have considered moving away from single predictions, but usually they do it within a multiple hypothesis framework, where instead of a single output, they have multiple different outputs. Usually, the number of outputs is fixed in this case. This work does not output multiple hypotheses, but rather regresses a distribution of plausible poses given the input image.

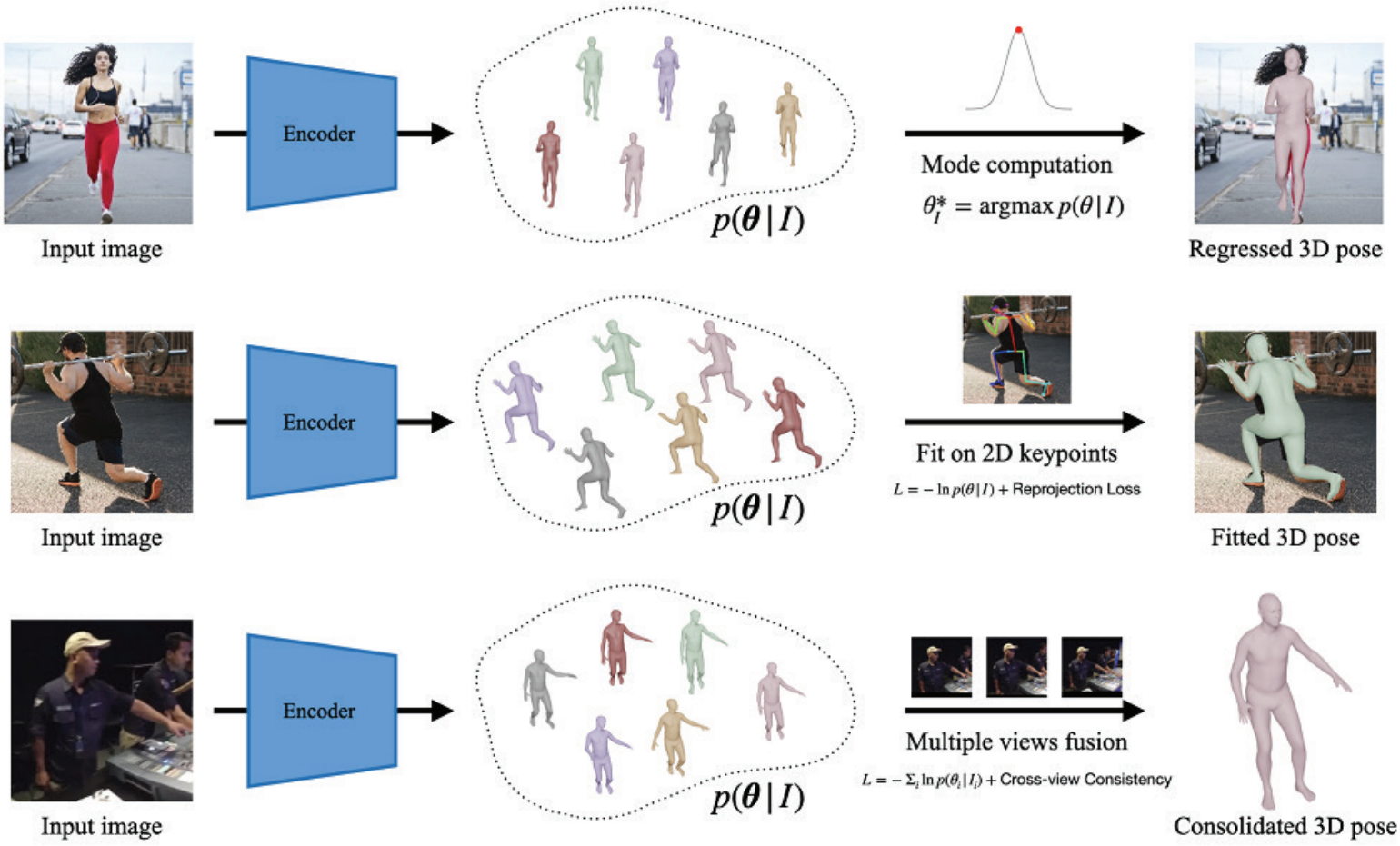
The key takeaway message from the paper is that by taking this one step further and showing that this learned distribution is indeed useful for a variety of downstream tasks, it can combine information from multiple sources and generate more accurate predictions.

"*The most difficult part of the work has been using these probabilistic models, normalizing flows for instance, in practice,*" Nikos tells us.

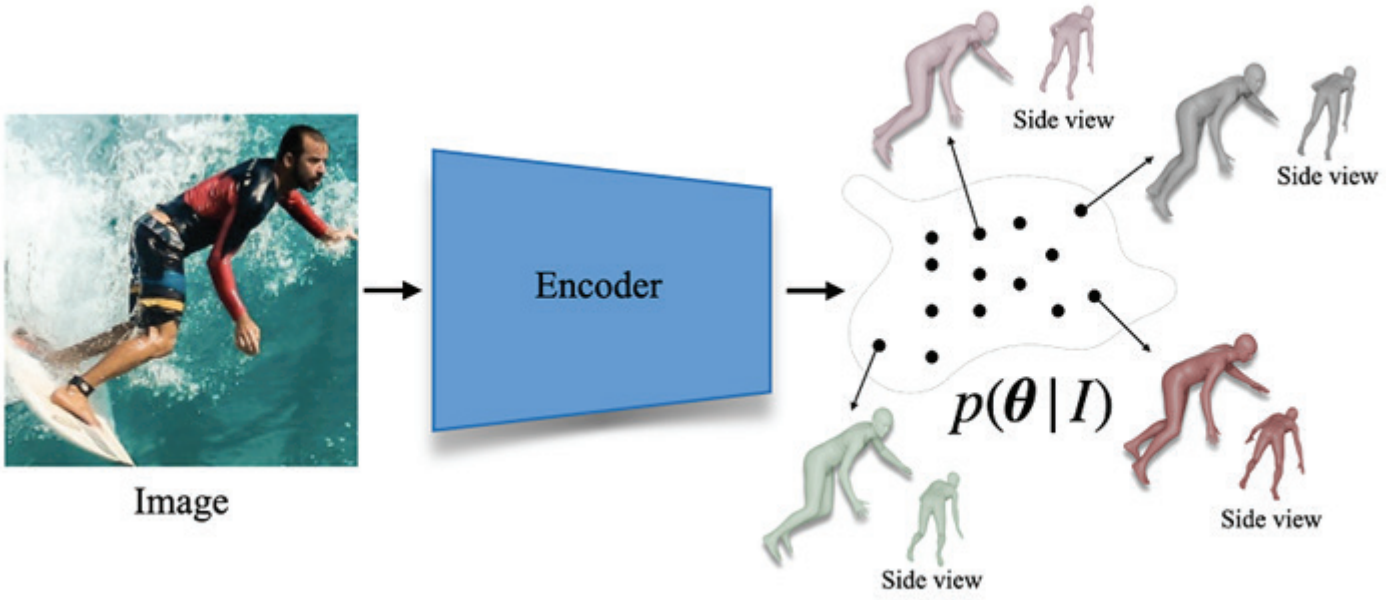
"*We were very familiar with the human pose estimation task, but the challenge was to incorporate this probabilistic model in there. It's different to read about them and what they do in principle. Making them work in practice takes a lot of experimentation.*"

"*I totally agree with Nikos,*" Georgios adds.

"*The problem of human mesh recovery is something that we are quite familiar with. We had some previous works recently, so we knew what was hard and what was interesting to do in that direction. However, using this toolbox of normalizing flows and all the advantages that it has, while it was very appealing to us as an idea, it was really hard to integrate properly. All credit goes to Nikos for making this work so nicely.*"



"We were very familiar with the human pose estimation task, but the challenge was to incorporate this probabilistic model in there. It's different to read about them and what they do in principle. Making them work in practice takes a lot of experimentation.



The work builds on recent advances in computer vision. It makes use of the **deep neural networks** that were designed for image classification a few years ago and capitalizes on that knowledge to adapt them to work for the problem of **human mesh estimation**. Neural networks do not work out of the box if you try to apply them to a new computer vision problem, especially a 3D computer vision problem. There is a need to incorporate domain-specific knowledge about 3D human mesh estimation.

"*We had been inspired by the previous computer vision works in this area and that's how we got the courage to work on this problem,*" Georgios reveals.

"*We took inspiration from those and tried to combine and capitalize on previous ideas both from a mesh recovery direction and from the normalizing flows literature to get the best of both worlds.*"

The work is one step towards achieving more accurate reconstruction from images, not only for 3D human mesh recovery, but much more generally, and people could apply this framework to other subfields in computer vision and machine learning.

We ask Georgios to tell us one thing that he has learned from working with Jitendra and Angjoo.

"Just saying one thing would be an understatement!" he laughs.

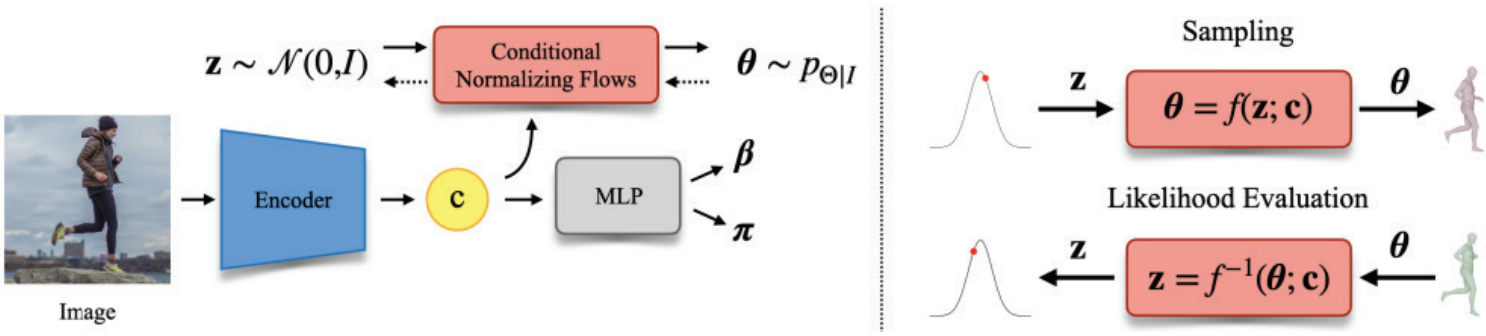
"Both are people I had huge respect for before coming here, which is why I was so excited to work with them in the first place. When we are talking about a problem, Jitendra has a fascinating way of visualizing the plan of what we need to do, not just in the immediate future, but in the few years ahead. He transfers this vision and his passion to you. That's something I really admire about him, and it has changed the way I think when I'm doing research now."

Nikos adds:

"I haven't met Jitendra in person, but I interacted with Angjoo at a poster session at ICCV 2019. I was really impressed when she asked key questions about our method that we had in our poster. It felt to me like she had a very deep knowledge of the subject."

Georgios continues:

"Angjoo is closer to my age and what has been very impactful is how she sees things through my prism. Her research advice has been super valuable to me, but her career advice on how to move forward as a young scientist is something that has greatly benefited me too. Also, her ***endless enthusiasm*** *is something that really captivates you and gets you excited about the work you're doing!"*



"These guys are golden. I was delighted that they were both interns at MPI and I'm always excited to see their new work. They are the perfect combination of nice and smart!"
@Michael_J_Black

Learning To Cut by Watching Movies

Alejandro Pardo is a third-year PhD student at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia, under the supervision of Professor Bernard Ghanem.

His work is the first to tackle an important video editing problem called ranking cut plausibility. It has been accepted as a poster and he spoke to us ahead of his live Q&A session.

Movies often record 10 times more footage than we see in the final cut. For a documentary, it can be 100 times more. So, **for every hour we watch, 100 hours of footage will have been recorded**. This is a big problem for editors.

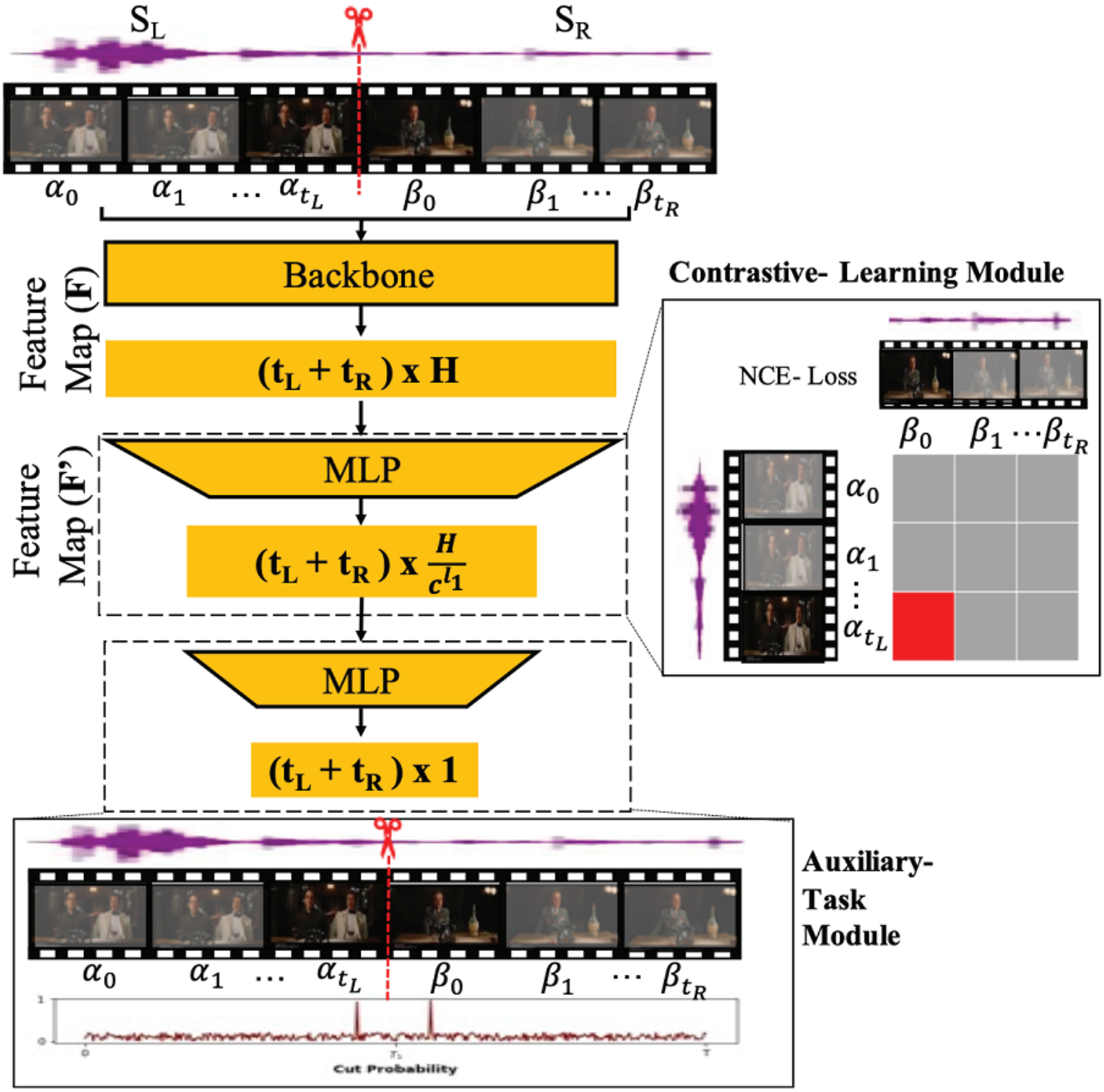
Film editors are presented with several shots taken from different cameras and have to **decide where to cut** in each of these shots to then put them all together to form a scene. **This process is very long and requires a lot of manual work.**

Currently there are software tools that help, but there is no automated process. A human has to do everything. Given a pair of shots from two different cameras, Alejandro and his team propose **the first automated solution to this problem**.

