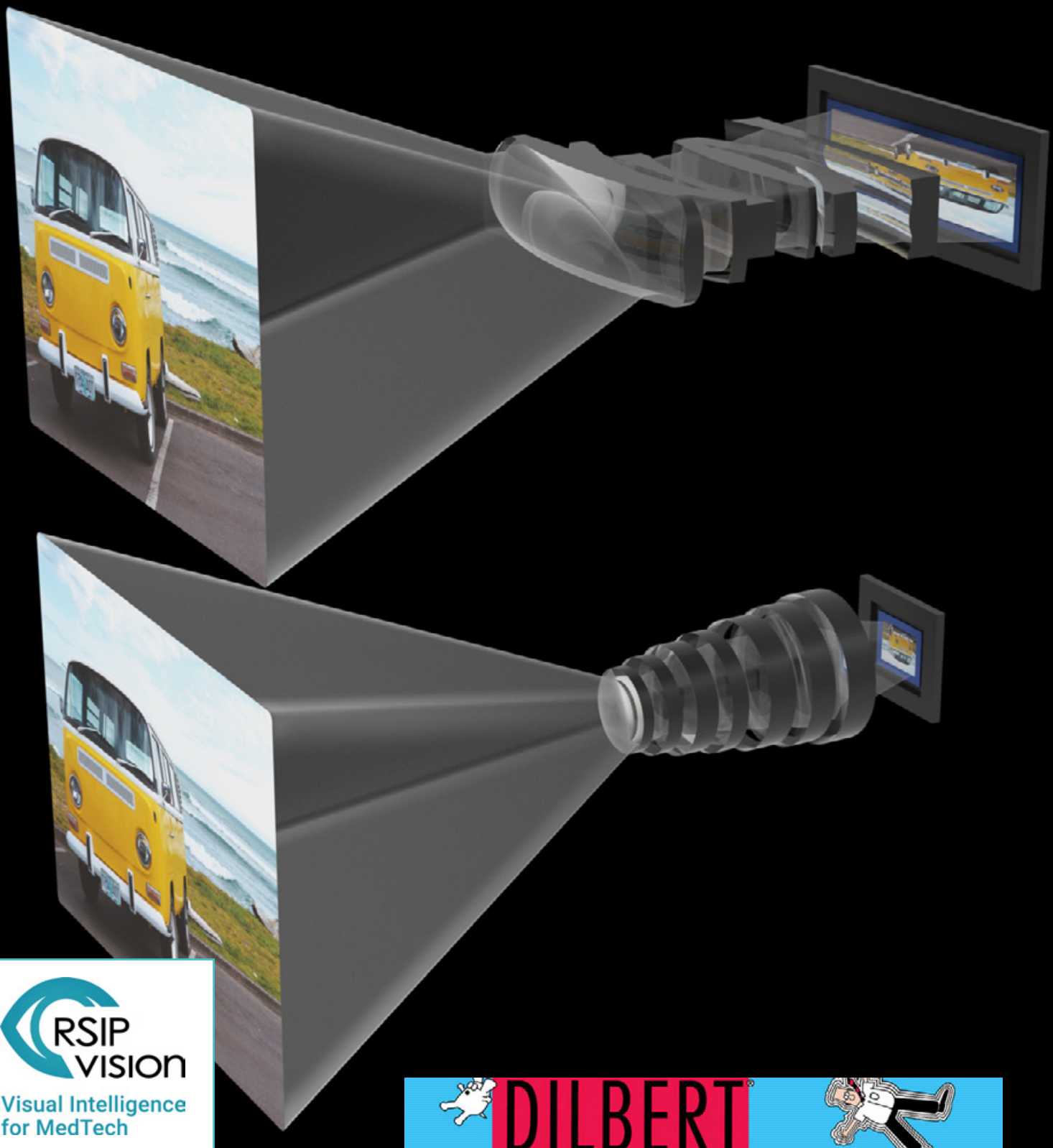
AUGUST 2022

Computer Vision News

& Medical Imaging News

The Magazine of the Algorithm Community

BEST OF MIDL 2022



This photo was taken in peaceful, lovely and brave Odessa, Ukraine.

Computer Vision News

Editor:
Ralph Anzarouth

Engineering Editors:
Marica Muffoletto
Ioannis Valasakis

Publisher:
RSIP Vision

Contact us

Free subscription

Read Previous Magazines

Follow Us

Dear reader,

It's been another huge month for the computer vision community! After a stunning **CVPR 2022 in New Orleans**, everyone has much to think about and build on. We'll be discussing the technology and innovation on display for a long time to come. Huge congratulations to all involved!

Between **CVPR Dailies** and the **BEST OF CVPR**, our magazine collected **more than a quarter of a million page views in only a few weeks!** Our audience and followers clearly appreciate our work reporting the most popular conference in our field for the 7th consecutive year. **We thank you a quarter of a million times!**

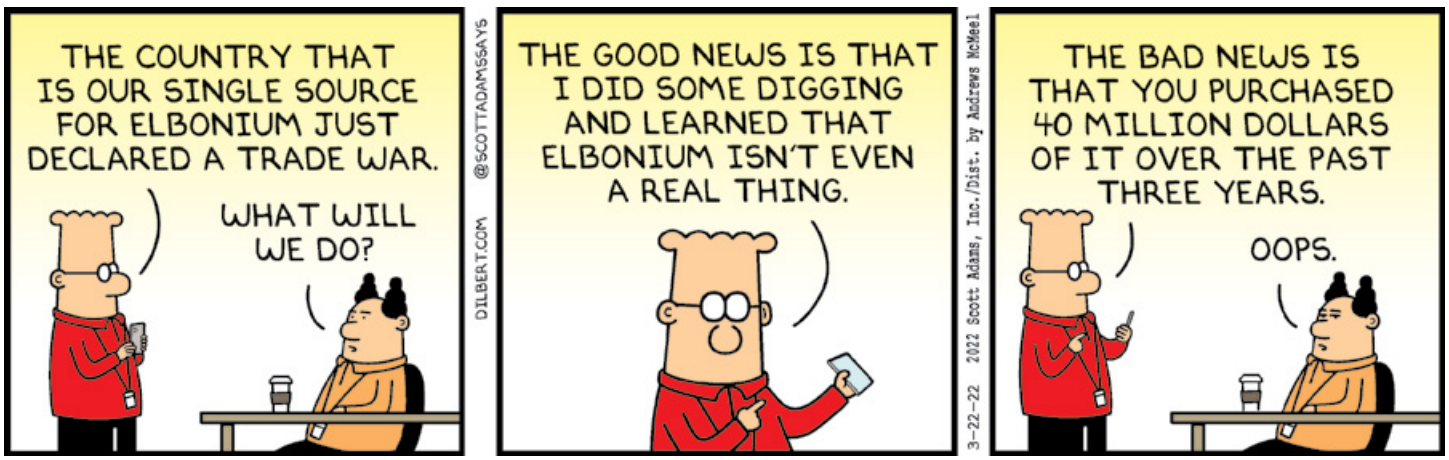
It's been a busy month for our new supplement **Medical Imaging News** with several exciting in-person events on the calendar! The MICCAI-endorsed **MIDL (Medical Imaging with Deep Learning)** conference was a big success and saw many enthusiastic participants on-site in Zürich. We've got three awesome reports featuring the **Best Paper Award** winner and two runners-up in our **BEST OF MIDL** special on pages 28, 32 and 36.

RSIP Vision have been out in force at the celebrated **Hamlyn Symposium** in London, absorbing fascinating surgical robotics learnings for our R&D work and inspiring content for our readers. We review the impressive winner of its **Best Innovation Prize** on page 44.

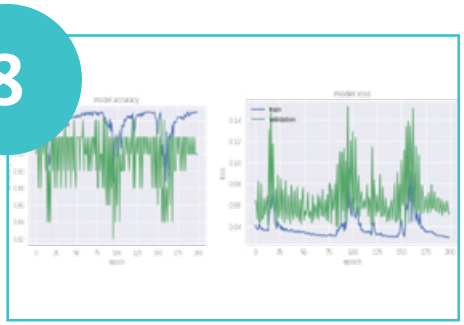
Last but not least, on page 48, you can find our report on the **International Workshop on Biomedical Image Registration (WBIR 2022)**. Next month, we'll publish features on the two works that won the **WBIR Best Paper and Audience Awards**. We'll also be preparing for **MICCAI 2022** in Singapore – back in person for the first time in three years!

We hope you enjoy this August issue of **Computer Vision News**. We'd love it if you could tell your friends about us, and don't forget to **subscribe for free!**

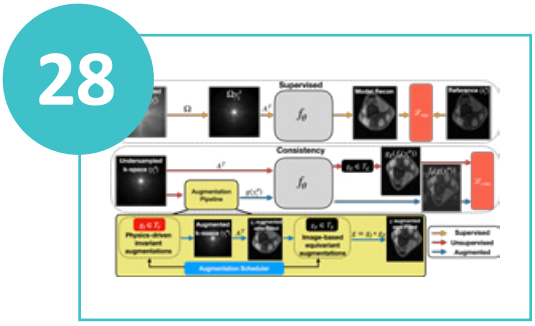
Ralph Anzarouth
Editor, **Computer Vision News**
Marketing Manager, **RSIP Vision**

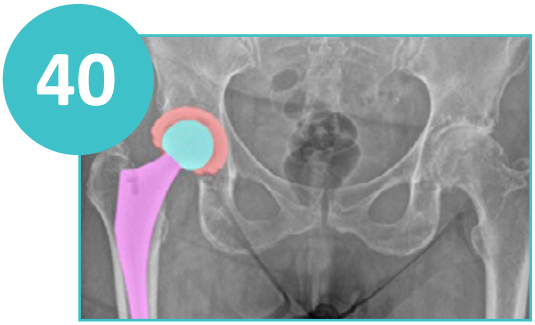


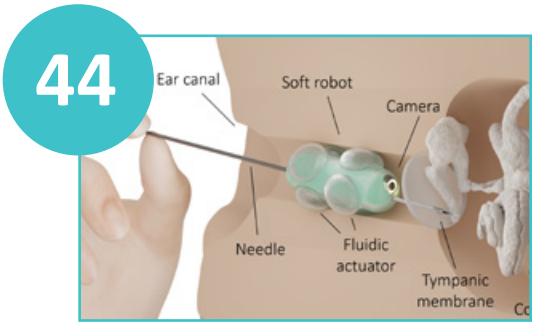
Computer Vision News

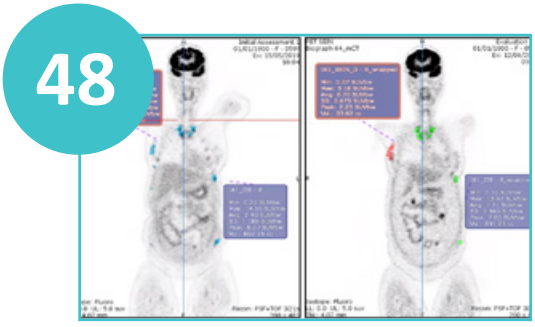


Medical Imaging News









BEST OF MIDL 2022

WINNER OF INNOVATION PRIZE HAMLYN

SEE EYE TO EYE

By Marica Muffoletto (twitter)

This month we are reviewing the paper entitled: **See Eye to Eye: A Lidar-Agnostic 3D Detection Framework for Unsupervised Multi-Target Domain Adaptation.** We deeply thank all authors (**Darren Tsai, Julie Stephany Berrio, Mao Shan, Stewart Worrall, Eduardo Nebot**) for allowing us to use their images.

We start with a question. What is a LIDAR? This word stands for laser imaging, detection, and ranging and it's used to describe sensors which determine ranges and distances from objects, using light properties. LIDAR sensors in combination with 3D detection techniques can be applied to different fields, such as the one this paper focuses on- autonomous vehicles. The performance of state-of-the-art 3D detectors across different lidars widely changes. And therefore, the authors of this paper are looking into **Unsupervised Domain Adaptation (UDA)** techniques which can bridge these performance gaps between lidars. According to the discussed state-of-the-art, Yang et al. beat previous methods, using a self-training approach which generates high-quality pseudo-labels. Unfortunately, this still suffers from a big limitation: it doesn't work on lidars with adjustable scan pattern. Hence, Darren Tsai and colleagues propose a UDA method called **SEE** that works on **both fixed and adjustable scan pattern lidars without requiring fine-tuning a model for each new scan pattern**. This is based on a scan pattern agnostic representation of objects to enable a trained 3D detector to perform on any lidar pattern.

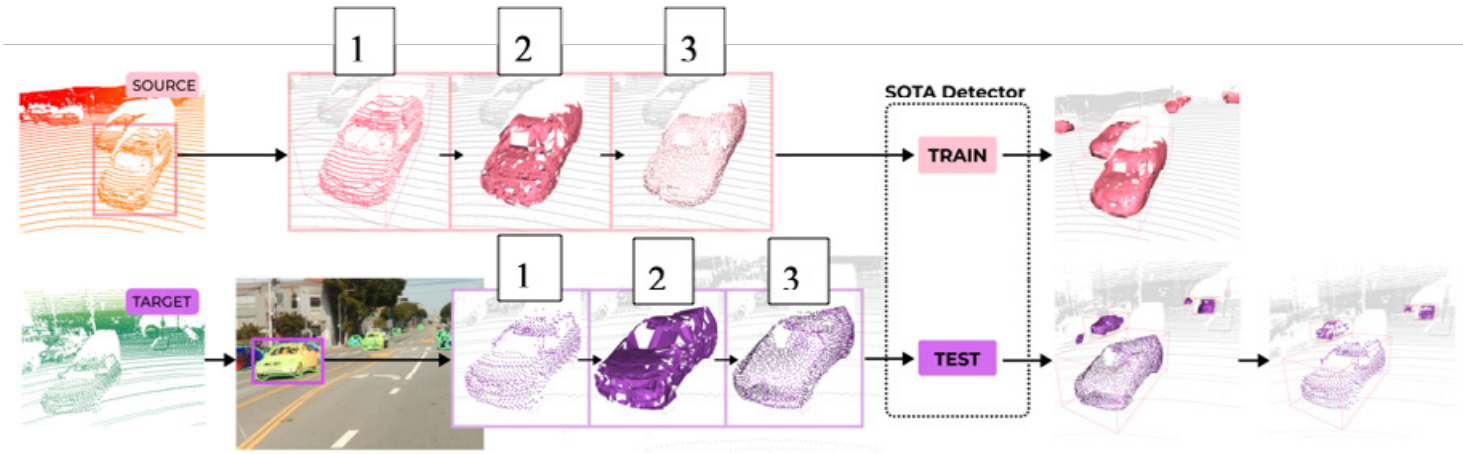


Figure 1: Overview of proposed method- SEE

The pipeline of SEE includes: 1. **Object isolation**. This step is different for the source domain- where ground truth boxes are available and can be used to crop the point cloud- and the target domain, where image instance segmentation with clustering is employed. This point seems to be an essential and particularly challenging one. This can be observed in Figure 2, where loosely fitted instance masks and calibration errors lead to inclusion of points which do not belong to the main object (vehicles). The solution to this seems to be having a well-calibrated lidar and camera pair with minimal viewpoint misalignment. Hopefully, this can approximate the ideal scenario of ground truth bounding boxes in the target domain, which is discussed later on in the article; 2. **Surface Completion**: the Ball-pivoting Algorithm is used to interpolate a triangle mesh and recover the geometry. This seems to work well addressing the issue of partial occlusion (most recurrent in driving datasets); 3. **Point sampling**. Since points at closer ranges typically have more confident detections, the triangle meshes obtained in 2 are upsampled using Poisson disk sampling. This is done by emulating the point density of objects at a closer range, which should generally improve performance unless there are errors in the corresponding isolation phase for that particular object.

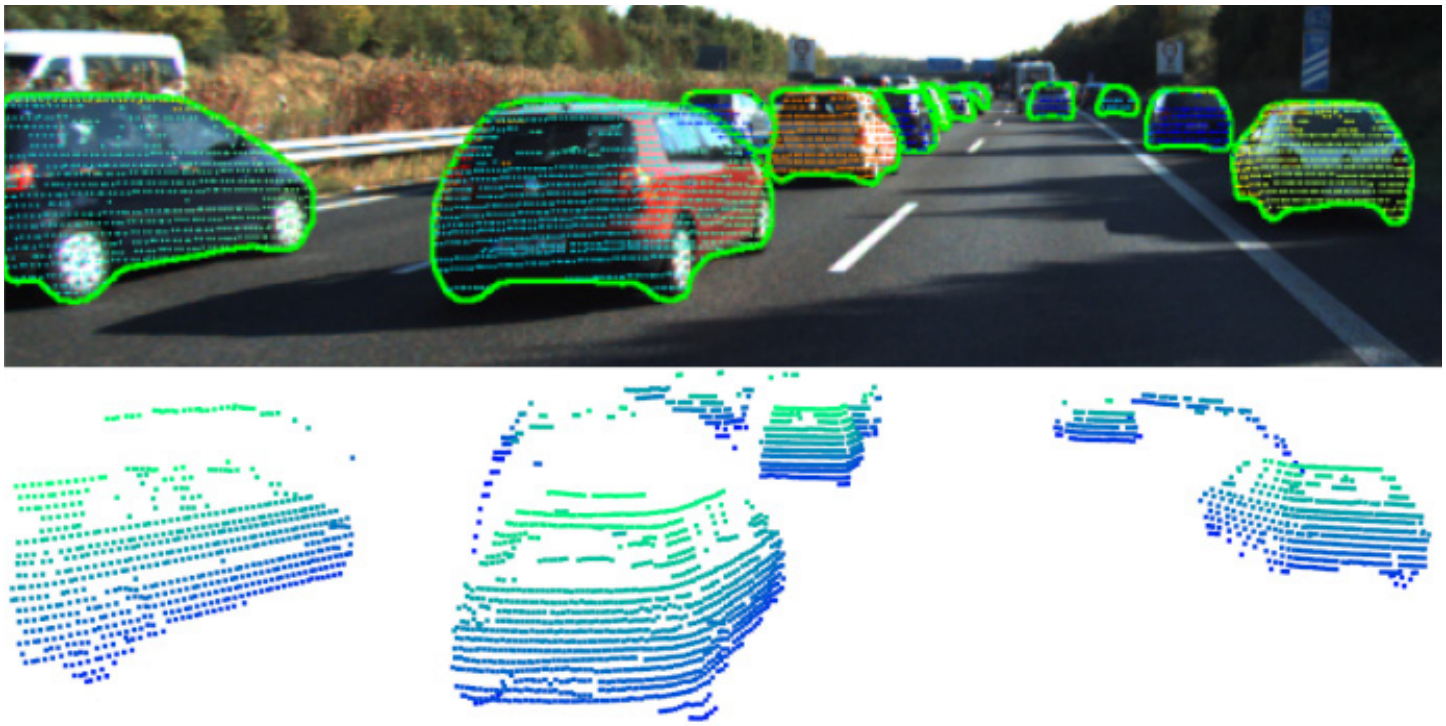


Figure 2: Issues with background points in KITTI dataset

The SEE method is validated on three public datasets (Waymo, KITTI, nuScenes) and a novel one (Baraja Spectrum-Scan™ dataset), on the “Car” or “Vehicle” class. The difference between the public datasets is shown in Figure 3, where it’s possible to observe the effect of different types of lidars and scan patterns on ring separation.

