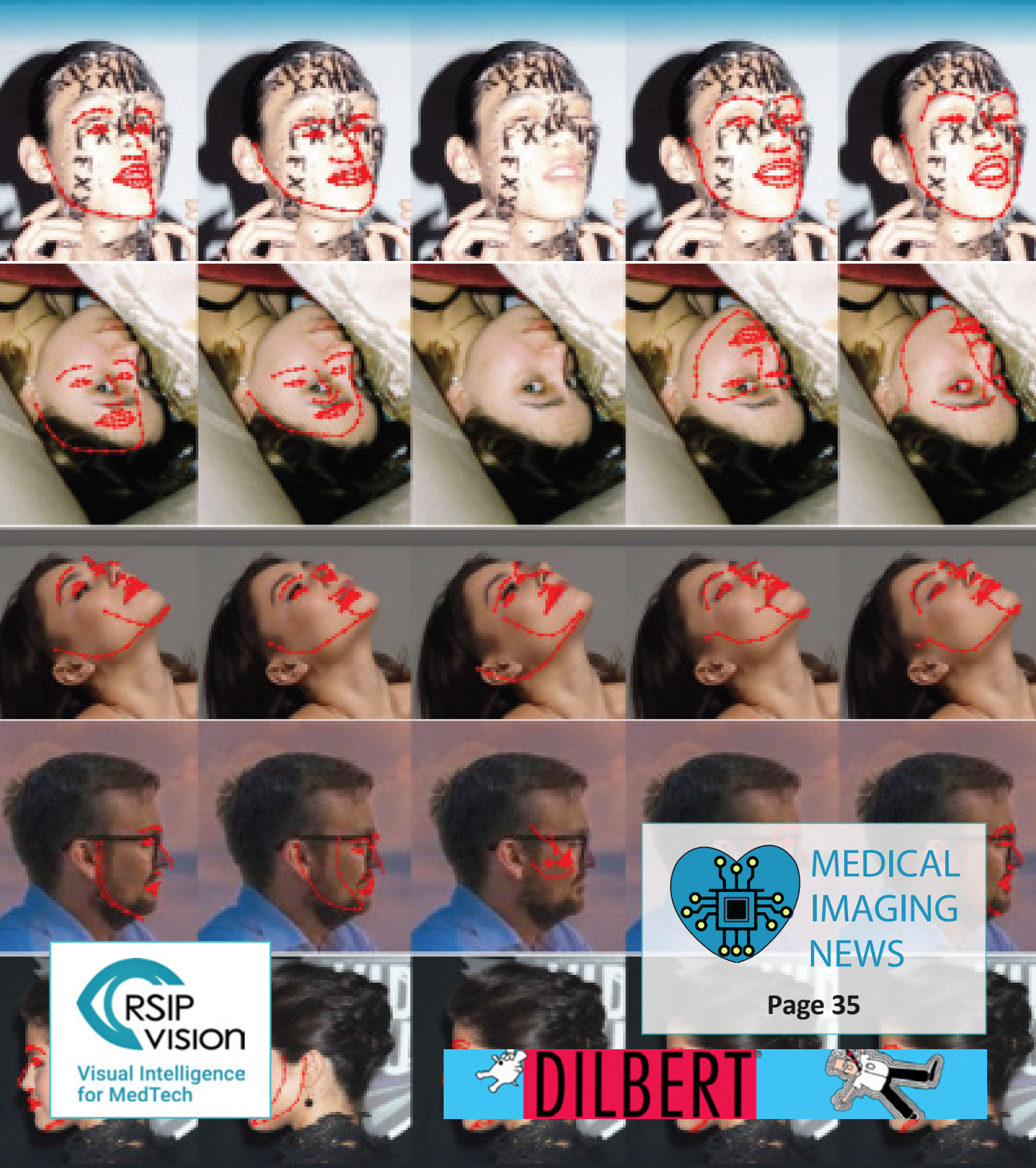
APRIL 2022

Computer Vision News

The Magazine of the Algorithm Community



MEDICAL IMAGING NEWS
Page 35

DILBERT

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,

CVPR 2022 is only a couple of months away, and after two long years apart, it has been announced that we will finally be reunited at **an in-person conference again!** Registration is open, and in this April edition of **Computer Vision News**, we can already review some of the exciting work set to be presented at the event.

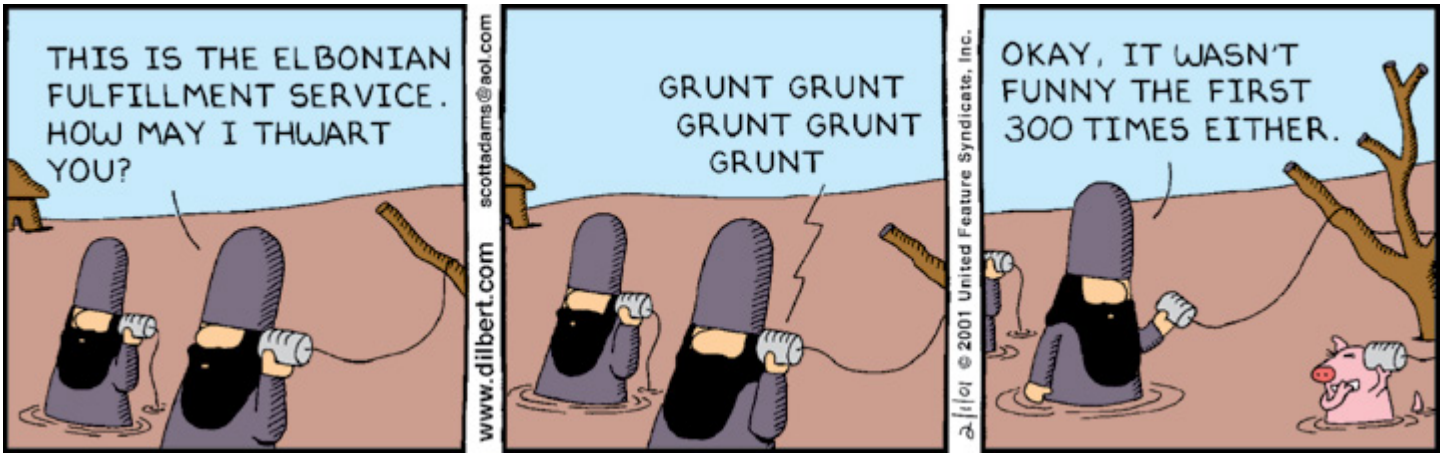
Over the page, you will find our first review of a CVPR2022 paper: **DAD-3DHeads - A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image.** We are especially happy to feature this fascinating work as Tetianka and most authors are from **Ukraine. Marica and I give their homeland and people a warm mark of support.** Read the review on page 4.

There are many workshops, challenges, tutorials, and much more to look forward to at CVPR. You will find a preview of this year's **Medical Computer Vision workshop** on page 46. Now in its 9th edition, it has well and truly earned its place as a classic feature on the CVPR calendar. The event is an opportunity for the community to hear about the latest progress and discuss new and future developments, with a focus on being as diverse and inclusive as possible in terms of speakers and attendees.

Away from CVPR, on page 10, Ioannis explores the **recognition of human faces from images** using deep learning and the model framework OpenCV 2, with **plenty of coding and Python experimentation** to get your teeth into!

One final word about CVPR 2022 before we go: Computer Vision News will be there to publish the official **CVPR Daily magazine for the seventh consecutive year**. We are already working with organizers to ensure the magazine is your first port of call for all the best picks from the conference meetings and presentations. **We look forward to welcoming you all to New Orleans in June!**

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



Computer Vision News

Medical Imaging News

DAD-3DHEADS: A LARGE-SCALE DENSE, ACCURATE AND DIVERSE DATASET FOR 3D HEAD ALIGNMENT FROM A SINGLE IMAGE

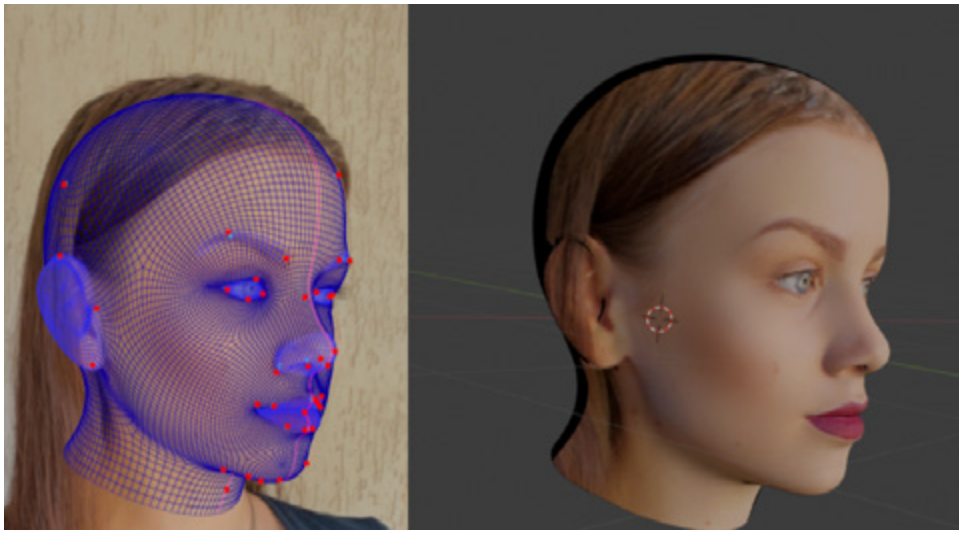
by Marica Muffoletto (twitter)

Hi everyone and welcome to a special review this month! We have a paper called **DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image,** recently accepted to CVPR 2022. Ralph and I are especially happy with choosing this paper, because it enables us to show **support to the main author's homeland and people**.

We deeply thank all authors (**Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Viktoriia Sharmanska, Jiří Matas**) for allowing us to use their images.

Introduction of a new dataset

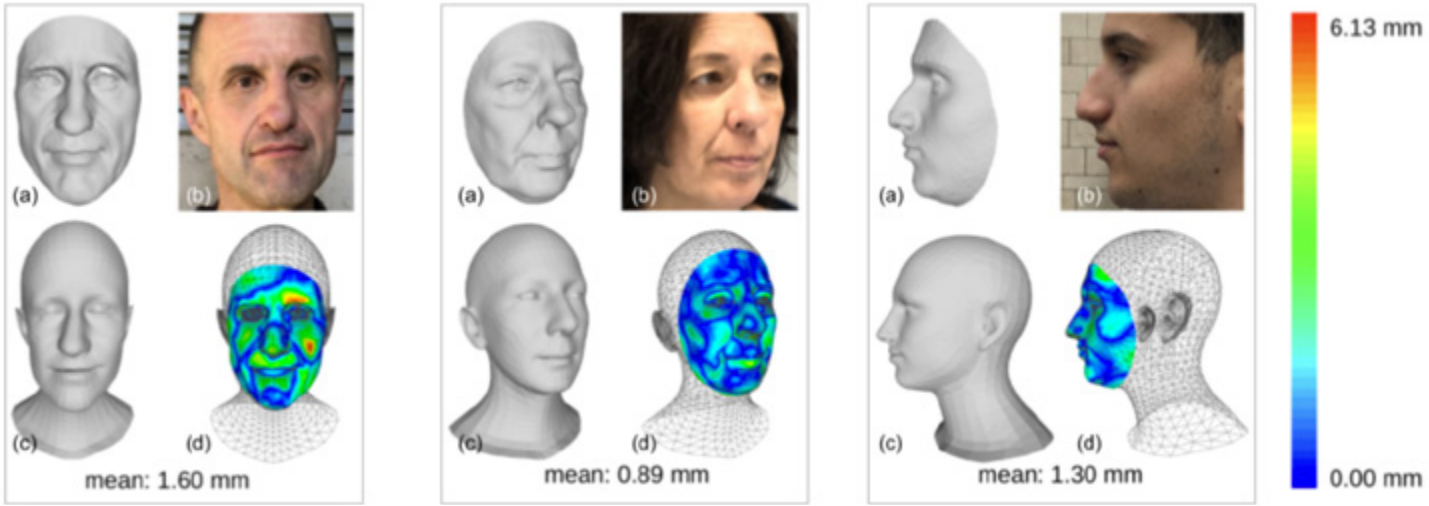
As suggested by the title, the main contribution of this paper is the DAD-3DHeads Dataset for 3D face analysis. Needless to say, many tasks in Computer Vision rely highly on the quality and quantity of the data employed, and therefore such a contribution might be essential to advance the state-of-the-art. This dataset aims at containing a variety of extreme poses, facial expressions, challenging illuminations, and severe occlusion cases. Moreover, as a big step forward, it proposes to include annotations of 3D landmarks directly from images, that are tested for accuracy and consistency compared to 4D ground truth scans. The authors argue that the lack of those and hence the employment of 2D landmarks in current datasets is one of the main challenges in the training of pose estimation and head alignment models.



To obtain this dataset, a novel annotation method is used, which employs a 3D modelling tool. A 3D Morphable model is fit to a human head image by picking anchoring points on the 3D mesh surface (head on the left). Based on the pinned points, the tool automatically modifies the mesh to optimise the 3DMM parameters, so that the “pin” reprojection error is minimized. Finally, the annotator can check the texture rendered onto the 3D mesh to verify that it is realistic.

The total dataset is made of 42,130 images from different sources. Each of these is provided with 5,023 vertices of which 3,669 (head subset) are proven to be accurately labelled. The flexibility of this dataset is given primarily by the extension of the landmarks, and one can choose from different subsets (see figure on the left: starting with 68 landmarks -> 191 -> 445).

The annotation accuracy can be observed qualitatively in the image below, where a) shows the ground truth scan provided by previous datasets, b) is the annotated result, while d) shows the alignment of the mesh and the GT scan.



The quality of the 3D novel annotations is also quantitatively measured through comparison with the manually labelled facial landmarks (found by 10 different annotators).

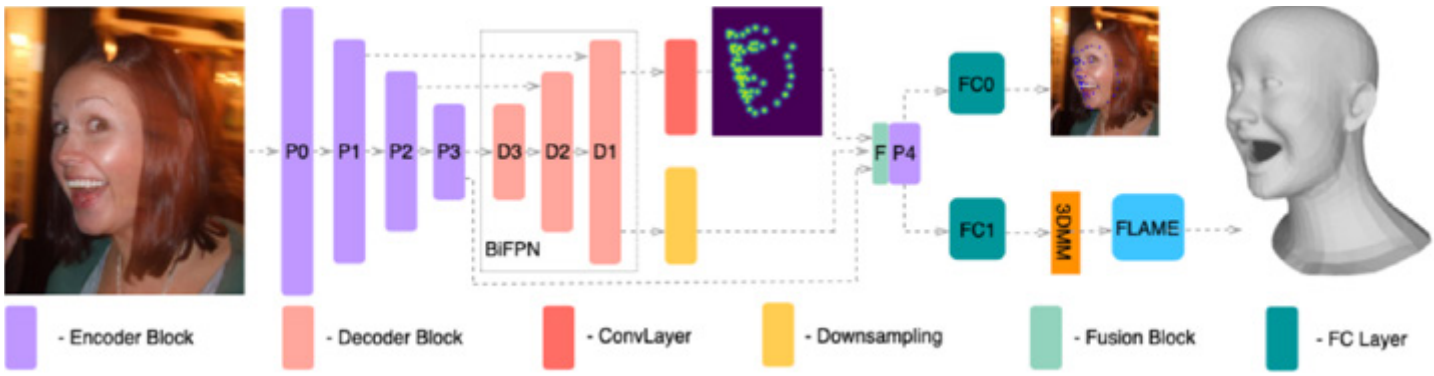
This is done by computing the quality score F_Q for each approach averaging across the images as a normalised mean error between each pair of labels, resulting in an average NME 45.8% lower in the 3D method compared to the manual one.

$$\mathcal{F}_Q = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{d_n}\cdot\frac{2}{m(m-1)}\sum_{i=1}^{m}\sum_{j>1}\left|\left|\vec{x}_n^i - \vec{x}_n^j\right|\right|_2$$

where N is a subset of 30 images, d_n is the head bounding box size, and $\vec{x}$ is an array of 68 labelled landmarks.