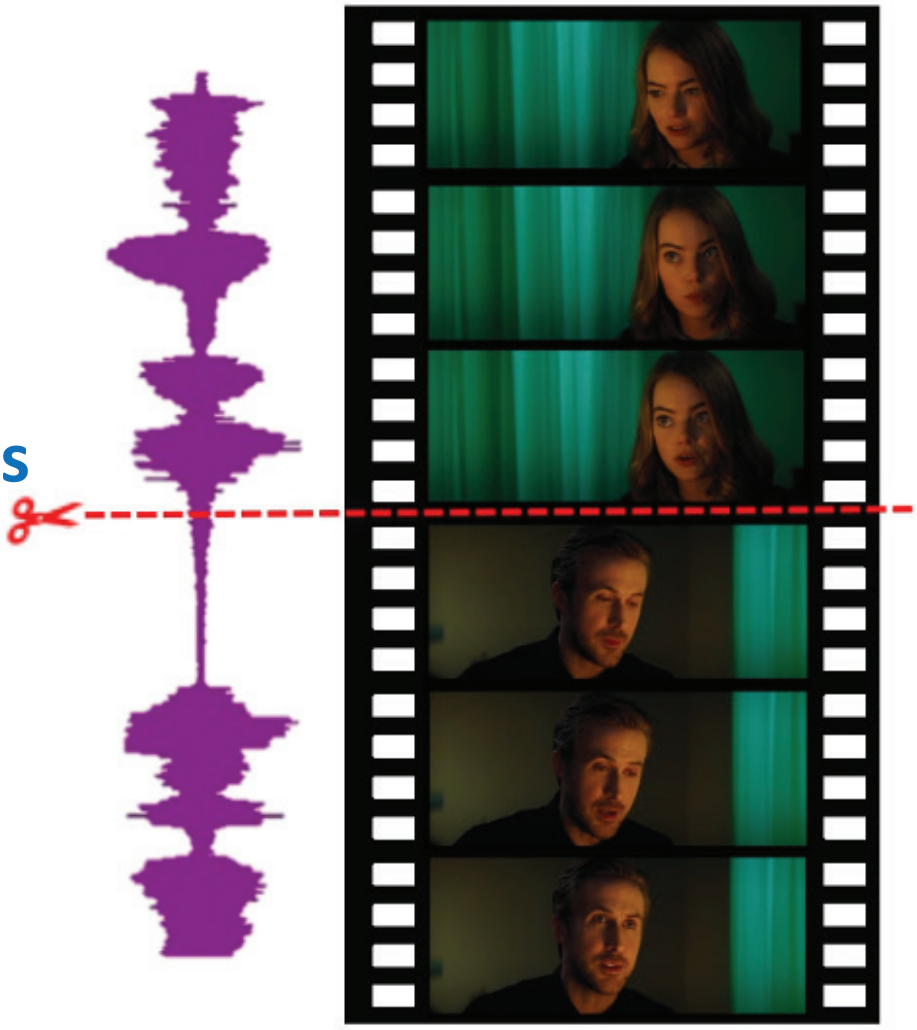
To do this, the team not only need movies, but they need all the raw recordings from the movies as well. Although there are lots of movie clips available on the internet, this **raw footage is harder to find**. Instead, they download scenes from YouTube and use a shot detector, which tells you when there is a camera change in a scene.

"We divide each of the scenes into shots, but these shots are already edited, so we only have the parts that were shown in the movie, not the parts that were cut out," Alejandro explains.

*"To solve this, we create a **proxy task**. That's a task which is similar to a real task, but not exactly the same. Let's say we have a pair of shots. The left-hand side shot is joined with its last frame to the first frame of the right-hand side shot. We're going to imagine that these shots were joined in different places and create artificial cuts. Then we're going to say, okay, give me the real one and all the other ones are negatives. We do this using **contrastive learning**. We create a lot of negatives and only one positive and are trying to learn embedding space that allows us to divide positives and*



Given a pair of shots from two different cameras, Alejandro and his team propose the first automated solution to this problem



negatives. We want things that are together in the original movie to be close to each other in that space, and far from the things that were artificially created."

The team create metrics for the proxy task to understand whether the method is working or not. By proposing several baselines, they show that the method works to solve the proxy task. However, they need to test it on real data. To do this, they use a website called **EditStock**, in which aspiring editors can test their abilities on real movies. It has raw data from professionally produced independent movies, which they can download and use to create edited scenes with their algorithm.

"*It is a difficult problem to measure because two different editors may come up with two different answers and both may be right,*" Alejandro tells us.

"*We decided to ask people whether the edited movies made sense or not, so we designed a human study and the results showed that our method, while still far from being a professional editor, indeed works compared to all the baselines and compared to doing ranking.* ***We have the first step towards a fully automated editor!***"

Now they have conquered the first step, the team are not stopping there. They have plenty of ideas for extending this work. Alejandro would like to explore the types of cuts that are made by editors, which depends on a number of factors, such as the editor themselves, the director, and the style of the movie. Horror movies will tend to employ different cuts to romances, for example.

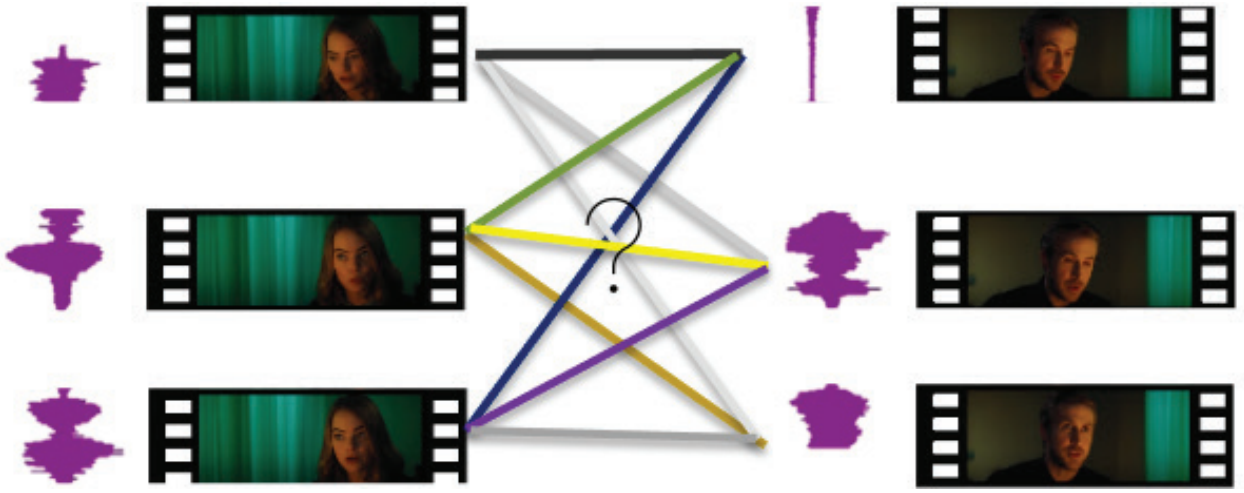
"*If we were able to condition the method so that it's not just general, but can learn some kind of artistic style, that would be awesome and would give us a much, much finer solution to the problem,*" Alejandro points out.

Originally from Colombia, Alejandro ended up studying in Saudi Arabia when the opportunity arose during his master's in 2017.

"*I heard there were opportunities to attend university in Saudi Arabia and I was just saying yes to any option that was out there!*" he laughs.

"*It just happened that Bernard was happy to have me as a visiting student. We developed something super nice. I applied and decided to come here for my PhD because I really like the environment here in KAUST. Life here is very chill and the people are super nice too.*"

As well as this poster, Alejandro's team are running the **AI for Creative Video Editing and Understanding (CVEU) workshop** at ICCV on Sunday, which will bring together people from around the globe who are working on creative video editing and understanding. Everyone is warmly invited to attend what promises to be an interesting day!

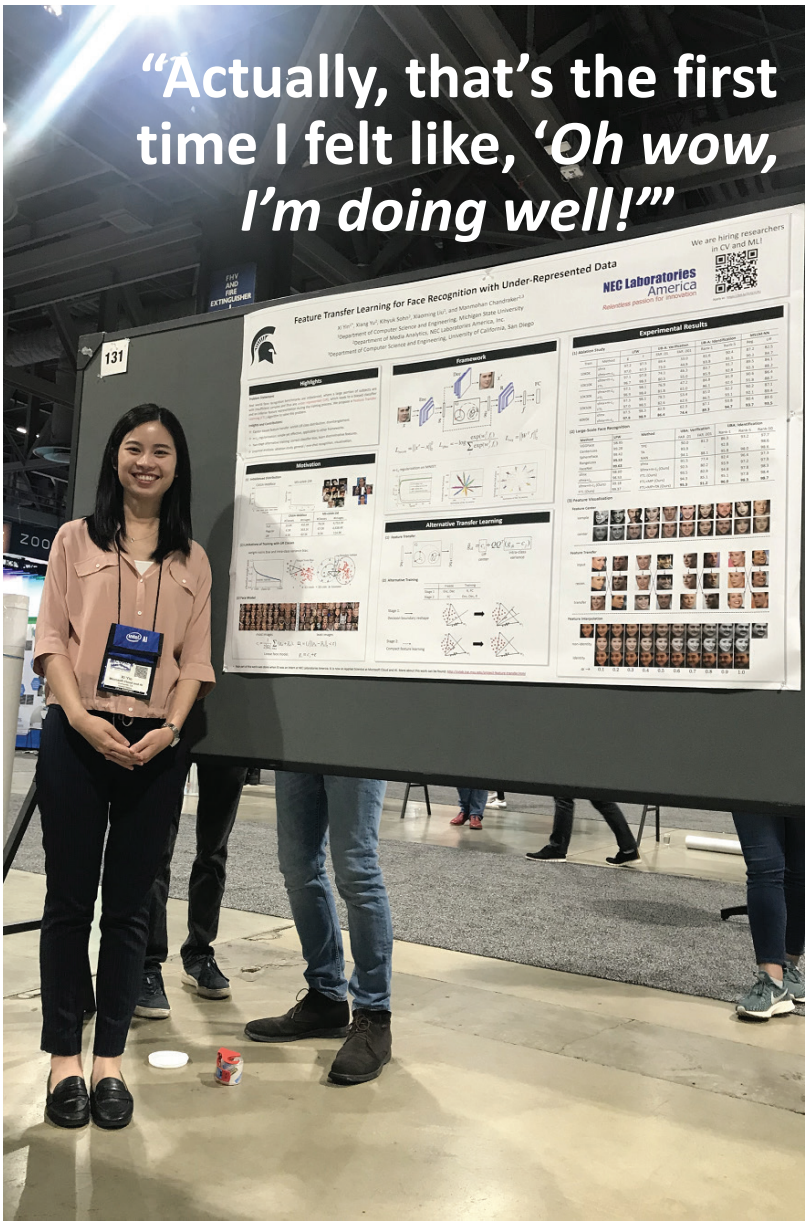


Xi Yin is a Research Scientist at Facebook AI Research. She graduated from Michigan State University.

More than 100 inspiring interviews with successful Women in Computer Vision in our archive

Xi, can you tell us about your work at Facebook?

I'm generally interested in computer vision and machine learning. My research background is in two different areas. One is the face. I worked on facial recognition for my thesis. After graduation, I spent two years at Microsoft and one year at Facebook. In recent years, I started to expand my research area to multi-modal understanding of vision and language. I find that very exciting. That's an area that I'd like to focus on more in the future.



What about this new field attracted you?

Vision and language are the two very fundamental human capabilities. Humans use the interaction between vision and language in almost everything we do in our daily life. Just imagine when the kids start to learn. They read storybooks with pictures and words. They learn to interact with language and how they visualize the world. For a machine to really have intelligence, it is crucial to understand this multi-modal. From a research perspective, vision and language data are easy to acquire. A lot of the images available online have those associated alt-text. That can easily scale to a lot of data compared to the vision supervised training where annotation data is very challenging. In recent years, we have seen that vision language data can help do vision tasks. That is very promising research.

What is the most challenging part of this field for you?

There are many challenges. For example, when we think of images

and videos, usually there is a lot of redundancy. For language, repetition is very sparse. We think: a picture is worth a thousand words, right? We need to use a lot of language to describe an image. When we put these two modalities together, it is interesting how they interact with each other. Using one modality to understand the other helps them to learn better. That's the real challenge.

We can make things even more challenging. You and I speak a lot with our hands. So you can put our hands, and our body language, also into the equation.

That's why video interaction is more than just speech. It all depends on the final application. There are tasks that try to explicitly use the hand gestures to understand the video better. In general, there are many other visual cues from video and images that needs to be implicitly learned from the data.

How do you learn that?

There are ways that we can explicitly add in human prior knowledge to guide the learning, to learn the way we want it to learn. There are also implicit ways. You have supervision. You have the video. So it will learn explicitly the pattern and figure out things from the data itself.

How did you get from Wuhan to Michigan?

I find that throughout my career, I had a lot of luck! I didn't really have a five-year plan. When I first entered college, I learned from others about

the opportunity to study abroad. I found that interesting! I wanted to give it a try. Then I started to prepare for English exams. Then in my last year of undergraduate, I joined a pattern recognition lab just by chance. Then I started to learn more about biometrics. That's also something I was very interested in, and I wanted to study that topic in graduate school. I applied to 10 different universities for the PhD program.

Is that a lot?

I think it's about average. People do that. But, for me, I only got admitted to one university. I'm very grateful. I didn't have a strong profile at that time. I am very grateful for my PhD. Professor Xiaoming Liu gave me tremendous support during those five years to prepare me for work after I graduated.

Do you think that it's sheer luck that brought you to this point? There must have been something intentional about your career path.

I have been thinking about that actually. I want to use a quote here from Linus Torvalds, who is the father of the Linux system. He has made a tremendous impact on the software industry. He said, "*There are people who are looking at the stars and saying I want to get there. Those are visionary people. I am looking at the ground, and I want to fix the pothole that's right in front of*

me before I fall in." I didn't mean to compare myself to Linus, but I'm also more of the second type of person. I work hard, and I'm very interested in learning more. I try to do well with the task at hand, and then opportunities will come.

You sound very confident!

It has worked well for me so far! I plan to continue this pattern.

Tell us about the University of Wuhan.