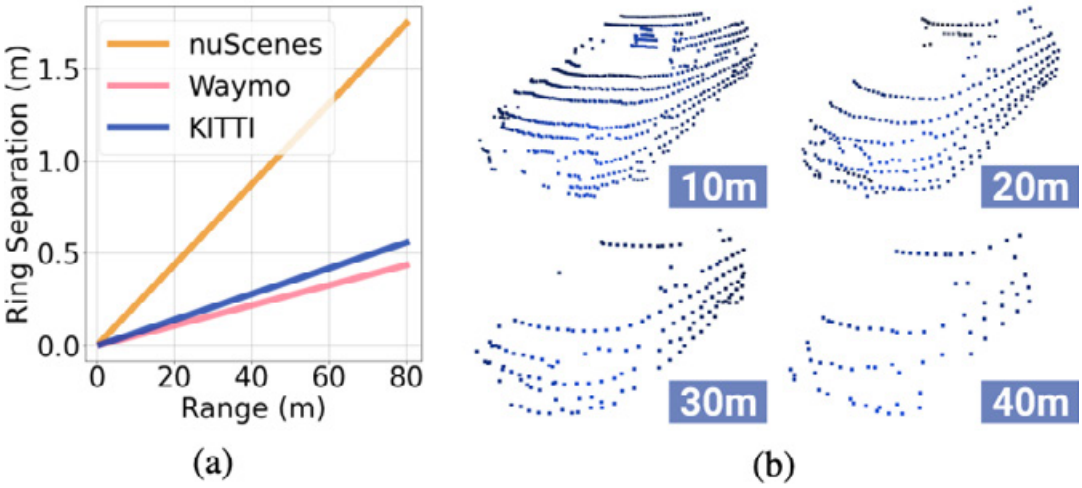


Figure 3: Difference between public datasets (a), KITTI ring separation (b)

The new dataset is manually labelled and uses a high-resolution lidar that can dynamically increase point cloud resolution around key objects and a camera Intel RealSense D435i, mounted directly below the lidar. The cars are labelled using the tool SUSTechPOINTS and only if they are visible in the image FOV.

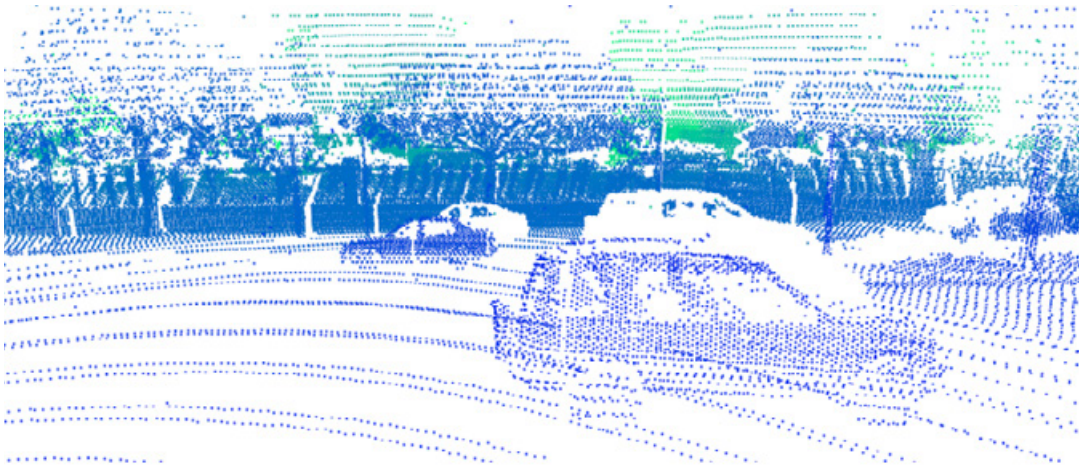


Figure 4: Example of Baraja Spectrum-Scan™ dataset

The experiments include:

- Domain adaptation between nuScenes -> KITTI:
  addressing difference in lidar ring numbers
- Domain adaptation between Waymo -> KITTI:
  addressing the use of multiple concatenated point clouds to single point cloud
- Domain adaptation between nuScenes/Waymo -> Baraja:
  ring-based to a uniform, interleaved scan pattern

On the above three tasks, the authors compare the following methods: 1) source-only (no domain adaptation), 2) ST3D (state-of-the-art method on 3D object detection using self-training, 3) SEE with segmentation algorithm for object isolation (step 1 in Figure 1), 4) SEE-Ideal with ground truth annotations to isolate the target domain objects, 5) Oracle, the fully supervised detector trained on the target domain. For validating the methods, the 3D detectors PointVoxel-RCNN and SECOND-IoU are used.

To perform the comparison, the metrics used is the average precision (AP) over 40 recall positions at 0.7 and 0.5 IoU (Intersection over Union) thresholds for both BEV (bird's-eye-view) and 3D IoUs.

In the first task (nuScenes -> KITTI), SEE closes the performance gap by 39.61 AP for SECOND-IoU and 24.49 AP for PV-RCNN in AP_{3D}. In the second task (Waymo -> KITTI), SEE closes the performance gap by 53.60 AP for SECOND-IoU and 37.85 AP for PV-RCNN in AP_{3D}. Here, the authors' approach outperforms ST3D with both SECOND-IoU and PV-RCNN in both IoU thresholds.

The table below shows the results of the last experiment on the new dataset, where green highlights the performance increase of models trained with SEE over the Source only method. Based on the lower performance in the nuScenes -> Baraja Spectrum-Scan™ dataset, it is possible to conclude that the domain gap between these is higher. This is due to the difference in scan patterns between the two lidars (Figure 3), which is drastically reduced in the Waymo dataset, which achieves much better performance in both Source-only and SEE methods.

Source	Method	SECOND-IoU				PV-RCNN			
		0.7 IOU		0.5 IOU		0.7 IOU		0.5 IOU	
		3D	BEV	3D	BEV	3D	BEV	3D	BEV
nuScenes	Source-only	1.02	5.12	6.53	7.10	10.85	13.74	14.46	14.50
	SEE	34.54	58.45	70.46	73.02	64.34	77.73	83.66	85.72
	Improvement	**+33.52**	**+53.33**	**+63.92**	**+65.93**	**+53.49**	**+63.99**	**+69.20**	**+71.22**
Waymo	Source-only	49.96	74.64	84.11	88.03	76.14	84.10	86.67	88.08
	SEE	73.79	84.74	90.36	92.53	79.13	87.79	93.05	93.17
	Improvement	**+23.84**	**+10.10**	**+6.25**	**+4.50**	**+2.98**	**+3.69**	**+6.38**	**+5.09**

The authors state that the performance between SEE and SEE-Ideal can be closed if the camera and lidar viewpoint alignment is minimised, as this reduces background points. Unfortunately, SEE doesn't run in real-time, but it seems the be the only method available which does not require new training on new lidars- especially if they have an adjustable scan pattern. This appears to be a large component of the domain gap. Through results and ablation studies, and due to the nature of the approach, it's concluded that SEE performs better with more points on the object, which anyways might be the general future direction in lidar manufacturing, and better surface completion methods.

We are already at the end of this one too! See you next month. If you have any suggestions for future articles or questions on the topics discussed, feel free to contact me.

DETECT GRAPHIC INTENSITY AND POWER IN VIDEOS USING CNN + LSTM NEURAL NETWORK

IOANNIS VALASAKIS, KING'S COLLEGE LONDON

@WIZOFE

Hello again! I hope that this month you'll be back for our amazing tutorial. I decided, after getting a few questions about video and CNNs and deep learning, to create a tutorial and modify using GitHub data resources, for detection

Play with the code, use new videos to train the network and send any questions you may have! Enjoy the article and coding ☺

Introduction

There are many instances where one may want to detect roughness in a video. A TV channel may be interested in that to protect its viewers, or for events happening in a public space. All systems today require human intervention and inspection to identify such cases. Using large datasets and deep learning we can train models which can automatically monitor and identify such videos with techniques such as object detection, tracking, action recognition, and legend generation.

Flowchart

The method consists of extracting a set of frames belonging to the video, sending them to a pretrained network called VGG16, obtaining the output of one of its final layers and from these outputs training another network architecture with a type of special neurons called LSTM. These neurons have memory and are able to analyze the temporal information of the video, if, at any time they detect violence, it will be classified as a violent video.

Helper Functions

We will use the function `print_progress` to print the number of videos processed and `download_data` to download the datasets

Load Data

Firstly, we define the directory to place the video dataset

```
in_dir = "data"
```

We set the url to download the dataset

```
url_hockey = "http://visilab.etsii.uclm.es/personas/oscar/FightDetection/HockeyFights.zip"
```

to download the dataset and decompress it:

```
download_data(in_dir,url_hockey)
```

Copy some of the data dimensions for convenience.

```
# Frame size
img_size = 224
img_size_touple = (img_size, img_size)
# Number of channels (RGB)
num_channels = 3
# Flat frame size
```

```
img_size_flat = img_size * img_size * num_channels
# Number of classes for classification (Violence-No Violence)
num_classes = 2
# Number of files to train
_num_files_train = 1
# Number of frames per video
_images_per_file = 20
# Number of frames per training set
_num_images_train = _num_files_train * _images_per_file
# Video extension
video_exts = ".avi"
```

Plot a video frame to see if data is correct

```
# First get the names and labels of the whole videos
names, labels = label_video_names(in_dir)
```

Then we are going to load 20 frames of one video, for example

```
names[12]
```

```
'fi191_xvid.avi'
```

The video has violence, look at the name of the video, starts with 'fi'

```
frames = get_frames(in_dir, names[12])
```

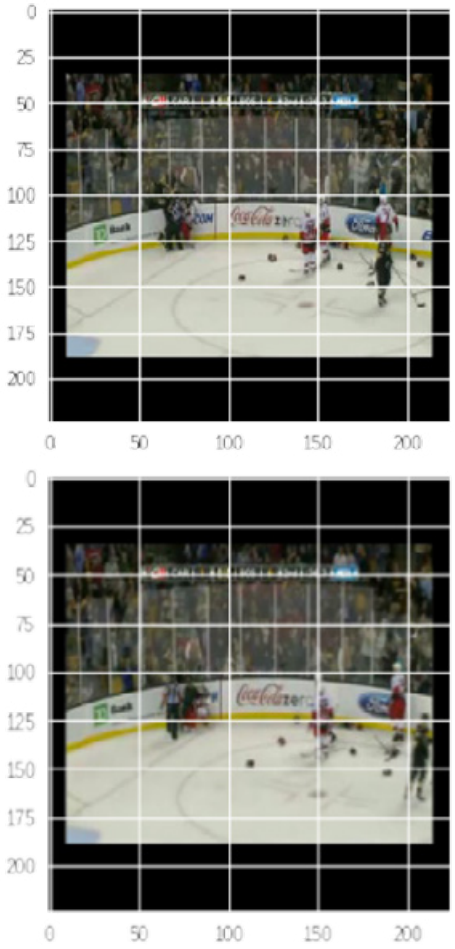
Convert back the frames to uint8 pixel format to plot the frame

```
visible_frame = (frames*255).astype('uint8')
```

```
plt.imshow(visible_frame[3])
```

```
<matplotlib.image.AxesImage at 0x7f37ef72fef0>
```

```
plt.imshow(visible_frame[15])
```



Pre-Trained Model: VGG16

The following creates an instance of the pre-trained VGG16 model using the Keras API. This automatically downloads the required files if you don't have them already.

The VGG16 model contains a convolutional part and a fully-connected (or dense) part which is used for classification. If include_top=True then the whole VGG16 model is downloaded which is about 528 MB. If include_top=False then only the convolutional part of the VGG16 model is downloaded which is just 57 MB.

```
image_model = VGG16(include_top=True, weights='imagenet')
```

```
image_model.summary()
```

We can observe the shape of the tensors expected as input by the pre-trained VGG16 model. In this case, it is images of shape 224 x 224 x 3. Note that we have defined the frame size as 224x224x3. The video frame will be the input of the VGG16 net.

```
input_shape = image_model.layers[0].output_shape[1:3]
input_shape
```

```
(224, 224)
```

VGG16 model flowchart

The following chart shows how the data flow when using the VGG16 model for Transfer Learning. First, we input and process 20 video frames in batch with the VGG16 model. Just before the final classification layer of the VGG16 model, we save the so-called Transfer Values to a cache file.

The reason for using a cache file is that it takes a long time to process an image with the VGG16 model. If each image is processed more than once then we can save a lot of time by caching the transfer values.

When all the videos have been processed through the VGG16 model and the resulting transfer values saved to a cache file, then we can use those transfer values as the input to LSTM neural network. We will then train the second neural network using the classes from the violence dataset (Violence, No-Violence), so the network learns how to classify images based on the transfer values from the VGG16 model.

```
# We will use the output of the layer before the final
# classification-layer which is named fc2. This is a fully-connected (or dense) layer.
transfer_layer = image_model.get_layer('fc2')

image_model_transfer = Model(inputs=image_model.input,
                             outputs=transfer_layer.output)

transfer_values_size = K.int_shape(transfer_layer.output)[1]

print("The input of the VGG16 net have dimensions:",K.int_shape(image_model.input)[1:3])

print("The output of the more select layer of VGG16 net have dimensions: ", transfer_
values_size)
```

```
The input of the VGG16 net has dimensions: (224, 224)
The output of the more select layer of VGG16 net has dimensions:  4096
```

Function to process 20 video frames through VGG16 and get transfer values

```
def get_transfer_values(current_dir, file_name):
    # Pre-allocate input-batch-array for images.
    shape = (_images_per_file,) + img_size_touple + (3,)
    image_batch = np.zeros(shape=shape, dtype=np.float16)
    image_batch = get_frames(current_dir, file_name)

    # Pre-allocate output-array for transfer-values.
    # Note that we use 16-bit floating points to save memory.
    shape = (_images_per_file, transfer_values_size)
    transfer_values = np.zeros(shape=shape, dtype=np.float16)
    transfer_values = \
            image_model_transfer.predict(image_batch)

    return transfer_values
```

A generator that processes one video through VGG16 each function call

```
def proces_transfer(vid_names, in_dir, labels):
    count = 0
    tam = len(vid_names)
    # Pre-allocate input-batch-array for images.
    shape = (_images_per_file,) + img_size_touple + (3,)
```

```
    while count<tam:
      video_name = vid_names[count]
      image_batch = np.zeros(shape=shape, dtype=np.float16)
      image_batch = get_frames(in_dir, video_name)

       # Note that we use 16-bit floating points to save memory.
      shape = (_images_per_file, transfer_values_size)
      transfer_values = np.zeros(shape=shape, dtype=np.float16)
      transfer_values = \
          image_model_transfer.predict(image_batch)

      labels1 = labels[count]
      aux = np.ones([20,2])
      labelss = labels1*aux

      yield transfer_values, labels
      count+=1
```

Functions to save transfer values from VGG16 to later use

We are going to define functions to get the transfer values from VGG16 with a defined number of files. Then save the transfer values files used from training in one file and the ones uses for testing in another one.

```
def make_files(n_files):
    gen = proces_transfer(names_training, in_dir, labels_training)
    numer = 1
    # Read the first chunk to get the column dtypes
    chunk = next(gen)
    row_count = chunk[0].shape[0]
    row_count2 = chunk[1].shape[0]

    with h5py.File('prueba.h5', 'w') as f:
        # Initialize a resizable dataset to hold the output
        maxshape = (None,) + chunk[0].shape[1:]
        maxshape2 = (None,) + chunk[1].shape[1:]
        dset = f.create_dataset('data', shape=chunk[0].shape, maxshape=maxshape,
                                chunks=chunk[0].shape, dtype=chunk[0].dtype)

        dset2 = f.create_dataset('labels', shape=chunk[1].shape, maxshape=maxshape2,
                                 chunks=chunk[1].shape, dtype=chunk[1].dtype)
         # Write the first chunk of rows
        dset[:] = chunk[0]
        dset2[:] = chunk[1]

        for chunk in gen:
            if numer == n_files:
                break

            # Resize the dataset to accommodate the next chunk of rows
            dset.resize(row_count + chunk[0].shape[0], axis=0)
            dset2.resize(row_count2 + chunk[1].shape[0], axis=0)

            # Write the next chunk
```

```
            dset[row_count:] = chunk[0]
            dset2[row_count:] = chunk[1]

            # Increment the row count
            row_count += chunk[0].shape[0]
            row_count2 += chunk[1].shape[0]
            print_progress(numer, n_files)
            numer += 1
def make_files_test(n_files):
    gen = proces_transfer(names_test, in_dir, labels_test)
    numer = 1
    # Read the first chunk to get the column dtypes
    chunk = next(gen)

    row_count = chunk[0].shape[0]
    row_count2 = chunk[1].shape[0]

    with h5py.File('pruebavalidation.h5', 'w') as f:

        # Initialize a resizable dataset to hold the output
        maxshape = (None,) + chunk[0].shape[1:]
        maxshape2 = (None,) + chunk[1].shape[1:]


        dset = f.create_dataset('data', shape=chunk[0].shape, maxshape=maxshape,
                                chunks=chunk[0].shape, dtype=chunk[0].dtype)

        dset2 = f.create_dataset('labels', shape=chunk[1].shape, maxshape=maxshape2,
                                 chunks=chunk[1].shape, dtype=chunk[1].dtype)

         # Write the first chunk of rows
        dset[:] = chunk[0]
        dset2[:] = chunk[1]

        for chunk in gen:
            if numer == n_files:
                break

            # Resize the dataset to accommodate the next chunk of rows
            dset.resize(row_count + chunk[0].shape[0], axis=0)
            dset2.resize(row_count2 + chunk[1].shape[0], axis=0)

            # Write the next chunk
            dset[row_count:] = chunk[0]
            dset2[row_count:] = chunk[1]

            # Increment the row count
            row_count += chunk[0].shape[0]
            row_count2 += chunk[1].shape[0]
            print_progress(numer, n_files)
            numer += 1
```

Split the dataset into a training set and test set

We are going to split the dataset into a training set and testing. The training set is used to train the model and the test set to check the model accuracy.

```
training_set = int(len(names)*0.8)
test_set = int(len(names)*0.2)

names_training = names[0:training_set]
names_test = names[training_set:]

labels_training = labels[0:training_set]
labels_test = labels[training_set:]
```

Then we are going to process all video frames through VGG16 and save the transfer values.

```
make_files(training_set); make_files_test(test_set)
```

Load the cached transfer values into memory

We have already saved all the video transfer values on disk. But we have to load those transfer values into memory to train the LSTM net. One question would be: why not process transfer values and load them into RAM? Yes is a more efficient way to train the second net. But if you have to train the LSTM in different ways to see which way gets the best accuracy, if you didn't save the transfer values on disk you would have to process the whole videos each training. It's very time-consuming processing the videos through VGG16 net.

To load the saved transfer values into RAM we are going to use these two functions:

```
def process_alldata_training():
    joint_transfer=[]
    frames_num=20
    count = 0

    with h5py.File('prueba.h5', 'r') as f:
        X_batch = f['data'][:]
        y_batch = f['labels'][:]

    for i in range(int(len(X_batch)/frames_num)):
        inc = count+frames_num
        joint_transfer.append([X_batch[count:inc],y_batch[count]])
        count =inc

    data =[]
    target=[]

    for i in joint_transfer:
        data.append(i[0])
        target.append(np.array(i[1]))
    return data, target

def process_alldata_test():
    joint_transfer=[]
    frames_num=20
    count = 0
```

```
    with h5py.File('pruebavalidation.h5', 'r') as f:
        X_batch = f['data'][:]
        y_batch = f['labels'][:]

    for i in range(int(len(X_batch)/frames_num)):
        inc = count+frames_num
        joint_transfer.append([X_batch[count:inc],y_batch[count]])
        count =inc

    data =[]
    target=[]

    for i in joint_transfer:
        data.append(i[0])
        target.append(np.array(i[1]))

    return data, target
data, target = process_alldata_training()
data_test, target_test = process_alldata_test()
```

The basic building block in a Recurrent Neural Network (RNN) is a Recurrent Unit (RU). There are many different variants of recurrent units such as the rather clunky LSTM (Long-Short-Term-Memory) and the somewhat simpler GRU (Gated Recurrent Unit) which we will use in this tutorial. Experiments in the literature suggest that the LSTM and GRU have roughly similar performances. Even simpler variants also exist and the literature suggests that they may perform even better than both LSTM and GRU, but they are not implemented in Keras which we will use in this tutorial.

A recurrent neuron has an internal state that is being updated every time the unit receives a new input. This internal state serves as a kind of memory. However, it is not a traditional kind of computer memory that stores bits that are either on or off. Instead, the recurrent unit stores floating-point values in its memory state, which are read and written using matrix operations so the operations are all differentiable. This means the memory state can store arbitrary floating-point values (although typically limited between -1.0 and 1.0) and the network can be trained like a normal neural network using Gradient Descent.

Define LSTM architecture

When defining the LSTM architecture we have to take into account the dimensions of the transfer values. From each frame, the VGG16 network obtains as output a vector of 4096 transfer values. From each video, we are processing 20 frames so we will have 20 x 4096 values per video. The classification must be done taking into account the 20 frames of the video. If any of them detects violence, the video will be classified as violent.