Experiments with the new dataset

To demonstrate the efficiency of the dataset, a data-driven DAD-3DNet model is trained. This is made of a **CNN encoder**, a **Landmark Heatmap Estimator** that predicts coarse locations of 2D landmarks, a **Fusion Module** that fuses the heatmap prediction with the CNN encoder features, and a **Regression Module** that predicts finer facial landmark's locations and 3DMM parameters vectors. **FLAME** Layer maps the 3DMM vector to 3D head model vertices.



The overall loss employed is a combination of 4 terms with respective weights:

$$L = 50L_{3D} + L_1 + 0.05L_{proj} + L_{AWing}$$

Here, L_1 is a Landmark Regression Loss while L_{AWing} is a Gaussian Heatmap Loss. The other two terms are novel: L_{3D} is a **Shape+Expression Loss**, which disentangles shape and expression from pose, measuring the discrepancy between the normalised subsampled (only head) predicted vertices and the ground truths in 3D. L_{proj} (**Reprojection Loss**) optimises the pose accuracy through a projection of the 3D vertices of the mesh into the image. The L1 loss is used to measure the difference between the reprojected subsampled vertices.

The encoder is pre-trained on ImageNet. The FLAME layer is fixed during training, and ADAM optimiser with learning rate = $1x10^{-4}$ which reduces when validation loss stops decreasing. No image augmentation is used.

Experiments focus on:

1. Evaluate the task of 3D Dense Head Alignment from an image
2. Test model generalization (on a range of tasks - Face Shape Reconstruction, Head Pose Estimation)
3. Test robustness to extreme poses

Evaluation Protocol

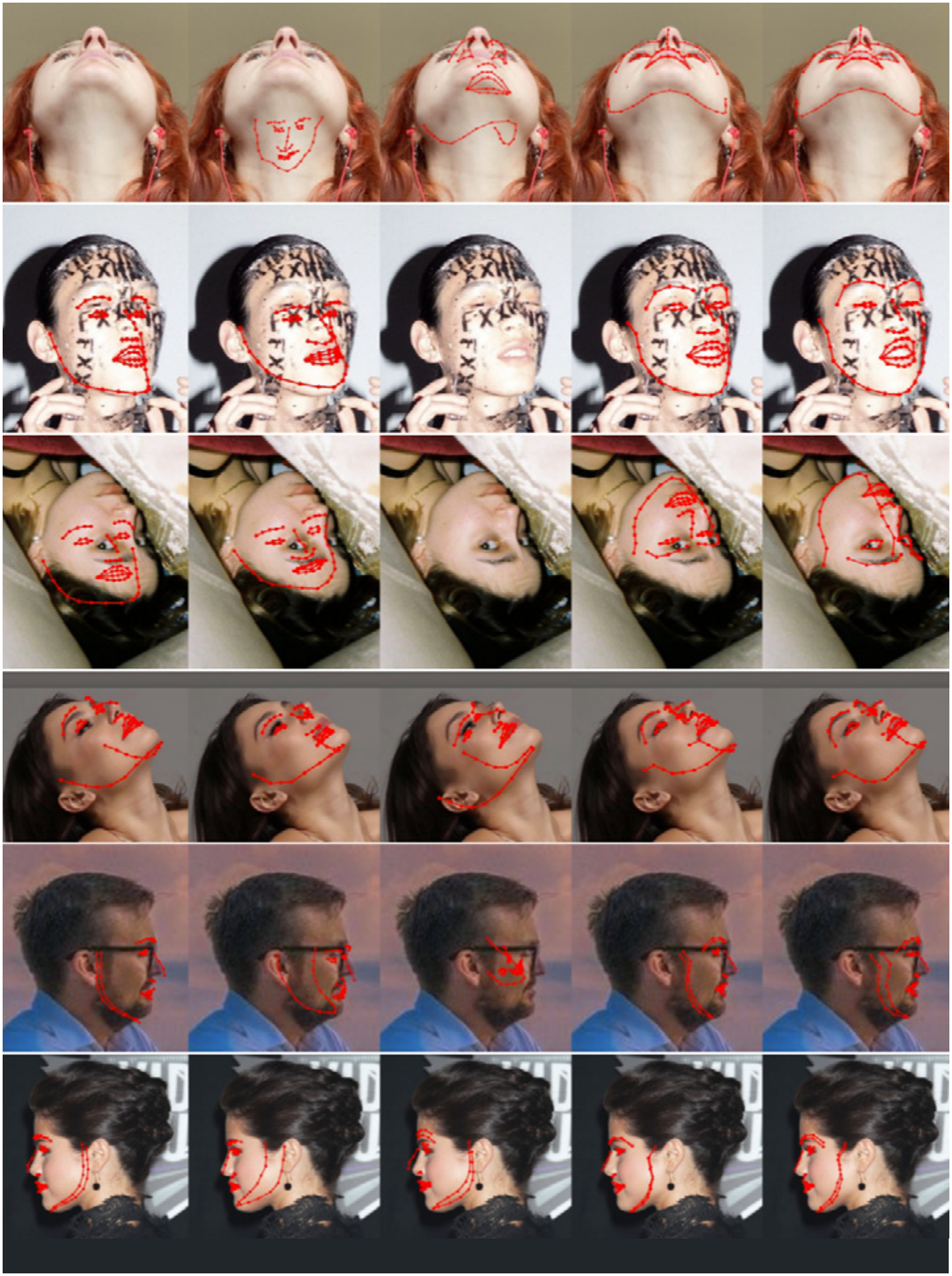
In this paper, two new metrics for 3D Head Learning tasks are introduced. These are the **Reprojection NME** and the Z_n **accuracy**.

The former computes the normalised mean error (NME) of the reprojected 3D vertices onto the image plane for 68 landmarks.

The latter finds K/N v_i closest vertices in the ground-truth mesh and calculates which of them is closer or further from the camera, finally it compares to the configuration for the predicted vertices w_i.

$$Z_n = \frac{1}{N}\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{K}\left(\left(v_i \geq v_i^j\right) == \left(w_i \geq w_i^j\right)\right)$$

Moreover, traditional metrics are employed: **Chamfer Distance**, for measuring the accuracy of fit on any number of predicted vertices and **Pose error** to measure accuracy of pose prediction based on rotation matrices.



Experiment 1) All these metrics are used to compare the state-of-the-art 3D Dense Head Alignment models on the DAD-3DHeads Benchmark. On the full test dataset, the NME for DAD-3DNet for this task is 2.302, significantly lower than 3.580 for the second-best of these methods (3DDFA-V2). Similar results are found on subsets of atypical poses, compound expressions and heavy occlusion. DAD-3DNet shows superior performance in all cases, which can be also observed in the qualitative results below. From left to right, we find 3DDFA-V2, FaceSynthetic, JVCR, DAD-3DNet.

Experiment 3) Similar analysis using all the metrics has been conducted on multiple subgroups (camera pose, age, image quality, occlusions, expressions, lighting) to show robustness of the proposed approach across various conditions (distribution shifts) in-the-wild.

Experiment 2) The tasks of 3D Head Face Shape Reconstruction and Head Pose estimation are also evaluated on standard benchmarks (the NoW Face Challenge, and the Feng et al. benchmark) showing advantageous or comparable performance to other state-of-the-art methods.

Conclusion

Through this work, the authors stress the importance of a diverse, accurate, dense, and in-the-wild dataset for 3D face analysis. The quality of the dataset is boosted by the efficiency of the loss components and the use of ad-hoc evaluation metrics, leading to the conclusion that incorporating information about the full head can improve the model stability, and that the landmarks regression and coarse heatmap estimation modules significantly improve the model performance.

We end on a good note for people who want to extend this work and further improve the performance of 3D Landmark localization models… No method is perfect! And even DAD-3DNet fails under challenging situations. Time to work on future advancements and thank the authors for their beautiful illustrations and the decision to make their dataset and models publicly available.

See you all next month ☺

Facial recognition

IOANNIS VALASAKIS, KING'S COLLEGE LONDON

@WIZOFE

Hi everyone! As I promised, this month we are going to explore the exciting topic of recognizing human faces from images, using deep learning and the model framework Open CV2. I hope you'll enjoy as always and please keep me updated of your new projects, creations or questions ☺

I expect that you like this change of topic from medical to computer vision but as always let me know what's your preferences and I'll try to commit to it next month. Without further ado, let's dive into a lot of coding and python experimentation!!

Facial Recognition using Open CV2

Let's start by importing a few libraries:

```
# Install VGGface for later use
! pip install keras-vggface
! pip install keras_preprocessing
! pip install keras_applications
```

Facial Image Classification using TensorFlow

We want to perform a face recognition which is the general task of identifying and verifying people from photographs of their face.

Step 1.1: Importing Libraries

Import libraries and define environment variables

```
import cv2
import numpy as np
import os
import math
from matplotlib import pyplot as plt
%matplotlib inline
print(cv2.__version__)
%matplotlib inline
cv2.startWindowThread()
from os import listdir
from PIL import Image
import warnings
warnings.filterwarnings(action='once')
import urllib.request
```

```
import keras
import keras_vggface
from keras.engine import  Model
from keras.layers import Flatten, Dense, Input
```

```
from keras_vggface.vggface import VGGFace
from tensorflow.keras import datasets, layers, models
from keras.optimizers import RMSprop, SGD
from keras_vggface.utils import preprocess_input
from keras_vggface.utils import decode_predictions
from numpy import asarray
from keras.applications.vgg16 import VGG16
from keras.applications.vgg16 import preprocess_input
from keras.preprocessing.image import load_img
from keras.preprocessing.image import img_to_array
from keras.models import Model
from matplotlib import pyplot
from numpy import expand_dims
from keras.preprocessing import image
```

Step 1.2: Load Data/Image

We have various ways to load images e.g. OpenCV, MatplotLib, PIL, etc.

Read and Write Images Example using OpenCV

```
python cv2.imwrite(file_path (str), image (numpy.ndarray))
cv2.imread(file_path (str), read_mode (int))
```

Read Modes

- 1 = cv2.IMREAD_COLOR
- 0 = cv2.IMREAD_GRAYSCALE
- -1 = cv2.IMREAD_UNCHANGED

Load a Sample Smith image using OpenCV library

```
# Read Image using OpenCV

import cv2

img =cv2.imread("/content/Facial-Recognition-MMAI844-Tutorial/sample/Smith.jpg",1)
img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
plt.imshow(img)
plt.show()
# Read Image using matplotlib

import matplotlib.image as mpimg
import matplotlib.pyplot as plt

img = mpimg.imread('/content/Facial-Recognition-MMAI844-Tutorial/sample/Smith.jpg')
plt.imshow(img)
```

Step 1.3: Detecting Faces with OpenCV and front face detector xml

```
detector = cv2.CascadeClassifier( xml_file_path)
face_coord = detector.detectMultiScale(image, scale_factor, min_neighbors, min_size, flags)
face_coord: Numpy array with rows equal to [x, y, width, height]
```

Let's import a facial image to detect the face

```
frame= cv2.imread('/content/Facial-Recognition-MMAI844-Tutorial/sample/channing_tatum.jpg')
name= "channing_tatum"
plt_show(frame)
```

Below is the loading of the OpenCv face detector module and the detection of the face coordinators on the image

```
detector = cv2.CascadeClassifier("/content/Facial-Recognition-MMAI844-Tutorial/xml/frontal_face.
xml")

scale_factor = 1.2
min_neighbors = 5
min_size = (40, 40)
biggest_only = True
flags = cv2.CASCADE_FIND_BIGGEST_OBJECT | \
            cv2.CASCADE_DO_ROUGH_SEARCH if biggest_only else \
            cv2.CASCADE_SCALE_IMAGE

faces_coord = detector.detectMultiScale(frame,
                                        scaleFactor=scale_factor,
                                        minNeighbors=min_neighbors,
                                        minSize=min_size,
                                        flags=flags)
print("Type: " + str(type(faces_coord)))
print("this is face coordinator in the picture is {}".format(faces_coord))
print("Face is successfully detected!! Let draw a box on the picture")
```

Define a function to draw a rectangle around the detected face

```
def draw_rectangle(image, coords):
    for (x, y, w, h) in coords:
        w_rm = int(0.2 * w / 2)
        cv2.rectangle(image, (x + w_rm, y), (x + w - w_rm, y + h),
                              (0, 0, 255), 8)

draw_rectangle(frame,faces_coord)
cv2.putText(frame, name,
                            (faces_coord[0][0], faces_coord[0][1]),
                            cv2.FONT_HERSHEY_SIMPLEX , 2, (66, 53, 243), 6)
plt_show(frame)
```

```
################################################################
# Wrap up the code into a face detector module for later use
################################################################

class FaceDetector(object):
    def __init__(self, xml_path):
        self.classifier = cv2.CascadeClassifier(xml_path)

    def detect(self, image, biggest_only=True):
        scale_factor = 1.2
        min_neighbors = 5
        min_size = (30, 30)
        biggest_only = True
        flags = cv2.CASCADE_FIND_BIGGEST_OBJECT | \
                    cv2.CASCADE_DO_ROUGH_SEARCH if biggest_only else \
                    cv2.CASCADE_SCALE_IMAGE
        faces_coord = self.classifier.detectMultiScale(image,
                                                       scaleFactor=scale_factor,
                                                       minNeighbors=min_neighbors,
                                                       minSize=min_size,
                                                       flags=flags)

        return faces_coord
```

Step 1.4: Cut Faces and resize faces

for (x, y, w, h) in faces_coord: cv2.rectangle(frame, (x, y), (x + w, y + h), (150, 150, 0), 8) plt_show(frame)

```
def cut_faces(image, faces_coord):
    faces = []

    for (x, y, w, h) in faces_coord:
        w_rm = int(0.2 * w / 2)
        faces.append(image[y: y + h, x + w_rm: x + w - w_rm])

    return faces
Cut_Face = cut_faces(frame, faces_coord)
plt_show(Cut_Face[0])
```

Define a function to resize the face back to 224x224 resolution

```
def resize(images, size=(224, 224)):
    images_norm = []
    for image in images:
        if image.shape < size:
            image_norm = cv2.resize(image, size,
                                    interpolation = cv2.INTER_AREA)
        else:
            image_norm = cv2.resize(image, size,
                                    interpolation = cv2.INTER_CUBIC)
        images_norm.append(image_norm)

    return images_norm
resize_faces = resize(Cut_Face)
plt_show(resize_faces[0])
```

Step 1.5: Standardize the image for fitting Neural Net Model

Normalizing image inputs: Data normalization is an important step, which ensures that each input parameter (pixel, in this case) has a similar data distribution. This makes convergence faster while training the network.