Wuhan is like my second hometown. I spent high school and college in Wuhan. Wuhan is famous for having a lot of universities. Wuhan University is a general university. It has all different majors in different domains. It has a very diverse background. People came here from all different cities of China. It's a really diverse environment with a lot of activities going on. For me, I was more into studying. I was preparing for my English exam. I was going to the lab to get some experience to prepare myself for studying abroad. From that perspective, I didn't really participate much in my university.

How did you start working at Microsoft and Facebook?

In my PhD, I was trying to work very hard in my research. There are certainly difficulties in the PhD process. I am generally trying to expand my knowledge and dive deep into problems. When I graduated, I started interviewing. It went very smoothly. I received offers at interviews on-site. Actually, that's the first time I felt like, "*Oh wow, I'm doing well!*" There were times when I was looking for an internship, but I didn't get the meetings. In the end, when I was looking for jobs, it got easier. Then I joined Microsoft remotely from the Bay Area due to family reasons. The team members were all in Redmond in Seattle. I joined them because I really enjoyed the interview process, having an interaction with the team and the manager and I really enjoyed my time there. I didn't plan to move, and I don't like to travel frequently. So after two years, I changed to Facebook and stayed in the Bay Area, although I haven't been to the office yet due to the COVID situation.

So you work for Facebook, but you have not been to Menlo Park yet?

"...they assume and call me Mr. Yin..."

I went once to visit friends, but I haven't worked from there yet.

During all these years in the States, have you ever thought about going back to China?

For my professional life, I don't have very high expectations. I think that is working well. I'm fine with that. Personally, during holidays, I miss my family and friends. I have very close friends in China, and they can get together frequently. I miss that. When my family goes through challenging situations, I think, "*Oh, I wish I was there.*" When I try to connect with friends and family, those are the times I think about whether I should go back to China. Professionally, I don't experience much of that.

"It has worked well for me so far!"

After Covid, it will be easier to travel. But still, you are going to be far away. Is your love for science enough to bear this separation?

Yes, for family and friends, we can just find ways to get together whenever possible. I try to keep a good balance.

What's your message to the community?

I would like to mention two points. I have been thinking a lot about the role of women in the workplace. There are times when I interact with people who do not know me. If they do not do their homework, they assume and call me Mr. Yin, for example. In an anonymous discussion, if people don't know the person behind the comments, they assume that it is a "he" or "him". That makes me uncomfortable as a woman in this field. On the other hand, there are also more opportunities for women. For example, for this interview, we don't have a Men in Science version. This counters the first point. For me, in my professional life, I don't feel discriminated against. I work with people who know me, which I think is good. People in the research community are generally very nice. I consider myself a scientist, not a female scientist. I understand that we need to promote more women to join the workplace. I hope eventually, we will not use a phrase like "female scientist" or "female CEO". That's probably when we will have real gender equality. People should be defined by what they love and what they do, rather than their gender, race, age, and so on.

Maybe one day you will interview me and ask me tough questions! [*both laugh*] You said you have a second thing to say.

Yes, I'm also very interested in different people's perspectives through interactions with people and through things that I read. We should always examine our beliefs. My Facebook training covered subjects not only limited to the Facebook infrastructure and the tools we are using, but also tips on how to work better with others.

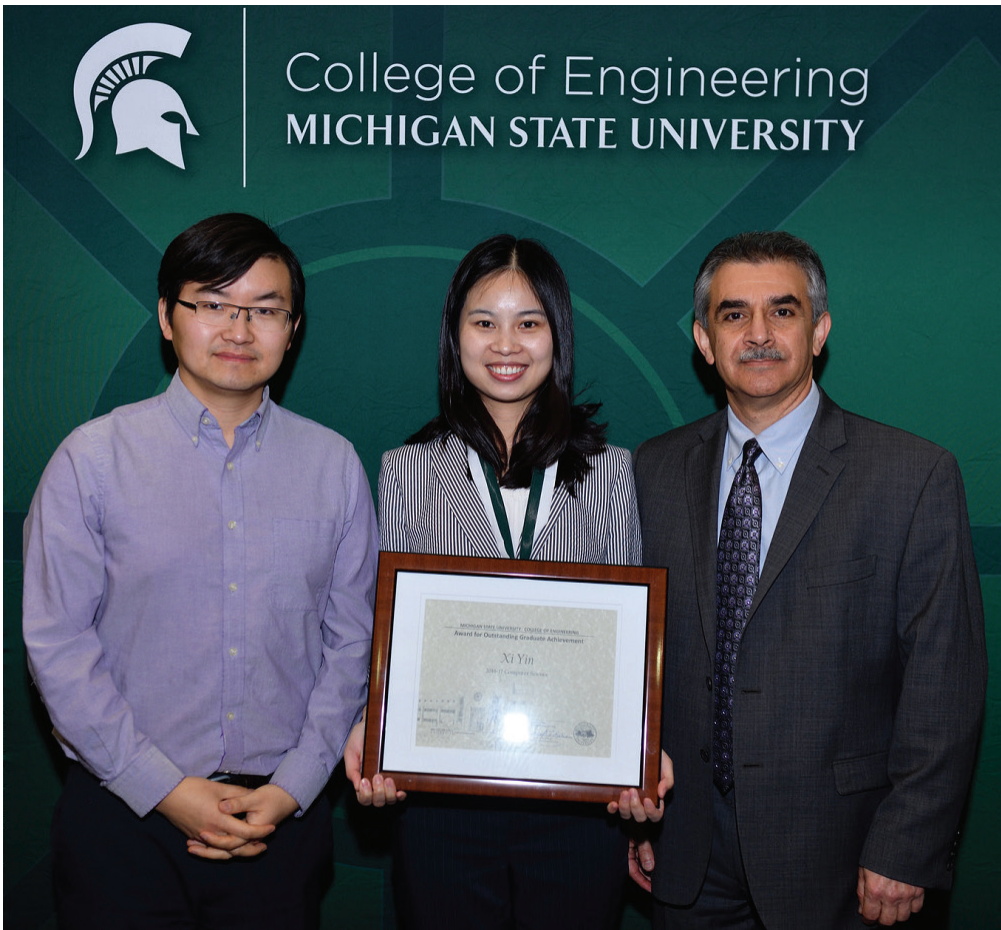
We used to believe that we should treat people the way we would like to be treated. In practice, it says that is not necessarily true. You should treat people the way they want to be treated. Although we are all human beings, we are fundamentally different. We have different perspectives. That encouraged me to be more patient when interacting with people, to have more emphasis on trying to understand each other better, in order to work together better. That's something I personally found very useful and wanted to share with the community.

Thank you for being patient with me! What did you learn in this interview?

You are a very nice person! All those things that you said in the beginning, that I am speaking clearly... You are very patient as well. I really enjoyed the conversation with you. Thank you so much for this opportunity, I really appreciate it.

Over 100 interviews with as many inspiring Women in Computer Vision - in our archive

"Professor Xiaoming Liu gave me tremendous support during those five years to prepare me for work after I graduated!"



42 Congrats, Doctor!

Natasha Jaques recently completed her PhD at MIT, where her thesis received the Outstanding PhD Dissertation Award from the Association for the Advancement of Affective Computing. She now holds a joint position as a Research Scientist at Google Brain and Postdoctoral Fellow at UC Berkeley. She is focused on building socially and emotionally intelligent AI agents that can flexibly learn from multi-agent and human-AI interaction. Her research aims to build the case that social learning can enhance RL agents' ability to acquire interesting behavior, generalize to new environments, and interact with people. Congrats, Doctor Natasha!

Learning from human-AI interaction: Ultimately, for AI to be as satisfying and useful to people as possible, we want to train agents that directly optimize for human preferences. However, manually training a device is cumbersome, and people exhibit low adherence to providing explicit feedback labels. Instead, passive sensing of the user's emotional state and social cues could allow the agent to learn quickly and at scale, enabling human-in-the-loop training without extra human effort. I have explored this idea in several papers related to learning from human sentiment and social cues in dialog [1,2] including with a novel Offline RL technique [3,4]. Experiments deploying these models to interact live with humans reveal that learning from implicit, affective signals is more effective than relying on manual feedback. My work also demonstrated that a recurrent art generation model could be significantly improved with less than 70 samples of people's facial expression reactions to its sketches [5] (see Figure 1). My goal is to demonstrate that learning from implicit social cues can enhance human-AI interaction, and guide AI systems to take actions aligned with human preferences.

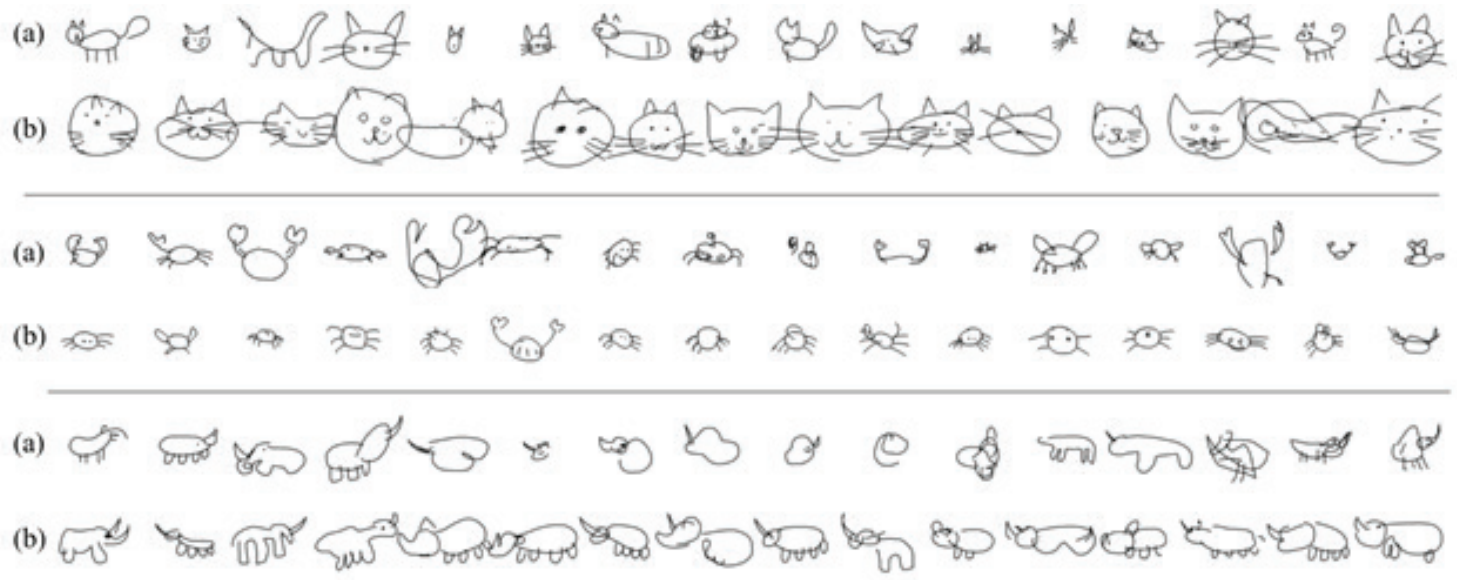


Figure 1: Samples of cat, crab, and rhinoceros drawings produced by (a) the original Sketch RNN, and (b) our model trained on a small amount of facial expression feedback.

Understanding human affective signals:
Learning from humans' affective cues requires recognizing them first. Throughout my PhD, I developed machine learning methods for automatically interpreting human data and recognizing affective and social signals such as stress, happiness, and conversational rapport (e.g. [6, 7, 8, 9]). We developed novel techniques for analyzing physiological sensor data such as Electrodermal Activity (EDA), accelerometry, and temperature [10, 11, 12], and built a popular open-source tool which deployed them (https://eda-explorer.media.mit.edu/). However, I noticed that the accuracy of our affect detection models appeared to be severely limited by the degree of inter-individual variability in emotion and wellbeing. What makes one person stressed can have the exact opposite effect on someone else. Therefore, my collaborators and I designed methods for personalizing machine learning models using multi-task learning, enabling the predictions for one person to gain statistical strength from the data of others, to the degree that it is relevant. In a series of papers [13, 14, 15], we showed that personalization via multi-task learning achieves large performance gains and state-of-the-art accuracy in predicting outcomes like happiness, stress, and health (see Figure 2).

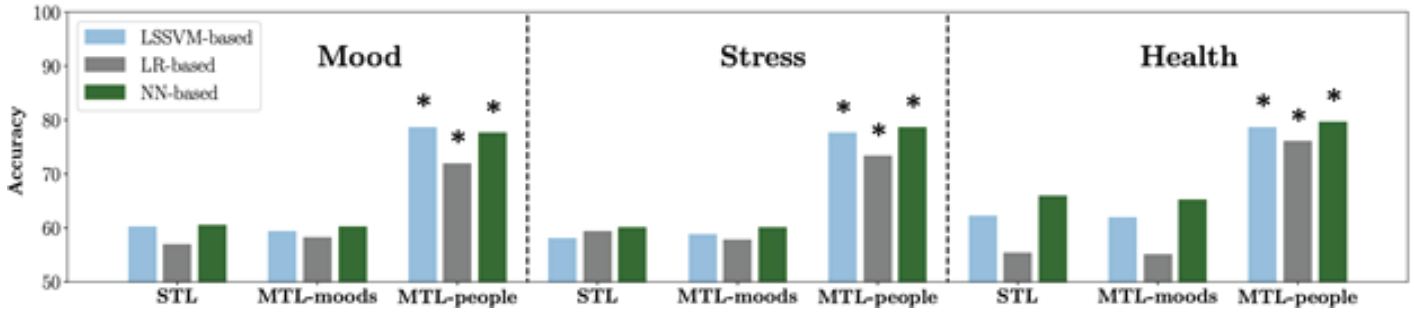


Figure 2: Emotion and wellbeing prediction performance improvements achieved through personalized multi-task learning (MTL-people).

Multi-agent coordination and communication:

Current RL systems require many samples to learn effectively, which makes learning from human data prohibitively expensive. Instead, we can use multi-agent reinforcement learning (MARL) to learn in simulation, as a way to pre-train agents to coordinate effectively with humans. Social influence [16] proposed a unified mechanism for achieving both coordination and communication in MARL. Agents use counterfactual reasoning over a model of other agents to compute and maximize the degree of causal influence they exert on other agents' actions (see Figure 3). This mechanism led agents to cooperate more effectively, because they learned to use their actions to communicate useful information to other agents in order to gain influence (for example, signaling the presence of resources outside the field-of-view of other agents). Our results demonstrate that influence, which can be computed in a fully decentralized manner, led agents to cooperate more effectively than existing state-of-the-art multi-agent methods which made use of centralized training, or privileged access to other agents' rewards or parameters.