The first input dimension of LSTM neurons in the temporal dimension, in our case it is 20. The second is the size of the features vector (transfer values).

```
chunk_size = 4096
n_chunks = 20
rnn_size = 512

model = Sequential()
model.add(LSTM(rnn_size, input_shape=(n_chunks, chunk_size)))
model.add(Dense(1024))
model.add(Activation('relu'))
```

```
model.add(Dense(50))
model.add(Activation('sigmoid'))
model.add(Dense(2))
model.add(Activation('softmax'))
model.compile(loss='mean_squared_error', optimizer='adam',metrics=['accuracy'])
```

Model training

```
epoch = 200
batchS = 500

history = model.fit(np.array(data[0:750]), np.array(target[0:750]), epochs=epoch,
                    validation_data=(np.array(data[750:]), np.array(target[750:])),
                    batch_size=batchS, verbose=2)
```

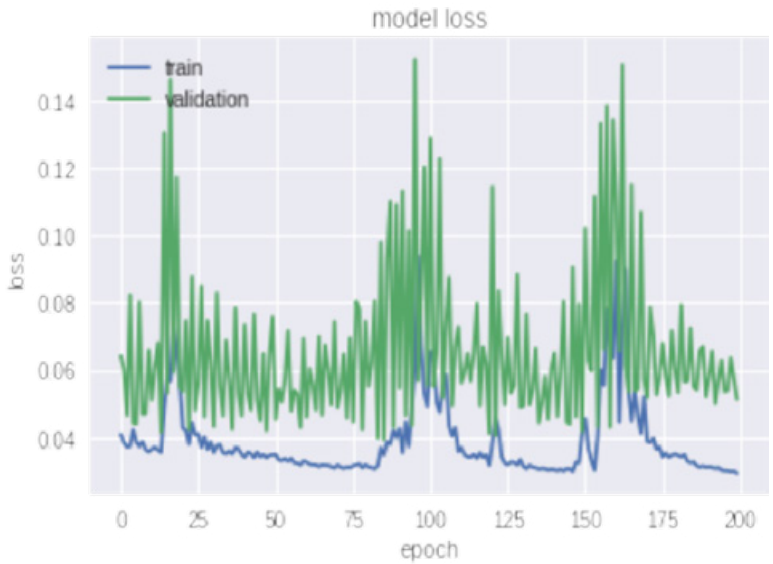
```
Train on 750 samples, validate on 50 samples
Epoch 1/200
 - 1s - loss: 0.0408 - acc: 0.9573 - val_loss: 0.0642 - val_acc: 0.9200
(...)
Epoch 200/200
 - 1s - loss: 0.0293 - acc: 0.9693 - val_loss: 0.0513 - val_acc: 0.9200
```

Test the model

We are going to test the model with 20 % of the total videos. These videos have not been used to train the network.

```
result = model.evaluate(np.array(data_test), np.array(target_test))
```

```
200/200 [==============================] - 0s 2ms/step
```



Thanks!

As always thank you for your great comments feedback and input! Let's keep this short and see you next month (as always please look over all the amazing magazine content!) ☺

Tom Bishop is the CTO of Glass Imaging, a start-up aiming to revolutionize the quality of cameras on smartphones. He speaks to us about its exciting plans for the future and takes us back to where it all began.

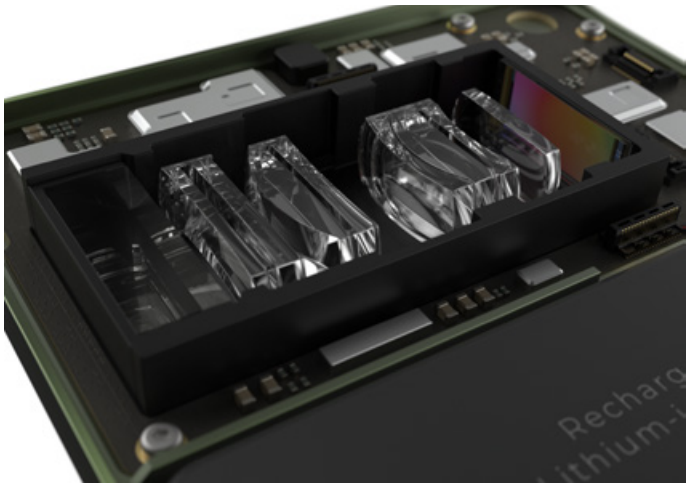
Tom Bishop and Glass co-founder **Ziv Attar** worked at Apple for several years, watching smartphone camera trends while developing the **iPhone camera's portrait mode**. They saw hardware improving and, more recently, the arrival of computational imaging in the form of multi-frame fusion methods that enhance the quality of the raw camera hardware.

In the last couple of years, they noticed that quality was starting to plateau. The different smartphones on the market were all similar in many ways, with the main hardware improvement being to squeeze in bigger cameras.

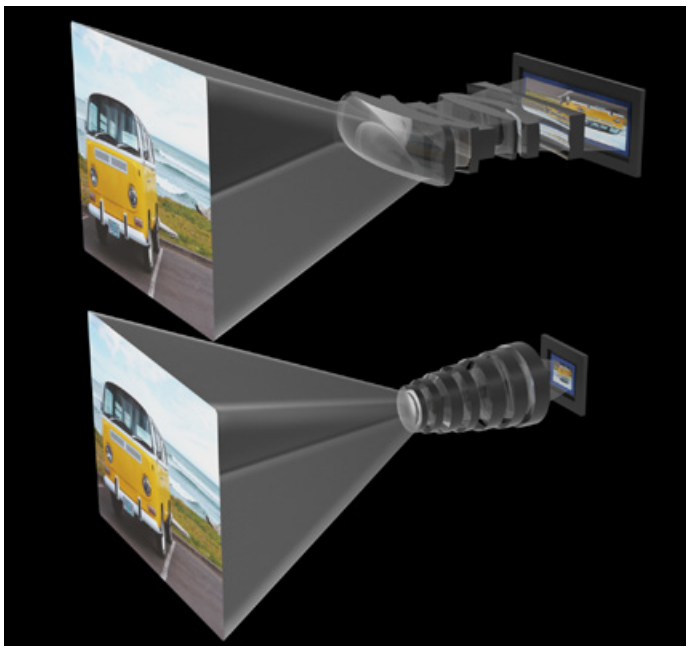
However, the difference in quality between what you get from a smartphone camera today and a professional DSLR or mirrorless camera is still huge. Photos look great on a small screen but blowing them up to 4k or 8k displays, massive TVs, or VR glasses with a wide field of view tells a different story.

"Our approach is to combine the hardware and software in a much more integrated way," Tom tells us. *"That's hard to do in a large company where teams are siloed, but we can work across the stack in a start-up, **integrating the hardware design and the software algorithms**. There is much computational power available in smartphones now and some very advanced computational photography algorithms, but we didn't know we had those when we designed the optics and the sensor in the camera module. Now, **we can tailor the hardware design so that it's combined with the software**, and they work hand in hand together. That allows us to explore a new range of hardware designs that previously wouldn't have been possible."*

Glass uses some of the latest developments in **computational imaging, machine learning, and AI techniques** to restore and enhance the images from the sensor. It is not just taking what the hardware



Render showing a cutaway of a Glass Imaging folded camera module with mirror, anamorphic lens elements and sensor, as would be integrated into a smartphone.



Render showing the difference between a Glass Imaging anamorphic camera module and a regular smartphone camera module; note the larger area and shape of the former.

gives it and trying to improve it but also thinking about how the optics can be designed differently. An image may be distorted, degraded, or corrupted, but **the AI algorithms work to get it perfect** at the end of the pipeline. That whole system working together creates the end image, and the neural networks are trained with knowledge of the optical design to do that all in one go.

How has Glass managed to get ahead of the crowd with many big players circling in this field?

"Other companies are trying to do things like this, but their teams are often disconnected across hardware and software, and it requires deep integration and a mindset of trying things that are quite different from what's done currently," Tom responds. *"Innovation in the big companies is somewhat incremental – let's take last year's design and improve the lens, the sensor, the algorithm, and the processor a little bit. No one's thinking, wait, is there a better design?"*

Glass discovered that the same real lens architecture camera design had been used for the last 150 years. It plans to change how the camera works, using **anamorphic lenses** and wide aspect ratio sensors on the hardware side, which are more akin to costly Hollywood systems. No one had thought of using them on a smartphone before.

*"We capture ten times more light than a traditional smartphone, which means we get **a higher signal-to-noise ratio**,"* Tom explains. "We're using image restoration approaches, such as blind deconvolution and super-resolution, to recover the possible detail according to the trade-offs we make in the lenses. Another interesting area is the end-to-end optimization of the entire system. **We can model the optical system using our deep learning approach in an integrated fashion, optimizing the lens parameters along with the neural network doing the image restoration**."

Further benefits include single-shot HDR – with larger pixels and a high dynamic range; extreme low-light imaging, so night photography and freeze motion; and a natural depth of field.

"When we created the portrait mode, we wanted to emulate the look from SLR cameras where you can blur out the backgrounds," Tom recalls. *"Software has done a great job there, but you still get artifacts around the borders if you have someone's hair against a complicated background. With our hardware optics, we can get a natural shallow depth of field. We can apply that to video as well. Apple released the cinematic mode not so long*

Render showing how a Glass camera module integrates into a smartphone - note the large entrance window

ago, a great attempt to take portrait mode into video, but it still suffers from those artifacts around the edges. We can do it naturally, just like in a Hollywood movie. You can shoot that on a smartphone with our technology!"

Glass is applying its knowledge of the whole manufacturing and supply chain, design, optics, sensing, algorithms, and what can be run on a smartphone chip. Tom, Ziv, and the team have accumulated much interdisciplinary expertise across those areas throughout their careers.

"That enables us to look at other trade-offs and design constraints that we can open up," Tom points out. *"Back at Apple, we had to push hard to go to a dual camera when single cameras were still being used in smartphones. We needed the stereo information to create the depth maps required for a portrait mode effect and to blur the background. They were very resistant to putting in more cameras. The executive eventually saw the benefits, but making that change was an uphill battle."*

Tom tells us he enjoyed attending the **Computational Cameras and Displays workshop** at CVPR this year. Metalenses was a hot topic, which he believes could complement Glass's work in time.

"The idea of using wave optics designs to make very thin lenses has been shown to have some great potential and is used in some research environments," he says. *"There are hopes to put that into commercial photography as well, but there are some practical constraints, such as scaling up the systems to large sensors and working across multiple wavelengths. That's going to be technology that takes longer to get*

to market. We have an innovative new design that we're focused on getting to market in the next couple of years with a few small tweaks to the existing lens design manufacturing system."

Tom expects its technology will initially hit the shelves in high-end smartphones, as companies are keen to have a marketing advantage for their flagship models. Glass has already developed proven prototypes, which have been well received. The next challenge is getting those into a form factor ready for mass manufacture.

"*We have to work with our supply chain partners and line up various customers, partners, and large-scale things in industry to deliver that,*" he adds. "*We need to work on the business front to license our designs. We're sure of the technology we've demonstrated so far, but there are some practical things to take care of next.*"

What is clear is that Glass is succeeding in punching above its weight. It is a small core team of around six people plus consultants, and it is hiring! If you are an experienced computer vision and machine learning engineer, particularly with optics and computational photography knowledge, or if you are a PhD intern working in those fields, you could be a part of this exciting new camera experience! PHOTO CREDITS Glass and Christopher Michel.

Comparisons between crops from images captured with a recent Glass Imaging prototype camera (left), showing much improved detail and quality compared to the iPhone Pro Max 13, wide and tele images respectively (right); captured from same distance

Example of a before (left) and after (right) image undergoing enhancement by Glass Imaging co-designed image restoration Neural Network.

Giovanni Maria Farinella is an Associate Professor, and Antonino Furnari is an Assistant Professor at the University of Catania, Italy. They play key roles in the International Computer Vision Summer School, which took place in Sicily last month – Giovanni is a director, and Antonino is the academic assessment chair. They are both here to tell us more about their successful event.

The International Computer Vision Summer School (ICVSS) is a collaboration between the **University of Catania** in Italy and the **University of Cambridge** in the UK. It was launched 15 years ago in 2007 with many big names on board, including **Yann LeCun** and **Serge Belongie**, and more than 100 students joining in.

ICVSS has run almost every year since. In 2017, it was awarded the **IEEE PAMI Mark Everingham Prize** in recognition of the benefits it has brought the community. However, like many other events, it recently suffered a two-year pandemic hiatus.

This year, with much excitement and anticipation, ICVSS was back! Over 600 people applied from 40 countries, whittled down to 180 students who just spent a memorable week together in beautiful Sicily learning and talking about science.

The participants are primarily first and second-year PhD students, some working

in academia, who stay together in a village, where they can get to know each other and establish connections.

"Networking is one of the most important parts of our school," Giovanni tells us. *"The students can interact and talk with the speakers anytime because everyone is in the same village. This makes the school different from other events, like conferences, where you don't have as much time to talk in a friendly way because everyone stays in different places, and it's difficult to catch up."*

Antonino adds:

*"This year, there was a different vibe than usual. We had some **third-year PhD students who had never had a poster session in person because of Covid**. They were very, very happy to be there. We had two poster sessions at night, and one ended at 2 am! There was no way these guys were going home because they were too excited about being able to talk about their work with other students, which they had not been able to do in the past few years."*

Giovanni tells us his highlights from the week included the reading group with **Stefano Soatto**, for which speakers mentored students, and the essay competition led

by **Fabio Galasso**, where students had to write an essay before the event about the social impact of computer vision research.

The scientific content was strong this year. Is Giovanni able to pick his favorite speaker?

"*Every day, I was saying, this is the best talk. No, this is the best one. No, this one is more interesting!*" he laughs. "*All the speakers were great. We have much material to study now. I'm biased because I work in egocentric vision, but the talk from* ***Dima Damen*** *was amazing for me. She presented a new data set for segmentation in egocentric vision, and I said, wow, we can make progress with this. But also, the talks from* ***Abhishek Gupta***, ***Michael Bronstein***, ***Laura Leal-Taixé***, *and* ***Andrea Vedaldi*** *were excellent. I don't want to say who was the best!*"

Science is a massive part of the event, but the social side is also significant, including dinners, drinks, and night-time swims.

"*Every night, after the lectures and poster sessions, people were jumping in the swimming pool from midnight to 4 am,*" Antonino reveals. "*All of them managed to be back for the lectures at 9 am! I think there was much more social contact than before because of the pandemic. Aside from the scientific part, people wanted to get together with others their age because they had been in their homes for the last three years.*"

Alas, the week had to come to an end, and parting is such sweet sorrow, as Giovanni makes clear:

*"We had a Telegram channel to communicate with all the students, and our last message was, 'Sorry, something is wrong. **We're missing you already!**' It's the highlight of the year for our team to work on this event for the community."*

Students occasionally return to the school if they come back with the industry or as a postdoc, but generally, **an entirely new group is given the opportunity each time**. Diversity in the selection process is critical. The team tries to provide access to as many groups as possible. A third of the students are female.

What is the one thing Giovanni would add to the school if he could?

"A laboratory," he answers immediately. *"This is something we can't do because of the number of students and the difficulty of arranging it. It's just too complicated. We're not in a university. We tried a couple of times in 2012 and 2013, but it wasn't manageable, and we had to stop it. We'd need more time and facilities to do it."*

Giovanni and his fellow directors, **Roberto Cipolla** and **Sebastiano Battiato**, are already thinking about the scientific content for **ICVSS 2023**. The planning will begin in earnest later this year.

Finally, we ask what the secret to the school's success is in the hope that it could be bottled for others who wish to replicate it. Antonino has the perfect response: *"Hold their school in Sicily – that helps a lot!"*