For image inputs we need the pixel numbers to be positive, so we might choose to scale the normalized data in the range [0,1]

```
resize_face_stad = resize_faces[0]/255.0
plt.imshow((resize_faces[0]/255))
```

Pre-trained Model in TensorFlow

Step 2.1 Install libraries

VGGFace and VGGFace2 Models

The VGGFace refers to a series of models developed for face recognition and demonstrated on benchmark computer vision datasets by members of the Visual Geometry Group (VGG) at the University of Oxford.

For more information, please visit the github link: **https://github.com/rcmalli/keras-vggface**

Step 2.2: Verify Data and extract faces from public dataset

```
#########################################################
# Define a function to load the image, then detect faces, and finnally cut face
#########################################################

def extract_face_from_file(filename, required_size=(224, 224)):
    # load image from file
    pixels = plt.imread(filename)
    detector = FaceDetector("/content/Facial-Recognition-MMAI844-Tutorial/xml/frontal_face.xml")
    # create the detector, using default weights
    faces_coord = detector.detect(image=pixels)
    faces = cut_faces(pixels, faces_coord)
    faces = resize(faces)
    return faces[0]

print("Laoding the sample piciure of Sharon Ston \n")

Sample= plt.imread('/content/Facial-Recognition-MMAI844-Tutorial/sample/sharon_stone1.jpg')
plt.imshow(Sample)

#####################################
# load the photo and extract the face
#####################################

extract_face = extract_face_from_file('/content/Facial-Recognition-MMAI844-Tutorial/sample/
sharon_stone1.jpg')
plt.imshow(extract_face)
```

Step 2.3 Check Model Input and output, and model summary

```
from keras_vggface.vggface import VGGFace
# create a vggface2 model
model = VGGFace(model='resnet50')
# summarize input and output shape
print('Inputs: %s' % model.inputs)
print('Outputs: %s' % model.outputs)
```

We can see that the model expects input color images of faces with the shape of 244×244 and the output will be a class prediction of 8,631 people. The input dimension is 4. This means that you have to reshape your training set with .reshape(n_images, 286, 384, 1)

prints out summary of model

The model expects input color images of faces with the shape of 244×244 and the output will be a class prediction of 8,631 people

```
#Printing out summary of model
model.summary()
```

Step 2.4 Prepare input and predict the image using Pre-train model

```
#########################################
# Prepare the input for feeding the model
#########################################

Face_array = asarray(extract_face,'float32')
Preprocess_face = preprocess_input(Face_array)
```

```
print(Preprocess_face.shape)
Preprocess_face_input = Preprocess_face.reshape(1, 224, 224, 3)
print(Preprocess_face_input.shape)
```

Run Prediction and convert prediction into names

```
# perform prediction
yhat = model.predict(Preprocess_face_input)

# convert prediction into names
results = decode_predictions(yhat)
# display most likely results
for result in results[0]:
    print('%s: %.3f%%' % (result[0], result[1]*100))
```

Draw a rectangle on the original image with the prediction label

```
def draw_rectangle_with_label(image, label):
    faces_coord = detector.detectMultiScale(image,
                                        scaleFactor=scale_factor,
                                        minNeighbors=min_neighbors,
                                        minSize=min_size,
                                        flags=flags)
    draw_rectangle(image,faces_coord)
    cv2.putText(image, name,
                            (faces_coord[0][0], faces_coord[0][1]),
                            cv2.FONT_HERSHEY_SIMPLEX , 1, (66, 53, 243), 3)

    plt.imshow(image)
    plt.show()
    for result in results[0]:
        print('%s: %.3f%%' % (result[0], result[1]*100))

draw_rectangle_with_label(Sample, results)
```

Step 2.5 Understand the hidden layer in VGG or CNN

Check the MaxPooling Layer

```
# load the model again
model = VGGFace(model='resnet50')

# redefine model to output right after the third hidden layer
model = Model(inputs=model.inputs, outputs=model.layers[4].output)
model.summary()
# load the image with the required shape
img = extract_face
# convert the image to an array
img = img_to_array(img)
# expand dimensions so that it represents a single 'sample'
img = expand_dims(img, axis=0)
# prepare the image (e.g. scale pixel values for the vgg)
img = preprocess_input(img)
# get feature map for first hidden layer
feature_maps = model.predict(img)
# plot all 4 maps in an 4x4 squares
```

```
square = 4
ix = 1
for _ in range(square):
    for _ in range(square):
        # specify subplot and turn of axis
        ax = pyplot.subplot(square, square, ix)
        ax.set_xticks([])
        ax.set_yticks([])
        # plot filter channel in grayscale
        pyplot.imshow(feature_maps[0, :, :, ix-1], cmap='gray')
        ix += 1
# show the figure
pyplot.show()
```

Check one of the Convolutional Layers

```
# load the model again
model = VGGFace(model='resnet50')

# redefine model to output right after the fifth hidden layer
model = Model(inputs=model.inputs, outputs=model.layers[6].output)
model.summary()
# load the image with the required shape
img = extract_face
# convert the image to an array
img = img_to_array(img)
# expand dimensions so that it represents a single 'sample'
img = expand_dims(img, axis=0)
# prepare the image (e.g. scale pixel values for the vgg)
img = preprocess_input(img)
# get feature map for first hidden layer
feature_maps = model.predict(img)
# plot all 4 maps in an 4x4 squares
square = 4
ix = 1
for _ in range(square):
    for _ in range(square):
        # specify subplot and turn of axis
        ax = pyplot.subplot(square, square, ix)
        ax.set_xticks([])
        ax.set_yticks([])
        # plot filter channel in grayscale
        pyplot.imshow(feature_maps[0, :, :, ix-1], cmap='gray')
        ix += 1
# show the figure
pyplot.show()
```

Wrapping up

I hope that you enjoyed another month of coding! Let’s keep as always connected and don’t forget to explore each month’s articles from the amazing colleagues, learn from the news and understand the scientists behind the brilliant publications and projects by reading our interviews.

VISUM
10 YEARS

Crowne Plaza, Porto, PT
July 10-16, 2022
@visumschool

APPLICATIONS OPEN UNTIL APRIL 14, 2022

DON'T MISS THIS OPPORTUNITY!
apply at visum.inesctec.pt

Basics | Lectures | Challenge | Mentors | Industry

Congrats, Doctor!

Christian Marzahl recently completed his PhD at the Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg. His research is focused on efficient high-quality and high-quantity dataset annotation and deep learning-based object detection on whole slide images. Furthermore, he conducted studies to analyse how algorithms and experts can efficiently work together and developed the open-source online annotation tool EXACT which is already applied in industry, research and academia. Congrats, Doctor Christian!

Christian would like to thank his advisors and collaborators Andreas Maier, Marc Aubreville, Christof Bertram and Katharina Breininger for their support during this PhD.

Bridging the inter-species gab:

Throughout history, humans have learned to treat human diseases by studying animal conditions. I have explored whether the same principle can be applied to deep learning. The generalised applicability of deep learning models between species could offer enormous scientific and economic value, especially for domains that lack appropriate training data due to privacy restrictions, data protection or the disease's rarity in certain species.

For this purpose, species-independent cell detection methods were developed and analysed in my dissertation, allowing the automatic quantification of pulmonary haemorrhage [1] and asthma [2] on cytological image data. The basis for developing these methods is qualitatively and quantitatively comprehensive data sets created online and interdisciplinary with the annotation tool EXACT I developed in my thesis.

To support research in inter-domain fields, I co-created and published the first fully annotated multi-species cytopathological pulmonary haemorrhage dataset [3]. This was made possible by working with an interdisciplinary team to efficiently create this novel data set by combining expert-algorithm collaboration and EXACT in a multi-step pipeline. Initially, an object detection model pre-trained on equine cytology whole slide images was applied to

human and feline samples. Afterwards, multiple clustering and manually screening steps were incorporated into the annotations to increase the overall dataset quality (Figure 1). My research showed that, for pulmonary haemorrhage detection, we were able to bridge the domain gap between equine and human samples [4].

Supporting pathologists with deep learning-based methods: Digitisation forms the basis for supporting the work of pathologists by utilising computer-assisted methods. Modern machine learning methods can accelerate previously time-consuming quantitative analyses, support reporting, and, therefore, standardise and improve the results of pathological examinations. Hence, I developed novel methods and software solutions that can be used to support both the analysis of image data and the cooperation between users. Figure 2 presents the results of the developed deep learning-based object detection method to quantify pulmonary haemorrhage on whole slide images by classifying macrophages into five corresponding grades, represented in the figure by bounding boxes with unique colours. Furthermore, my work indicated that experts are biased towards accepting precomputed annotations, which has important implications for using AI in clinical routines in the future [5,6]. Finally, to support reproducible scientific research, all code developed during my Ph.D. and the datasets used are publicly available on GitHub [7].

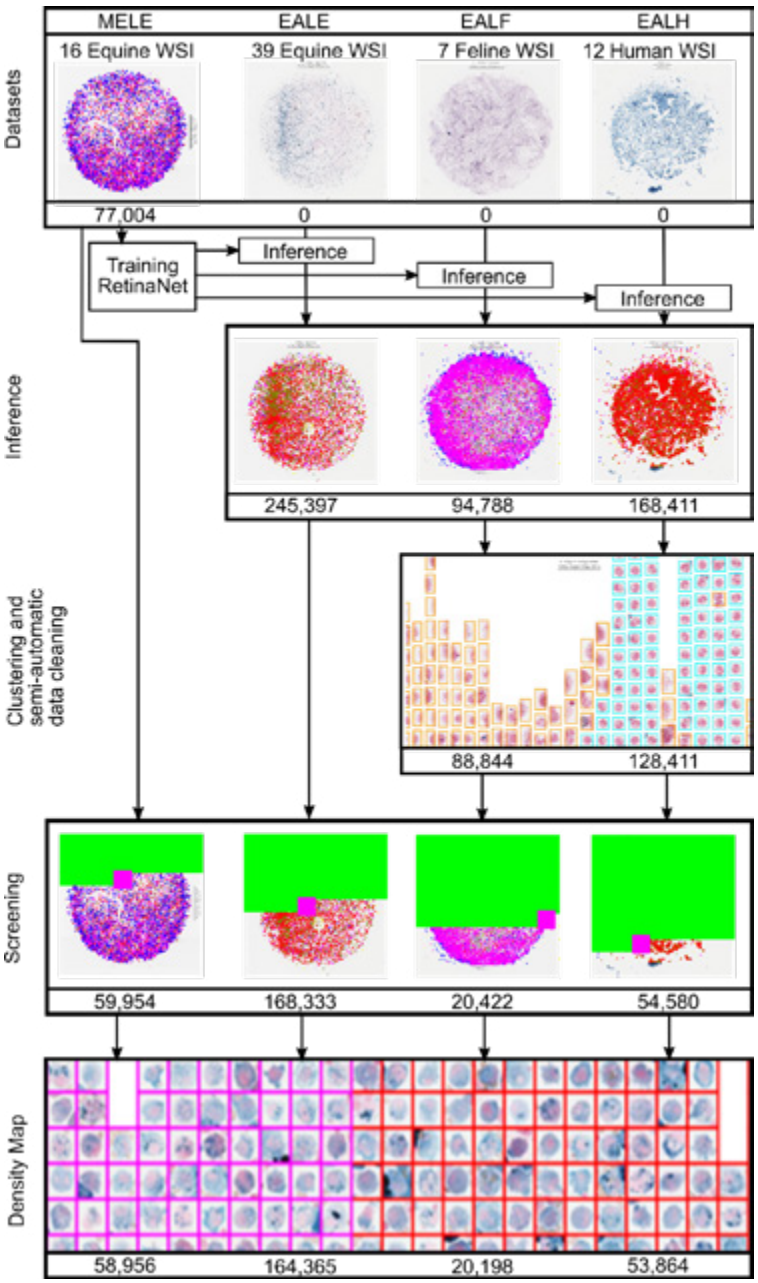


Figure 1. Visualisation of the developed pipeline to create high-quality multi-species datasets [4]

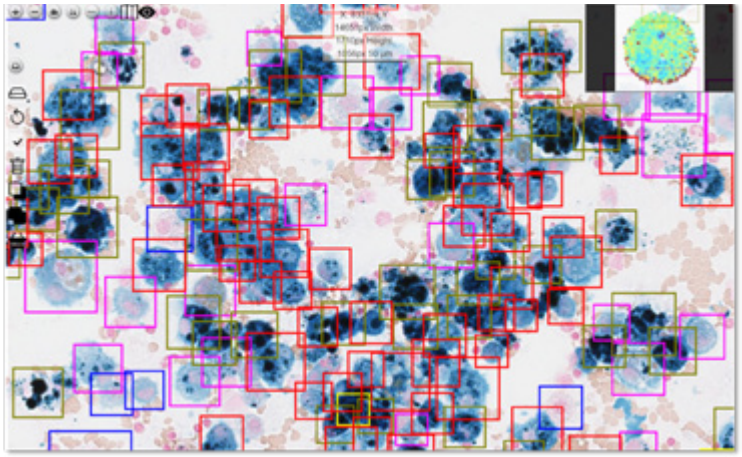


Figure 2. AI object detection and annotations of pulmonary haemorrhage on whole slide images.

CLAIRE VERNADE HAS BEEN A RESEARCH SCIENTIST AT DEEPMIND IN LONDON, UK SINCE 2018. SHE RECEIVED HER PHD FROM TELECOM PARISTECH IN OCTOBER 2017, UNDER THE GUIDANCE OF PROF. OLIVIER CAPPÉ.

Claire, can you tell us about your work?

I work on bandit algorithms and, in general, sequential learning. It's a niche part of the broader field of reinforcement learning and machine learning, in which we are concerned with theoretical guarantees and theoretical approaches to reinforcement learning. In particular, with bandit algorithms, we are specifically interested in the question of exploration versus exploitation trade off, trying to understand how to explore options while still trying to maximize rewards. This is a central question in reinforcement learning, which is better addressed when you can clean the problem up and try to address easier models.

Some say that maximizing rewards on reinforcement learning may lead to problematic consequences. Think at an autonomous vehicle trying to maximize the reward to seek the safest route. It doesn't mean the safest way for this vehicle will lead to the safest outcome for other vehicles. Maximizing reward could be challenging. Would you agree?

Yes, absolutely - in the case of autonomous driving. In many cases, in real-world problems, there are constraints. You can't just maximize rewards. You have to take into account other constraints, other typical aspects of your rewards. In my case, for instance, I looked into non-stationary environments. I still maximize rewards, but I focus on the situation where the environment is not always giving us the same reward. Sometimes one action does the best, and then two hours later, it will be another action. I want to track the non-stationary, this kind of volatility of the environment. In the case of autonomous driving, there is a safety constraint. Sometimes there are also diversity constraints, and you don't want to always take the same actions. You also want to maximize diversity in your recommendations or in your propositions. Integrating new constraints modifies the machine learning problem. My job, or the one of my team here at DeepMind, is to try and boil it down to the simplest problem, where we can actually say something that is theoretically valid, and perhaps bring light to problems in more complex environments. We cannot solve autonomous driving altogether, by looking at the entire problem. Sometimes it's good to make blocks. To solve autonomous driving, you need to first split it into many different problems. Our job is to create the building blocks of that tower.