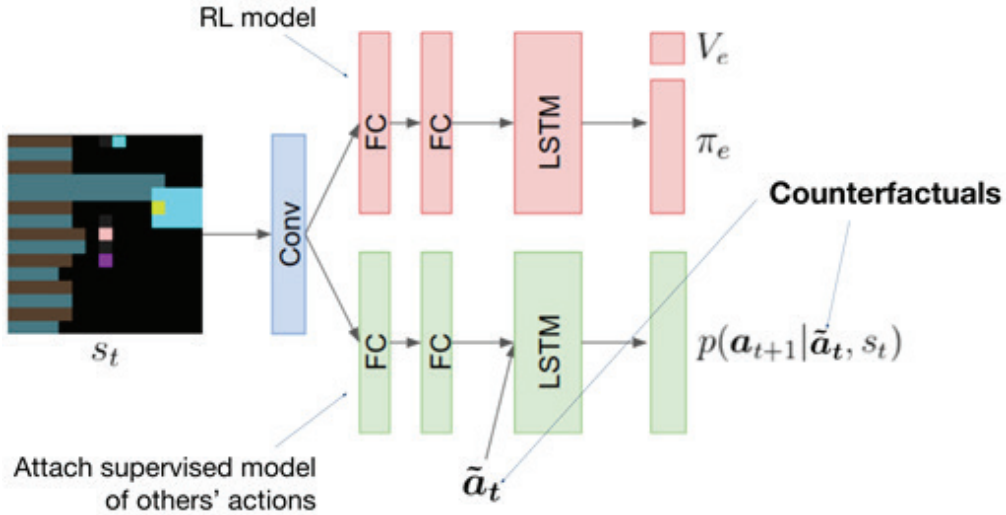


Figure 3: Model used to compute counterfactual causal influence of the agent's action on other agents; requires both RL and supervised components

COMPUTER VISION EVENTS

Innovation Summit SF
Hyatt Regency SF Airport, CA
1-3 Nov

TCT
Virtual Orlando, FL and online
4-6 Nov

AI for Good
Seattle, WA
10-11 Nov
Postponed!

DigiHealthDay 2021
Pfarrkirchen, Germany
12 Nov

SIPAIM 2021
Campinas, Brazil
17-19 Nov

AM Medical Days
Berlin (Germany) and Virtual
22-23 Nov

BMVC 2021
Online
22-25 Nov

IOT TECH EXPO EUROPE
Amsterdam, NL
23-24 Nov

Hamlyn Winter School...
online
6-10 Dec

NeurIPS
virtual-only
6-14 Dec

Due to the pandemic situation, most shows are considering to go virtual or to be held at another date. Please check the latest information on their website before making any plans!

MEDICAL IMAGING NEWS

The new supplement to Computer Vision News

October 2021

Daniel Rueckert
@DanielRueckert

Replying to @maier_ak and @RSIPvision

Keep up the great work

Daniel Rueckert
Professor, TU Munich and Imperial College London

3D Measurements in Robotic Assisted Surgery:

The ability to accurately measure during **surgical procedures** can significantly improve outcomes. It is a common practice to use sizing jigs during surgeries – in orthopedic procedures the implant fitting is assessed using demo-implants, in cardiology, valve sizing is achieved using sizers, etc.

Robotic assisted procedures use video guidance and adding a 3D module – stereoscopic imaging – is standard in these platforms. The 3D perspective adds the depth dimension to the image and allows better spatial orientation throughout the procedure.

This technology is the baseline for allowing accurate measurements in images. Using well established methods for camera calibration (e.g., calibration jig), a coordinate system can be overlayed on the image, providing **accurate mapping of the field-of-view**. The current use of stereo camera is increasing rapidly in many different areas, and it is integrated in many medical devices. The ability to view and estimate the depth and size of objects opens new possibilities during surgery.

There are several benefits for this application.

1. **Standard Measurements:**
   a. many procedures require **accurate tissue excision**. In orthopedics, it is vital to saw the bone accurately to fit the implant without causing limb length discrepancy. During soft-tissue excisions, to avoid perforation of nearby organs or blood vessel it is essential to know how deep into the tissue the tools are. Gastric bypass procedures often require tedious measurements of the gastrointestinal tract, which can be executed efficiently using stereoscopic measurements.
   b. Diagnostic procedures often rely on measurements to **assess severity**. Cancer staging is determined by the size of the tumor. This can be accurately measured, provide adequate staging, and reduce complications due to misdiagnosis.
   c. Another aspect where accurate measurements surpass current practice, is **surgical precision**. If all incisions are performed using a defined measurement it ensures that surgeries are performed in a similar manner, making the procedure repeatable.

2. **Buffer Zones:**
   a. When removing tumors, healthy tissue from around the tumor is also removed to verify a "negative margin" and prevent recurrence. Verifying that excision occurs at a constant distance from the tumor is challenging, therefore an **objective measurement** tool can assist in this task. Additionally, this buffer can be overlayed on the image given a manual segmentation of the tumor.

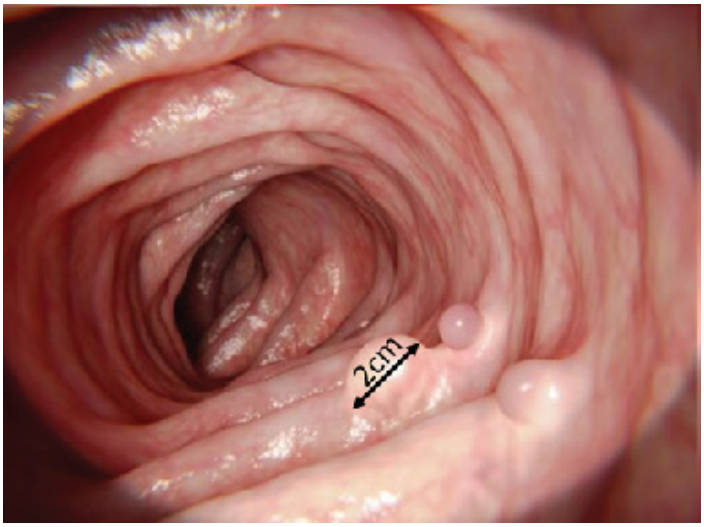
3. **Temporal/Spatial Measurements:**
   a. During stitching, applying too little or too much tension on the sutures can have devastating effects. Assessing tension based on images only is difficult, and there is no tactile feedback in RAS. Strain is defined as:

$$\epsilon(t) = \frac{L(t) - L(0)}{L(0)}$$

   b. Where L is the length of the suture as it is continuously measured throughout the stitching process. Positive strain means the suture is under tension, whereas negative strain means the suture is loose. **Upper and lower strain limits can be defined** to provide proper feedback to the surgeon.

   b. Other procedures can also benefit from detection of **dimensional changes over time**. Ablation procedures require total ablation of the region with minimal harm to the surrounding tissue. Analyzing the ablation radius (and radius change rate) in real time can assist in determining when the ablation is successful and prevent damage to surrounding tissue.

Accurate measurements are an objective assessment method to the success of a procedure. In real life we have intuition regarding the size of objects. In the surgical setting this intuition is difficult to achieve and implementing AI modules can replace the lack of intuition, ultimately **improving procedural outcome and increasing patient safety**.



Read more about Robotic Assisted Surgery: https://www.rsipvision.com/surgical-robotics/

MEDICAL IMAGING NEWS

Anomaly detection in medical imaging with deep perceptual autoencoders

IOANNIS VALASAKIS, KING'S COLLEGE LONDON
@WIZOFE

Hi everyone, how are you doing? Let's have another article to read. This month's review is "Anomaly detection in medical imaging with deep perceptual autoencoders" by Nina Shvetsova et al.

This is a pre-print which was uploaded this month, but it already has 15 citations and you can easily find it on Arxiv, to read it in full.

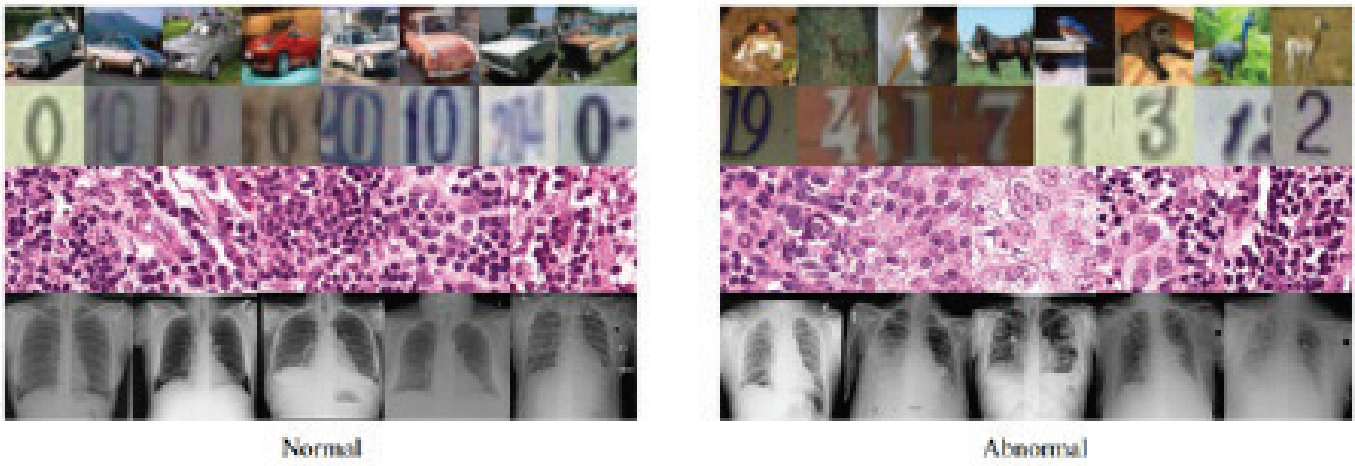
Anomaly Detection

Anomaly detection is a task with significance, especially in the deployment of machine learning models. The knowledge of a "a normal" data sample would be used to compare -in a sense of a ground truth- to an "abnormal" one. To identify less often occurrences is another application where anomaly detection is useful and with the method proposed in this paper, the efficacy of autoencoders for anomaly detection is utilized.

The authors compared here the three strongest SOTA anomaly detection methods in two challenging medical tasks: **Chest X-rays** and **H&E-stained histological images**. In the manner of open science (Richard Stallman would call the free, oh well!), the source code of all our experiments to facilitate the development of anomaly detection in medical imaging is shared online. The solution was sequentially compared with both datasets with SOTA methods which were outperformed.

Even though **Computer Vision News** and **Medical Imaging News** focus on computer vision and medical learning, there is a wide range of fields where anomaly detection

has been extensively studied. Domains including fraud detection, cyber-intrusion detection, as well anomalies in videos and financial analytics are some of the latest fields where research is active. Distribution based methods try to predict if the new example lies in the high-probability area or not.



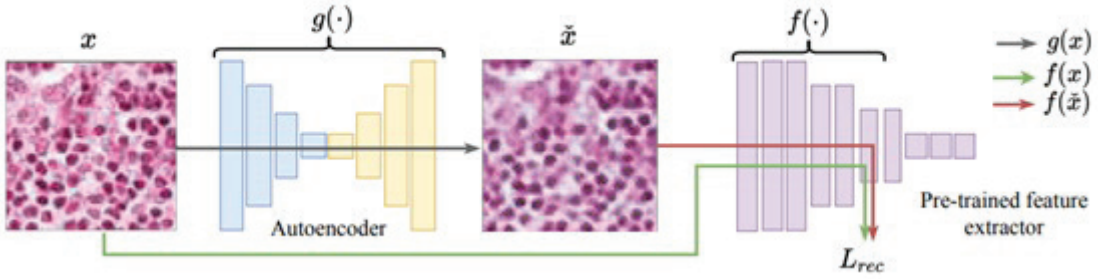
Examples of dataset with abnormality vs normality comparisons. CIFAR10 dataset images, with SVHN digit dataset in the row below, followed by histopathological and chest X-ray images from the NIH dataset.

DEEP PERCEPTUAL AUTOENCODER

Autoencoder-based approaches rely on the fact that they can learn shared patterns of the normal images and then restore them correctly. The key idea of the proposed method is to simplify the learning of these common factors inherent to the data, by providing a loss function that measures "content dissimilarity" between the input and the output.

The progressive training idea is explored further to improve the expressive power of the autoencoder. The reasoning is that the pipeline gradually grows the "level" of the "perceptual" information in the loss function. The paradigm has many hyperparameters, therefore tuning is essential to ensure detection quality. Labels can be used during the model setup to create a weakly-supervised training paradigm, where a low number of labelled anomalous examples of a limited variation is available.

MEDICAL IMAGING NEWS

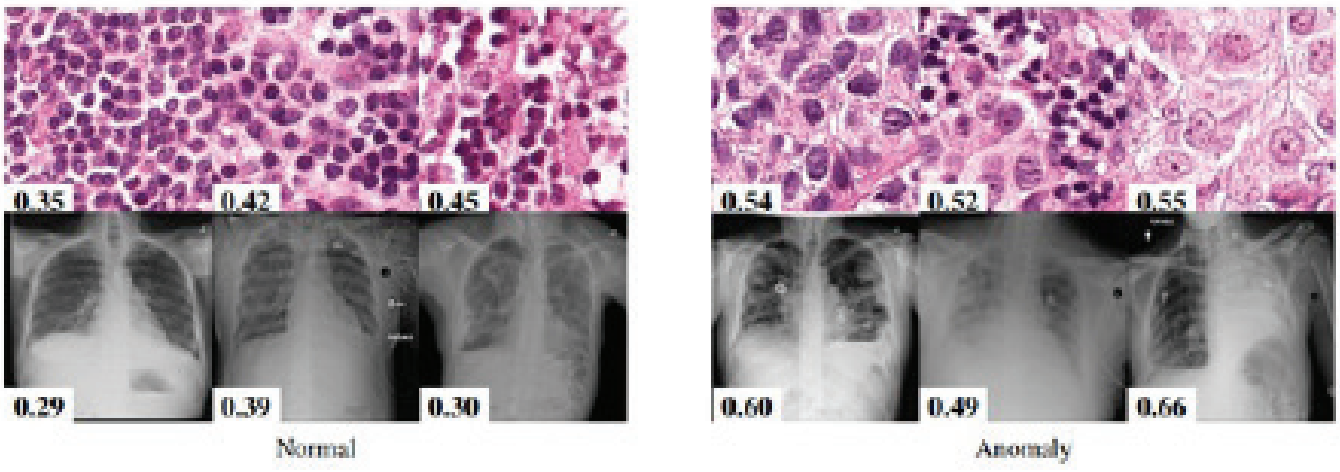


The autoencoder architecture for image anomaly detection, where g is the network, f the feature extractor, x^ the reconstructed image while the reconstruction L calculates the difference between $f(x)$ and $f(\check{x})$

SOTA baselines were used for the evaluation in the problem statement of novelty detection and the training data were taken as “anomaly free” to examine two challenging medical problems with different image characteristics and appearance of abnormalities: metastases detection in digital pathology and chest X-Rays.