ICVSS 2022 in Sicily

COMPUTER VISION EVENTS

SIGGRAPH
Vancouver, Canada
8-11 August

Summer School on Imaging for Medical Applications
Oradea, Romania
September 5-9

3DV
Prague, Czechia
September 12-16

TCT 2022
Boston, MA
September 16-19

MICCAI
MEET US THERE
Singapore
September 18-22

TechEx Europe
Amsterdam,
The Netherlands
September 20-21

AI in Healthcare Summit
Boston, MA
13-14 October

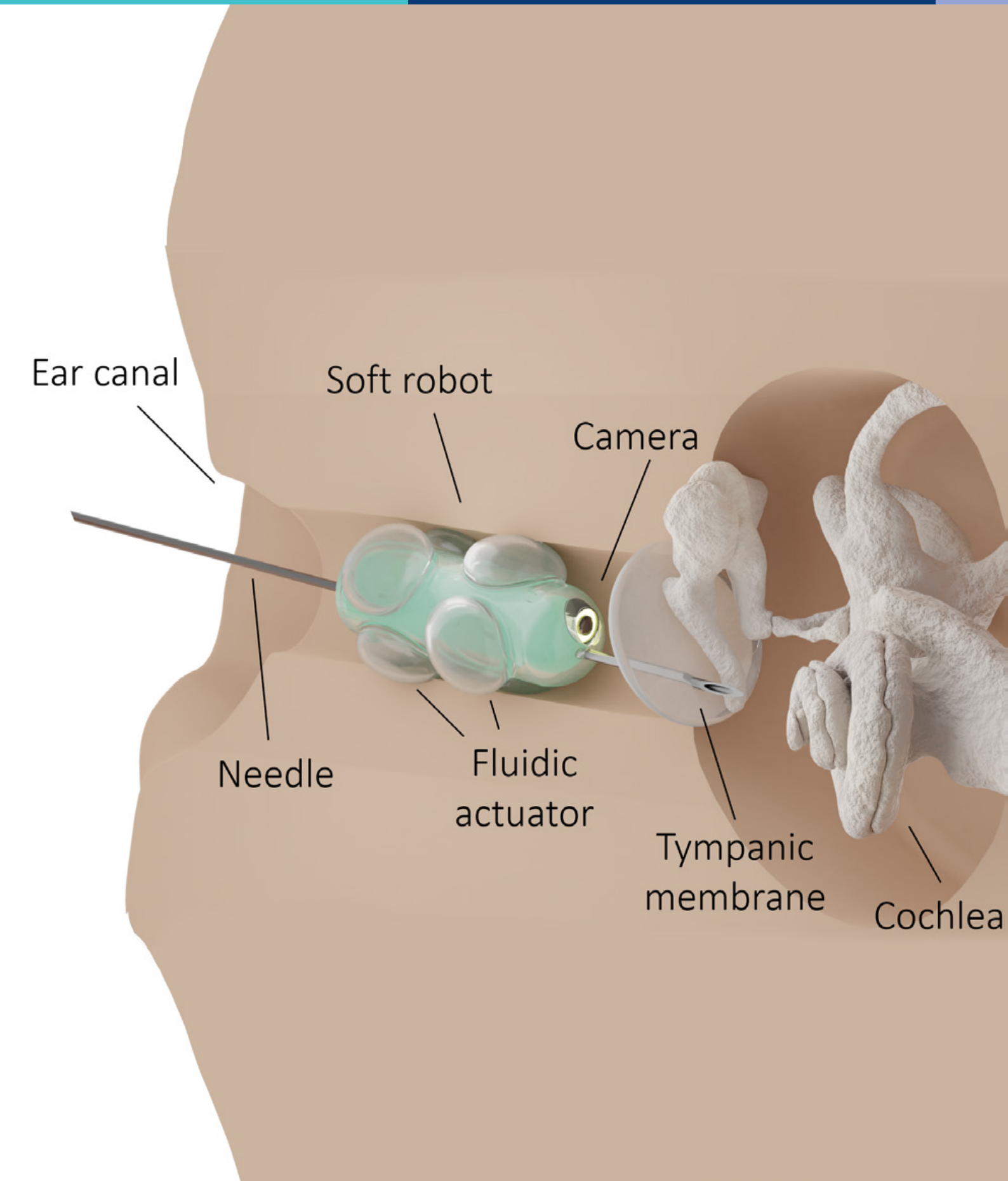
ICIP
Bordeaux, France
16-19 October

ECCV 2022
MEET US THERE
Tel Aviv, Israel
23-27 October

MedTech Conference
Boston, MA
24-26 October

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

MEDICAL IMAGING NEWS
AUGUST 2022

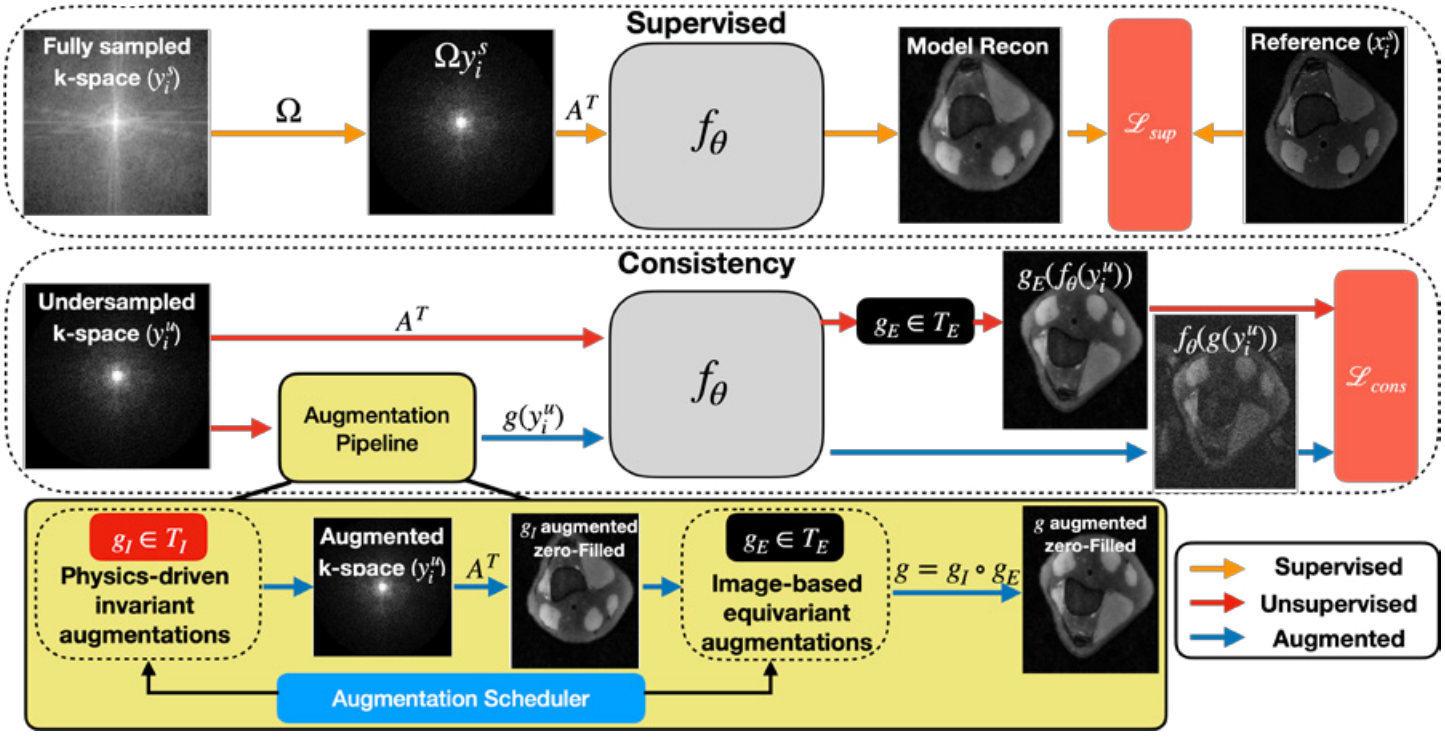


VORTEX: PHYSICS-DRIVEN DATA AUGMENTATIONS USING CONSISTENCY TRAINING FOR ROBUST ACCELERATED MRI RECONSTRUCTION

Arjun Desai

Beliz Gunel

Arjun Desai is a PhD student at Stanford University, working with Akshay Chaudhari and Christopher Ré. Beliz Gunel is a fifth-year PhD candidate in electrical engineering at Stanford University, advised by John Pauly. Arjun and Beliz speak to us fresh from winning the Best Paper award at MIDL 2022 for their work on accelerated MRI reconstruction and robustness to physics-driven perturbations.



MRI is a powerful non-ionizing imaging technology commonly used in clinical practice, but its downside is that **it can take some time to acquire**. Accelerating the scans requires dropping data points, which results in poor quality or degraded images. When data is dropped, it is referred to as being **undersampled**. **Accelerated MRI reconstruction** aims to recover high-quality images from this undersampled raw data.

"***Most reconstruction methods are fully supervised***, *which requires many fully sampled scans to train the deep learning models,*" Arjun explains. "*In clinical practice, we typically acquire undersampled scans. Also, deep learning and traditional iterative methods for reconstruction are sensitive to perturbations in the acquisition process.* ***MRI, like other forms of imaging, is rooted in the physics of the hardware****, and there are certain perturbations in the acquisition. We may experience noise in the image, or if a patient moves, that will cause some artifacts in the data.*"

The reconstruction process is highly sensitive, and **noise and motion corruption are two of the perturbations most prevalent in clinical practice**. Alongside a lack of fully sampled scans, these are the fundamental limitations of deploying deep learning-based methods in clinical practice and are the motivation for this work.

Prior to the deep learning era, people attempted to develop solutions and algorithms tailored to a small subset of perturbations. Arjun, Beliz, and the team are trying to **create a generalized framework robust to all these perturbations**.

"*Generally, with iterative methods, they're sample-specific, but with machine learning, if you play your cards well, you can generalize to unseen things,*" Beliz points out. "*That's very powerful. Here, we can teach our models to be prepared for what's to come and leverage our physics knowledge about MRI acquisition. In machine learning methods, you usually give large amounts of data to the model and expect it to learn and*

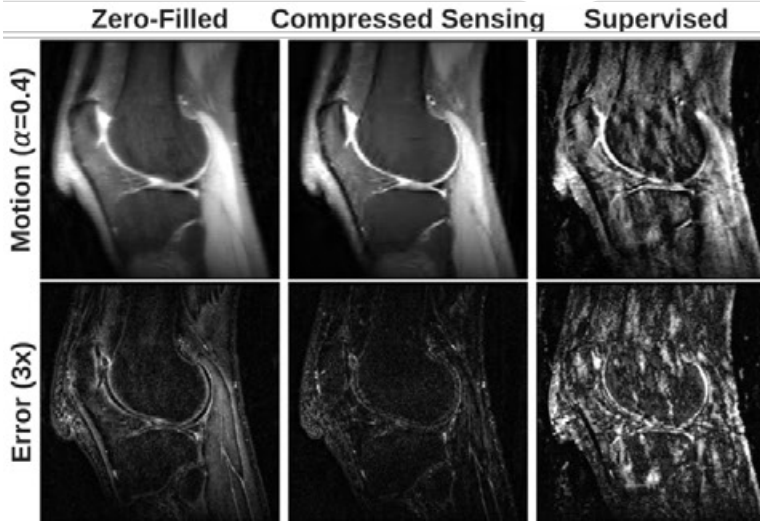
> *"Medical imaging and computer vision go hand in hand."*

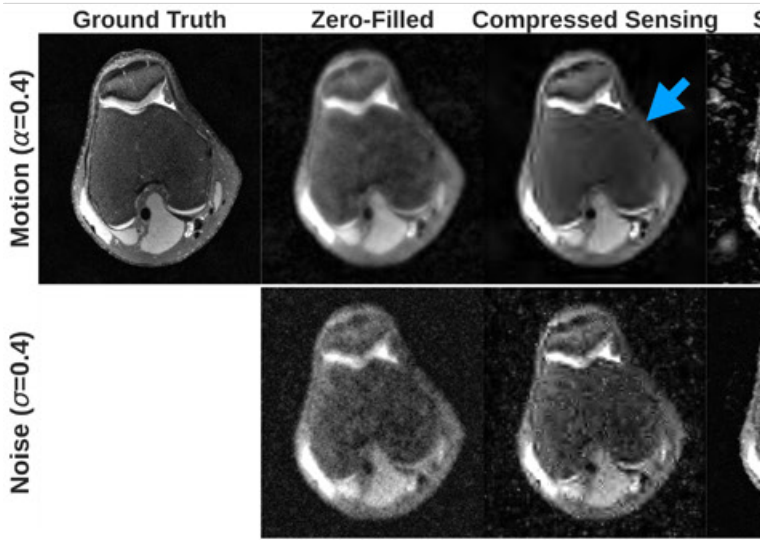
be prepared for everything, so how can we use that data while integrating our physics knowledge into the model? That's where the noise and motion modeling comes in."

Taking home the **Best Paper award at MIDL** is no mean feat, but the fact that the **VORTEX** framework is rooted in **the physics of how imaging is done** could be what most impressed the judges. Traditionally, machine learning has relied upon collecting millions of examples and training a model. This work is driven by the process of how images are acquired. It is focused on MRI, but you can imagine a scenario where it could be generalized to CT or ultrasound.

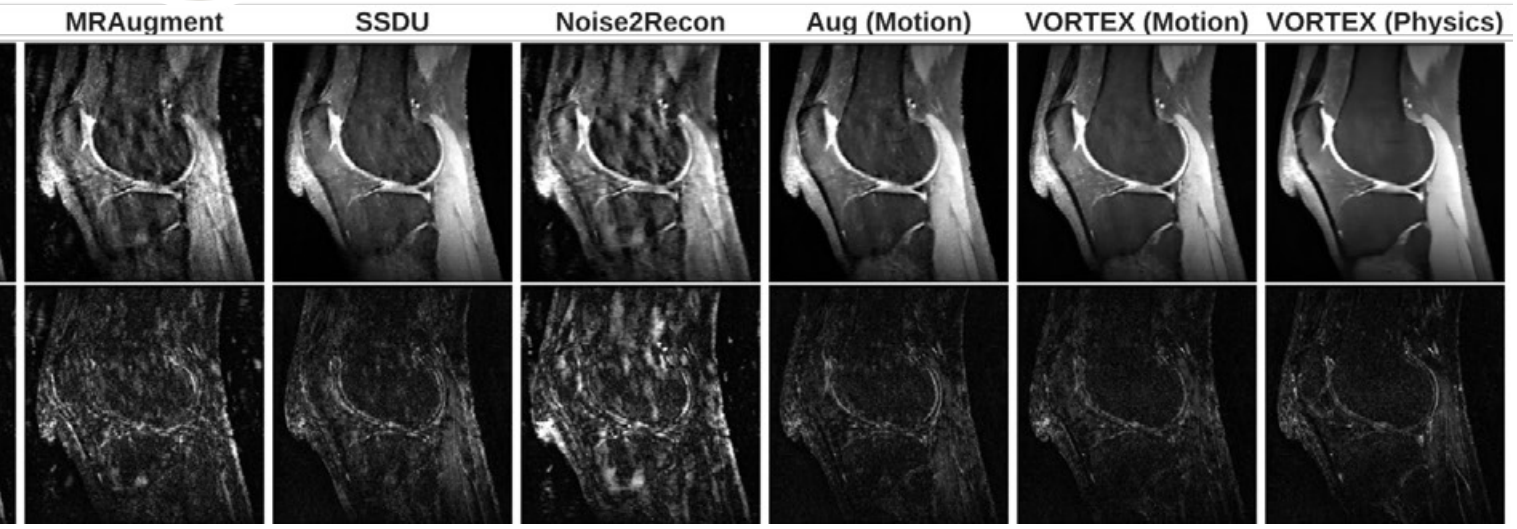
"*The medical field is often limited by the amount of publicly available data that we can perform experiments on, or even if we're just trying to build the best model, the amount of available labeled data that we have,*" Arjun notes. "*This is going back to one of the limitations of standard supervised methods in deep learning. A takeaway for the MIDL community and beyond is that it's important to consider how deep learning can be motivated or almost hybridized with the standard analytical and signal processing fundamentals we have known for centuries! That was probably one of the pieces that appealed to the judges.*"

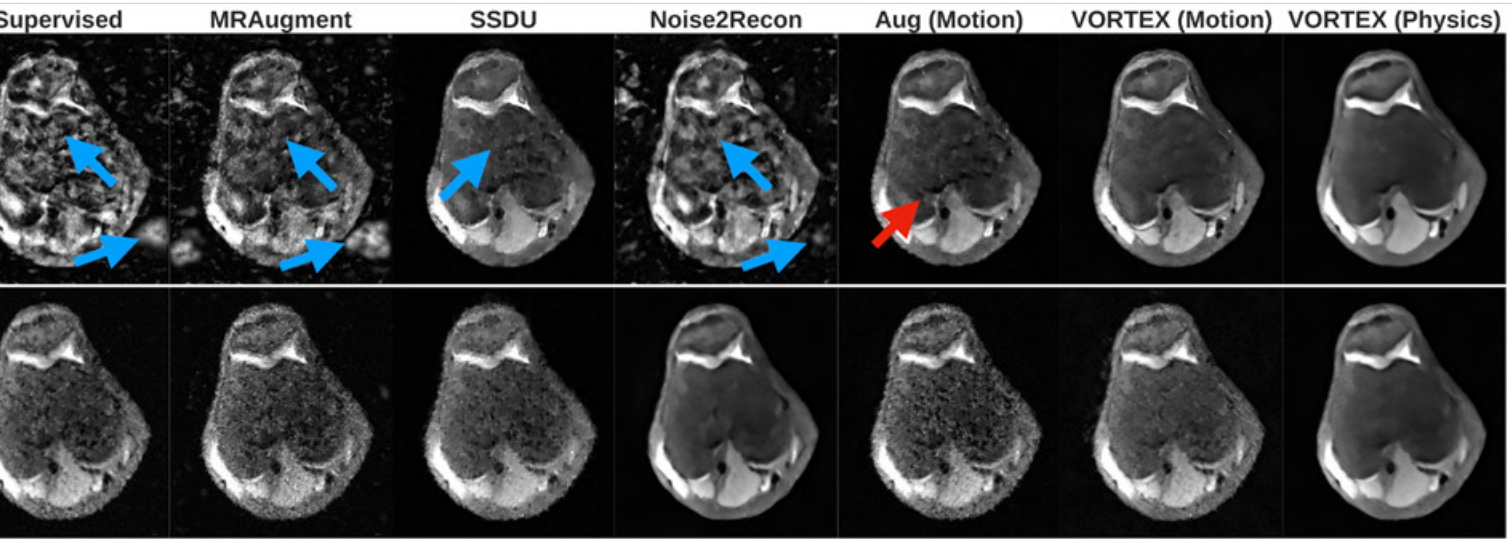
Beliz adds: "***Medical imaging and computer vision go hand in hand. In many computer vision tasks, you go from image to label, whereas in most medical imaging tasks, at least within reconstruction, you go from image to image. This framework can be modeled agnostic to other image-to-image tasks. As long as you root your augmentation in the physics of the process, this sort of framework can work.***"





If you want to know the secret to their success so that you can create a winning paper yourself, Arjun and Beliz tell us they have been fortunate to be surrounded by talented folk and that it is all about choosing the right team. Try to pick people





who know something you don't know as that diversity can be incredibly helpful in discussions.

"It's always nice to make the work available beyond a paper," Arjun adds. *"Keep speaking to the community. Make your methods easy for people to try out and reproduce. We have some tutorials and open-source code that we made available as part of this paper, which really helps broaden its accessibility to people who might not have the set-up we are fortunate enough to have. I hope the community keeps pushing for that moving forward."*

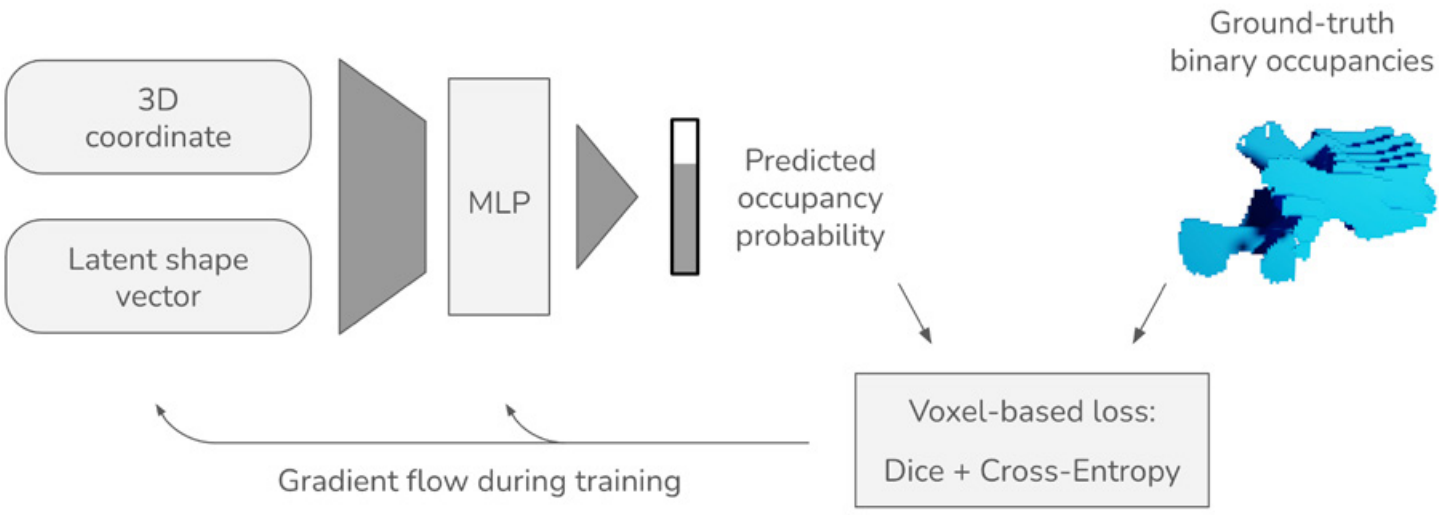
As we're finishing up, Arjun reflects on their path to producing this award-winning work and is keen to express their appreciation for those who supported them along the way.

"We'd like to thank MIDL, the reviewers, and everyone who was part of the process that helped us get here," he tells us. *"I also appreciate your effort, Ralph, to help understand the high-level concepts behind this work. What you've covered is fantastic. If anyone has any technical questions, please read our paper or reach out to us!"*

LEARNING SHAPE RECONSTRUCTION FROM SPARSE MEASUREMENTS WITH NEURAL IMPLICIT FUNCTIONS

Tamaz Amiranashvili is a PhD student at the Technical University of Munich, with close ties to the University of Zurich and the Zuse Institute Berlin, under the supervision of Bjoern Menze and Stefan Zachow. His paper exploring the task of reconstructing full or high-resolution shapes from sparse or partial measurements has just won a Best Paper Runner Up award at MIDL 2022. He gives us an insight into why it impressed the judges.

Anatomical structures usually have very distinct shapes, and the distribution of shapes that naturally occur within a population can be learned and applied to various tasks in medical imaging. What distinguishes the model described in this paper from related works is that **it can learn on sparse measurements, particularly**



segmentations with large slice distances.

If you are given three orthogonal slices, it is impossible to know what the output should be if you do not know beforehand what you are looking at. The measurement on its own is insufficient to reconstruct the shape.

Take a lumbar vertebra, for example. A radiologist would know what lumbar vertebrae look like in general and could guess the shape. This model tries to mimic that behavior with **a neural network learning the distribution of vertebrae shapes** that naturally occur and using this knowledge to help perform these reconstructions.

"We leverage so-called implicit functions in this work, and the special property of this shape representation is that it is continuous and not discrete like meshes or voxel-based representations," Tamaz explains.

"This continuous representation helps us to work with different kinds of sparse training data inputs and reconstruction target resolutions. It essentially allows us to unify these different kinds of discrete measurements that we have."

Although the training data is sparse, and the network has never seen a structure's full shape, **the model can reconstruct smooth and natural-looking shapes from just three slices**. Does Tamaz think this is the aspect of the work that stood out for the judges?

"When you first look at it, it seems impossible, and it's a bit of a surprise that you can train this model on such sparse data – but it works!" Tamaz reveals.

"That is likely what caught people's attention. I didn't believe it myself at first. I tried many things to check if it was doing what it looked like it was doing. Luckily, all the experiments were successful. It was

a nice example of where a good idea translated into practice and worked from the start – with a few tweaks here and there, of course."