Tell us about practical applications in real life with non-stationary situations.

The practical applications that you may think of are typically recommending content on platforms. That could be books on Google or audiobooks on Audible. It can be routes, if you want to try to find the route to your work. The best route is

I want to track the non-stationary, this kind of volatility of the environment.

not always the same. Sometimes there is traffic. Sometimes there is a trash truck in one street, so you want to go to another street. The best option is changing with time. The idea is to detect that and adapt. There are real applications to non-stationary time series or non-stationary environments. Personally, I don't directly work on one application in particular. I'm giving practitioners tools to think about if they should address their problem this way, or that way. It depends on the conditions that I can formulate in my papers. But I don't go and solve one particular problem.

Does this mean that you enjoy solving a problem more than seeing your work translated into real-world applications?

There are two things. There is the eureka moment, but I would say that this is the last part. The first part that I enjoy a lot is looking at real-world problems and asking the question: How can I write it in math? How can I translate it into a model that I can solve? I know that we have the tools in statistics, in mathematics, and in learning theory to address this. I've been interacting with people from YouTube and with people from Google Books via DeepMind applied research: these people have concrete product constraints. They come to me and say, "*We have this problem. We want this, and we want that. We have this constraint and that constraint*." I'm like, "*Okay, we need to sit down and write down what these things mean in math. Then we can write down the equation and find an algorithm that is going to work.*" I really enjoy this modeling part. This is what I have been working on since the beginning of my PhD. I was working with Criteo and with Peugeot in France to help them formulate their problem. Then I came up with a theoretical problem, which is kind of a toy problem but models the situation. Then the eureka moment comes at the end when we try to prove that our algorithm actually makes sense and works. Passing this knowledge into application takes much more time.

Do you ever get out of the office and see something that does not work as it should and think, "*How do I translate this into a*

math problem that I can solve?"

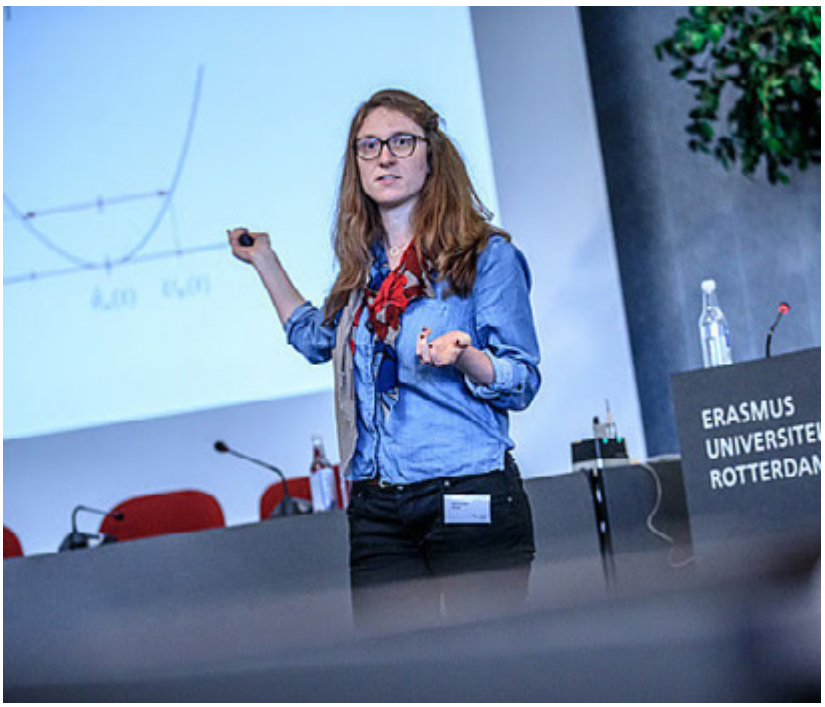
Ha! I'd say yes, it happens to me, especially during Covid. You had all these vaccine slot booking services where you had to go on the platform and book your slot. Then it will tell you where to go, but you realize that this is not the place you wanted to go. Then you go back and the appointment that you wanted is no longer available.

It's because you are competing with other internet users who are trying to book the same thing at the same time. If you hesitate, the slot is no longer available!

It's a typical computer science problem to have this queuing system and memory allocation trying to optimize peoples' path for a several-stage system. That's typical computer science. There are lots of results in scheduling tasks. It sounds like common sense, but it's just that formulating and teaching this theory to people allow us to have more efficient systems since the first run. The problem with Covid was that people didn't have time! I guess, in the long run, we would have the knowledge to improve the systems, but the challenge is to have the system work as quickly as possible. These are super interesting problems. There are so many!

How does your work answer your drive to solve problems?

DeepMind has a very long-term strategy. My team looks at very basic theoretical problems. I mean basic in a good way; fundamental. Our team is called Foundations, so it's really foundational, fundamental problems. They really give us the time we need, and they value our



I focus on the situation where the environment is not always giving us the same reward

work. We try, as much as possible, to have a lot of visibility of other teams. When they come, they ask us questions. They have challenges in their work, and we can help. On our side, we feel no pressure to have strict deadlines to deliver products. This is working super well here.

You live in London working for DeepMind. Can you tell us where you come from?

I am French. I grew up in France, I went through the classic French path at the Grandes Écoles and I studied engineering at Telecom ParisTech. I did everything in France, from the beginning until the end of my PhD, except for one short internship at Adobe in California with Branislav Kveton, which was a really great time. I learned a lot about the interactions between industry and research. I had a roughly one-year, part-time Postdoc at Amazon in Berlin. After this

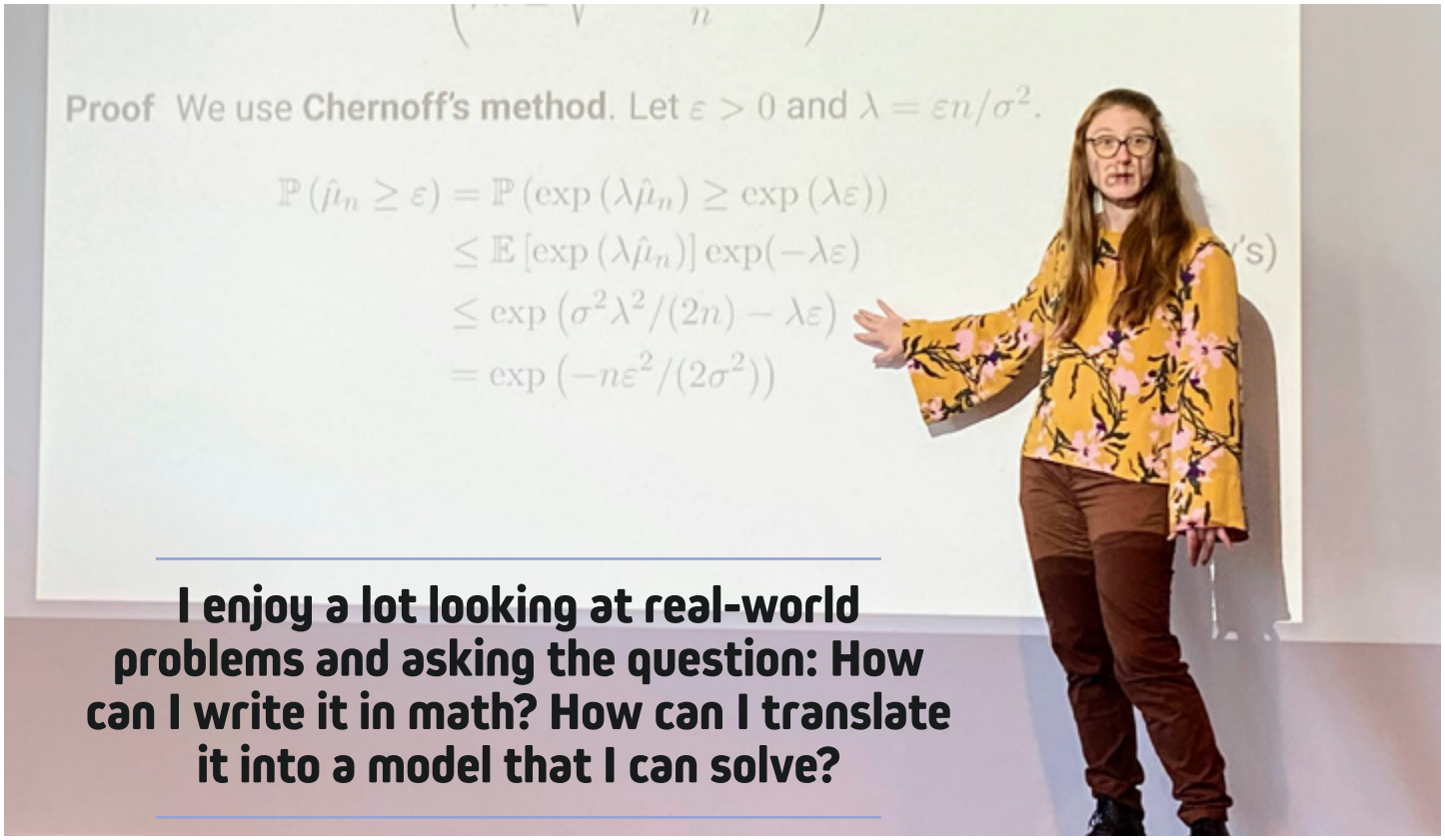
time, I got the position at DeepMind, and I moved to London about three years ago. This feels so short after the last two years.

Tell us something about DeepMind that we don't know.

There are lots of things that people don't know. The food is very good! [*laughs*] That's not surprising... One thing I've realized at DeepMind: internally, there is a community keen to exchange their passions and interests. We have an internal Slack, which is not common. Google doesn't have Slack, I think. At DeepMind, we do and there is a channel about everything! In particular, I learned a lot about gender studies and feminism. We have a very lively community of people who know about these things and are happy to debate and exchange. I read a lot of interesting discussions on the Slack channels, and how gender studies and feminism impact us, as a company that tries to do good. We are like a philosophical discussion group, kind of independent of the corporate structure. It's a self-organized group. This is for Gen Fem.

There are many more self-organized groups that freely discuss important philosophical or political topics that affect work. We are at the center of attention as a technology company. It's also important that they have these discussions about social topics. It is refreshing to see people sit down and have proper discussions about topics that are so hard and so important.

You said it's easy to share your passions and interests in that setting. Can you share one passion or interest of yours that we would not suspect?



I enjoy a lot looking at real-world problems and asking the question: How can I write it in math? How can I translate it into a model that I can solve?

Over the past two years, I've become very keen to learn about the economy. It's in the news, and everyone talks about it. The currencies and everything are so technical. There is so much to learn about the economy.

What would you change in economics?

The question is very broad. It's very political. People should never forget that everything is political. This is one thing that I've seen looking at economic theories. Sometimes they tend to forget that some of the decisions or models affect peoples' lives. These are decisions that, in democracies, we have to make together.

Do you have a final message to share with our community?

Wow! There has been a lot of debate lately in machine learning, very broadly, about the quality of evaluation, the validity of theoretical results, even just the validity of empirical results. I find that people don't dare ask too many theoretical questions because they sound like fundamental theory questions in machine learning, that people should not ask about. In fact, this is a very difficult learning theory: this is a whole part of machine learning. There are people working on it full time and it's difficult. We notice at DeepMind that opening a Slack channel called *"Ask Foundations"* has had so much success! People come and ask very basic questions. I want to say that people should feel comfortable asking questions even when they seem basic, even fundamental questions that sound like graduate school questions. Very often they aren't. Let's exchange! Let's ask questions!

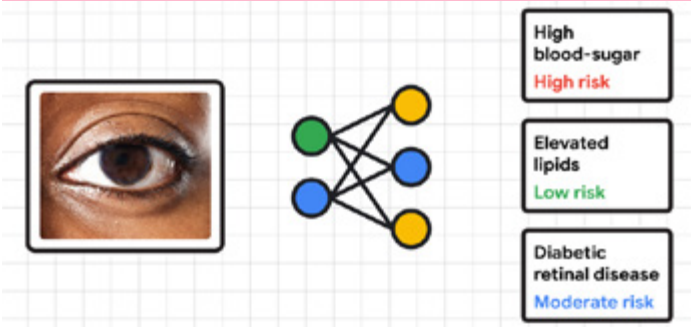
Computer Vision News has found great new stories, written somewhere else by somebody else. We share them with you, adding a short comment. **Enjoy!**

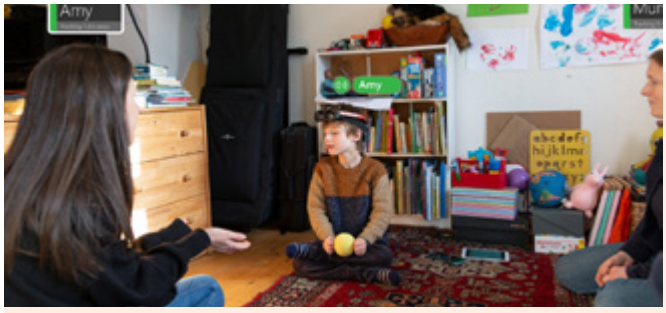
NVIDIA Research Turns 2D Photos Into 3D Scenes in the Blink of an AI

We have already told you about **NeRF ("neural radiance fields")**, an award-winning pioneering technique launched a couple of years ago at **ECCV** to map the color and light of 2D photos: you read about it first here and here. The challenge is to turn a collection of still images into a digital 3D scene in a matter of seconds. The NVIDIA Research team has developed an approach that almost instantly reconstructs a 3D scene from a handful of 2D images taken at different angles. The result is impressive **NVIDIA Instant NeRF**, that increases speed, ease and reach of 3D capture and sharing. **Watch the Video**

Google - Recording and Translating Heart Sounds with Smartphones

In different news, **Greg Corrado - Head of Health AI at Google** - has published a new blog relating their latest **health AI developments**. Close to 100,000 patients have been screened for **diabetic retinopathy** in project ARDA: in addition to eye disease, these fundus images can reveal cardiovascular risk factors, such as high blood sugar and cholesterol levels, with assistance from **deep learning** and using existing tabletop cameras in clinics. Moreover, smartphone's built-in microphones can record heart sounds when placed over the chest to help clinicians detect heart valve disorders, such as aortic stenosis. Read More



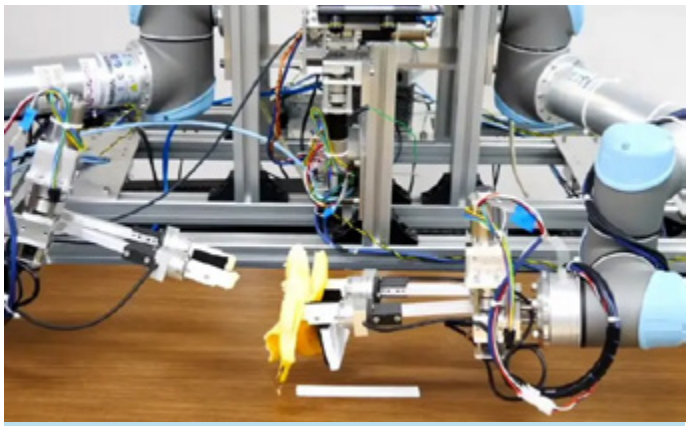


PeopleLens: Using AI to Support Social Interaction Between Children Who Are Blind and Their Peers