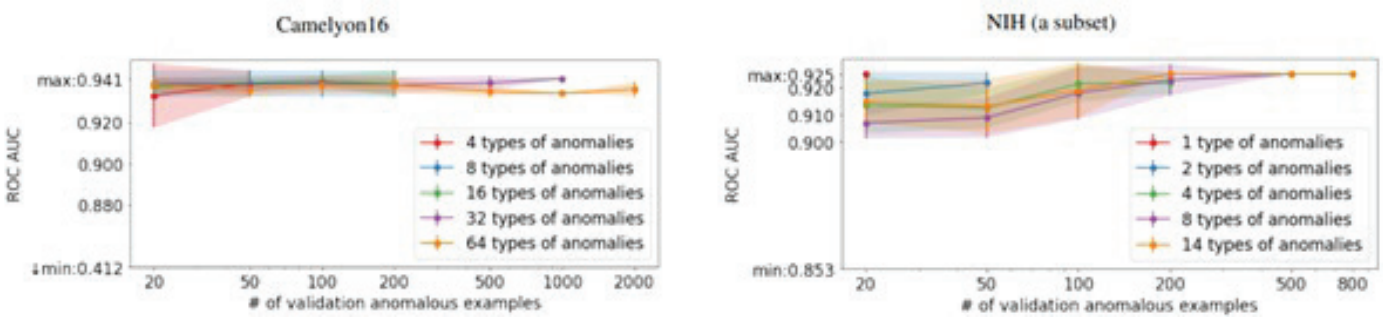
The task is to **detect metastases** in H&E stained images of lymph nodes in the Camelyon16 challenge by training anomaly detection models only on healthy tissue aiming to identify tissue with metastases. Tissues exhibiting metastasis may differ from healthy types only by texture, spatial structure, or distribution of nuclei.

Chest X-ray is one of the most common examinations for diagnosing various lung diseases and the task was to recognize 14 findings, such as Atelectasis or Cardiomegaly, on the chest X-rays in the NIH dataset (ChestX-ray14 dataset). To create the training dataset, the existing one was split into two parts with only posteroanterior (PA) or anteroposterior (AP) projections.



This is an example from the Camelyon 16 challenge (top) and the NIH dataset (bottom). Next to the image the anomaly prediction score by the proposed method. The higher score signifies an anomaly. It’s interesting to note that borderline cases are correctly evaluated by the model.

The network was also evaluated on two different natural image benchmarks: the CIFAR10 and SVHN. The official train-test split dataset was used, each of 10 classes, where one abnormal class was randomly on the train set, to use as a validation set with abnormal images.



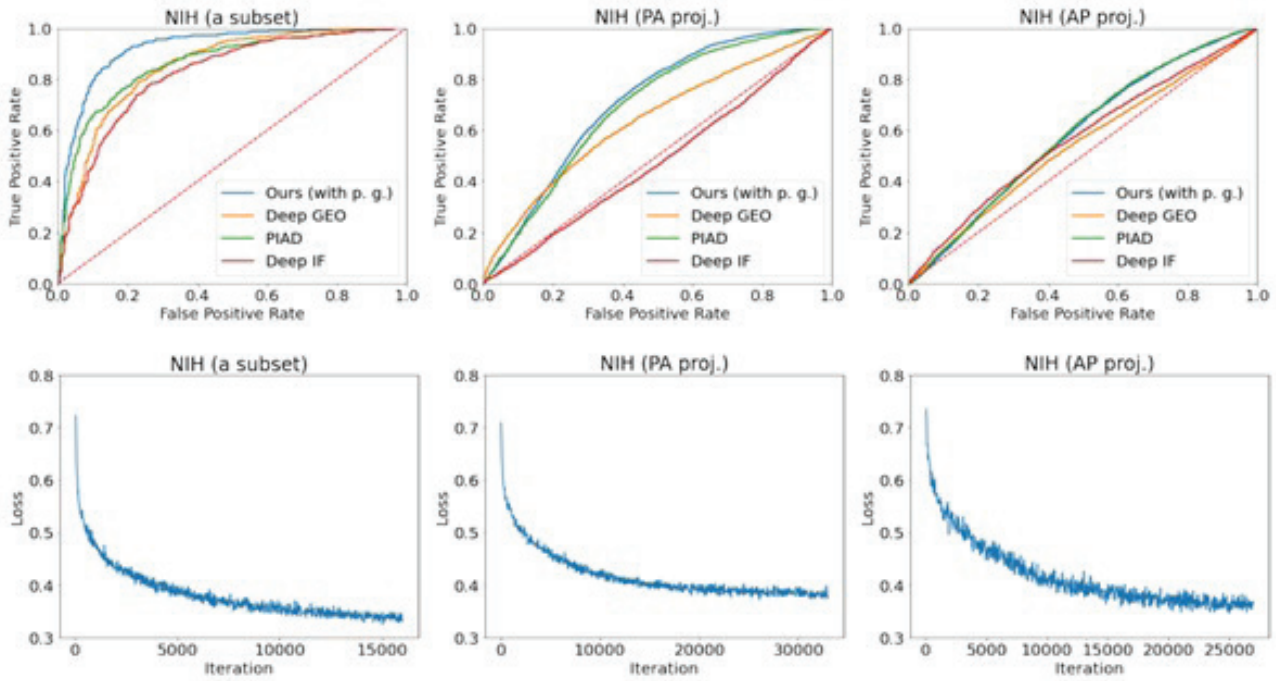
Here is shown the dependence of the quality of anomaly detection on the number of anomaly examples and their variability in the validation set.

Which was the evaluation metric used? Area Under the Curve of Receiver Operating Characteristic (ROC AUC) and the reason is that it integrates the classification performance (normal vs. abnormal) for all the different thresholds for the decision. That helps as there is no need to have a threshold for the predicted abnormality scores which allows to assess the performance of the models "probabilistically" and without a bias.

Some experiment specific settings are following! The SOTA baselines were chosen based on their efficiency on different paradigms: Deep GEO, Deep IF and PIAD. Additionally, for natural images AnoGAN, GANomaly, DAGMM, DSEBM and OCGAN methods. The Deep IF and PIAD approaches were implemented by using extensive descriptions provided by the authors. For GANomaly and Deep GEO the official code was used and the results of DAOL and OCGAN methods were obtained in the corresponding papers. Hyperparameter search was performed by maximizing average ROC AUC over 3 "folds".

The Deep GEO approach has been used with great success on the SVHN dataset to distinguish individual digits (and from themselves), while the Deep IF approach shows the second-best result. There is showing that the experiment that even a

small number of abnormal samples of one type of anomaly is enough to reject inferior hyperparameter configurations. In the two experiments considered, having 20 abnormal examples of the same type of abnormality proved sufficient to select the hyperparameters within the 2% margin of the optimal configuration.



Here's shown the training performance curves of the proposed method on a subset of the NIH, NIH (PA proj.), and NIH (AP proj.) datasets.

An ablation study shown on the Table 1 was used to signify the importance of the individual components proposed in the paper. The different scenarios are shown in the Fig. 5, where the use of perceptual loss outperforms the mere L1 norm (1) with a large margin. It is also important to note that the anomaly detection performance is affected by the method of selecting the hyperparameters as it reveals a subset of anomalies of confined variability. The proposed method excelled on the medical datasets but not as much for the nature image baselines of SOTA. One reason for that may be that the high diversity present in the natural data lead to overgeneralization of the autoencoder.

	Camelyon16	NIH (a subset)	NIH (PA proj.)	NIH (AP proj.)
(1) L1 + unsupervised	21.1 ± 1.4	70.8 ± 0.6	66.5 ± 0.1	52.4 ± 0.1
(2) PL + unsupervised	87.9 ± 0.6	89.3 ± 0.2	68.9 ± 0.1	56.4 ± 0.2
(3) PL + weakly-supervised	92.7 ± 0.4	92.0 ± 0.2	70.3 ± 0.2	**58.6 ± 0.1**
(4) PL + 1· adv + weakly-supervised	79.4 ± 4.0	64.4 ± 7.8	52.3 ± 3.3	51.5 ± 3.4
(5) PL + 0.1· adv + weakly-supervised	90.8 ± 0.7	82.2 ± 2.6	59.2 ± 1.4	55.4 ± 0.9
(6) PL + 1· L1 + weakly-supervised	75.3 ± 1.6	91.7 ± 0.4	70.7 ± 0.2	57.3 ± 0.1
(7) PL + 0.1· L1 + weakly-supervised	93.0 ± 0.3	92.0 ± 0.1	70.6 ± 0.2	58.5 ± 0.1
(8) PL + 1· L1 + 1· adv + weakly-supervised	57.5 ± 6.3	59.3 ± 5.0	50.1 ± 2.0	51.7 ± 0.8
(9) PL + 0.1· L1 + 0.1· adv + weakly-supervised	90.6 ± 1.0	78.2 ± 1.0	60.8 ± 1.8	55.5 ± 0.4
(10) PL + weakly-supervised + progressive growing	**93.4 ± 0.3**	**92.6 ± 0.2**	**70.8 ± 0.1**	58.5 ± 0.0

Ablation study with ROC AUC in percentage standard deviation.

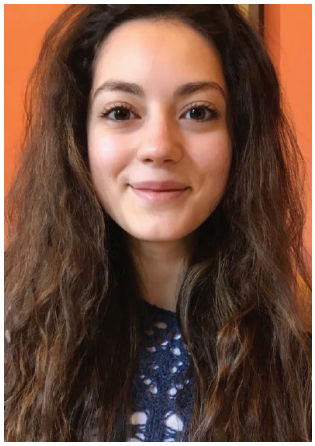
Overall, a range of state-of-the-art image anomaly detection methods was explored and a proposed novel method using an autoencoder to understand the representation of the normal data was shown. This could influence a real-life clinical workflow, where generation of reproducible results by an efficient experiment design can be achieved.

Wrapping up!

I hope you enjoyed again this month's article. Even though the topic was a little bit different than usual, being able to get inspiration and example from other subjects is more than worthy. Sometimes, similar problems have been solved in other disciplines in various ways, and it would be un worthwhile to ignore this!
Thank you, Nina, for providing access to the paper and the results. It's great to discuss the latest on imaging techniques, especially ones that are valid in different domains 😊 Remember to always refer to the original article for more discussion, details, and the full references for all the material used!
Take care for yourself and the people around you! See you next month 😉

MEDICAL IMAGING NEWS

Python scripting for 3D Slicer



by Marica Muffoletto

Welcome all to a new article dedicated to a medical imaging tool.
Today, instead of reviewing something completely new, we will take a step back to our September article from last year, and dive more into the potentialities of the 3D Slicer software.

As a powerful open-source platform dedicated to the analysis and visualization of medical images, 3D Slicer also allows the users to customize their experience and extend the modules' behavior through the interaction with Python scripts. Today we are going to explore how to do that in two different ways!

Option 1: Start a Python interactor window pressing on CTRL+3 (Windows/Linux), or cmd+3 (Mac).

This option allows the user to directly insert code in a new rectangular window which appears at the bottom of the GUI, and simultaneously interacts with it.
Installing a library with pip can be done both from terminal using:

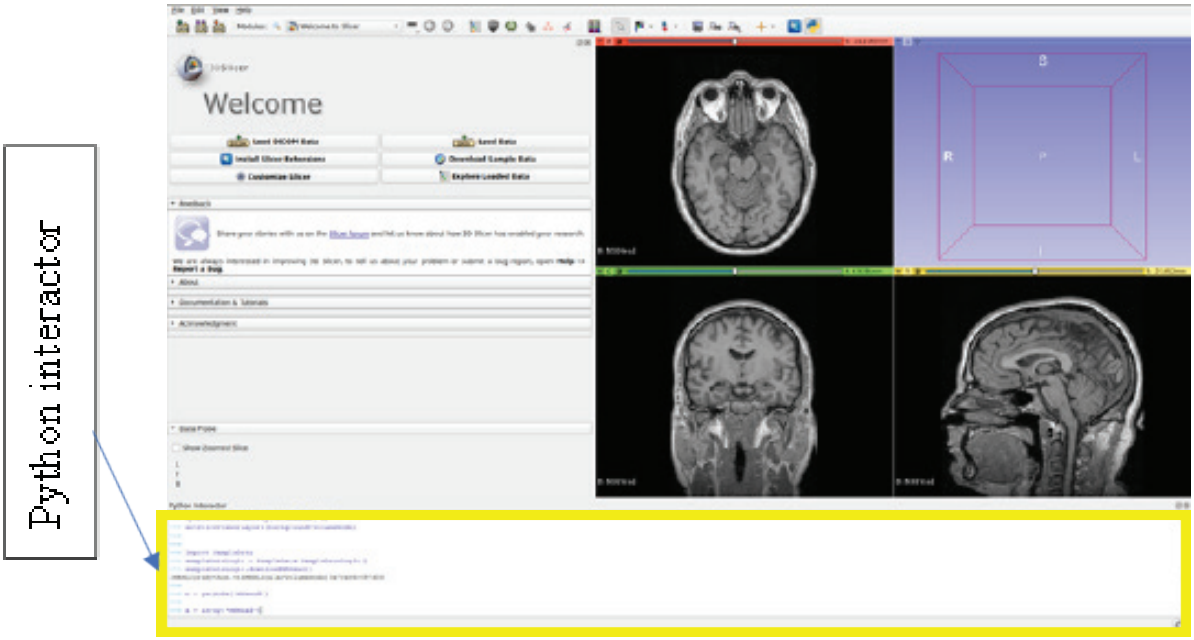
```
/path/to/Slicer-X.Y.Z-plat-arch/bin/PythonSlicer -m pip ...
```

Or from the window itself:

```
pip_install('scipy')
```

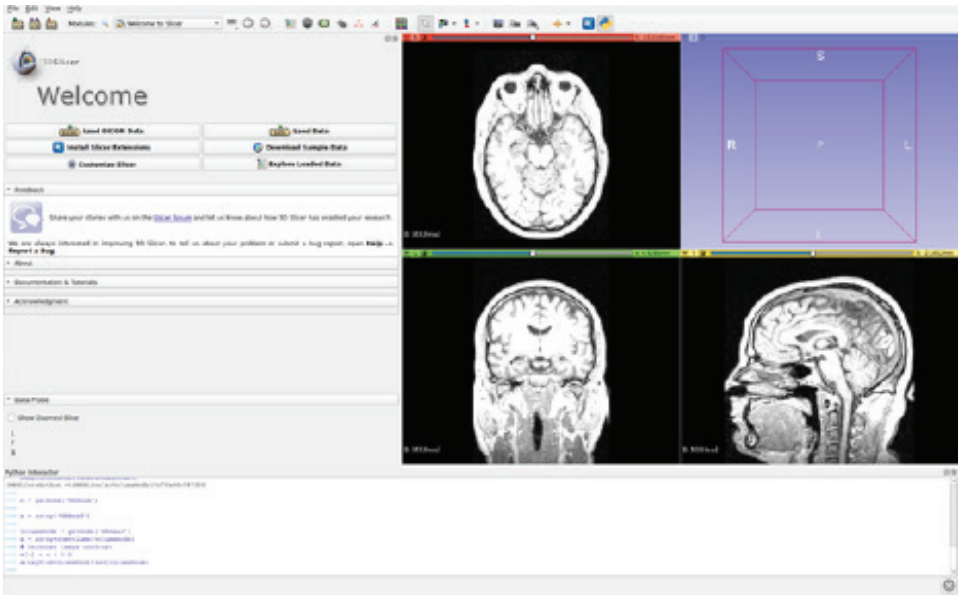
Any snippet of Python code can be tried out here but note that to properly interact with the software tools, you will need to call the right classes and functions from Slicer. The following lines of code, for example, load a sample image from a downloaded brain MR dataset.

```
import SampleData
sampleDataLogic = SampleData.SampleDataLogic()
sampleDataLogic.downloadMRHead()
n = getNode('MRHead')
```