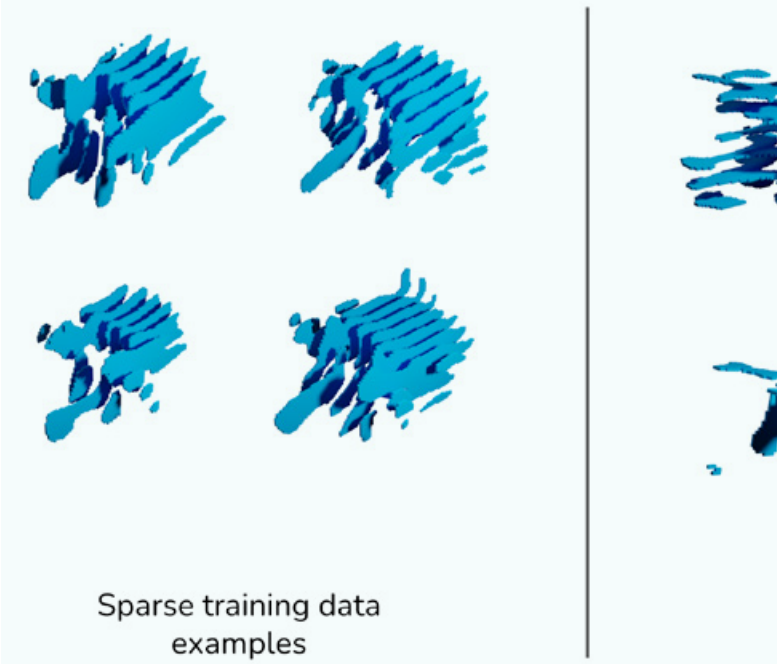
Thinking about what comes next, Tamaz tells us he would like to consider the different applications of this generative model. It has been shown to work well on reconstruction tasks, but could this shape analysis be applied to other scenarios in the medical domain, or even outside of it?

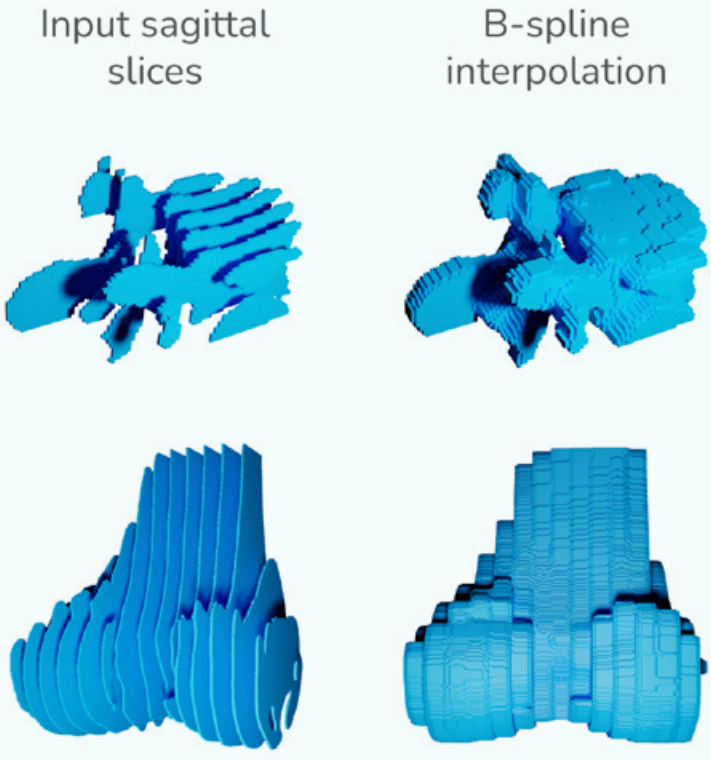
"*The kind of data we use for training occurs naturally in the medical domain,*" he responds.

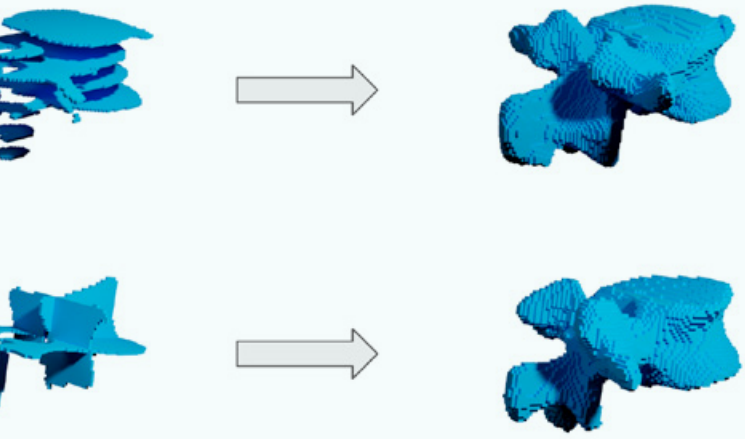
"*In computer vision, you typically would have sparse measurements from* ***LiDAR scanners****, for example, where you get point clouds from only one side of the object. Certainly, it has been applied there, but it's a different setting.*"

Before CNNs and deep learning, people built shape prior models to solve the segmentation task because image-based models were very simple. Specific statistical shape models were used extensively to regularize the segmentation process because it was hard to make sense of those images.

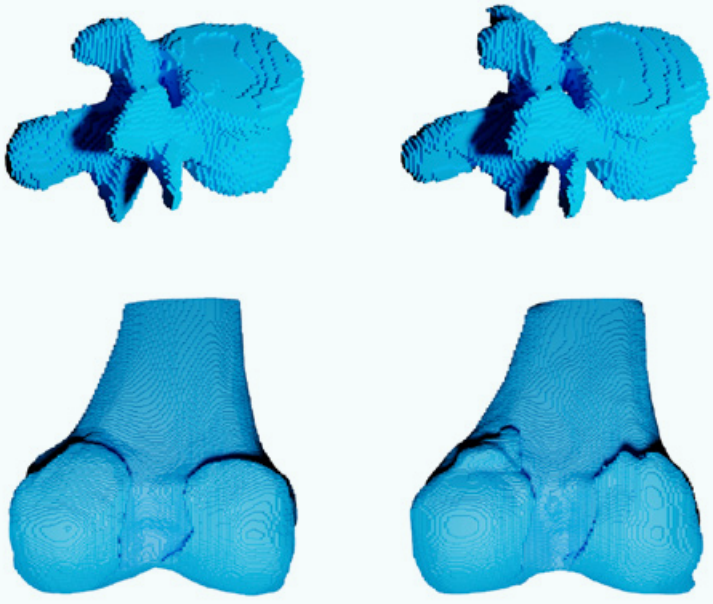
It used to be tedious to build statistical shape models, with manual







Reconstructions from various sparse inputs after training



annotations required to establish dense correspondence between the training shapes. Nowadays, there are plenty of segmentations because the segmentation process can be automated, and these shape prior models can be built automatically – even from sparse data, which is the contribution of this work.

"*These days, the segmentation task often performs well enough,*" Tamaz points out.

"*We have the **nnU-Net**. We have the contrastive self-supervised pre-training on large unlabeled data sets. We're going backward and saying, okay, now we have lots of segmentations and shapes, can we learn a probabilistic generative model of those shapes or a shape prior?*"

Reflecting on his time at MIDL, Tamaz leaves us with an anecdote.

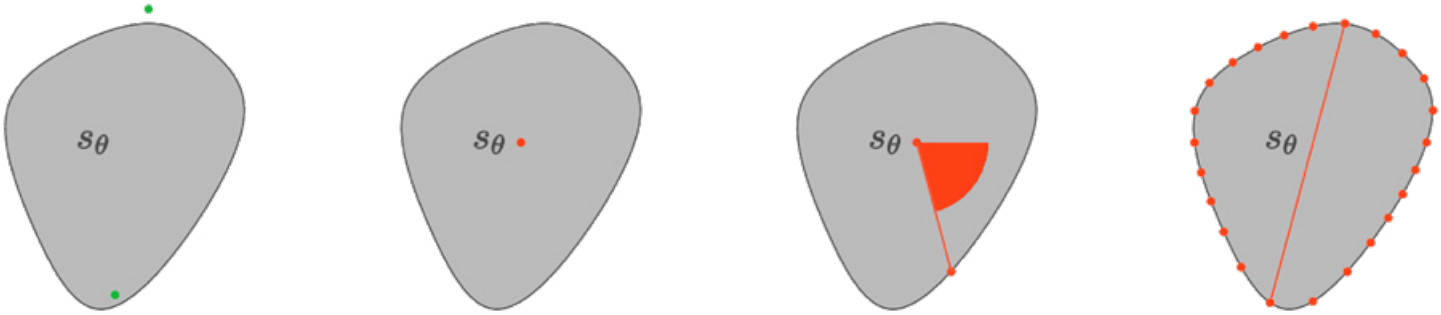
"*It was funny that the conference is called Medical Imaging with Deep Learning,*" he laughs.

"*In my work, **I haven't used a single medical image!** I only work with segmentations because it's about shape reconstruction and building shape priors, which is a very important and useful topic in medical image processing. In the end, what do we do the segmentation for? It's to obtain the shape of an anatomical structure. I hope that our work attracts more people into the shape analysis world!*"

DIFFERENTIABLE BOUNDARY POINT EXTRACTION FOR WEAKLY SUPERVISED STAR-SHAPED OBJECT SEGMENTATION

Robin Camarasa is a third-year PhD student in the Biomedical Imaging Group Rotterdam, part of the Department of Radiology of Erasmus MC, under the supervision of Marleen de Bruijne and Daniel Bos.
He speaks to us about his paper on weakly supervised segmentation, which has just won the second Best Paper Runner-up Award at MIDL 2022.

Carotid artery atherosclerosis is the thickening or hardening of the carotid artery caused by a build-up of plaque on its walls which can rupture. The artery's diameter is measured in different locations to assess this, and the ratio of those diameters is computed to obtain a biomarker.



In this paper, Robin and the team try to **predict a carotid artery segmentation** based on its diameter annotations. The novelty of the work is a differentiable loss going directly from the output of the segmentation network to the boundary point coordinates and later in the pipeline to the diameter, which allows optimization downstream of a segmentation network based only on diameter annotation.

"*Usually, in computer vision literature, you have a problem, and you create more and more models that are more and more complex using more and more data,*" Robin explains

"*We wanted to use prior knowledge in our model to reduce this complexity and make sense of our methods.*"

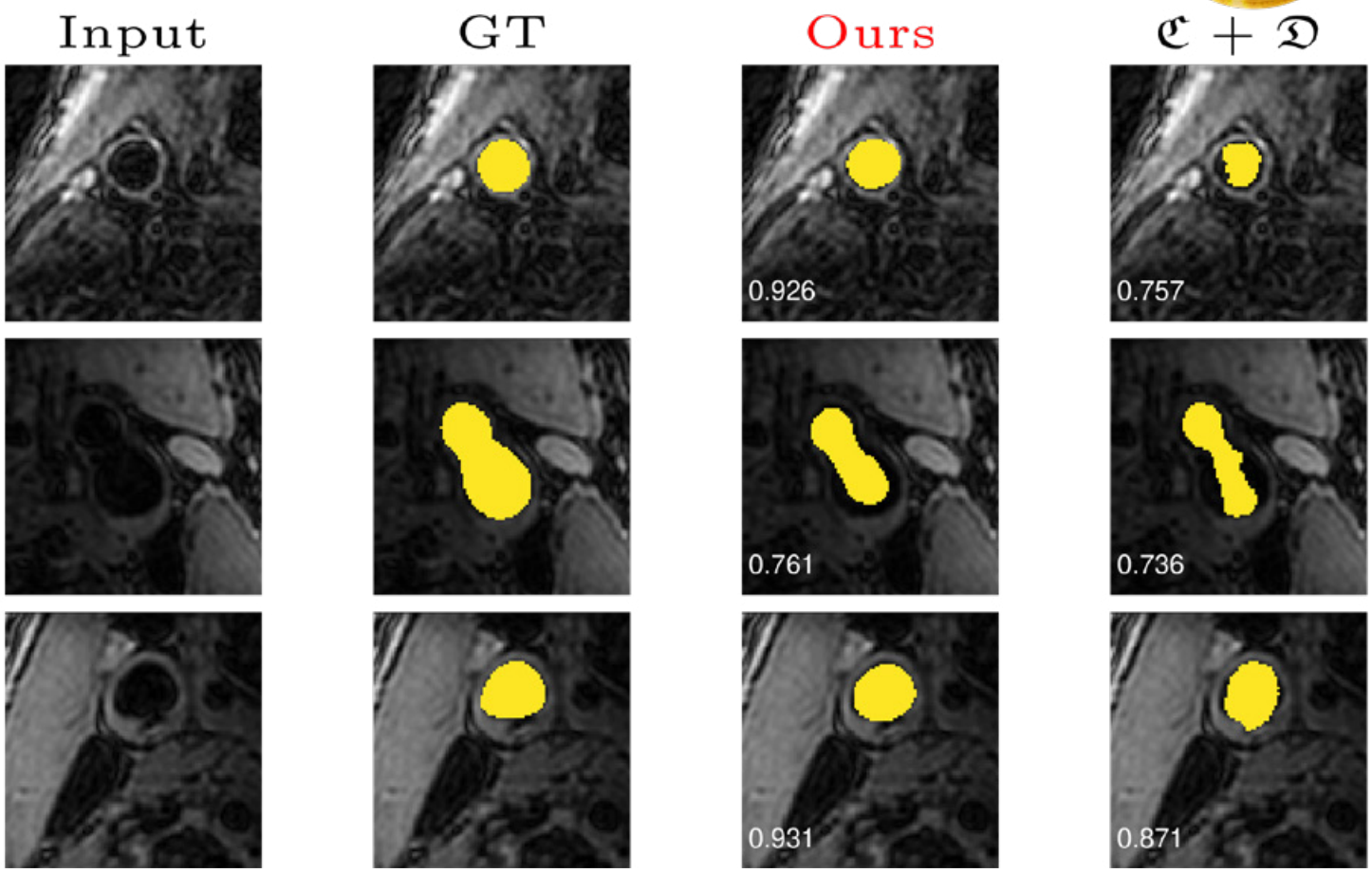
The model helps you to understand what you are optimizing based on the mathematical model integrated within the framework by adding this prior knowledge, which makes it more interpretable.

"*Before, the literature was more about trying to find some hacks to make this work within the current framework,*" Robin adds.

"*I had this idea that if we get the boundary points, we will be able to get the diameter. It was a lot of work on blackboards, just trying to derive the equations to arrive at something we were happy with. Ultimately, that led towards convergence because if you have a strong mathematical model, you have more chance of converging.*"

Throughout this project, Robin used a combination of **MLflow** for tracking experiments, **PyTorch Lightning** to ensure a standardized approach, and **MONAI by NVIDIA**, a toolbox for building computer vision models targeted more toward medical imaging. MONAI will be very familiar to our readers, and we must congratulate **Prerna Dogra**, **Stephen Aylward**, and the many scholars who have brought it to where it is today as such a valuable tool for scientists and engineers. Robin says **it streamlined the process of creating the model and made the research phase of this work much faster**.

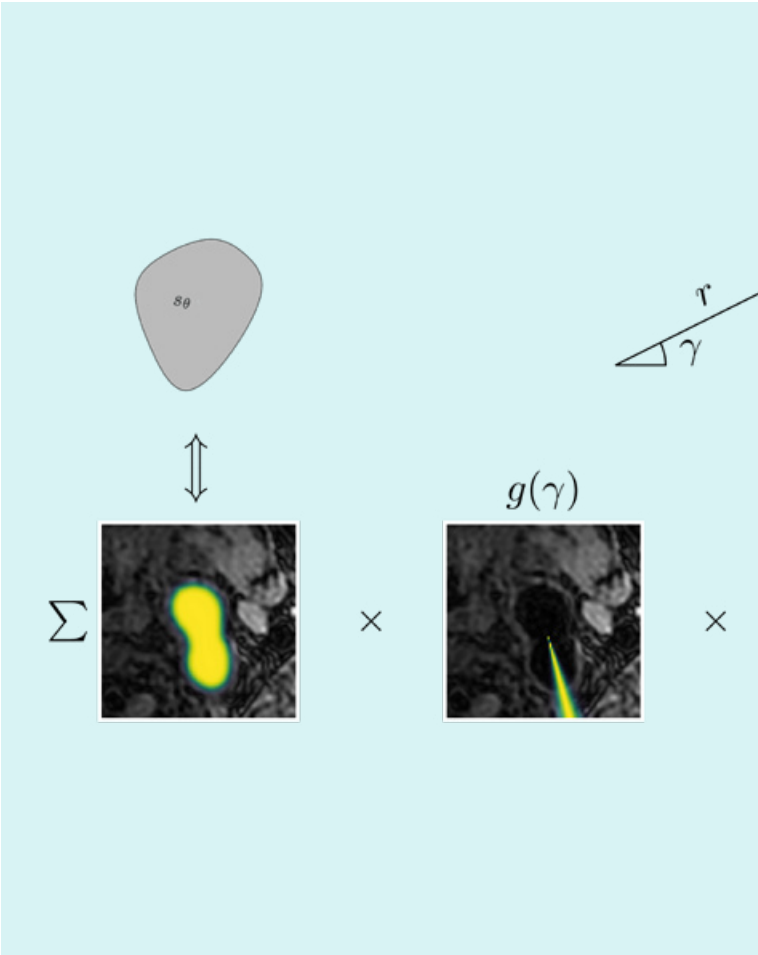
Best Paper MIDL 2nd Runner-up



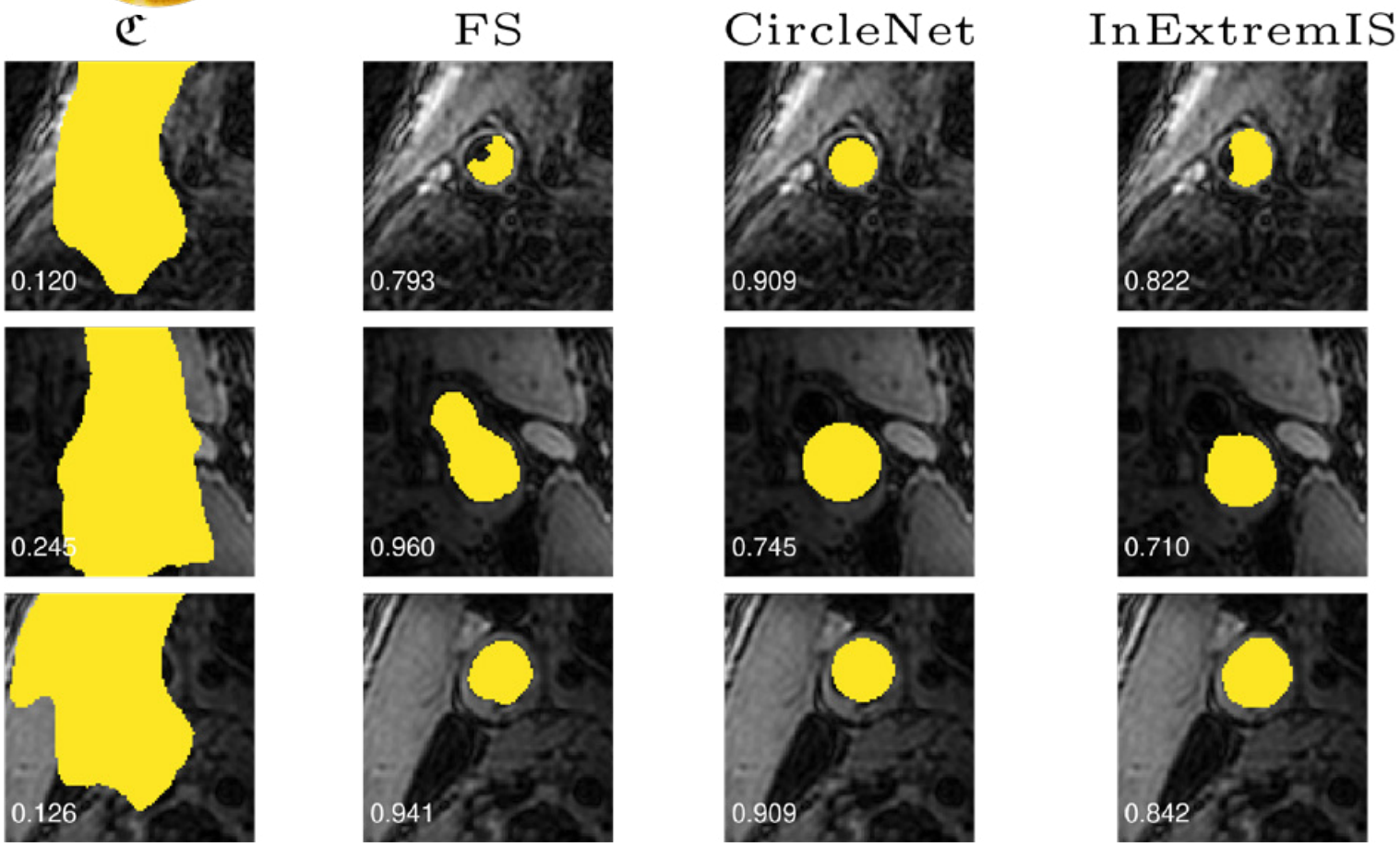
The team has been working with synthetic data and is still developing the model to make it work on a more realistic set-up before tackling the stringent regulations necessary in medical imaging.