Microsoft researchers have started **PeopleLens**, a new research technology that helps young people who are blind and their peers interact more easily. It's an open-ended AI system that uses a head-mounted augmented reality device in combination with **four state-of-the-art computer vision algorithms to continuously locate, identify, track, and capture the gaze directions** of people in the immediate social surroundings. It creates a map of people around the user and reads out the names of people that the child looks at in spatialized audio to give him a sense of the respective positions and distances of the people around. **Watch the Video**

insideBIGDATA: The Decade of Synthetic Data is Underway

The folks at **insideBIGDATA** publish a nice tribune (signed by Datagen's CTO), where it is declared that the 2020s will be remembered as the **"Decade of Data"** for AI, and - even more - the **"Decade of Synthetic Data"**. As we have ourselves explained a short time ago, deep learning-based systems require sufficient data for proper training and reliable testing. A solution is to **generate data synthetically**: the nice thing is that it works, and it has gained widespread acceptance across the research and enterprise communities. Probably because it is much simpler/faster than spending hundreds of hours on fine-tuning AI algorithms and models. **Read More**



Watch a Robot Peel a Banana Without Crushing It into Oblivion

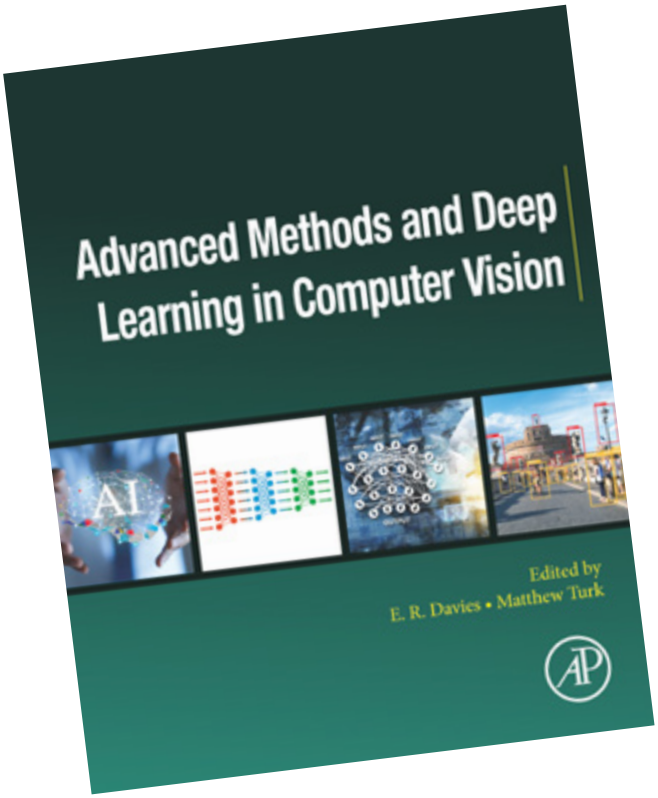
Handling eggs and soft fruit is tricky for robots, but apparently, after training with hours of data from a human operator, a **machine learning algorithm** developed the motor skills needed to correctly peel a banana. Not always, but most of the times and in less than 3 minutes. The effort was done by **Heecheol Kim** at the **University of Tokyo**: the machine-learning system that he and his colleagues have developed powers a robot, which has two arms and hands that grasp between two "fingers". 13 hours of **deep-imitation learning** were enough to train the robot.

Watch the Video

Building a Hybrid AI Able to Beat the World's Best Bridge Player

For the son of **two terrific bridge players** (and a very poor player myself), this is quite special: a French private AI lab called **Nukkai** has been working on an AI that may just be able to **beat the world's best bridge player**, something that has already been made possible in other games like Chess and Go. This is no simple matter: in chess and go, competitors play **with complete information and must react to the behavior of a single opponent** at a time, while everyone is in possession of all the information. In bridge, the opposite is true: multiple opponents and only partial information - a scenario much closer to human decision-making. **Watch the Video**

Computer Vision Book



Roy Davies and Matthew Turk are speaking to us about a new book they have co-edited: *Advanced Methods and Deep Learning in Computer Vision*. Roy Davies is Emeritus Professor of Computer Vision at Royal Holloway, University of London, UK. Matthew Turk is the president of the Toyota Technological Institute at Chicago (TTIC), an independent philanthropically endowed academic computer science institute.

What can you tell us about your new book?

MT: The book covers advanced computer vision methods, emphasizing machine and deep learning techniques applied to computer vision and imaging, particularly those that have become more frequent, common, interesting, and essential in the last five to ten years. There's been a tremendous amount of progress in the field, and we felt it was the right time to have a book that was neither an introductory text nor a collection of very detailed research, but something in the middle that provides some depth across several important topics.

RD: Indeed, we thought that it was important to emphasise principles and new methodologies rather than dwell excessively on details of ongoing research: in that way we should be able to help readers obtain a significantly deeper and longer term understanding of the subject.

Who is the target audience for the book?

The Editors: The book is aimed at people studying computer vision – graduates and undergraduates – and current researchers working in the field who want to look in more detail at an area they haven't focused on much. Practitioners who wish to gain a better insight into specific areas of computer vision will find it helpful too.

If a newcomer to the field has your book in their hand, what advice would you give them to best take advantage of it?

MT: A great place to start is with the first chapter, written by Roy, which is an overview of the basic concepts of computer vision and sets the stage for the rest. After that, stand back and look at all the other topics. Decide which ones are most interesting to you. You may end up with two or three chapters. For each, start with the introductory sections and if you need more background on the topic follow the references that those bring up.

Roy Davies

Matthew Turk

That will give you an understanding of the bigger picture, and then you can delve bit by bit into the details and explore those references in depth if you want to go even further.

RD: While following this overall plan, it is also relevant to think carefully not only of the approach to be adopted but also how the input data (often in the form of many millions of images or image patches) should optimally be managed so as to most effectively train the final system.

My aim in writing the rather long first chapter was to ensure first that readers were brought up to speed on legacy computer vision work; second, to open the doors to deep learning and to show how it can successfully be applied to computer vision; and third, to demonstrate that even when applied to the familiar rather basic subject of texture analysis that substantial changes are needed in the old ways of thinking.

Am I right to assume that seasoned scholars might have an advantage in perusing the book?

The Editors: Yes, absolutely. For any given chapter, only a handful of experts spend their entire time focused on the topic, but everyone in the field can learn something. If you're an advanced person in the field but not necessarily an expert and want to learn more about a specific topic, you will benefit from this book.

Is it possible that you learned things in the process of writing this book?

MT: For sure. I was an editor, I didn't write any of the chapters, but I spent a lot of time reading very carefully, providing feedback to authors, asking questions, and thinking through things myself to make sure I was understanding. Any author or editor learns a lot in the process, and I certainly did.

RD: As an editor, I found it stimulating finding how best to guide eminent authors

to produce chapters that were even better than their initial drafts – and of course more didactic and beneficial for the eventual readers!

What was it like working together?

MT: Believe it or not, we still haven't met face to face, but we've been able to meet online. Roy is a very organized person and easy to interact with. We ended up complementing each other well in terms of our styles and focus. I didn't know **his prior textbooks** in detail before this, so it allowed me to get to know them. His most recent introductory computer vision textbook is an excellent book to have on your shelf and use for classes.

RD: I found it a remarkable experience. As Matthew says, we had not met in advance of working together, but it soon became clear that our different past experience usefully led to different global and local slants on the writing. I should also remark that I was overwhelmed that this was the same Matthew Turk who had long ago invented the impressive eigenface approach to face recognition!

With computer vision changing so much in the last few years, how do we balance the fundamentals of the field with the things that are most important now?

The Editors: That's a question people are still struggling to answer. Knowing the fundamentals of computer vision and image processing has been crucial in making all this tremendous progress, and many people are worried we'll have a whole new generation coming along who don't know the basics. Part of the answer is that the field will probably start to break up more than it has in the past. It's hard for one person to be an expert in everything. Some will focus more on image processing and low-level computer vision, while others will focus more on machine learning and other aspects of computer vision. There will always be overlap, but maybe a little more intentional separation than a single computer vision track of education.

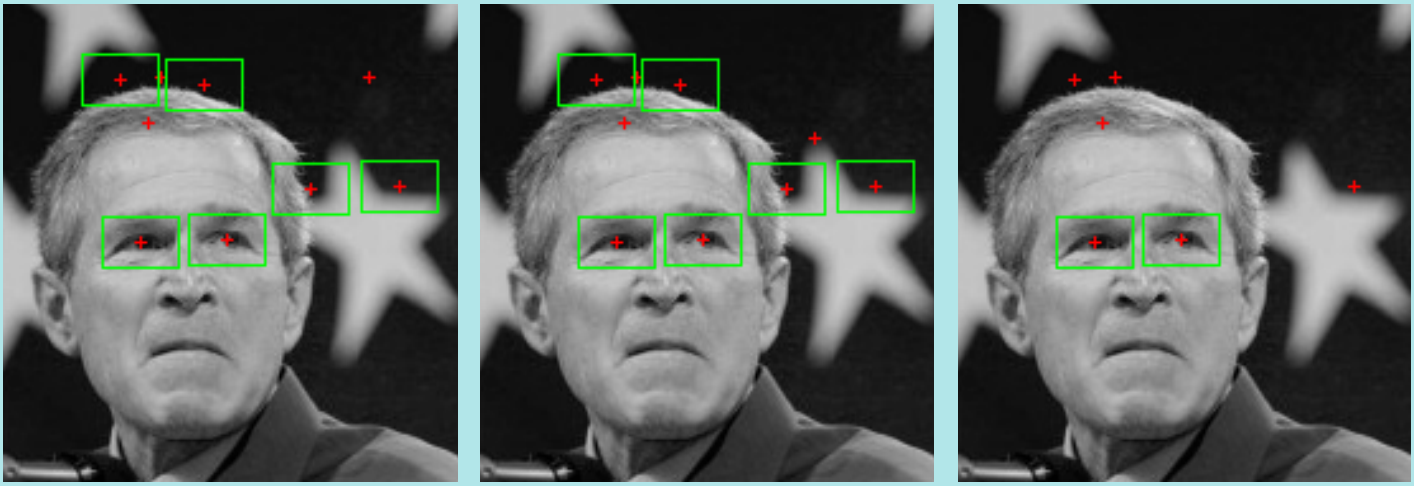
Of all the novelties and innovations you have seen in the last few years, can you pick the one you think experts and scholars should be most careful not to neglect?

MT: It would be hard to choose just one, but something that strikes me as important is the general limitations of what we know about machine learning. Machine learning has made incredible progress, but much of it has been pragmatic in the sense that people have tried many things, and some have been successful, but there's a lot we don't know fundamentally or theoretically about limits. For example, how much data is needed for a given problem? How do we know whether there is enough variety in the data? There is helpful best practice and guidance, but still a lot of missing parts.

The whole subfield of adversarial learning is an example where people have found situations that don't work so well. People are exploring that field to get a pragmatic view of these limitations, but we need to know theoretically and fundamentally what the limitations and possibilities are and the appropriate guidance.

RD: I think Matthew is absolutely right about the need to be able to theoretically ascertain the amount and variety of data needed in any application. Indeed, that is the 1-million-dollar question that previously put paid to the old neural networks of the 1980s.

Stages in the segmentation of moving objects: current frame, running median, result of background subtraction followed by morphological processing. The final result shows that higher level processing is still needed to disentangle the shadows. Note the intrinsic noise problems involved with such approaches.



This figure shows the highly variable outputs of a multi-parameter algorithm for detecting pairs of eye features. In fact, because of the huge variations in the Bush LFW dataset, the algorithm finds correct pairs of eyes in only about 50% of the first 30 database images. A more satisfactory eye detector would need to use a substantial number of highly trained neural detection masks.

Location of white road markings using RANSAC. Note that the three sets of red lines do not converge perfectly to a single spot on the horizon, because the white markings are slightly curved and candidate lines were taken to require a minimum of 6 points.

Matthew Turk presenting at TTIC

With the community having been separated for a couple of years, some young scientists have never attended an in-person conference. What would you say to them?

MT: You're raising a vital point. When I was a graduate student and a young professional in the research community, interacting with my peers face to face, going out for dinner at a conference, chatting late into the night, and meeting famous people in the field were all important to me. Many of those early relationships I built are still going strong today. I have worried about missing out on that for the last couple of years.

On the other hand, there are positives. One of them is that many people have attended conferences and meetings that they might not have been able to attend otherwise. It's all been virtual, but they can hear presentations, ask questions, and get answers they just wouldn't have been able to years ago. Also, senior people have been able to give more invited talks because they only need to commit an hour or two rather than a few days. People have been creating more opportunities – albeit online, which misses some important aspects. When we start to have in-person conferences again, we need to make sure we're thoughtful and conscious about building those interactions we haven't been able to do so well over the past couple of years – not just letting them happen if they do but being thoughtful and intentional about them.

RD: Matthew's answer doesn't just reflect a need felt by graduate students: I would say that it is one that exists at all levels, from senior staff downwards. In fact, talking directly to people is a *human* need that in my experience exists down to primary

Rama Chellappa at BMVC 2019 in Newcastle, UK. He and his co-authors contributed Chapter 7 of the book.

school level, and its resolution must not simply be left to chance.

Thank you for a fascinating interview, Matthew and Roy. Do you have a final message for our readers and the wider community?

MT: This field is moving very quickly. When you look back over the last decade, there's been a massive amount of progress and a significant change in focus. Still, these fundamental pedagogical textbooks, like Roy's book and Richard Szeliski's book, are critical to the field. Having that broad background of what has built this field over the years is still important. Our book and other books like it, which go into a handful of topics in some depth and breadth, will be relevant for a long time to come. The field may be moving quickly, but that doesn't mean people should focus solely on the latest conference papers. Those are important, but collections like this book or the basics of the introductory textbooks are still vital and relevant for our field.

RD: It is interesting that the field changed so radically around 2012, and this raises the question of whether advances will now once again proceed at a much steadier pace. In fact, some chapters in this book seem to indicate not: in particular, chapters 12, 13 and 15 raise the relevance of cognitive theories, self-aware systems and 'adversarial' methods. Readers would do well to keep such new directions in mind and maybe wonder whether further developments in these directions will take us into totally new worlds that reflect the parts of our brains that are not so narrowly focussed on vision: after all, vision is only one aspect of our cognitive evolution.

COMPUTER VISION EVENTS

ICLR 2022
Virtual
25-29 April

World Summit AI Americas 2022
Montréal, Canada and online
4-5 May

TechEx North America
S.Clara, CA
11-12 May

Robotics and AI 2022
Prague, Czech Republic
13-14 May

Int. Conf. and Expo on Robotics and AI
London, UK
16-18 May

Embedded Vision Summit
Santa Clara, CA
17-19 May

Image Analysis and Processing ICIAP
Lecce, Italy
23-27 May

CARS 2022
Tokyo, Japan
7–11 June

CAOS 2022
Brest, France
8–11 June

CVPR 2022
MEET US THERE
New Orleans, LA
19–24 June

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

APRIL 2022

CHALLENGES IN VIDEO FOR ROBOTIC ASSISTED SURGERIES

Page 40


VISION
SURGERY
AI

Karim Tamir is the President, CEO, and co-founder of Vision Surgery AI, a software and technology company using the power of computer vision and deep learning to track movement and objects in the operating room. He speaks to us about its work and how it plans to have an impact on healthcare and robotics to ultimately improve patient outcomes.