You can also convert an image to an array and then increase its contrast.

```
a = array('MRHead')
volumeNode = getNode('MRHead')
a = arrayFromVolume(volumeNode)
# Increase image contrast
a[:] = a * 2.0
arrayFromVolumeModified(volumeNode)
```



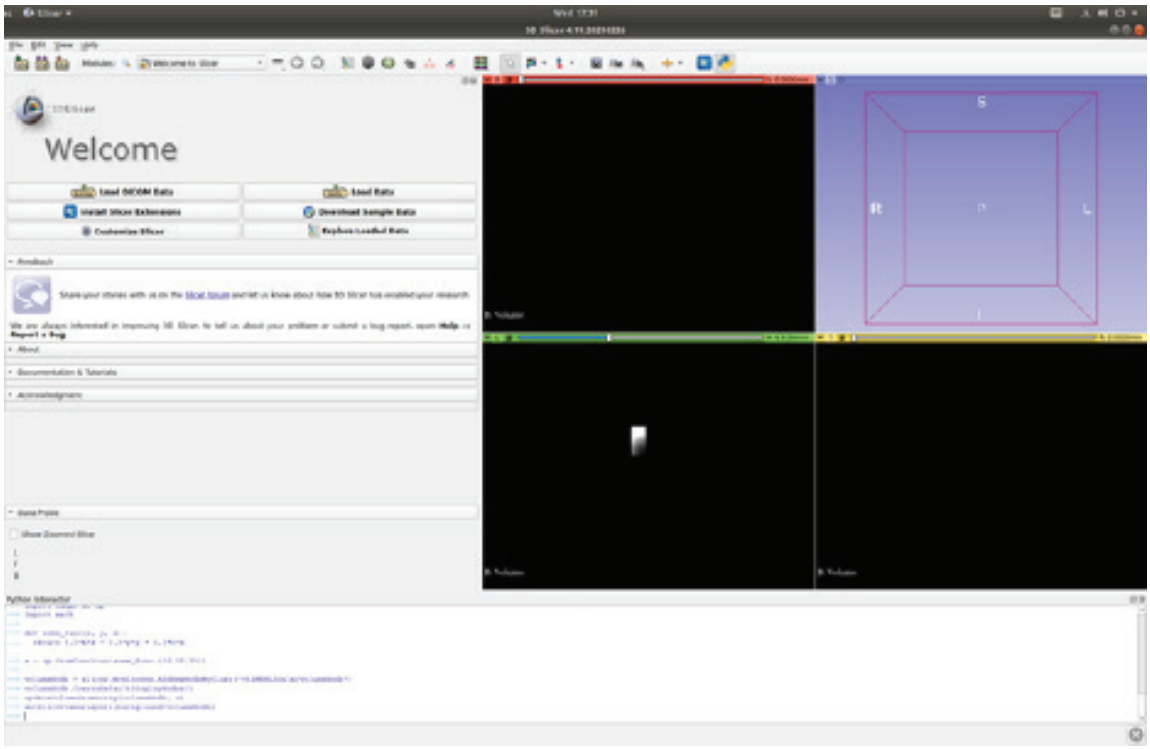
Or finally, import standard python libraries and create user-defined functions which match your needs:

```
import numpy as np
import math

def some_func(x, y, z):
  return 0.5*x*x + 0.3*y*y + 0.5*z*z

a = np.fromfunction(some_func,(30,20,15))

volumeNode = slicer.mrmlScene.AddNewNodeByClass('vtkMRMLScalarVolumeNode')
volumeNode.CreateDefaultDisplayNodes()
updateVolumeFromArray(volumeNode, a)
setSliceViewerLayers(background=volumeNode)
```



Option 2: Run Slicer operations from a batch script.

This allows the user to use the powerful toolset from Slicer without any need to even open the GUI. Everything can be done just by terminal and python scripting. First, we need to get into the directory where Slicer is installed and then run the software with the parameters –no-main-window and –python-script as below:

```
cd /usr/local/Slicer-X.XX-XXXXXXXX-linux-amd64

./Slicer --no-main-window --python-script
/home/user/Desktop/batchreg_3dslicer.py
```

The following python script is created to automatize the task we manually performed in the previous article. This consisted in selecting landmark points to register two images, and then save them to re-apply a rigid registration between the segmentation and the reference image. This can be very useful in itself, but what if we need to do it for multiple files or an entire folder? Of course, the first part - extraction of landmarks - needs to be manual, but, assuming we already have the landmark points saved in a specific directory, this python script can handle the interaction with 3D Slicer to automatically replicate all steps and save the new registered images.

```
def batch_reg(filename_dir,transform_dir):

    # exclude missing cases
    N = 100 #number of cases
    for i in range(1, N):

        if (i == 11 or i == 12):
            continue

        itemCT_filename = filename_dir
+ 'data/non_registered_cts/ ct'+str(i)+'.nii'
        itemMR_filename = filename_dir
+ 'data/non_registered_mrs/ mr'+str(i)+'.nii'
        itemSEG_filename = filename_dir
+ 'data/non_registered_segs/seg'+str(i)+'.nii'
        itemTransform_filename = filename_dir + \
            '/case' + str(i) + '/Transform.h5'

        itemMR = slicer.util.loadVolume(itemMR_filename)
        itemSEG = slicer.util.loadVolume(itemSEG_filename)
        itemCT = slicer.util.loadVolume(itemCT_filename)
        itemTransform
= slicer.util.loadTransform(itemTransform_filename)

        outputSEG_filename = filename_dir+'registered_segs/seg
+ str(i) + '.nii'
        SEGresampledNode
= slicer.vtkSlicerVolumesLogic().CloneVolumeGeneric(slicer.mrmlS
```

```
cene, itemSEG, 'out', True)
        SEGparameters = {'inputVolume': itemSEG,
'referenceVolume': itemMR,
'outputVolume':SEGresampledNode,'warpTransform': itemTransform,
'interpolationMode': 'NearestNeighbor'}
        slicer.cli.runSync(slicer.modules.brainsresample, None,
SEGparameters)
        slicer.util.saveNode(SEGresampledNode,
outputSEG_filename)

        outputCT_filename = filename
_dir+'registered_cts/ct' + str(i) + '.nii'
        CTresampledNode
= slicer.vtkSlicerVolumesLogic().CloneVolumeGeneric(slicer.mrmlS
cene, itemCT, 'out', True)
        CTparameters = {'inputVolume': itemCT,
'referenceVolume': itemMR,
'outputVolume':CTresampledNode,'warpTransform': itemTransform,
'interpolationMode': 'Linear', 'defaultValue':-1025}
        slicer.cli.runSync(slicer.modules.brainsresample, None,
CTparameters)
        slicer.util.saveNode(CTresampledNode, outputCT_filename)
        slicer.mrmlScene.Clear(0)

        print(i)

base_dir = '/home/user/Desktop/RegProject/'
transform_dir = base_dir + 'reg_transforms'

batch_reg(base_dir,transform_dir)
exit()
```

We are at the end of another article and it's time for you to try out your own Python scripts on Slicer ☺.
If you want to explore further this subject or you would like to initiate a discussion/ review on another software, please feel free to get in touch with the editors. Meanwhile, enjoy the rest of the November mag!

Did you read the BEST OF ICCV? It includes an awesome medical imaging article!

Go and read it on page 24!

Dominik Rivoir is a PhD student at the National Centre for Tumor Diseases, under the supervision of Stefanie Speidel.

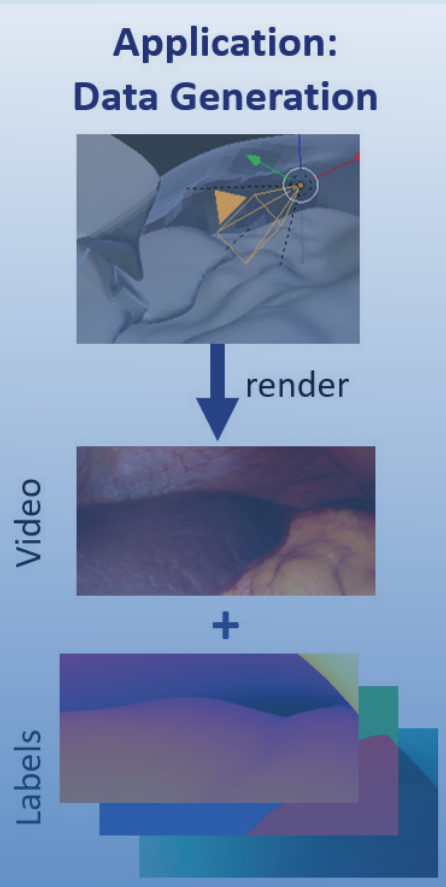
His work proposes a new approach to providing simulated 3D data in a surgical setting.

It has been accepted as a poster this year and Dominik spoke to us ahead of his live Q&A session.

Computer-assisted surgery involves **analyzing video feed from the surgery to provide feedback**. This may be to **identify structures**, like organs or unsafe regions the surgeon should stay away from, or to **recognize instruments** to understand what the surgeon is doing.

One of the biggest challenges in the surgical setting is the lack of labeled data. Obtaining labeled video data in the surgical setting, such as point correspondences over time, or training 3D information like depth or 3D positions, is especially difficult. This work aims to solve that by providing synthetically rendered, but realistic looking data, with the goal of **achieving view-consistency, or long-term temporal consistency, in the generated videos**.

"*This work is a mix between **GAN-based translation**, which has been a big focus for a few years now, and the more recent field of **neural rendering**,*" Dominik explains.



My summer internship on semi-supervised learning at UCL

by Christina Bornberg

Christina Bornberg recently started an Erasmus Mundus MSc in medical imaging and applications. She completed her undergraduate in electronics, was working on a fluorescein angiography classification project at the Medical University of Vienna and did a summer internship at UCL on semi-supervised vessel segmentation.

Over the summer, I had the opportunity to do a research internship at the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at University College London (UCL) under the supervision of Sophia Bano and Francisco Vasconcelos. The WEISS center is located at Charles Bell House and is home to engineers, clinicians and computer scientists, who are working together on developing technology for patients. My task was to perform semi-supervised semantic segmentation of vessels in fetoscopy images.

Introduction to semi-supervised learning

I would like to start with a short introduction to semi-supervised learning (SSL) for everyone who isn'tcompletelyfamiliar with the concept, the others can skip to the next section. SSL is a technique to combine a (small) labelled dataset with a (larger) unlabelled dataset. The combined dataset helps the model to generalise better and it has proven to be useful in the medical image analysis field where manual image annotation is very costly and requires experts.

Most SSL solutions are based on assumptions such as smoothness, consistency, clustering, or low-density. Neighbouring data points are expected to belong to the same class, alternatively, the decision boundary lies in a low-density region.

There are multiple approaches to tackle SSL:

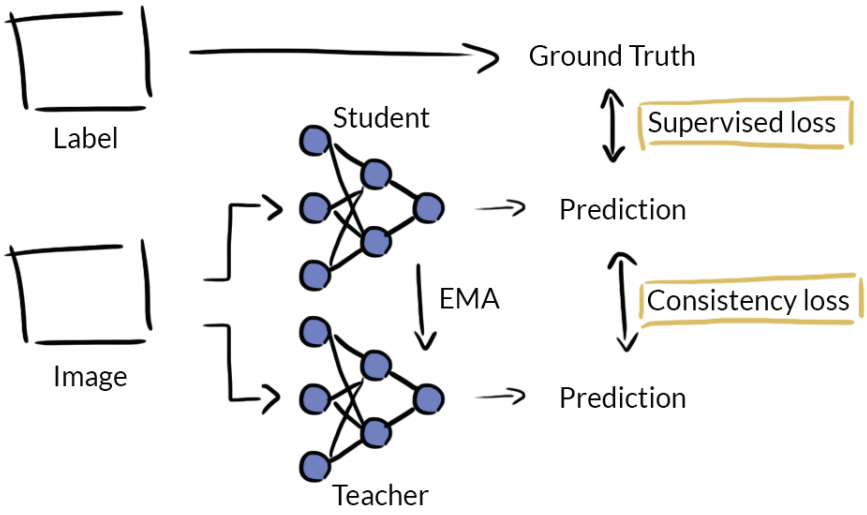
- Proxy-labelling as ground truth (e.g. self-training, co-training)

Proxy-labelling has the goal to generate pseudo-labels for unlabelled data in order to enhance the dataset and get more training samples.

Self-training, which I implemented, uses a model that is trained on labelled data in the first instance. Then, pseudo labels are iteratively generated for (portions of) unlabelled images and then used for the next training iteration.

Co-training uses two (or more) models, which are trained simultaneously and generate pseudo labels for each
other's unlabelled data after the initial training on labelled data. The models are supposed to agree on predictions and disagree on errors.

•Consistency regularisation (e.g. temporal ensembling, mean teacher)
Consistency regularisation follows the assumption that different perturbations produce the same output. Instead of considering predictions as ground truth, the distance of outputs is minimized to achieve consistency.

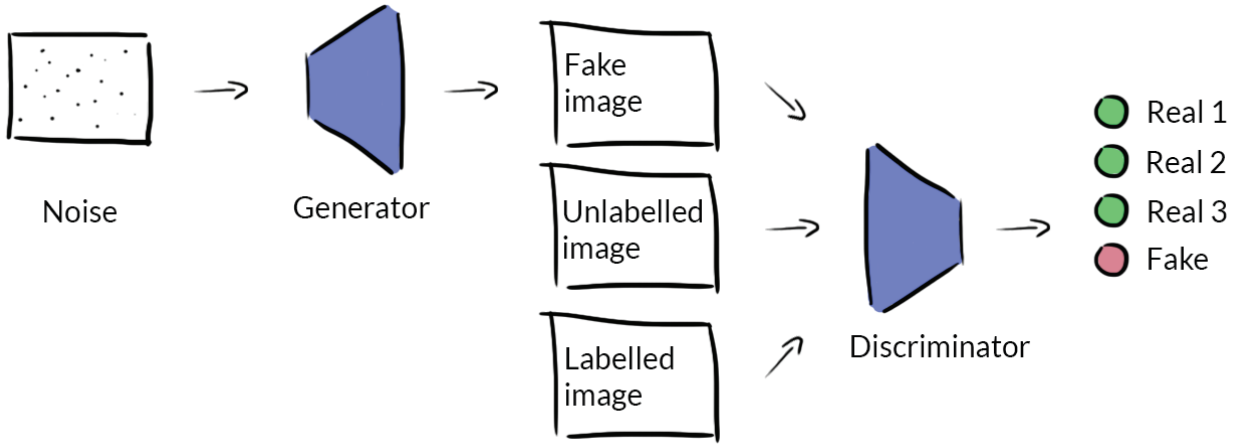


Semi-supervised learning using a student-teacher approach
with consistency loss and exponential moving average (EMA).

•Proxy-labelling with hybrid methods (e.g. MixMatch, FixMatch)
MixMatch combines entropy minimisation and consistency regularisation.
FixMatch first predicts pseudo labels from weakly augmented data and later uses these as the ground truth for the same images, but strongly augmented.

• GAN (e.g. SGAN):
Semi-Supervised Generative Adversarial Networks (SGAN) consist of a generator network and a discriminator network. The generator is

supposed to generate samples that look similar to the real data while the discriminator needs to learn to distinguish between real and fake samples. This is done to learn features, later on the discriminator is trained in a supervised manner to predict classes of the real examples. Alternatively, the discriminator can be trained simultaneously to predict K+1 classes, including K real classes and the fake class.