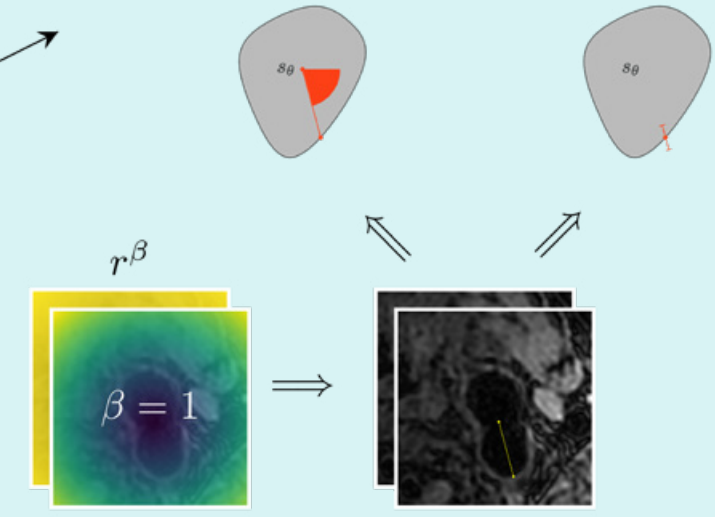
"We're working on getting a more realistic set-up for the diameter," Robin tells us.

"We considered the maximum diameter, but that's a bit limited. We'd also like to see if we could achieve a good result on other structures, such as tumors. We already have biomarkers derived from tumor diameters, such as **RECIST**."



Scooping an award at a conference like MIDL is something to be proud of. We ask Robin what he thinks is the secret behind





its success. He points to the alchemy and collaboration between the team, with everyone complementing each other to create a piece of research that stands out from the crowd. Epidemiologist and radiologist **Daniel Bos** brought good insight from a medical perspective; **Marleen de Bruijne** is one of the leaders in the field of medical imaging; and **Hoel Kervadec**, who won the Best Paper award at MIDL 2021, helped take the paper to another level.

"I think this paper was relevant because we come from a fully data-driven area, and right before it was fully model-driven, so we're trying to do something in between that," Robin reveals.

"We want to combine those two things to get something even better!"

TOTAL HIP REPLACEMENT (THR)

THR is a common orthopedic procedure, where the damaged bone and degenerative cartilage of the hip joint are replaced by **prosthetic components**. More than 450,000 procedures are performed annually in the USA alone, and as life-expectancy increases, more procedures are expected. There are several approaches for THR: lateral, posterior and direct anterior that gained popularity in recent years. The procedure is initiated with an incision in the hip, removal of the diseased areas of the bone and cartilage, implantation of the socket prosthesis into the pelvic bone, and insertion of the stem prosthesis into the femur bone, topped with a "ball" to conclude the ball-and-socket joint. These components are made of metal, ceramic, or polyethylene, and anchored with screws and/or cement depending on design. This is a well-established procedure with high success rates. However, **recent developments in surgical robotics and artificial intelligence (AI), can simplify the procedure and reduce complications (Chen et.al, 2018).**

Procedure planning

Adequate planning prior to surgery increases success rates. Currently the surgeon reviews different imaging scans of the patient - usually radiographs (X-ray), and if needed CT or MRI -

and assesses the proper approach for the surgery. Recent advancements in computer-vision (CV) and deep-learning (DL) allow **accurate, automatic segmentation of the hip and femur bones from either/all imaging modality, and reconstructing a 3D model from CT, MRI, and even from 2D X-ray images**. This 3D model can be used to test the fitting of implants (i.e., need for high offset), calculate bone-removal quantity needed for proper fitting, and recover any other patient-specific information (i.e. need for specific implant design). This model can be also used for 3D printing for hands-on practice. Specific MRI protocols allow soft-tissue segmentation for cases which require more attention and unique planning (obese patients, soft-tissue malformations, etc.).

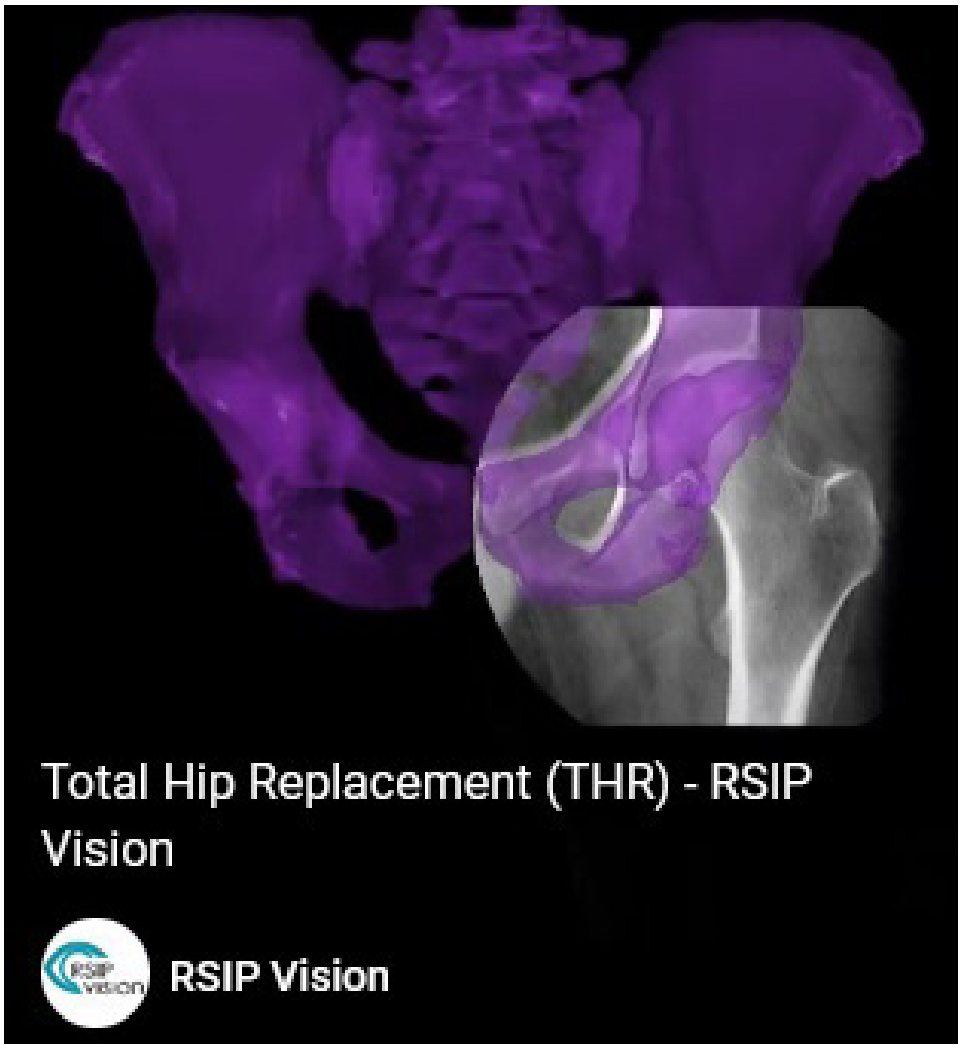
If the procedure is conducted with the assistance of a robot, **a specific surgical plan can be prepared and uploaded to the robot**. All procedural steps will be guided, and any deviation from the plan will be detected by the robot and the surgeon will be alerted.

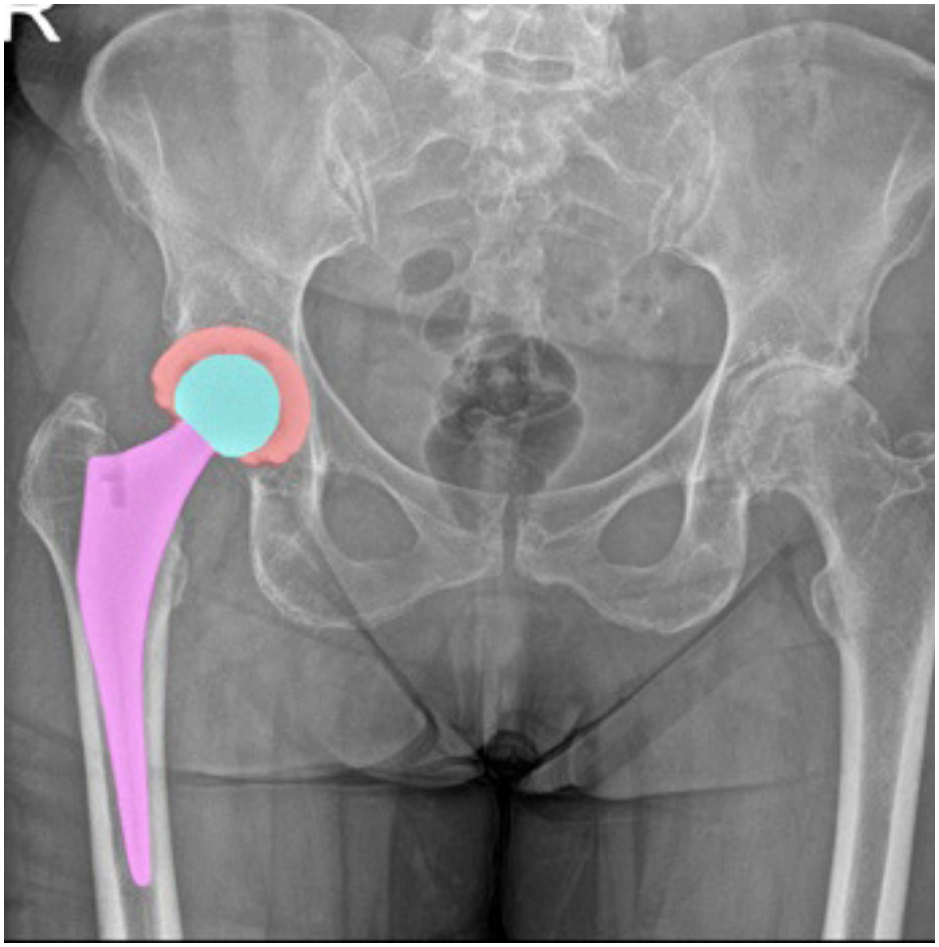
Real-time guidance

To verify the plan proceeds accordingly, **real-time tracking and registration of the surgical tools** is required. This can be achieved using manual registration or 3D tool tracking using the images from the operating room (OR), or other tracking equipment. Dedicated AI algorithms can register the live image with the pre-op scan and using augmented reality the tool position can be overlayed on the surgical plan.

Preparing the bone for implant insertion requires removal of the diseased bone and cartilage. Removing too much or too little can reduce procedural success rates. **Robotic control of the tools in combination with accurate registration of the surgical scene with the pre-op plan, prevents any alterations from the original plan**.

One of the most difficult steps in THR is implanting the stem in **the correct angle in**



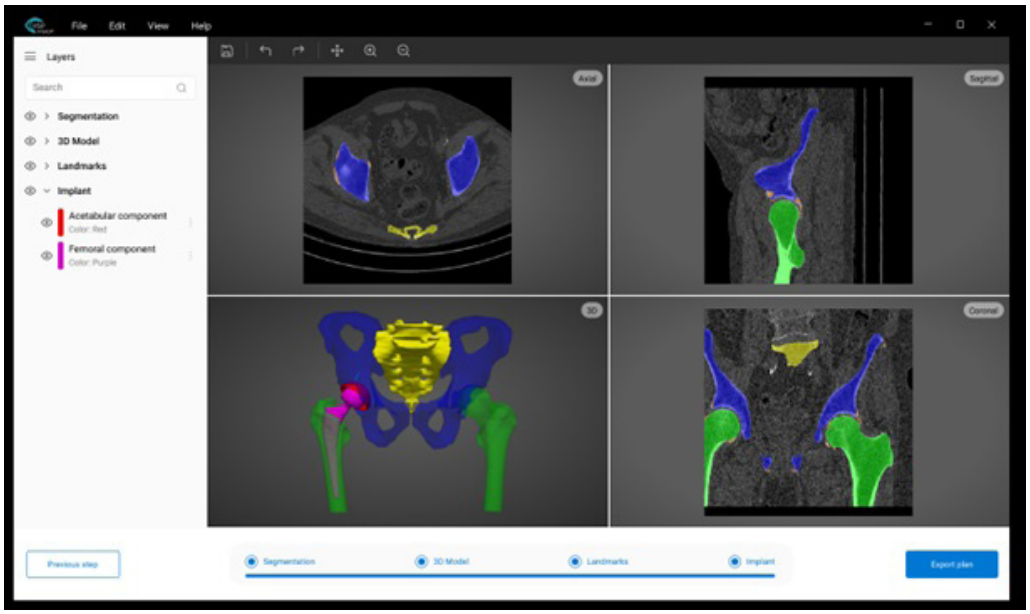


Increasing procedural success

The complete success of THR can be examined only after the patient recovers from the surgery. To correct surgical errors, additional procedures are required. The surgeon's experience and skill-set determine the chances for success. **Robotic and AI assistance can significantly increase these chances**. **RSIP Vision** can assist in implementing advanced algorithms and speed the assimilation of AI into robotic systems in THR, ultimately improving surgical outcomes.

the femur. Any diversion from the correct angle can result in significant limitation to leg rotation, and inaccurate stem depth could result in leg length discrepancy and pain to the patient. **Registration of the tool with the pre-op scan allows accurate measurement of the implant angle, providing assurance to the surgeon and reducing error-rate**.

Registration of the surgical tools with the pre-op plan can verify the **ideal position of each implant and incision**, resulting in the ideal procedure.



RSIP VISION
Visual Intelligence for MedTech

UCL

BAY VISION VIRTUAL
MEETUP

AI-Assisted Surgery for Next Generation Interventions

Recent trends in AI and surgical data science have shown promising technical advancements in imaging, surgical navigation and robotic intervention. This talk will highlight AI applications in various surgical procedures and where we stand in terms of their clinical translation as we head towards the next generation of surgical interventions.

GUEST SPEAKER

Sophia Bano

Senior Research Fellow at the Surgical Robot Vision Research Group, Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London.

HOSTED BY

Moshe Safran

CEO of RSIP Vision USA.

Defining & developing innovative Medical Visual Intelligence solutions, in partnership with MedTech industry leaders.

Missed the Meetup?

Don't miss the video!

Be sure not to miss next time :-)

TOWARDS SOFT ROBOT - ASSISTED NEEDLE INSERTION IN INTRATYMPANIC STEROID INJECTIONS

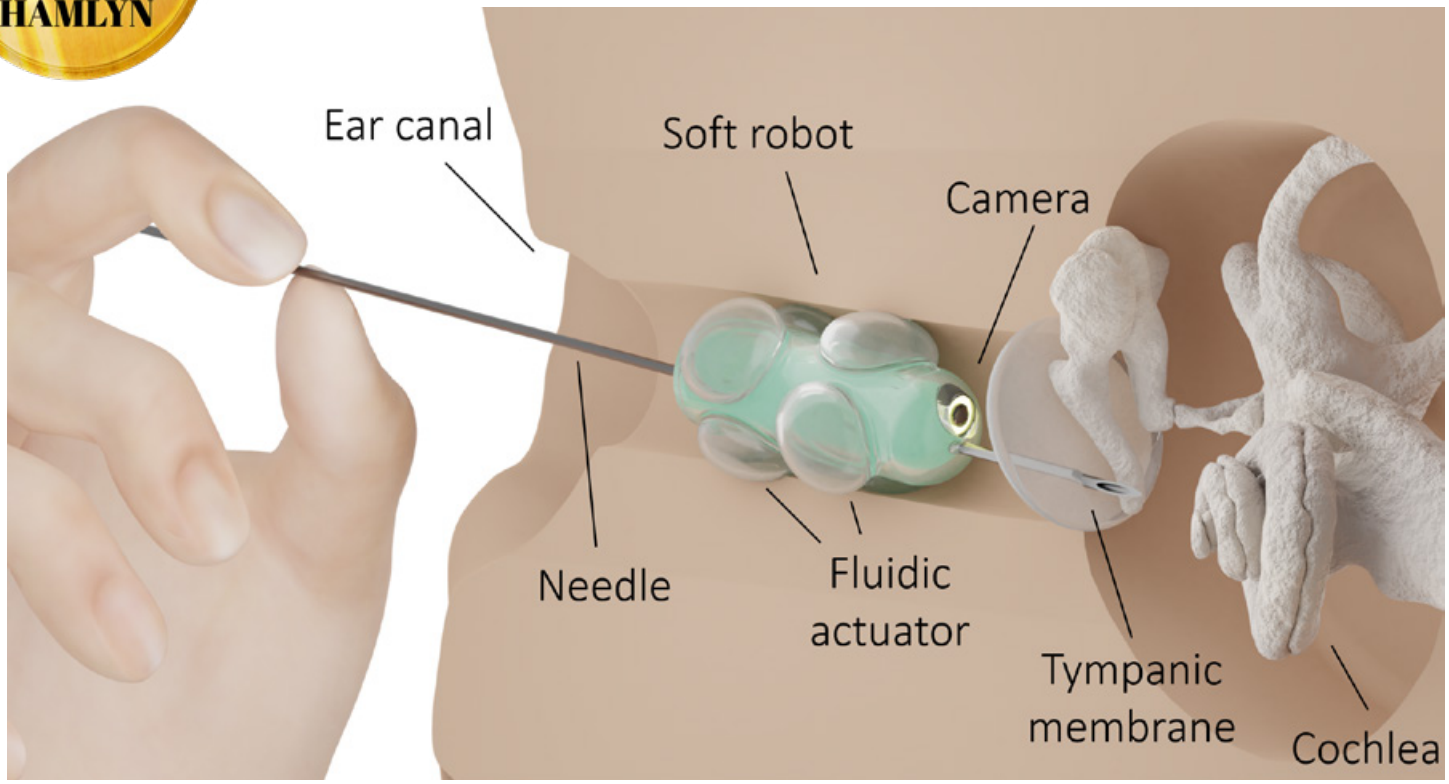
Lukas Lindenroth is a senior postdoctoral research fellow in the Surgical Robot Vision Group at UCL WEISS. He speaks to us about his work developing a soft robotic solution for ear interventions, which has just won the Best Innovation prize as part of the Surgical Robot Challenge at the Hamlyn Symposium on Medical Robotics 2022.

Intratympanic steroid injections are commonly performed by ear, nose, and throat (ENT) surgeons to treat sensorineural hearing loss, a problem becoming increasingly prevalent in the aging population. The procedure involves inserting a needle through the tympanic membrane (the eardrum) to deposit a steroid gel in the middle ear cavity.

Lukas proposes **a soft robot that helps guide the needle to the desired target** on the eardrum to avoid delicate anatomies while reducing needle motion by the surgeon.