Vision Surgery AI has developed a surgical AI platform that uses computer vision and deep learning techniques, including image and video data processing, to track landmarks and key points for human pose estimation and perform object detection and classification for instrument recognition.

Karim tells us the focus is not on patient-specific risk or any direct assessment of surgeons but instead on gathering surgical data to explore risk areas related to the surgical environment. High-performance cameras installed in the operating theatre capture as much data as possible, which is used to detect anomalies in movement and behavior and analyzed to provide insights into how successful the surgery has been.

The data is completely anonymized, meaning the team can avoid compliance issues around healthcare data.

"Nobody will give you data unless you can demonstrate they are anonymized, they are not going to be sold, and will not be misused," Karim explains.

"We wanted to avoid needing doctor sign-off, as we'd get nowhere fast. It's like the autopilot in a self-driving car from Tesla or Google. The program tracks as much data as possible so the algorithm can identify where to go, not go, and how to avoid an accident. If these companies asked people on the streets or taxi drivers, they wouldn't get anywhere. Just gather as much data as you can! AI and machine learning can give you a much better assessment of it. We have the tools and the science to do it, so let's use them."

Vision Surgery AI is working with **MIT**, the **University of Munich**, and **Charité Berlin**, one of the biggest university hospitals in Europe, and is looking to collaborate with others worldwide. Together, Karim says

The Team

they are one of the first groups to be using computer vision to look at anomalies in big data.

The team has several target groups for this information. **Insurance companies** selling hedging products to hospitals and surgeons can use it to **understand the risks in the operating room**. **Hospitals** will have an idea of the correlation between anomalies and errors in the operating theatre to support them in **mitigating risk, increasing efficiency, and reducing costs**. **Universities** and future clinicians will be able to use it to **reproduce surgical scenes for their work**. And **industry** can use the anonymized data available to **build better robotics**.

"It's a massive piece of work with a huge amount of surgical data to analyze," Karim tells us.

"We're using deep learning and convolutional neural networks to develop an algorithm that can detect what went well and what didn't go so well. The goal is to have an algorithm that provides a high degree of accuracy. We'll compare it to what happens in surgery so that the correlation is 99%. Then we'll have reached the level where we can say our

Karim Tamir

algorithm predicts what could happen in the operating room. Charité Berlin will be the first university hospital in Europe to allow us to use them as a pilot."

Thinking about what might go wrong, Karim reflects on the analogy of the autopilot driving the car and the fact that insurance companies are unwilling to insure those cars because they do not yet achieve 100% accuracy.

"In our case, we're not going to have a robot that will perform a surgery just yet – it will take a few years," he points out.

"But the information we're gathering will be a guidance tool for surgeons as to what can contribute to failure in surgery. Drivers don't know how likely they are to have an accident and what will cause it. Hundreds of things can impact a driver's behavior and lead to an accident. We're not talking about linear models. It's the same for surgery. Human beings alone cannot isolate the risks that lead to failure."

If this article has piqued your interest in the company, you might be pleased to hear that Vision Surgery AI is hiring! Currently, it has eight staff but is looking to be 20-30 people by the end of the year.

CHALLENGES IN VIDEO FOR ROBOTIC ASSISTED SURGERIES

by Oren Wintner, RSIP Vision

Asher Patinkin

"Our long-term goal is autonomous surgery. Robots can be very accurate, they don't get tired, and they can adapt quickly when necessary. We are still far from autonomous surgery, so we start with smaller steps," says **Asher Patinkin**, experienced algorithm developer at RSIP Vision.

The imaging modalities used during **Robotic Assisted Surgeries (RAS)** are numerous, but perhaps the most informative one is the video feed. Surgical videos can be obtained from laparoscopic or operation-room (OR) cameras, and they are the "eyes" of the surgeon.

Keeping in mind the end-goal, the first step is to implement surgical phase recognition by applying **artificial intelligence (AI) and computer vision (CV) methods**.

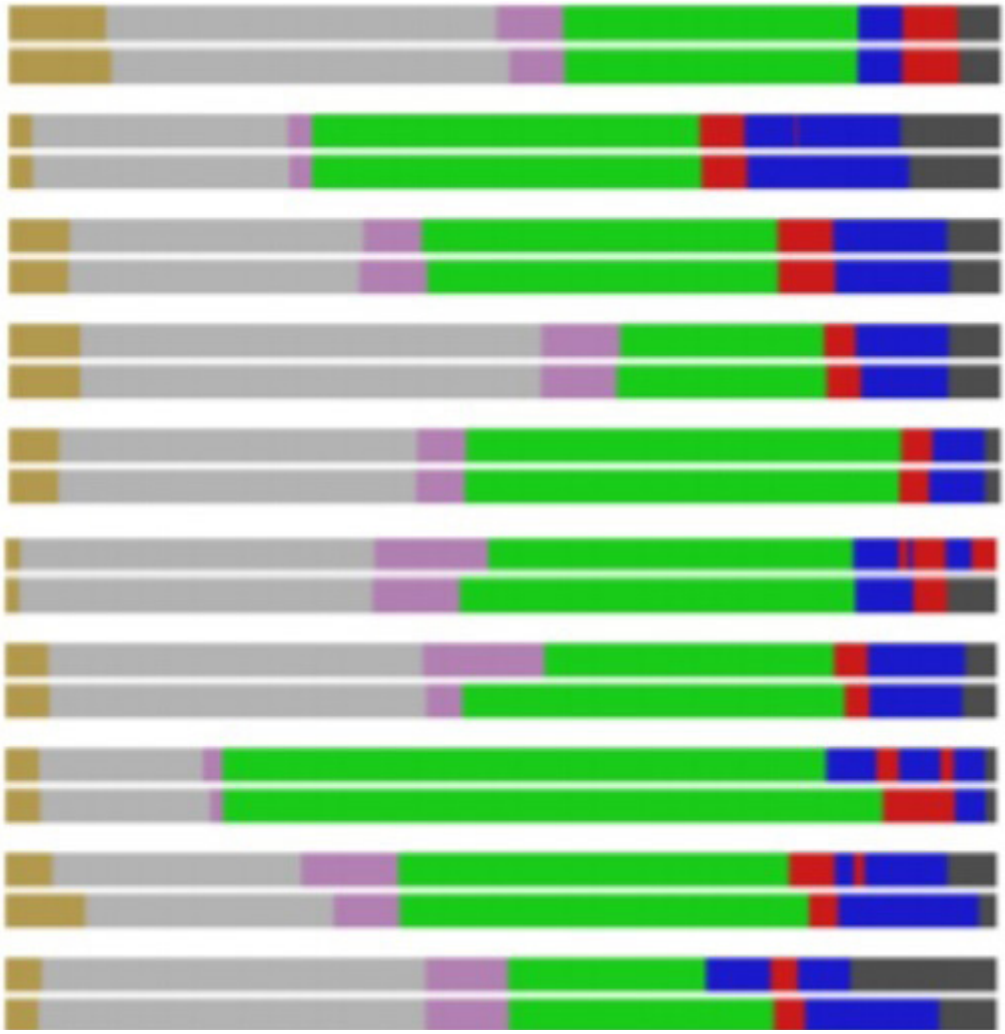
It is essential to detect at every time-point what procedural step is currently conducted to understand what tools are needed, what risks are relevant, and to alert the staff. Initially the team attempted using **convolutional neural networks (CNN)** to classify the procedural phase by a single frame. This method gave decent results in some of the frames. For example, it is relatively easy to understand that suturing occurs when there is a needle in the frame. However, when the needle is not visible in the frame, the CNN will be misled to classify the frame to be part of a non-suturing phase. Furthermore, some phases (e.g. exploration phase), *"can be detected only from a series of frames, as*

the phase is not associated with a single tool," Asher explains.

Therefore, more advanced **deep learning (DL) techniques** were applied. A combination of convolutional neural networks (CNN) with **long-short term memory (LSTM) networks** which take into account the time dimension of the video. Results were significantly improved and the final algorithm can detect the different phases of the surgery successfully.

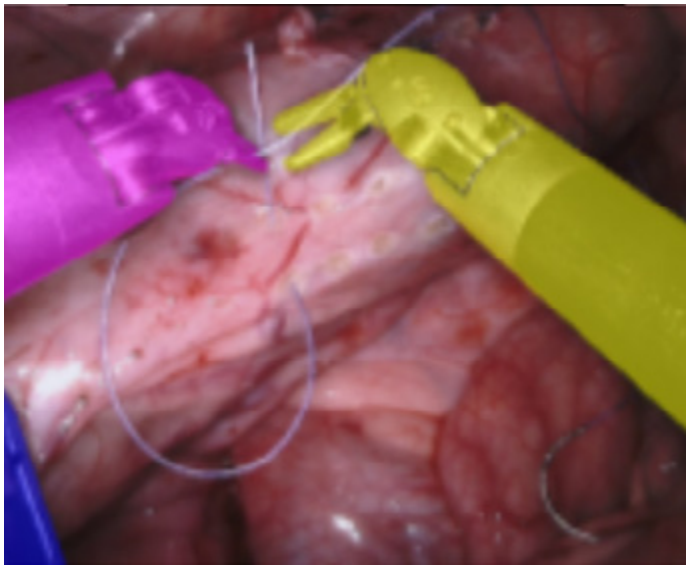
The next step towards autonomous procedures requires **tool identification**. As the tools are remarkably different from their background, this task is quite simple. Detecting and segmenting tools is feasible, but recognizing a specific tool requires previous knowledge, so the system must be trained on specific tools to allow that.

Once the tools are recognized, their position needs to be calculated. This is essential when attempting to conduct tasks like suturing - the position of the suturing needle must be calculated with high accuracy. The best method to obtain absolute coordinates from a video feed is by using **stereo imaging - two cameras recording simultaneously from different**



Deep Learning result paired with Ground Truth (top vs bottom)

Tool Segmentation

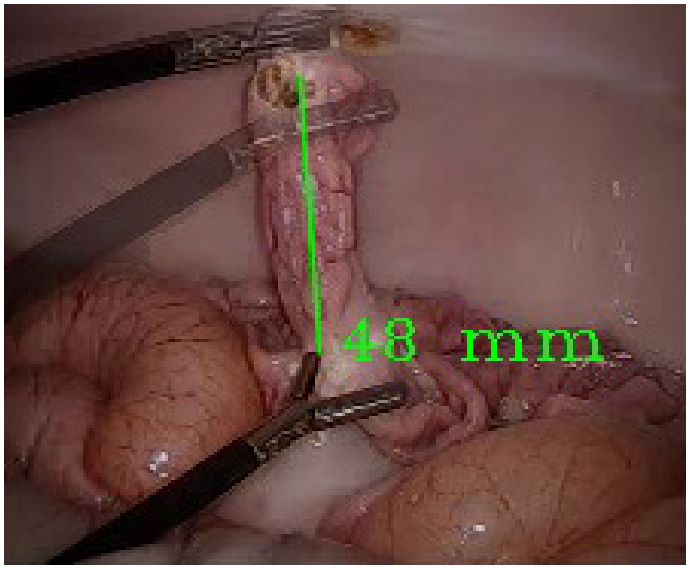


Segmentation and Tool Classification

angles. Most robotic systems today come with stereo imaging so obtaining this data should not be problematic. *"We were able to calculate position even from monocular imaging, however, stereo depth is much more accurate than monocular depth,"* Asher clarifies.

By segmenting the tools in each image and calculating the disparity (=difference of the location in the 2 frames) of the segmentation, the 3D position of the tool was reconstructed.

Generalizing this method to all the pixels in the image for **complete scene 3D reconstruction** is difficult, since the anatomy, unlike the surgery tools, is very monotonous and lacks distinguishable key-point features. Therefore, deep learning architectures - that separately compute features for each of the frames and calculate the disparity of each pixel in the frame - were implemented. This in turn

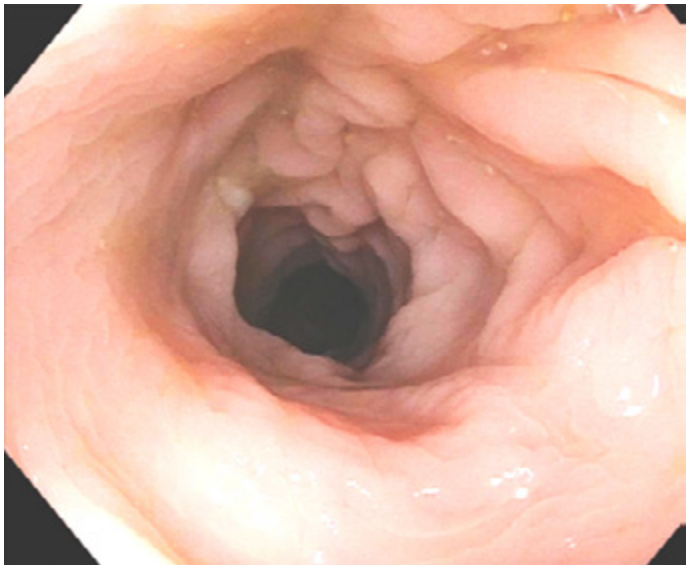


can be used to reconstruct an accurate 3D map of the full scene.

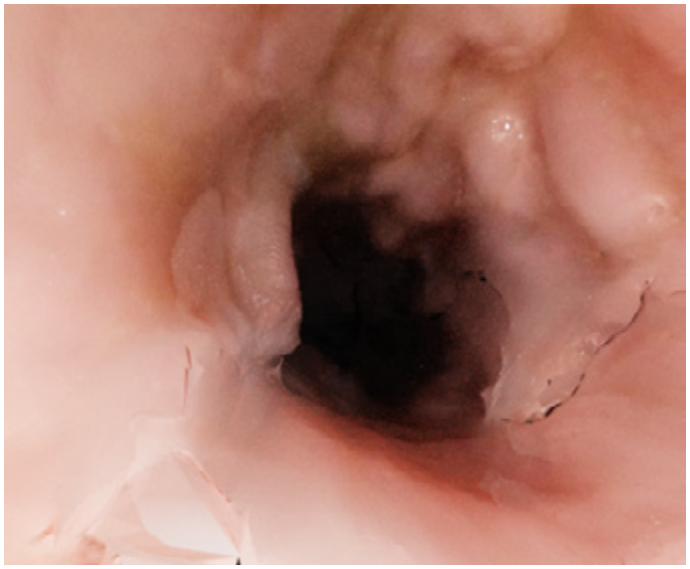
Obtaining videos of surgical scenes paired with accurate ground-truth (GT) is time-consuming and expensive, so typically only a small amount is available. In order to train the CNN we need to supplement the ground-truth data using **creative data augmentation methods, as well as synthetically generated data**.

Another significant feature which assists in RAS, is the ability to perform real-time measurements on the image from the video feed. Prior knowledge of camera calibration and tool dimensions is used to address this need. Using deep learning, key points on the tools are detected, and the image is calibrated based on the prior knowledge. This allows accurate 3D measurements within the field of view, even without stereo vision.