Semi-supervised learning using GAN.

- Graph-based (e.g. label propagation)

Label propagation is a transductive algorithm that finds groups in a graph and therefore is able to label unlabelled data based on labelleddata. The assumption followed is that data points close to each other belong to the same group.

If you are interested in learning more, I'd recommend the paper "**An Overview of Deep Semi-Supervised Learning**" which is available on arXiv. Additionally, both Papers with Code: Semi-Supervised Semantic Segmentation and Semi-supervised learning for medical image segmentation: SSL4MIS on GitHub show recent advances in the semi-supervised world.

How to deal with certainty?

For the methods using pseudo-labelling or consistency regularisation, certainty is taken into account to filter or weight labels. Certainty is information that can easily be calculated however it is often ignored. To get the model certainty for a sample (not to be confused with the probability output), approaches like Monte Carlo Dropout (dropout at test-time) or test-time augmentation (TTA) can be used. In both cases, an image is put through a network multiple times and leads to different outputs due to the randomly dropped nodes or different

augmentations. By using the standard deviation of the mean of all outputs for one image, the certainty
can be derived.

Dataset

The dataset I used is publicly accessible here: https://weiss-develop.cs.ucl.ac.uk/ and is provided by the WEISS center itself. It consists of manually annotated frames and frames with predictions obtained with leave-one-out cross-validation and was combined with additional in-house unlabelled images. I used the annotated images for the initial training and then generated pseudo masks for the unannotated frames to enhance the dataset for further training the self-teaching model and for teaching the student model.

Example fetoscopy image with manually annotated ground truth mask.

My implementation

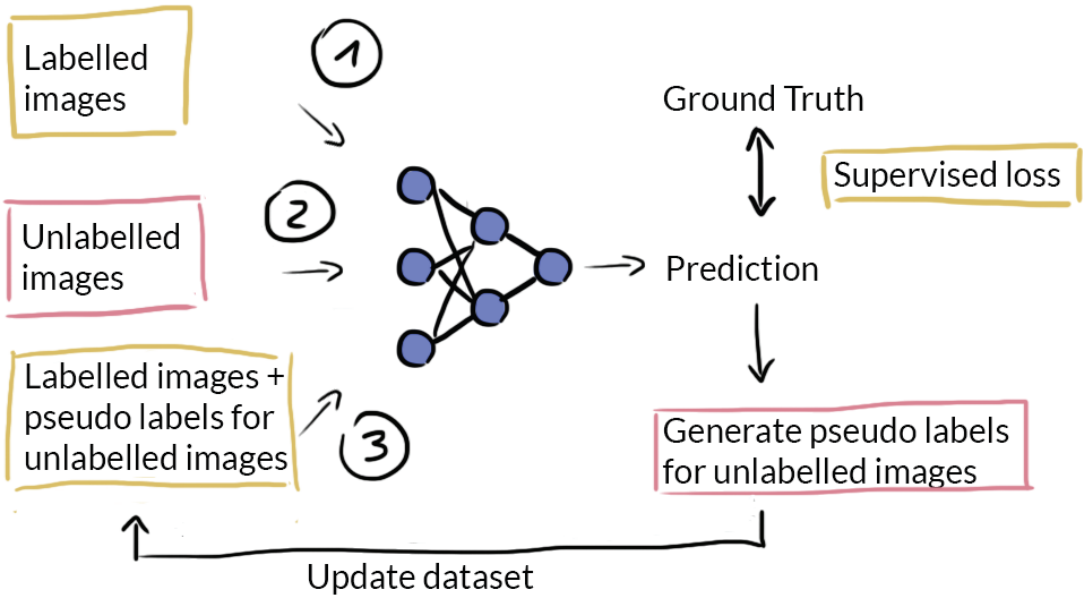
I implemented two simple semi-supervised versions, a self-teaching network, and the second one followed the student-teacher approach. I used the U-Net architecture with a ResNet-50 backbone for PyTorch which is available on GitHub.

```
import segmentation_models_pytorch as smp
self.model = smp.Unet(
        encoder_name="resnet50", # choose encoder, e.g. mobilenet_v2 or
        efficientnet-b7
        encoder_weights="imagenet", # use `imagenet` pre-trained weights for
        encoder initialization
        in_channels=1, # model input channels (1 for gray-scale images, 3 for
        RGB)
        classes=1, # model output channels (number of classes in your dataset)
        )
```

For the self-teaching network, features are initially learned in a supervised way while only using manually annotated data until it reaches an f-score above a certain threshold. ThenI used the unlabelled data to generate pseudo-labels, while using test-time augmentation to accomplish certainty estimation and majority voting while only using masks with a certainty above a threshold. For the continuation of the training, both manual and pseudo labels are used, and pseudo labels are added/updated iteratively.

```
for iteration in range(1, configs.num_epochs+1):
        # train initially on labelled (1) and further on combined dataset (3)
        train_mode.reload_data(configs) # labelled images and pseudo labels
        for unlabelled images
        fscore = train_mode.train(configs, iteration)

        if fscore > configs.threshold:
        # generate pseudo labels (2)
        pseudo_mode.reload_data(configs) # unlabelled images
        pseudo_mode.generate_labels(configs, iteration)
        # set high certainty pseudo labels for unlabelled images
        pseudo_mode.update_dataset(configs)
```



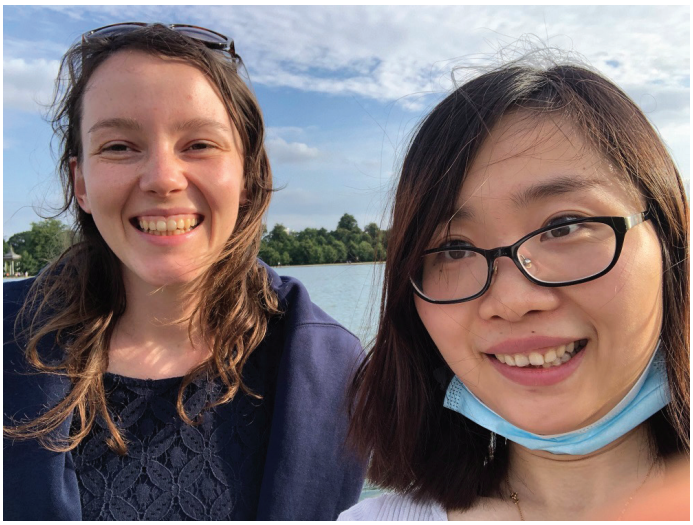
Self-training network: (1) Initial training on labelled images. (2) Generate pseudo labels for unlabelled images. (3) Re-train on combined dataset.

For comparison, I implemented a simple student-teacher approach using two equally sized models. Network A is trained on labelled images and pseudo labels for network B are generated by network A with the highest f-score.

Something I found really useful while organising the code was creating block diagrams and flowcharts and would recommend everyone to do it if it's not already part of your workflow.

Living in London

I want to end my article a bit off-topic with recommendations if you get the chance to live in the beautiful city of London! I can totally recommend musicals like *"Back to the Future"*, get street food in Camden market, walk along Regent's Canal or Hyde Park and watch out for free public events like open-air cinemas or concerts. During my time here, I was able to meet up with **Chen Chen**, whom I met online at MICCAI 2020. I really enjoyed learning about her work in the medical image analysis field at ICL and sharing my progress on my vessel segmentation project. Last, but not least, I would like to thank Sophia Bano, Francisco Vasconcelos, and the rest of the WEISS team for supporting and welcoming me!



MEDICAL IMAGING NEWS

Pietro Mascagni is a medical doctor, concluding a PhD in surgical data science between the Catholic University of Rome; ICube, University of Strasbourg; and IHU Strasbourg. His supervisors are Professor Guido Costamagna in Rome from the clinical side and Professor Nicolas Padoy in Strasbourg from the computer science side. He speaks to us about his work at the intersection of the two communities.

Pietro has a bold mission:
"I want to bridge the gap between surgical need and computer science to innovate for patients, surgeons, and healthcare systems."

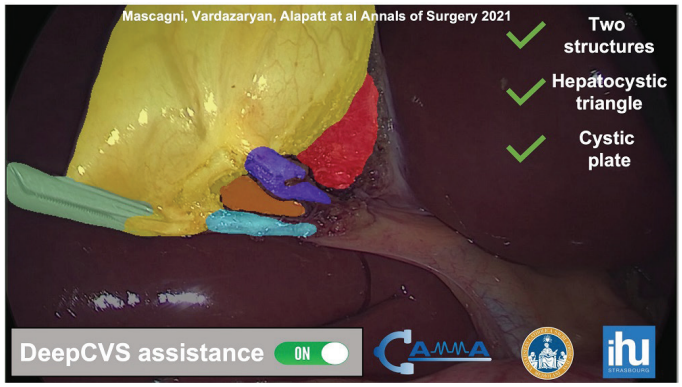
But how can innovations the computer science community has fostered over the years be integrated into the clinical workflow to bring value to healthcare? Pietro tells us **both sides need to communicate, collaborate, and cross-pollinate**. This calls for people who have a foot in both camps and can translate clinical need into scientific questions to be answered by means of computer science. Also, multidisciplinary efforts are needed to develop datasets and metrics that work for both communities.

Nassir Navab told us that he insists engineers in his lab meet regularly with clinicians to facilitate the exchange of information and culture.

Pietro Mascagni

"Professor Navab is a pioneer in this field and fully grasps the need for this reciprocal contamination," Pietro attests. *"Computer scientists must understand how their work fits in the broader picture of clinical care if they want to have an impact. This works both ways. Physicians often over or underestimate what can be done or have completely the wrong idea about technology. Visionaries like Professor Navab and computer science communities like MICCAI, who have just launched their first clinical day, are recognizing that we need to team up."*

The surgical community is taking a step towards computer science too. Half of the members of the **European Association of Endoscopic Surgery (EAES)**, of which Pietro is on the technology committee, are engineers. The same is true for the **AmericanSociety for Gastrointestinal Endoscopy (ASGE)**.



"Modern surgery has become a complex socio-technical process, with multiple highly specialized professionals coordinating and taking decisions in seconds, interpreting signals from highly technological devices," Pietro explains. *"The process of surgical care, when it goes smoothly, is highly effective, but it's also highly error prone, and the complexity can be described as a problem of information flow. We have analytics that are great at modeling this kind of unstructured information and could be used to deliver the right information to the right person at the right time."*

This is the vision that motivated computer scientists and surgeons to team up in recent years to establish the field of surgical data science, which has huge potential for **improving preoperative, intraoperative, and postoperative care**. Pietro is focused on translating that vision into something concrete by developing software that supports surgeons intraoperatively in operating rooms. 65% of medical error happens there, and an estimated 50% of complications are preventable.

"The fact is we know very little about what happens in the operating room because it is kind of a black box," he points out. "The only direct information on operating room activity is operator-written reports, so I'm using videos to shine a light on what happens."

The practical test case for this is **cholecystectomy** - the surgical removal of the gallbladder. It is a very common abdominal surgical procedure performed with a minimally invasive laparoscopic approach in 98% of cases. An endoscopic camera is inserted into the abdomen and a video guides the surgeon's movement, which is digital information that could be analyzed by deep learning and computer vision methods.

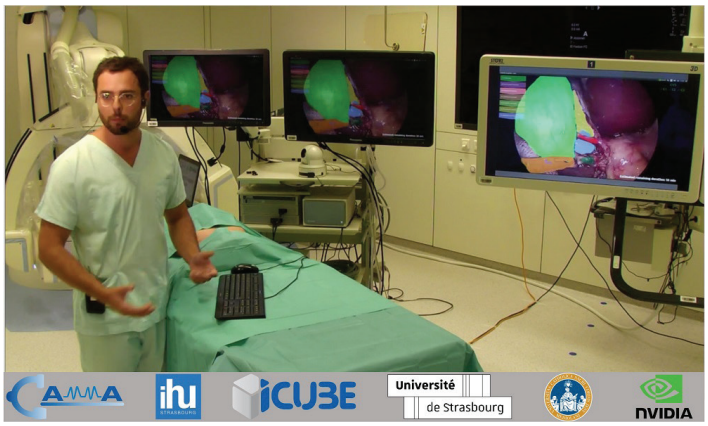
"We want to improve safety in cholecystectomy, in particular the risk of bile duct injury," Pietro tells us. *"The gallbladder sits underneath the liver, connected by the biliary tree to the gut. You need to detach the gallbladderfrom the biliary tree to remove it. Every clinical outcome improved with the introduction of the minimally invasive technique, except the rate of bile duct injury. When there is a bile duct injury it's not usually a technical mistake.*

What happens is the visual perception illusion. The procedure is guided by 2D or 3D images and the human brain is wired in a way that it tends to see what it expects to see, so if the common bile duct is retracted towards the gallbladder, then the surgeon can think they're cutting the cystic duct while actually dividing the common bile duct."

This complication happens three times more in **minimally invasive surgery** than it did in **open surgery**, where there would be haptic as well as visual feedback. Even though the incidence of this adverse event is low, at between 0.3% and 1.5%, given the high number of procedures performed each year, this translates to a huge number of patients. Patients experiencing a bile duct injury have a three-fold mortality increase at one year. This is devastating for surgeons too and is estimated to cost $1 billion every year in the US alone.

"A solution already exists," Pietro reveals. *"In 1995, surgeon Steven Strasberg proposed a secure target identification technique often used in hunting – you don't shoot an animal before you know for sure it's an animal. The same type of concept applies in surgery. You don't clip the cystic duct until you're sure it's the cystic duct."*

Strasberg called it the **critical view of safety (CVS)** and it is defined by three anatomical landmarks. However, despite the fact there is a large surgical consensus around it, a recent review found the rate of bile duct injury was not decreasing.



The reason for this lies in the fact it is a low-incidence problem, so it tends not to be a priority in the surgeon's mind when operating. Video-based assessment shows CVS is achieved in only 9-20% of the cases. In 99% of cases, the procedure is still successful, but the error trap lies in the other 1%. Surgeons must perform CVS systematically so that it is there when they need it.

"Moving continuously between the clinics and computer science labs has been the most exciting part of the work we've been doing," Pietro tells us. *"I must stress the 'we' because this is thework of a surgical data science team.*

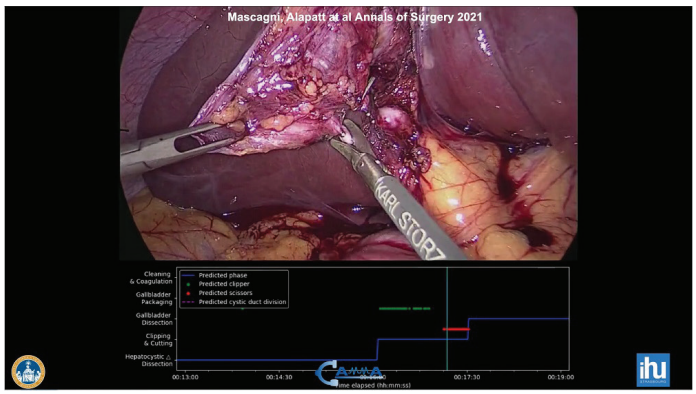
On most of my publications I have co-first authors who are engineers and

computer scientists. We do different jobs, but we work together, and it's highly rewarding. It's been a big learning opportunity for me, but I hope also for Deepak Alapatt and Armine Vardazaryan, who are the two main research engineers I work with under Nicolas's supervision."