"Generally, this procedure is perceived as painful by patients despite local anesthesia, and one of the reasons for that could be the mechanical movement of the needle by the surgeon," he tells us. *"Through a robotic solution that guides the needle to the target,* ***we can suppress this tremor in the hand of the surgeon*** *and make the procedure more comfortable for the patient."*

The essential anatomy parts to be avoided are primarily **the auditory ossicles or middle ear bones**. In the worst-case scenario, if these are harmed, it can lead to permanent hearing loss. Ultimately, the hope is to de-risk the procedure by



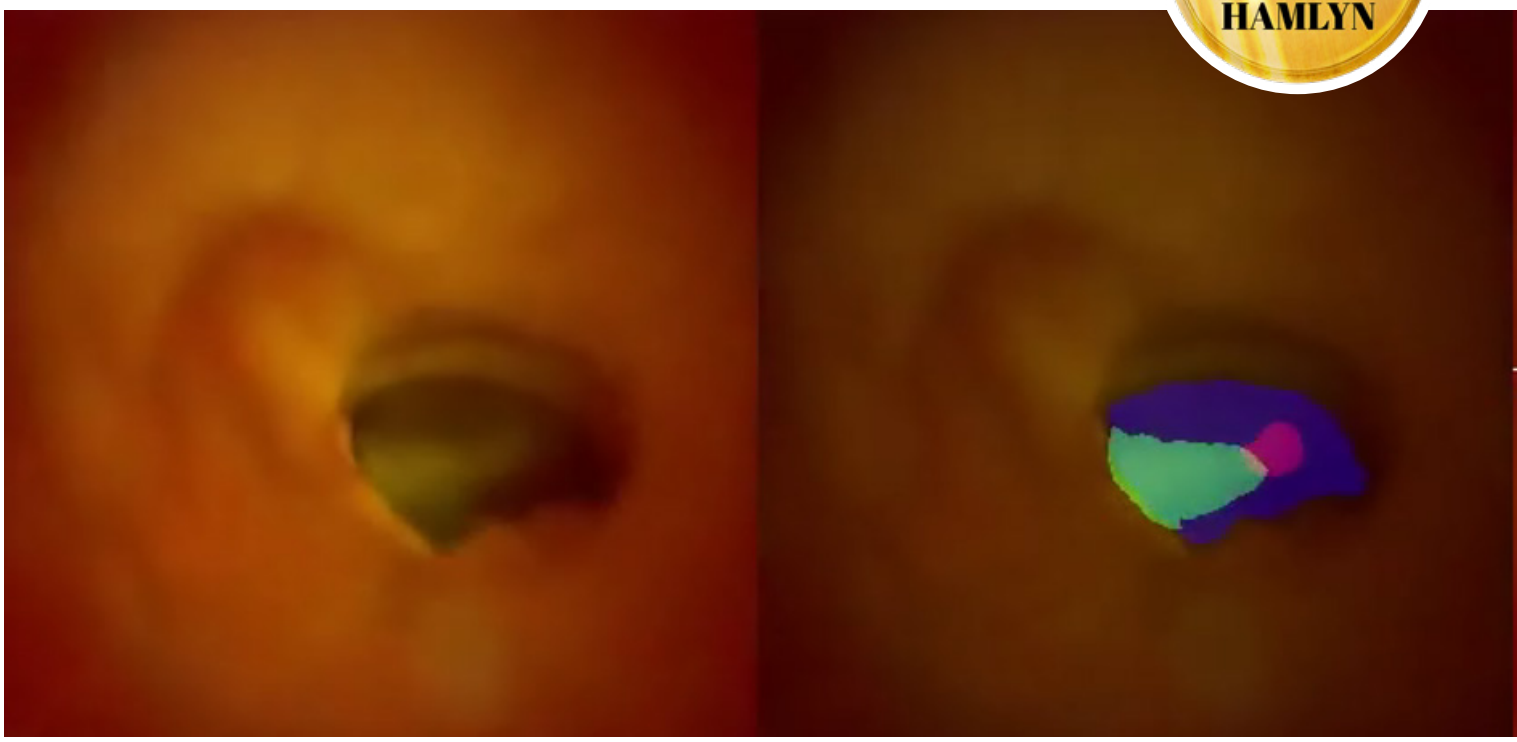
developing a platform to deliver treatment closer to the desired anatomies behind the eardrum.

The team is working with soft rubber materials and has been seeking to understand better how these materials behave and find optimal ways of actuating them.

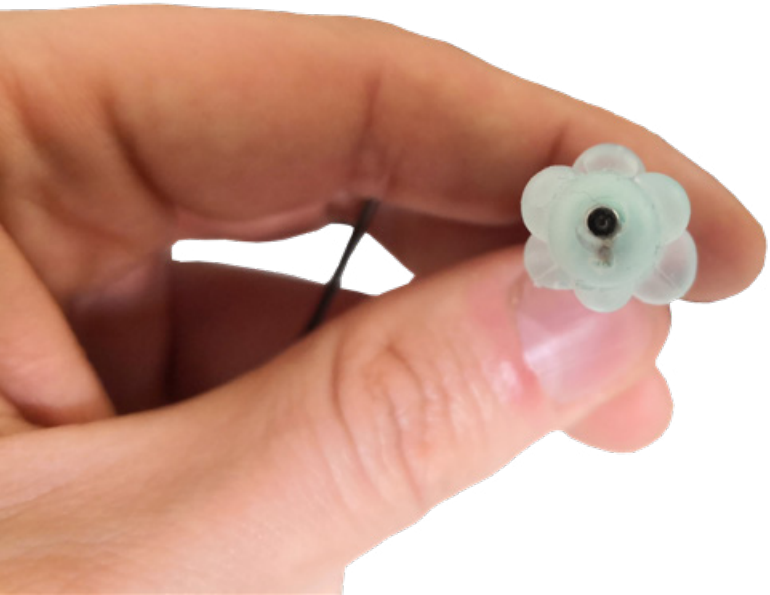
"*The way we pump liquid into the actuators that move the robot is not very well researched in soft robotics,*" Lukas points out. "*There's a lot to figure out, and some research is happening parallel to the project's overall design. Repeatability is also a big issue. We're developing calibration devices to do that on the fly.*"

Another issue concerns the miniaturization of the prototype, which is currently at two times its envisioned scale. The silicone rubber part can be scaled down, but embedding sensing and a camera require working with specialized distributors. When this procedure is performed by a surgeon manually, **an external microscope guides them toward the target**. However, the soft robot inflates in the ear canal, blocking the entire view, meaning a front camera is needed.

"*With vision, we can determine the anatomical regions we want to avoid, and then infer from these anatomical landmarks where we want to insert the needle,*" Lukas explains. "*We perform semantic segmentation of the image in real-time and identify the anatomical regions we're trying to avoid, like the malleus, which you can see protruding through the tympanic membrane, and then at the tip, you have the umbo.*"

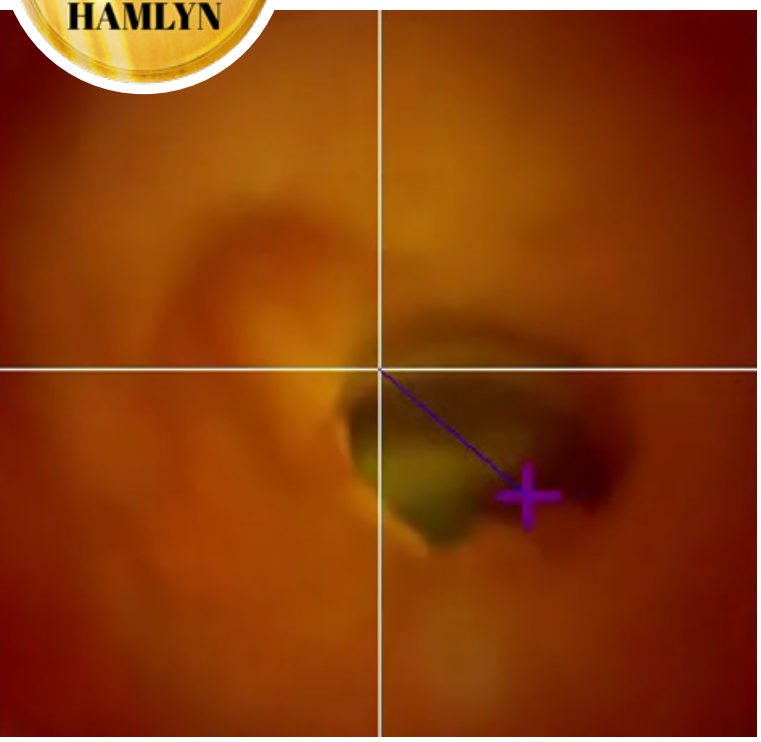


The team started out using **microscope images**, primarily. Now, they are looking into **domain adaptation** because they have an endoscopic view of the tympanic membrane, which is not easily transferable. Also, working with **custom-made medical phantoms**, they need to find ways to make the model more generalizable from the microscope views they get from publicly available data sets. They have found a data set that is more applicable to their work and are investigating how that improves the algorithm's performance.



*"**We are a big, multifaceted, interdisciplinary team working on this project**,"* Lukas tells us. *"We have all this expertise and close collaboration with the UCL Ear Institute, which helps us immensely in the design process. Your readers know Sophia Bano and Dan Stoyanov, so I think they'll know how good and well-established they are in the field of vision. **Jeref Merlin** is a research assistant with a biomedical engineering background, and he's working with me on the robotic side, the hardware development and integration, and the fabrication techniques we're using to build the robot. We also have our clinical collaborators, **Nishchay Mehta** and **Joseph Manjaly**, who work at the UCL Ear Institute and are very experienced in this field."*

From a clinical translation perspective, this work has a relatively linear validation process because there is no need to puncture the tympanic membrane initially – the robot can be deployed in the ear canal and perform some movements. The team can acquire imaging data and see how the



segmentation performs in a realistic ear canal before turning this into a clinical trial.

Could this linear pathway to clinical practice be the key to why the jury picked this paper for a prize?

"*I think it has great potential to be deployed quickly because of the simplicity of the device and the target application,*" Lukas replies. "*It's simple enough to be easily translatable into the clinic. Compared to other contenders, we're early in the project's progress, but what I think stood out was that we solved the problem differently. There are very established solutions to this problem of needle positioning with traditional robotics, but we're using* ***soft robotics****, which is not very well established in clinical environments.*"

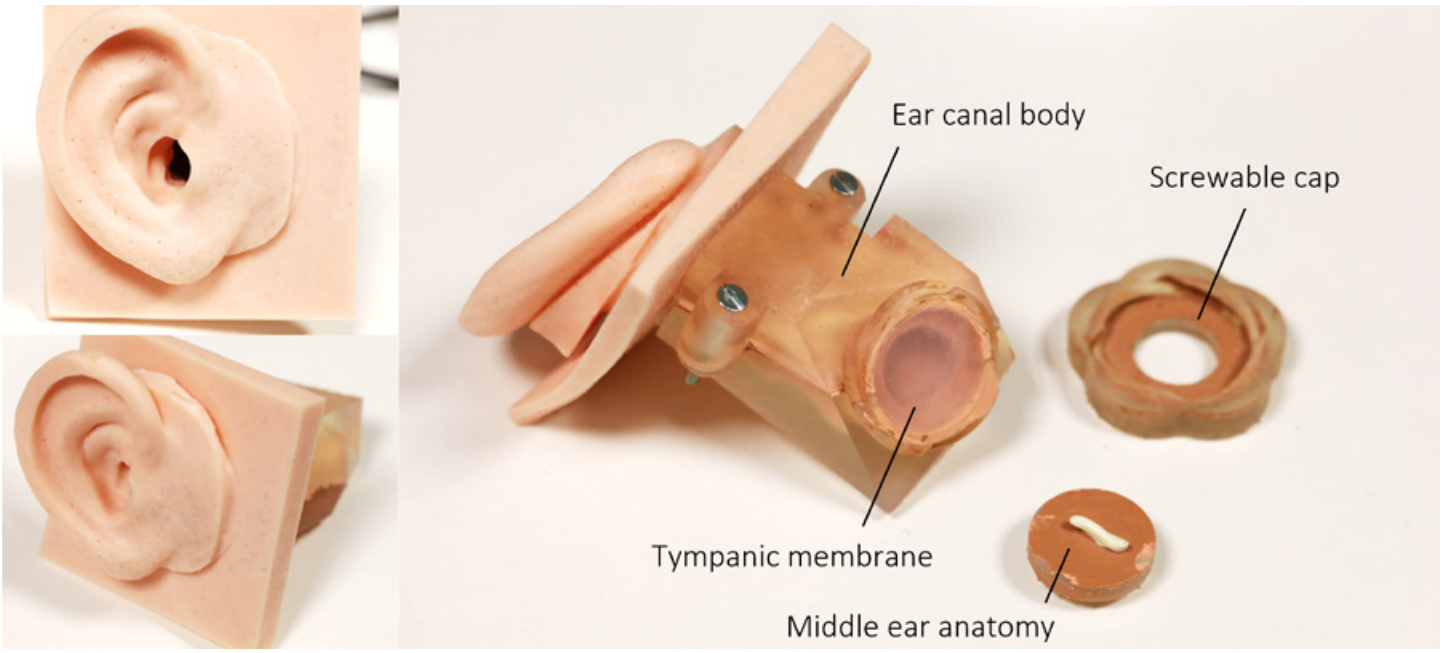


Image Registration Workshop

Mattias Heinrich is a Professor of Medical Informatics at the University of Lübeck. Alessa Hering is a PostDoc in the Diagnostic Image Analysis Group of the Radboudumc in Nijmegen, the Netherlands, and a Scientist at Fraunhofer MEVIS in Lübeck. Alongside Julia Schnabel and Daniel Rückert, they were two of the General Chairs at last month's successful WBIR 2022 event. Mattias and Alessa are here to tell us all about it

The Workshop on Biomedical Image Registration (WBIR) has just held its 10th edition, having been running for more than 20 years. It's a MICCAI-endorsed event focused on a specific subfield of medical image analysis that deals primarily with image registration.

Like so many other conferences this year, it was hybrid. Back in person in Munich, but also open to virtual participants, who could meet in Gather.Town.

"We really enjoyed holding an in-person event again," Mattias tells us.

"It was a very Bavarian experience! We had a gala dinner in a beer garden, pretzels at the workshop, and a nice mix of people from all over Europe, maybe even further away."

Mattias Heinrich

Alessa Hering

Alessa adds:

"We got much positive feedback that people enjoyed the conference in this hybrid setting. They interacted in Gather.Town, as well as at the posters and all the sessions. I think it's a nice thing that people can also join virtually."

The difference here to other conferences is that **virtual participants were asked to present their work live**, following feedback that people prefer this option to pre-recorded talks. Presenters were visible on the big screen next to their presentations. This set-up helped bridge the gap between the physical and virtual presenters and allowed remote participants to interact with and take questions from the in-person audience.

In terms of the science on show, there was a mixture of learning-based methods, plus some more classical ones. **Image registration is still a field where mathematical optimization plays an important role**, and two of the three shortlisted presentations for Best Paper were deeply into the mathematics of image registration. The winner was '**Deformable Image Registration uncertainty quantification using deep learning for dose accumulation in adaptive proton therapy**' by **Andreas Smolders**. You will read a full review in our magazine of September.

"Andreas included uncertainty into the propagation of planning of radiotherapy," Mattias explains.

"This might sound very technical, but it was a nice link of something theoretically interesting – uncertainty estimation – that could potentially impact clinical applications. I think that was great."

No successful event goes off without a hitch, and Mattias is already laughing as he recalls what happened in the first session, so it falls to Alessa to fill us in.

"*We had some technical problems, in the beginning, getting the Zoom room shown on the screen,*" she tells us.

"*Then the projector just turned off. For a brief period, the Zoom participants had a better experience than the in-person participants because we had no screen, so no presentations.*"

Mattias continues:

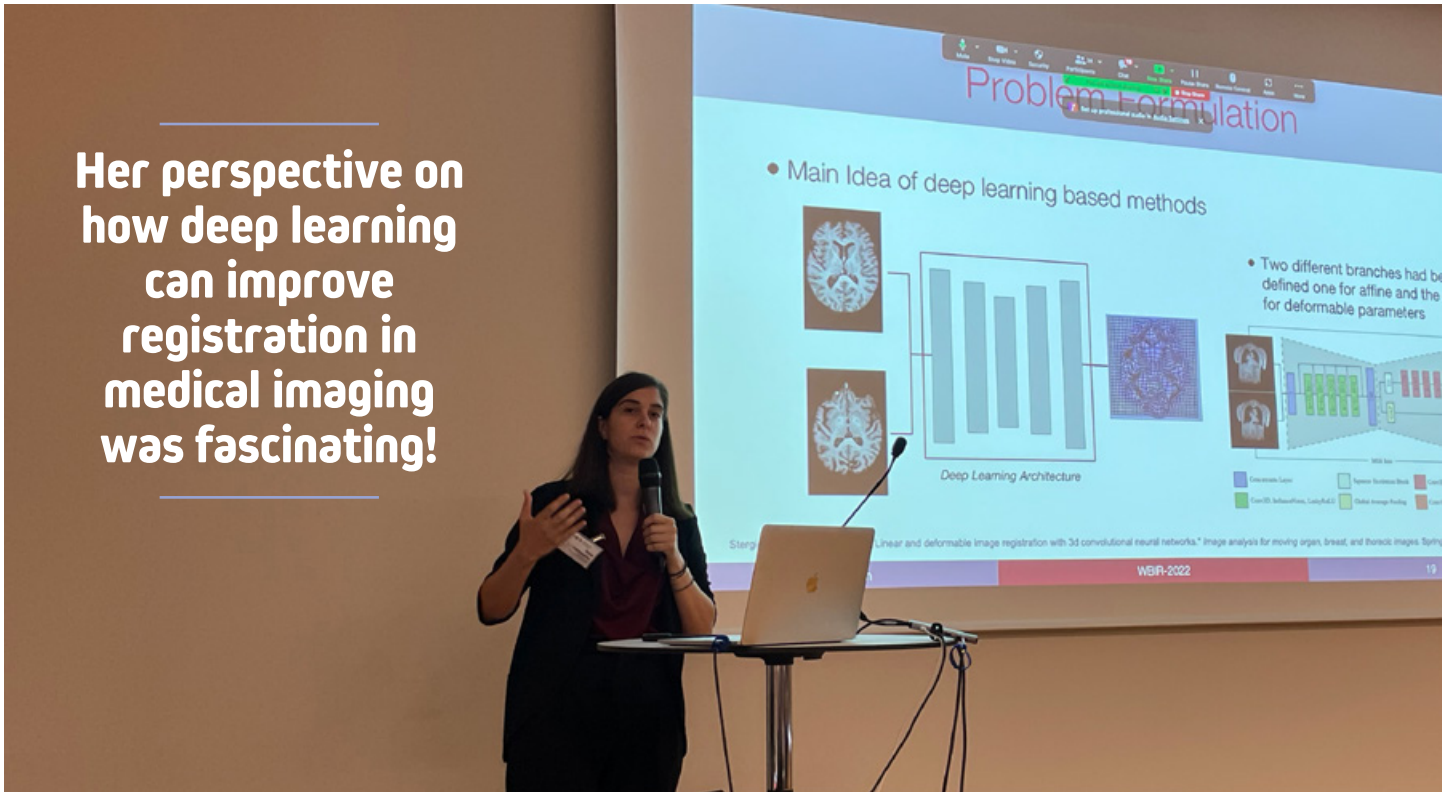
"*The funny part was that the virtual session chair, Matthew Toews, who probably had the best overview of what was going on because his connection was always there, filled the gaps when we tried to figure out the technical issues by playing the harmonica! Everyone enjoyed that.*"

WBIR is a small community, so there was plenty of time for interaction between participants outside the seminar room, who were happy to be back together. They took the opportunity to network during the day and went out together in the evening. The meeting size is beneficial for young scientists, who can more easily reach out to everyone, be it a professor or a fellow PhD candidate, in a smaller setting than some of the larger events.

WBIR has settled into a biennial format, with the next event planned for 2024.

"*We were able to convince Maria Vakalopoulou, one of our keynote speakers, to organize it together with **Marc Modat**, so we have a new location in France,*" Mattias reveals.