Despite the advancements in this field, there is still a long way to go. **Different surgeries present different challenges for the above applications**, and ensuring proper function in real-time is another obstacle which needs to be overcome. "*This field has important clinical implications, and as a computer vision specialist it presents interesting and fulfilling challenges,*" Asher concludes. "*It will definitely redefine the way surgeries are performed, and help patients get better results in non-invasive surgeries*". **Read More**



Real Data



Synthetic Data

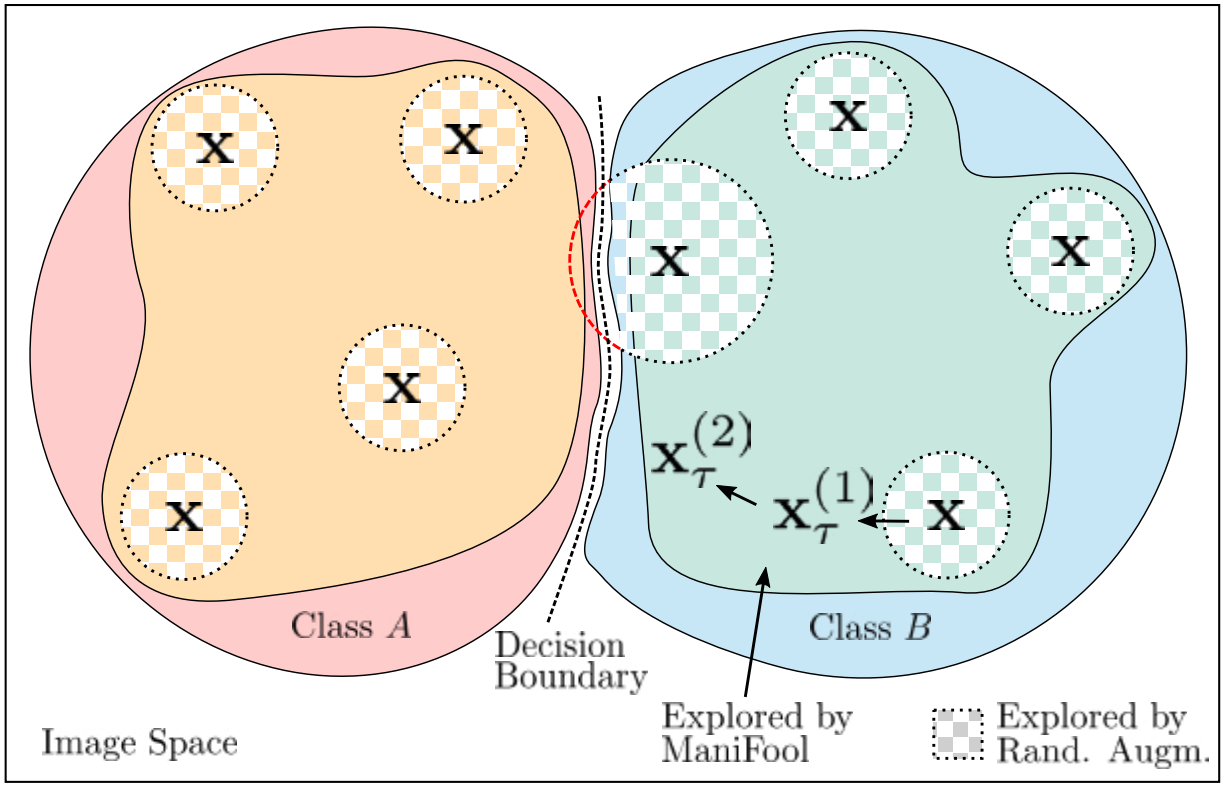
Congrats, Doctor!

Magda (Magdalini) Paschali recently completed her PhD at the Technical University of Munich at the Chair for Computer Aided Medical Procedures. Her research during her PhD focused on improving and evaluating the robustness of deep neural networks (DNNs) for medical imaging applications. Magda continues her research as a Postdoctoral Scholar in Stanford University at the Computational Neuroimage Science Laboratory (CNSLAB), where she focuses on machine learning models that can improve the understanding, diagnosis, and treatment of neuropsychiatric disorders. Congrats, Doctor Magda!

Computer-aided diagnostic systems powered by deep learning models need to not only perform well on limited known datasets but also generalize to unseen samples and be robust to challenges such as outliers and artefacts and threats like adversarial attacks. To this end our research aimed at developing methods that improve and thoroughly evaluate the robustness of machine learning models for medical diagnosis.

Improve Model Robustness

Training deep learning models on limited data can prevent models from generalizing to unseen samples. Data augmentation is an established way of combatting overfitting and improving model generalization. However, applying random affine transformations to training samples is limited to exploring the immediate vicinity of training samples [*see figure following*]. To solve this problem, we introduced a novel data augmentation technique that utilizes manifold-exploring affine geometric transformations that create samples that lie on the border of the manifolds between two-classes, maximizing the variance the network is exposed to during training [see fig below]. Our method improved model robustness against affine and projective transformations and increased model accuracy on fine-grained skin lesion and breast tumor classification. Finally, we proposed a metric based on geodesic distance that quantified the robustness of classifiers by measuring the distance of the augmented samples to the model decision boundaries.



Improve Training Dynamics

3D Convolutional DNNs are widely used for volumetric medical imaging data such as MRI and CT Scans. However, compared to 2D DNNs, such models have more training parameters and are more prone to overfitting when trained with limited data. To that end, we proposed 3DQ, the first ternary quantization method for 3D DNNs. 3DQ performed weight quantization and utilized two trainable scaling factors and a normalization parameter to increase model capacity while maintaining compression. 3DQ managed to not only reduce the model size by 16 times but also enhanced the training dynamics and increased the Dice Score achieved by large volumetric models for hippocampus and whole-brain segmentation trained on limited scans. 3DQ constitutes a solid approach for space-critical applications, like patient-specific models or model weight transfer for Federated Learning.

Model Evaluation with Adversarial Examples

Model robustness evaluation is necessary for DNNs deployed on critical applications. Thus, we proposed a novel benchmarking strategy that utilized adversarial examples to evaluate state-of-the-art models for classification and segmentation. Our method highlighted that models that achieve similar or identical performance on clean test data had substantial differences regarding robustness to adversarial attacks. That could be attributed to notable differences in the models' exploration of the underlying data manifold, resulting in varying robustness capabilities.

Yuyin Zhou is an Assistant Professor of Computer Science and Engineering at UC Santa Cruz.
Mathias Unberath is an Assistant Professor of Computer Science at Johns Hopkins University.
Together, they speak to us as co-organizers of this year's Medical Computer Vision Workshop at CVPR in June.

Yuyin Zhou

Mathias Unberath

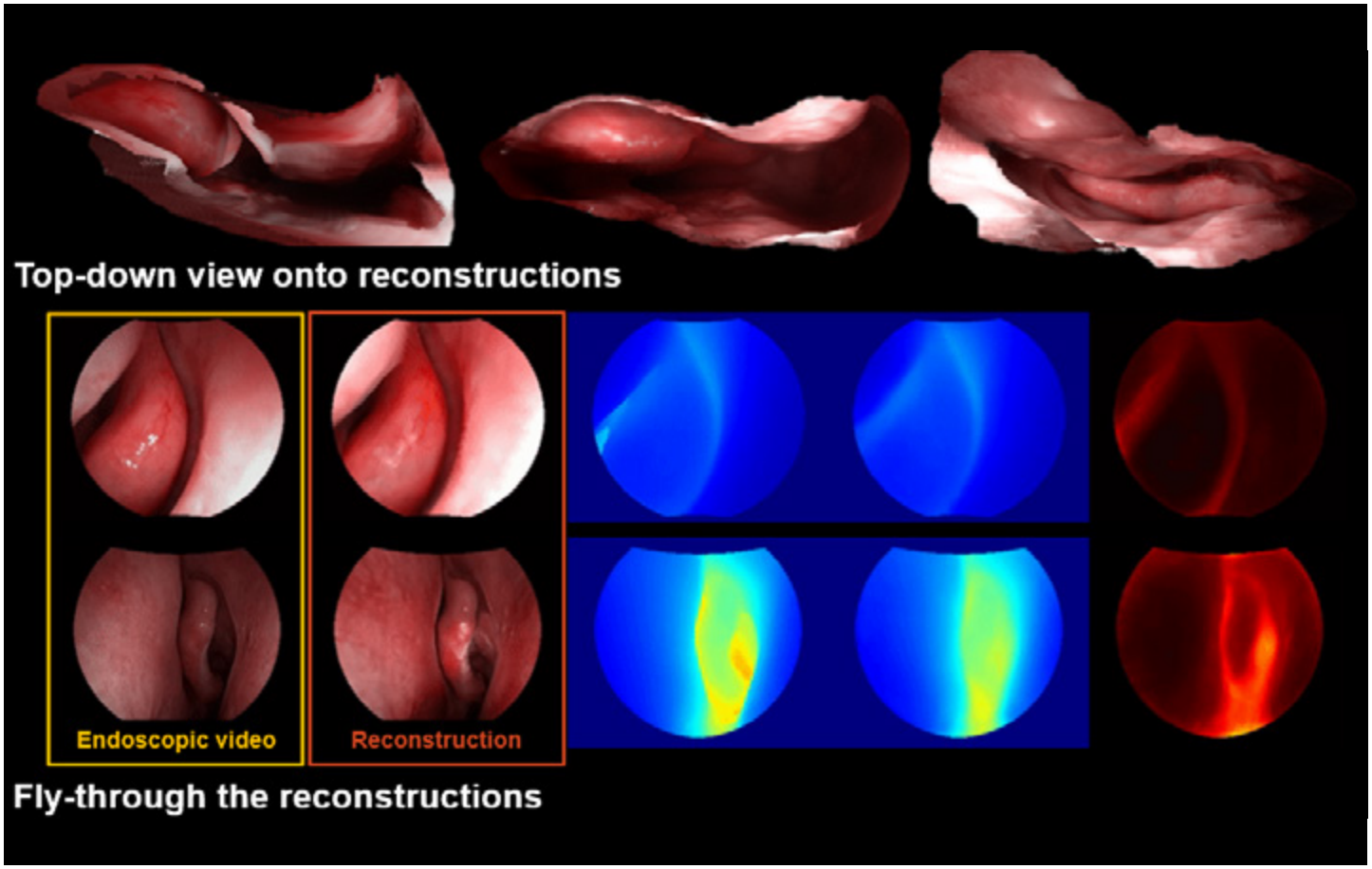
A subfield of computer vision, **medical computer vision** has seen rapid advances in recent years, with new ideas and novel technology flooding the field. **Its data representations, dataset sizes, privacy and safety issues, problems, and constraints on solutions can be very different to general computer vision**. Also, ethical concerns which are now finding their way into the wider field have been central to medical computer vision for some time, given that most clinical datasets require some form of ethical approval.

In its 9th edition, the **Medical Computer Vision Workshop at CVPR** offers an opportunity for the community to hear all about the latest progress and discuss new and future developments.

"One of the main things differentiating the Medical Computer Vision workshop from other workshops is we don't accept submissions," Mathias tells us.

"Our speakers appear by invitation only. In selecting them, we try to be as diverse and inclusive as we can regarding intellectual and demographic diversity and academia and industry representation. We do our best to balance all these considerations."

This year, the team has invited a broad range of researchers from academia, industry,



A look at Mathias' work

[Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G. D., Taylor, R. H., & Unberath, M. (2020, October). Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 3-13). Springer, Cham.]

and the fields of MIC and CAI. The cast list includes **Xiaoxiao Li**, Assistant Professor at the University of British Columbia, speaking about fairness and federated learning, and Professor **Polina Golland** from MIT, who has broad expertise in many different clinical problems. Also, several other professors working in medical image analysis and general computer vision, including **Ben Glocker** from Imperial College London and **Pablo Arbelaez** from the University of Los Andes (Colombia). With more speakers from industry to be confirmed, it looks set to be a stellar line-up. Of course, our readers already know many of these researchers.

The team is also seeking a diverse range of attendees by promoting the event within industry and inviting MD-PhDs who have a practice in addition to being researchers.

"It is critically important for researchers in this area to be proficient engineers and translators between the language and requirements of the end-users of our product," Mathias explains.

"If we're developing in a sandbox without appreciating the needs in the clinic, what impact will we ever have?"

Before organizing this year's event, Yuyin and Mathias took part in last year's workshop as speakers. Yuyin was also

featured in our magazine in October 2021 when she was a Stanford postdoc! They are joined on the organizing team by other workshop stalwarts bringing with them a rich set of experiences, including **Nicolas Padoy** (University of Strasbourg, France), **Tal Arbel** (McGill University, Canada), **Qi Dou** (The Chinese University of Hong Kong), and **Vasileios Belagiannis** (Universität Ulm, Germany).

The exact format of the event is still being ironed out and is likely to feature some hybrid elements. With the community separated for the past two years, some people in attendance will have never been to an in-person CVPR workshop before.

"What we're seeing is that the community hasn't been together, but people have also abandoned the idea that it is strictly necessary to be at every conference," Mathias points out.

"People are thinking twice about whether they need to travel now. Of course, many people are eagerly waiting to be back in person – including me!"

With the pace of innovation in this field, are we likely to see the Medical Computer Vision Workshop remain a feature of CVPR for a long time to come?

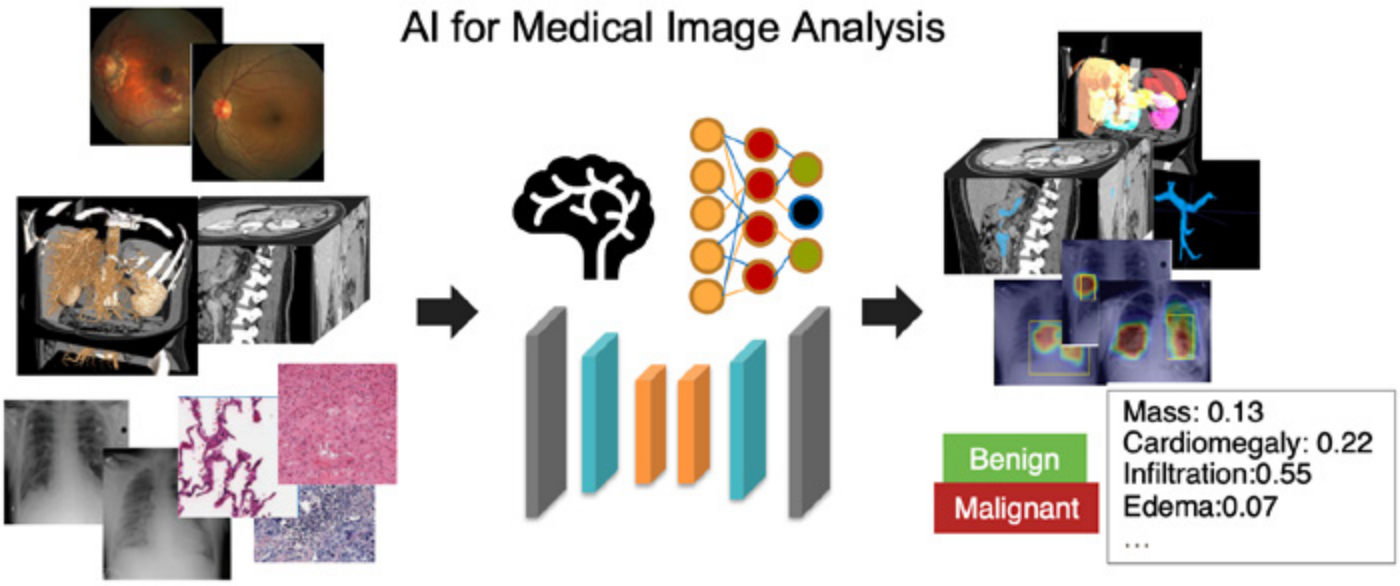
"Yes, and I think there should be more workshops of this type," Yuyin responds enthusiastically.