Pietro says both sides have had to converge in order to speak the same language.

"I had to learn how to code because it forces you to better understand what you're trying to do," he points out. *"It allows you to be more autonomous in running these models, analyzing them on your own, and following, understanding, and improving the computer science research a little bit as well. The point isn't to devise these tools myself – I don't want to steal the work of computer scientists and the computer scientists I work with are not interested in doing what I'm doing!"*

Four years ago, Pietro ended up in Strasbourg by chance with a mobilityscholarship from the European Union.



He says he got "super passionate" about the research being performed at IHU Strasbourg and was lucky enough to win a research grant from EAES. At that point, Nicolas offered him the opportunity to do a PhD with his computer science team.

"My clinical supervisors were quite puzzled at first!" he reveals. *"They were afraid it would take me away from my profession because at the end of the day, I'm a medical doctor."*

But Pietro took a chance, and four years later, he is very glad he did. "I'm super, super happy!" he grins. *"My supervisors, who were initially skeptical, are happy now too. I'm even more motivated than before to go back to clinical practice. In November, I will be starting a residency and will eventually be translating what we've been developing in the lab to the operating room."*

Pietro will be defending his PhD thesis in the spring, and he promises to return to our pages to continue this discussion then.

He teases: *"Next time, I will tell you a story of how a clinical need drives innovation in computer science, and how observation done with computer science methods are fed back to the surgical department to improve the quality of the care they deliver!"*

Amelia Jiménez-Sánchez recently completed her Ph.D. at **Pompeu Fabra University**. Her research aims at **learning representations** for **medical image diagnosis** facing common medical imaging **data challenges**, namely limited data, class-imbalance, noisy annotations and data privacy. She holds a degree in Telecommunications Engineering from the University of Granada and a Master of Science in Biomedical Computing from the Technical University of Munich. She's been awarded the mention "Cum Laude" for her Ph.D. **Congrats cum laude, Doctor Amelia!**

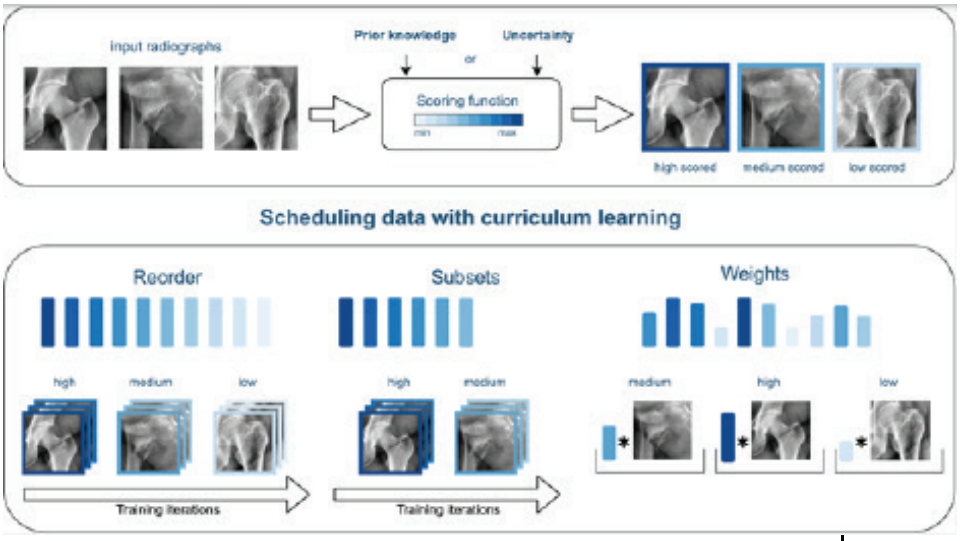
In this thesis, we investigate **two key aspects** to learn **feature representations** leveraging Convolutional Neural Networks (**CNNs**) from medical images for Computer-Aided Diagnosis (CAD) tasks. First, we explore the role of **architectural design** in dealing with spatial information. Second, we design **curriculum training** strategies to control the order, pace, and number of images presented to the optimizer.

Capsule networks

CNN's requirement for big amounts of data is commonly justified by a large number of network parameters to train under a non-convex optimization scheme. We argue, however, that part of these data requirements is there to cope with their poor modeling of **spatial invariance**. **Capsule networks** were introduced as an alternative deep learning architecture and training approach to model the spatial/viewpoint variability of an object in the image. We experimentally demonstrate that the **equivariance** properties of capsule networks **reduce the strong data requirements**, namely limited amount of data and class-imbalance, and are therefore very promising for medical image analysis.

Ordering and pacing

In a typical educational system, learning relies on a curriculum that introduces new concepts building upon previously acquired ones. The rationale behind, is that humans and animals learn better when information is presented in a meaningful way rather than randomly. We design **Curriculum Learning (CL)** strategies for the fine-grained classification of **proximal femur fractures** according to the AO standard. Our novel framework reunites three strategies consisting of individually **weighting** training samples, **reordering** the training set, or **sampling subsets** of data. We define the scoring function from domain-specific **prior knowledge** or by directly measuring the **uncertainty** in the predictions.

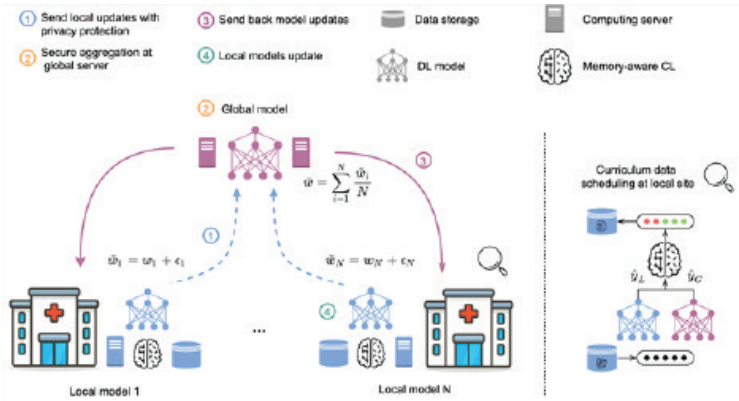


Training a CNN with CL.

Curriculum federated learning

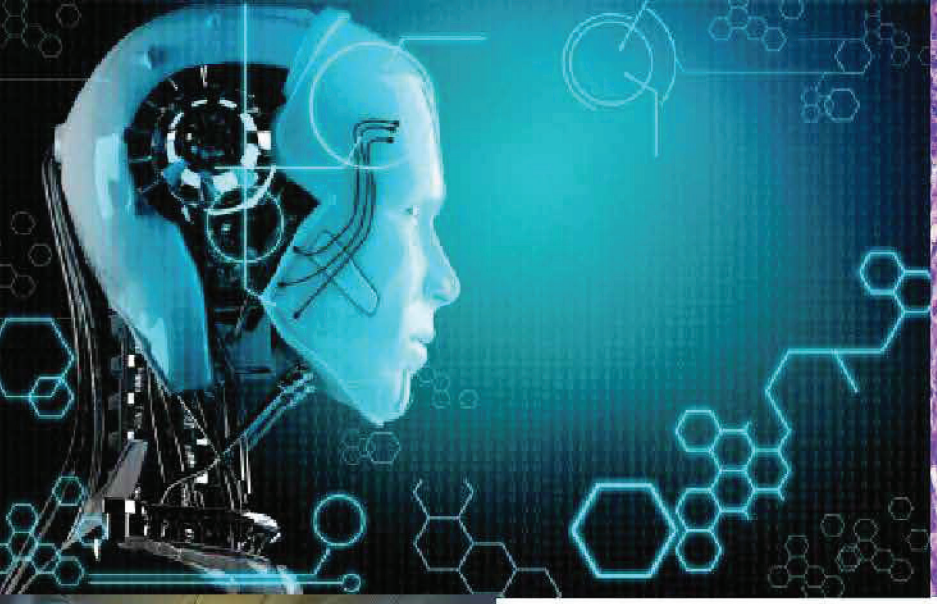
A potential solution to mitigate the frequent class-imbalance problem and to increase the size of annotated medical datasets is to employ data coming from multiple institutions. **Federated Learning (FL)** aims to train a machine learning algorithm across multiple decentralized nodes holding samples locally. Training such a decentralized model in a FL setup presents three main challenges: (i) system and statistical heterogeneity, (ii) data protection, and (iii) distributed optimization.

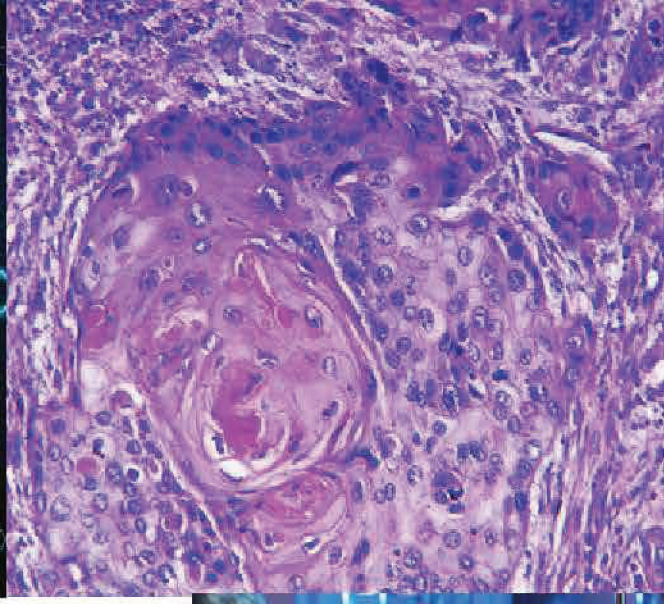
(i) We employ **federated adversarial learning** to deal with the alignment between the different domains. (ii) We leverage **differential privacy** to handle data protection. (iii) We propose a **novel CL** strategy for the FL setting (iii). We show that, by **monitoring the local and global classification predictions**, we can schedule the training samples to boost the alignment between domain pairs and improve the **breast cancer** classification performance.



Memory-aware curriculum federated learning framework with data privacy protection.

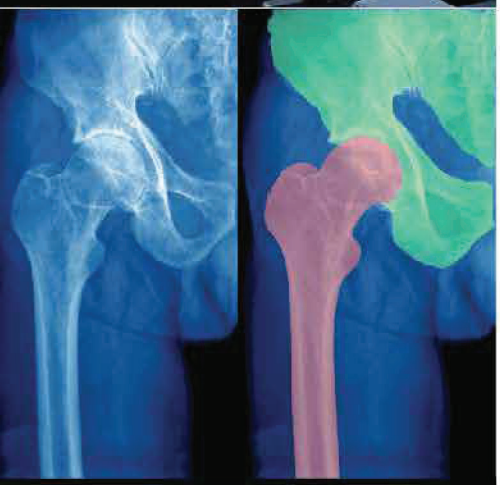
(*) Amelia would like to thank her advisors Gemma Piella and Diana Mateus for their support and guidance throughout these years. The work of this thesis would not have been possible without great collaborators and funders (EU's Horizon 2020 Marie Skłodowska Curie programme and "La Caixa" Foundation).





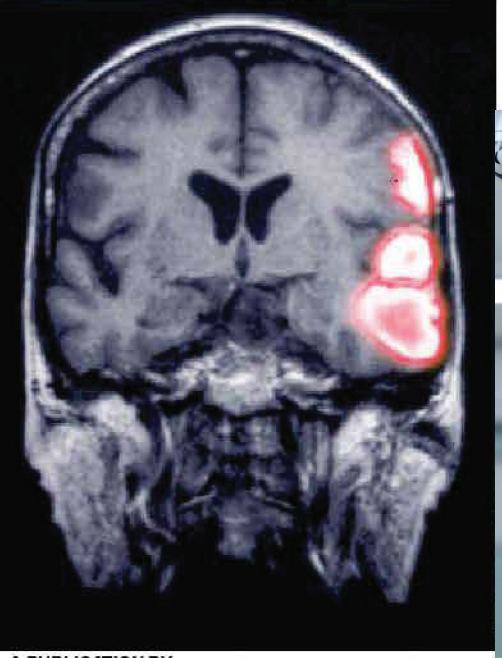
IMPROVE YOUR VISION WITH **Computer Vision News**

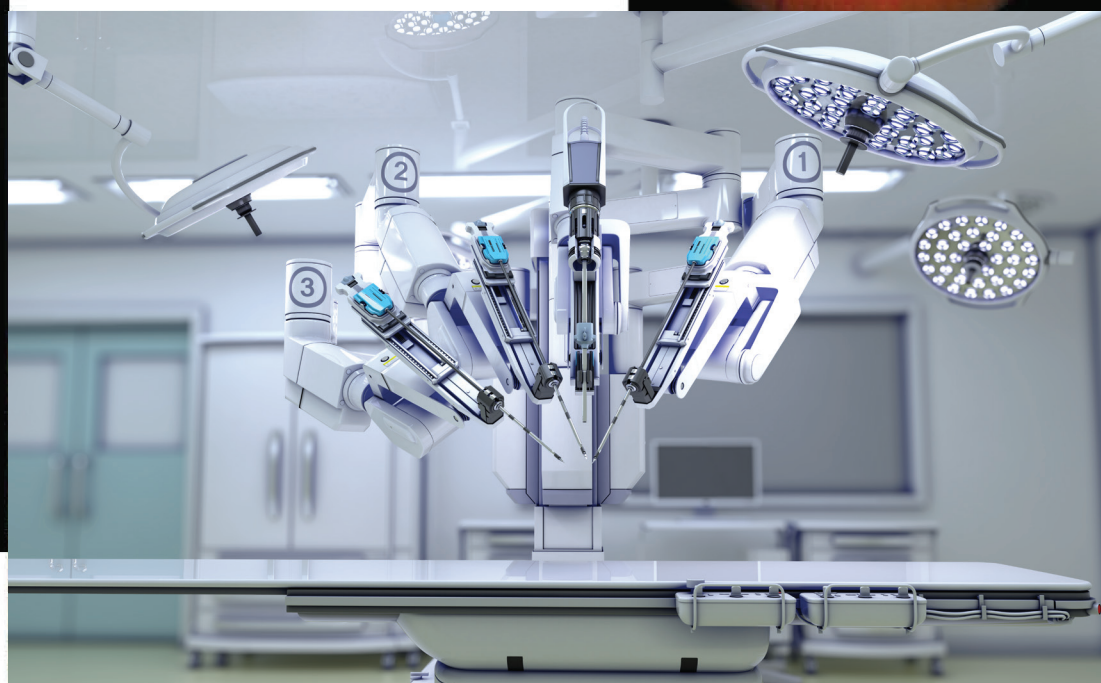
SUBSCRIBE
CLICK HERE, IT'S FREE



To the magazine of the algorithm community

and get also the ***New*** **supplement Medical Imaging News!**





A PUBLICATION BY

RSIP VISION

Global Leader in Computer