"*Maria is a great keynote speaker because **she connects the world of computer vision with the world of medical imaging**. She did a brilliant job showing how **classical feature extractors used primarily in vision***



can also be used in medical imaging. *Also, her perspective on how deep learning can improve registration in medical imaging was fascinating.*"

Alessa agrees:

"*We had a great deal of positive feedback about Maria's keynote. Everyone who worked in image registration said they learned something.*"

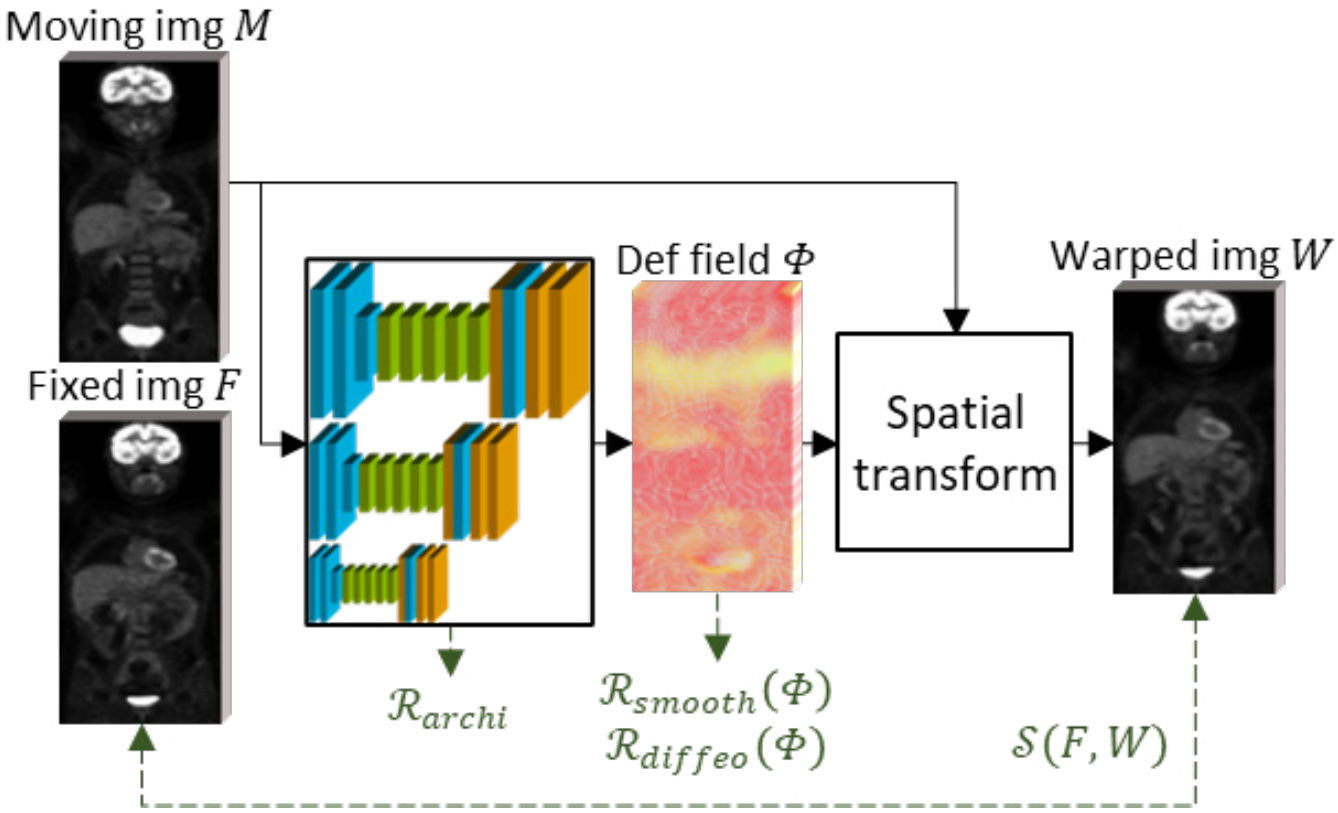
Another speaker, **Wolfgang Wein** from **ImFusion**, gave a memorable keynote, where he talked about the fact that in many real-world applications, you don't have to have one elegant end-to-end solution, but rather it can be a combination of many tried and proven modules.

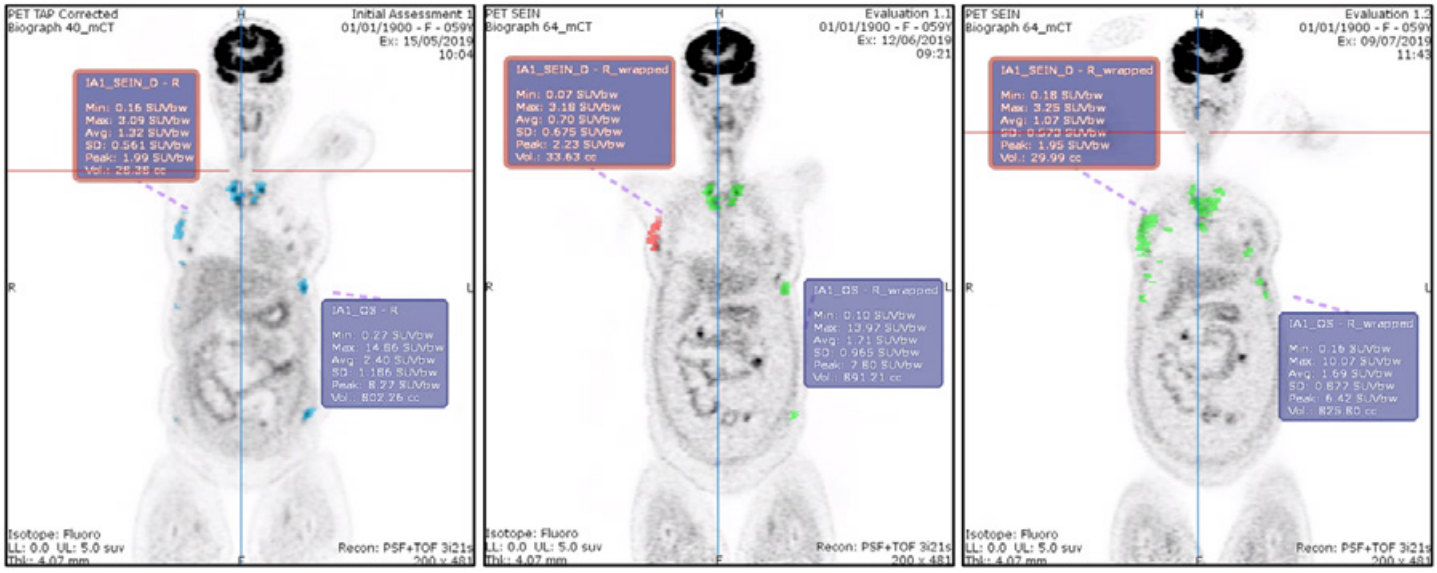
"*It shows that the tools many groups have developed over the last couple of years have already joined together to make a fantastic clinical product,*" Mattias points out.

"*That was a good takeaway message!*"

Constance Fourcade recently completed her PhD at the LS2N lab in the Ecole Centrale de Nantes, in collaboration with the company Keosys and the Institut de Cancérologie de l'Ouest (ICO) in Nantes, France. Her research focused on the use of medical image registration methods to monitor the evolution of breast cancer metastasis on Positron Emission Tomography (TEP) images. In the future, she aims to pursue her work on medical image processing to help understand and monitor diseases, such as heart infarct. Congrats, Doctor Constance!

Metastatic breast cancer is a cancer of poor prognosis that requires constant monitoring. During the follow-up care of a patient, **PET is regularly acquired** to assess the evolution of tumors over time. When interpreting the images, physicians follow specific guidelines such as **Positron Emission tomography Response Criteria In Solid Tumors (PERCIST)** to decide whether or not the treatment should be adapted or changed. However, these guidelines tend to focus only on a selection of lesions representing tumor burden, or in the case of PERCIST on only one lesion (the one showing the highest uptake). Assessing the total tumor burden is today a real challenge.





The objective of this PhD thesis was to assist physicians monitor metastatic breast cancer patients with longitudinal PET images and improve tumor evaluation by providing them tools to consider all regions showing a high uptake (aka hot regions). This thesis describes three contributions in this direction:

– Our first contribution is a method for the **automatic segmentation of active organs** (brain, bladder, etc.) based on a combination of superpixels and deep learning [Fourcade et al., 2020, *Combining Superpixels and Deep Learning Approaches to Segment Active Organs in Metastatic Breast Cancer PET Images*].

– Our second and main contribution formulates the **segmentation of lesions in the follow-up examination as an image registration problem**. The longitudinal full-body PET image registration problem is addressed first with conventional optimization-based methods and second, with the more recent deep-learning (DL) approaches. In particular, in this thesis, we developed a novel method called **MIRRBA (Medical Image Registration Regularized By Architecture)**, which combines the strengths of both conventional and DL-based approaches within a Deep Image Prior (DIP) setup (Fig. 1). We validated the three types of approaches (conventional, DL and MIRRBA) on a private longitudinal PET dataset obtained in the context of the EPICUREseinmeta project. Our proposed method performed better than all conventional approaches [Fourcade et al., 2020, *Deformable image registration with deep network priors: a study on longitudinal PET images*].

– Finally, the third contribution is the **evaluation of the biomarkers extracted from lesion segmentations** obtained from the lesion registration step. We proposed a protocol to evaluate tumor response in a case with multiple lesions. Our method provides a new visual tool for the monitoring of metastatic breast cancer (Fig. 2) [Fourcade et al., 2022, *PERCIST-like response assessment with FDG PET based on automatic segmentation of all lesions in metastatic breast cancer*].

Marwa Mahmoud is an Assistant Professor at the University of Glasgow. She is also a Visiting Fellow at the University of Cambridge, where she completed her PhD and worked for the past ten years.

Marwa, you're not Scottish and you're not English.

No, I'm originally from Egypt. I moved to Cambridge in 2010. I came here to do my PhD. Since then I've been in the UK.

Can you tell us about your work in Glasgow and Cambridge?

My main research area is in computer vision and machine learning for human behavior understanding and animal behavior understanding. I'm interested in multimodal signal processing, and I focus on applications in affective computing and social signal processing.

Do you want to give us one example?

I'm interested in applications using computer vision. For example, video analysis by looking at and analyzing human facial expressions, gesture analysis, all these nonverbal signals. In our daily interactions, we decode lots of these signals when we talk together. But machines are still not great at that. The idea is to build machines that can have social intelligence, that can make interpretations from facial expressions, body motion, tone of voice, this kind of multimodal signal, and context. These are useful for applications, for example, in robotics and virtual assistants to interact in a natural way with humans. These machines are everywhere, right? They can be in a watch. They can be in Alexa or any of these voice assistants. In the future, there will be more of these systems embedded in our world. The idea is to build models that can help in understanding human behavior and make interpretations automatically in the same way that we humans might make these interpretations. I mentioned animals as well because I've also recently been applying and devising vision techniques and computer vision for animals, automatically analyzing the movements and facial expressions of animals. This also can tell us a lot about their emotions. But, also for early diagnosis of diseases that can be painful and in general for animal welfare. It's all about using computer vision on these applications.

Shouldn't we be afraid of this? You're saying that robots will learn how to understand poses and expressions like humans. Your robot will make more judgments than a regular human!

Not really, basically there is some kind of a gold standard. What are we comparing with? On what basis? How do we know whether we can understand the person in front of us properly or not? The robot

Marwa, we are not in Giza anymore!

could be a machine. It could be a personal assistant. It could be an app on the phone, right? It can be anything. All these kinds of models can be embedded in them, and the models are usually based on data. So what is the ground truth?

Yes, our goal is to understand how we humans interact with each other, how we can understand each other better and respond in a natural way. That's what we want the machines to be able to do, to augment human intelligence and help us in our daily life. There are these assistive robots, for example, that can help provide care for the elderly. The models can also transfer the experience of an expert. For example, I'm interested in applications in mental health. I build data-driven models that are based on expert judgment related to psychological distress or depression and the aim is to make these models accessible as the first line of support.

I would say that most of this is data-driven modeling, and the main aim is really to have this natural interaction. These agents would help humans in the future.

Don't you sometimes look at a smile saying, *"Will my machine understand that?"* How do you train your software so that it understands, even if it doesn't look like a full smile?

I would ask you then, "*What do you mean if you, as a human, see this as a smile?*" How do you see it? I think the question would be why you, as a human, decide that this person is smiling.

Toddlers learn very early to recognize smiles.

But computer vision is getting better and better and actually can do that. I mean, I ask you how we humans do it because I think it's a bit similar. By having many samples or maybe a few samples from the same person, a smile can be detected. Some of these have been already discovered. The smile, for example, is not a very difficult task anymore. There are lots of cameras that can detect a smile now.

When we teach a child to draw a smile, we show a curved mouth, and it's enough. But we detect a smile much better from the eyes than from the shape of the mouth. So it's very complex, and you say that computer vision can do it simply. In my eyes, it's a very complex field.

It is complex because, as you mentioned, now the model will also learn something other than just the mouth, right? It will learn the whole face structure. It will also learn the context. Before, it was only looking at just some landmarks, movement, and very simple features. Now there are much more complex features, especially for the face. I'm interested in developing similar models also for other behaviors that are not as well studied as the face. Gestures, for example. It is a difficult problem, in general, but computer vision is getting better and better.

Some people have chronic depression, for instance, but they're not open about it. They may hide it. You are teaching a computer system to still detect depression and offer help?

Exactly. You can think about it if the person doesn't want to go to a specialist. Maybe they might be happy to chat with a chatbot or just have a conversation with an avatar on the computer. There are lots of open questions about that. People may act normally in their day-to-day life. It's a very big problem and not solved yet. The main aim is to help. Can we have some tools that can be helpful, that can make use of this kind of knowledge that I can encode in a model to be able to suggest interventions? It is still a hard problem.

The idea that a model can tell a person you're depressed has lots of implications. We need to think about how accurate the models are. I won't say that a machine would make a diagnosis of the person. Maybe in the future, but I think that's part of the feeling that machines will not replace humans. There are lots of uncertainties about false positives and false negatives and the evaluation criteria.

I am sure that you are fighting false positives and false negatives with all your might!

Even the expert will not always be perfectly accurate. For this kind of work to be deployed on a bigger level, the main thing would be lots of collaborations with people who are actually on the front line. I'm a computer scientist, but I read a lot on psychology and psychiatry because that's part of the work on these applications. I collaborate a lot with people thinking about all the ethical issues and how these interventions can be really useful to humans in the future... So lots of collaborations and translational research.

How could young Marwa leave such a warm, wonderful country like Egypt and go to cold countries far away - only for science?

I'm still not used to the cold weather here. The short days, especially in the winter, are kind of depressing. The idea came because I did my Master's at the American University in Cairo, and may work was also related to computer vision and affective computing. I got fascinated by this field at that time, and I wanted to push my research. When I started my PhD, my research proposal was ready. I didn't apply to many universities. I just applied to Cambridge because I somehow was like, this is what I want to do, and let's see if I can do it. I come from a conservative family as well, so it wasn't easy to just go and live on my own and say, *"Bye-bye, I'm moving to the UK!"* I think for me, I was like, okay, I have the research idea. I really want to do that. So I applied,

I got accepted, and I came to do my PhD. Somehow, I was also kind of lucky in a way to do what I wanted to do at that time. Also, other than work, I was keen on the experience of living in another country and knowing a different culture. I used to travel before that, but relocation is completely different. You find that you suddenly just do everything on your own, and the country's new and everything. I learned a lot from the experience. I recommend it.

What would have happened if Marwa, twelve years ago, had decided to do something other than science?

I never thought of that question. I don't know. I'm kind of a person that when I want to do something, I usually do it. I would have found a different way to pursue the same dream.

Do you think it was worth doing it?

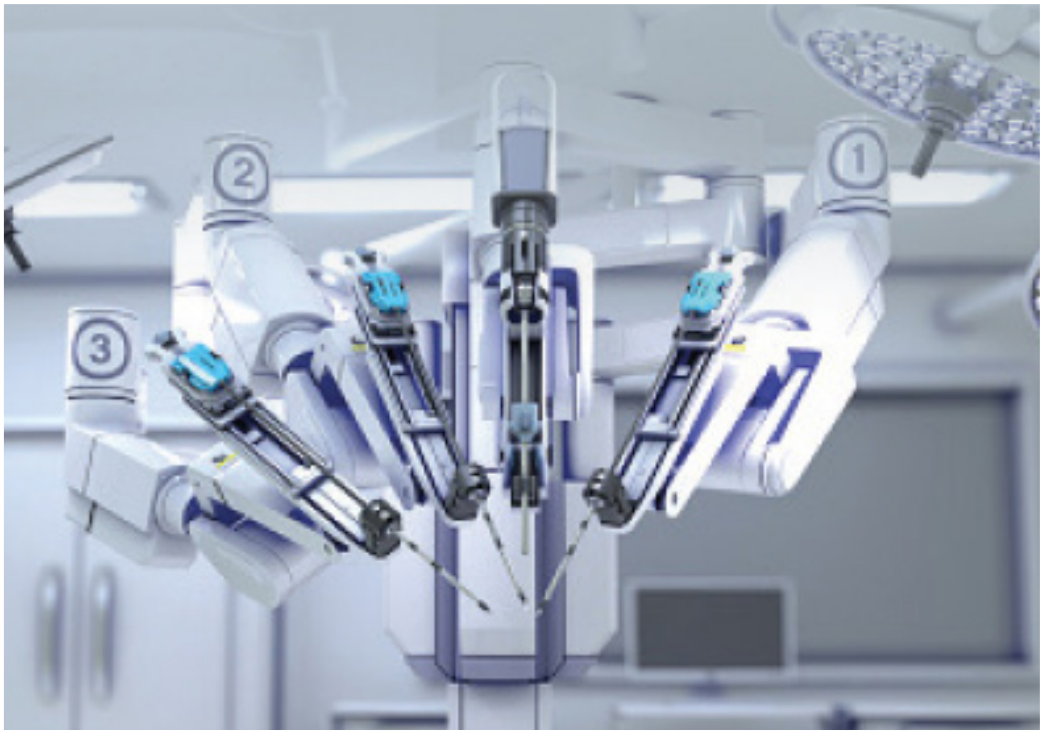
Of course. I didn't ever doubt the decision. Maybe at the very beginning when I came here, and it was raining all the time. I arrived in October, and I came to Cambridge. It was at night, and it was dark and raining and cold. And I was like, what am I doing?

It's like: Marwa, we are not in Giza anymore!

To be honest, the UK has been not too bad as well because it's only a four-hour flight to Egypt. For a long period, I used to go back three times a year, and it's not that bad. I still miss the food and the weather though.

Ah, the food and the weather in Egypt, I love them too. What's your message for the community?

I want to talk to young women who might be thinking about doing something like relocating or pursuing their dreams in a different country or taking any challenge really, and thinking about it, if it's worth it or not. I just want to tell them it is worth it and go for it. It pays off in the end. It's not an easy kind of journey. It won't be all kinds of rainbows, but it's really worth it. Also, don't listen too much to people who might make you doubt your decisions. If you really want to do something, just go for it and just keep going!



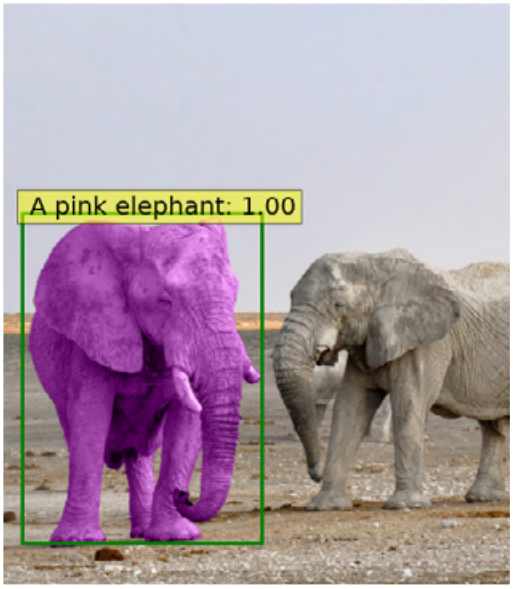
IMPROVE YOUR VISION WITH

Computer Vision News

SUBSCRIBE

to the magazine of the algorithm community

and get also the new supplement Medical Imaging News!



RSIP VISION