Mathias agrees:

"Most people have been focusing on a small subset of problems that exist in healthcare, for which we have solutions that work reasonably well. But there are many other areas that we haven't even begun to explore yet. Once we get there, we'll need a whole different set of solutions."

Yuyin adds, finally:

"In the future, maybe we will encourage paper submissions in addition to the keynote talks to address these remaining problems. Issues like ethical concerns need further discussion too. There is plenty still to talk about for many years to come!"



A look at Yuyin's work

20TH

ANNUAL MEETING
INTERNATIONAL SOCIETY FOR COMPUTER ASSISTED ORTHOPAEDIC SURGERY

SAVE THE DATE!

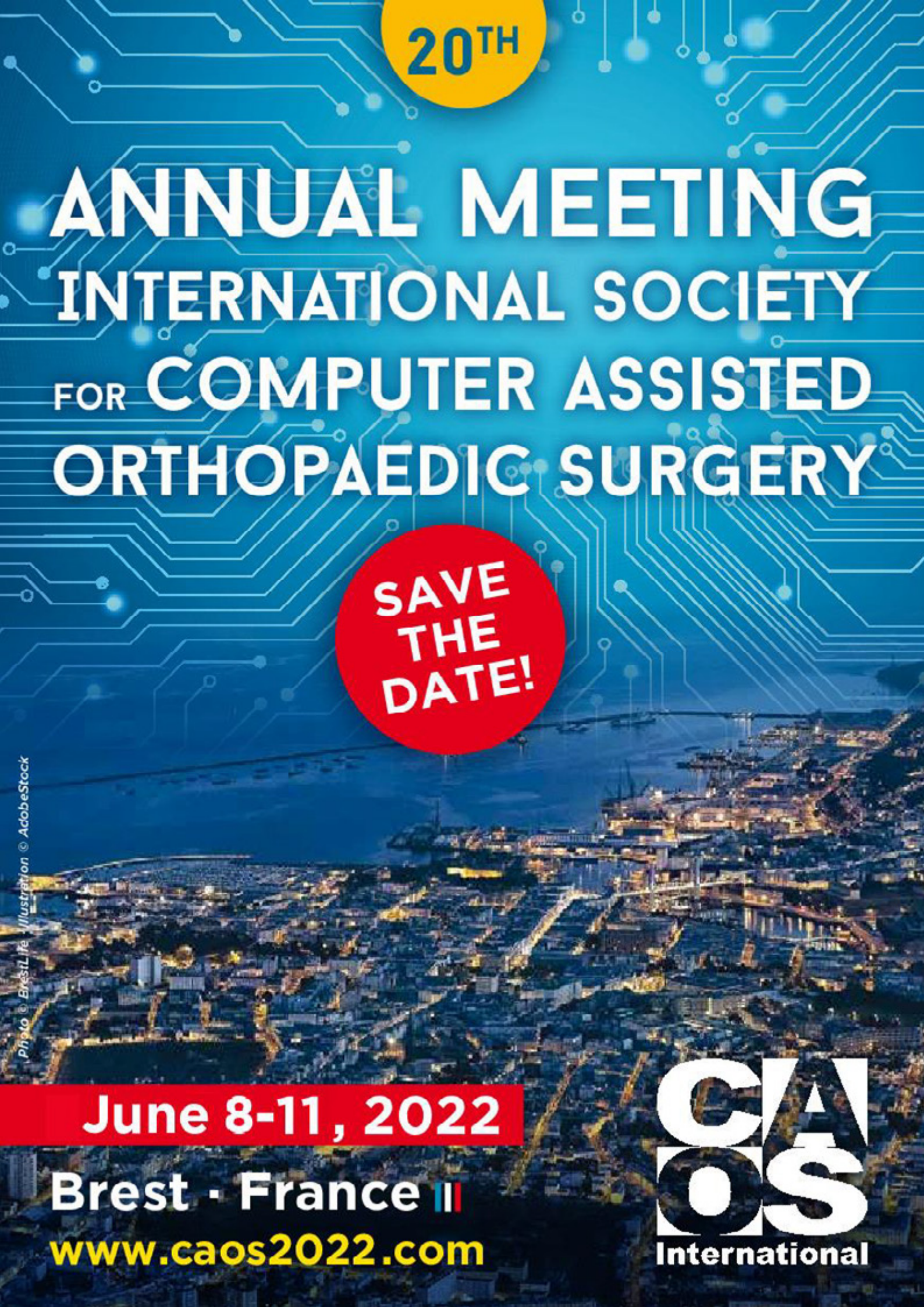


Photo © BrestLife Illustration © AdobeStock

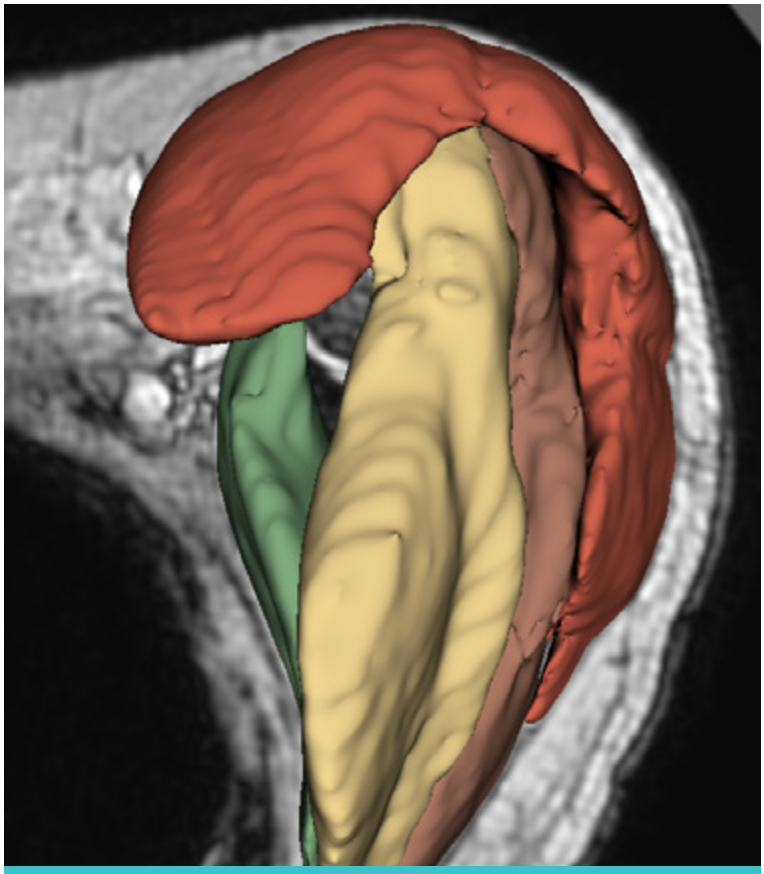
June 8-11, 2022

Brest · France

www.caos2022.com

CAOS International

SUMMER SCHOOL ON DEEP LEARNING FOR MEDICAL IMAGING



Courtesy Thomas Grenier - Shoulder Segmentation MRI

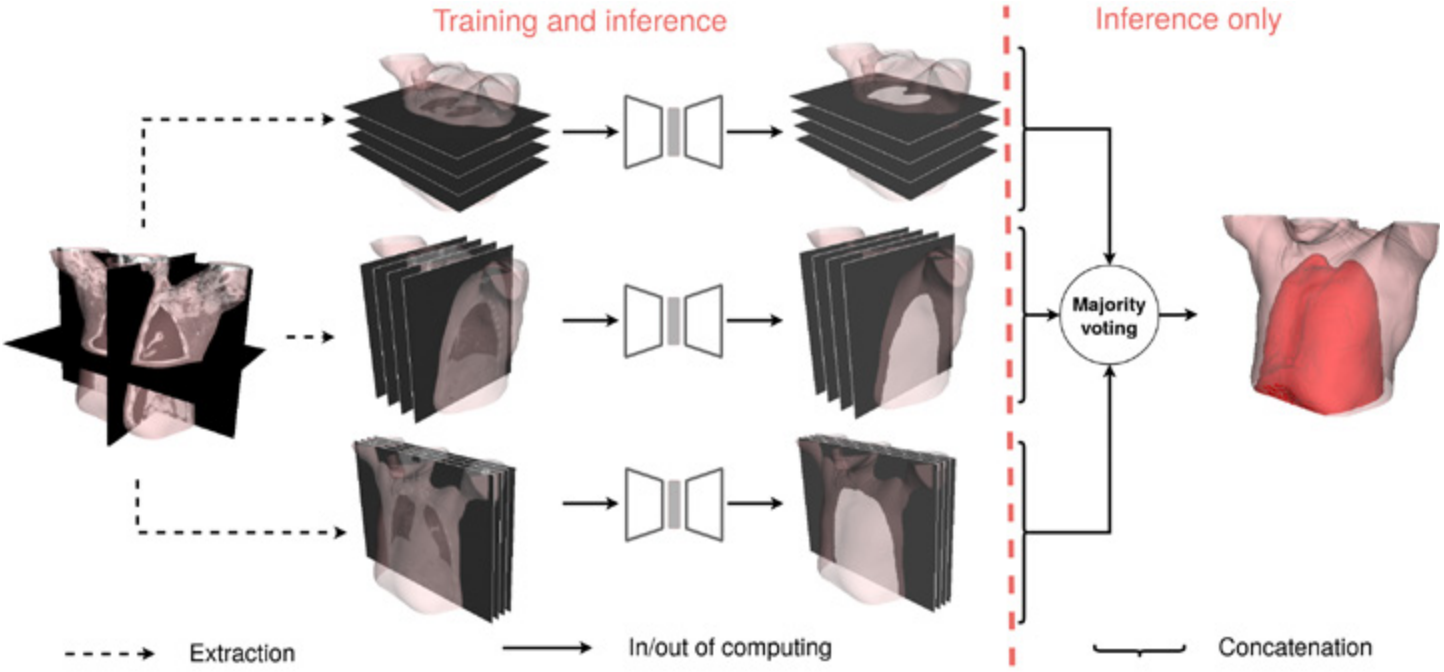
Pierre-Marc Jodoin

Pierre-Marc Jodoin is a Full Professor at the University of Sherbrooke in Canada and the co-founder and Chief Artificial Intelligence Officer at Imeka Solutions, a neuroimaging company dedicated to developing software for gauging the quality of the brain's white matter to improve treatments for neurodegenerative diseases. On top of this, Pierre-Marc is co-organizing this year's Summer School on Deep Learning for Medical Imaging (DLMI) in Montréal. He is here to tell us what we can expect from the event in July.

Now in its third edition, the **Summer School on Deep Learning for Medical Imaging** was launched in 2019 in Lyon, France. Its second edition was held virtually last year.

Its goal is to bring people working globally in medical imaging together for an interactive deep dive into the world of AI. It is aimed at those who have little or no knowledge in that area but are keen to learn the fundamentals.

"University classes can be 45 hours long and span 15 weeks," Pierre-Marc tells us.



Courtesy Emmanuel Roux - **Lung Segmentation**

"Here, it's just one week, which is great for people in industry or students who only have a week off. It's surprising how far we can get in just five days. Everything has been carefully chosen to give people a ***good starting point for navigating AI in medical imaging papers*** *and playing around with libraries like* ***PyTorch*** *or* ***TensorFlow****."*

The school has both theoretical classes and hands-on sessions. People who attend will receive an introduction to machine learning and the basics of deep learning before getting into the more advanced aspects of deep learning applied to medical imaging. Special presentations will be on popular topics, including **generative and adversarial methods, explainability, and weakly supervised deep learning.**

With over 180 people already booked to attend on-site and virtually, the team has made arrangements with a **GPU server farm** to ensure participants have the resources needed to train state-of-the-art deep neural networks.

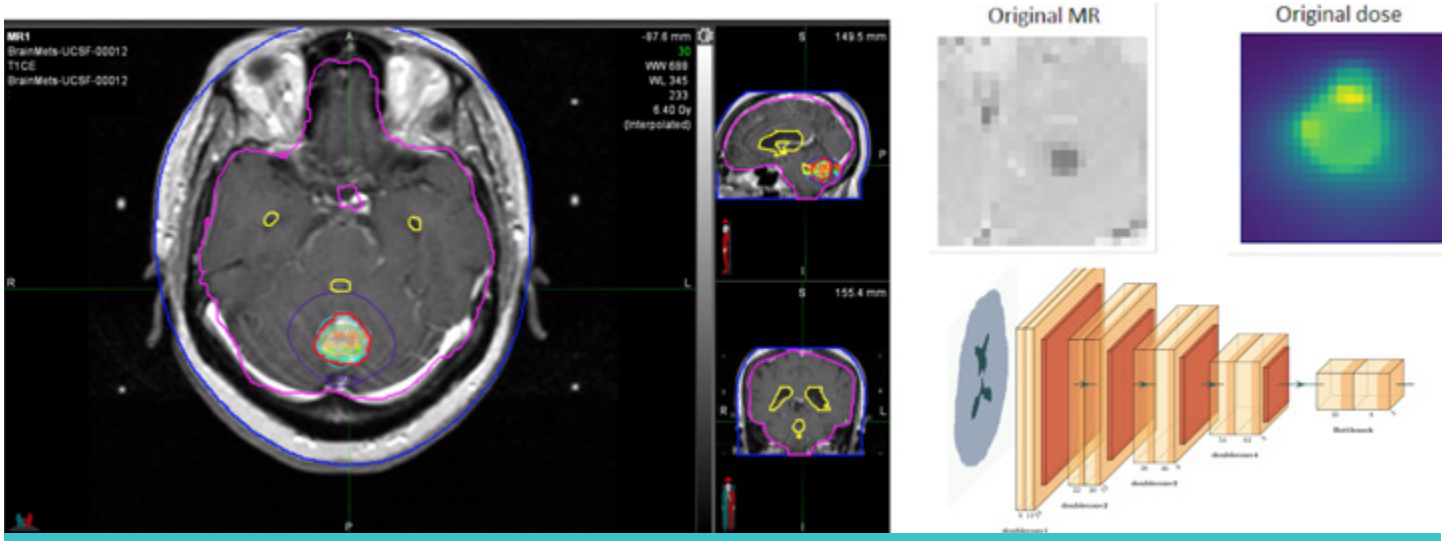
"*Most people say their goal isn't to become an AI expert or develop new models; they want to apply known models to their data," Pierre-Marc points out.*

"With AI in medical imaging, there are specific challenges you will not see in other fields. Medical data is huge, but ***databases can be small, with around 100 to 150 subjects****. There are limits to AI, but even with small datasets, there is still plenty you can do if you do it right. That's the overarching objective of the classes."*

The week will feature a **round-table discussion:** 'Why so much AI in research, why still so few AI in clinic?'

"*AI is everywhere in research, but if you walk into a hospital, you won't see a lot of AI," Pierre-Marc explains.*

"Why is that? How come AI doesn't seem to percolate easily in a clinical setting? On the panel, we have a doctor, the CEO of a software development company, a medical expert, and a research scientist from industry, who will all give their views.



Courtesy Martin Valliere - MRI Brain Analysis

It goes beyond the science to underline how powerful AI is, but how difficult it is to penetrate the market."

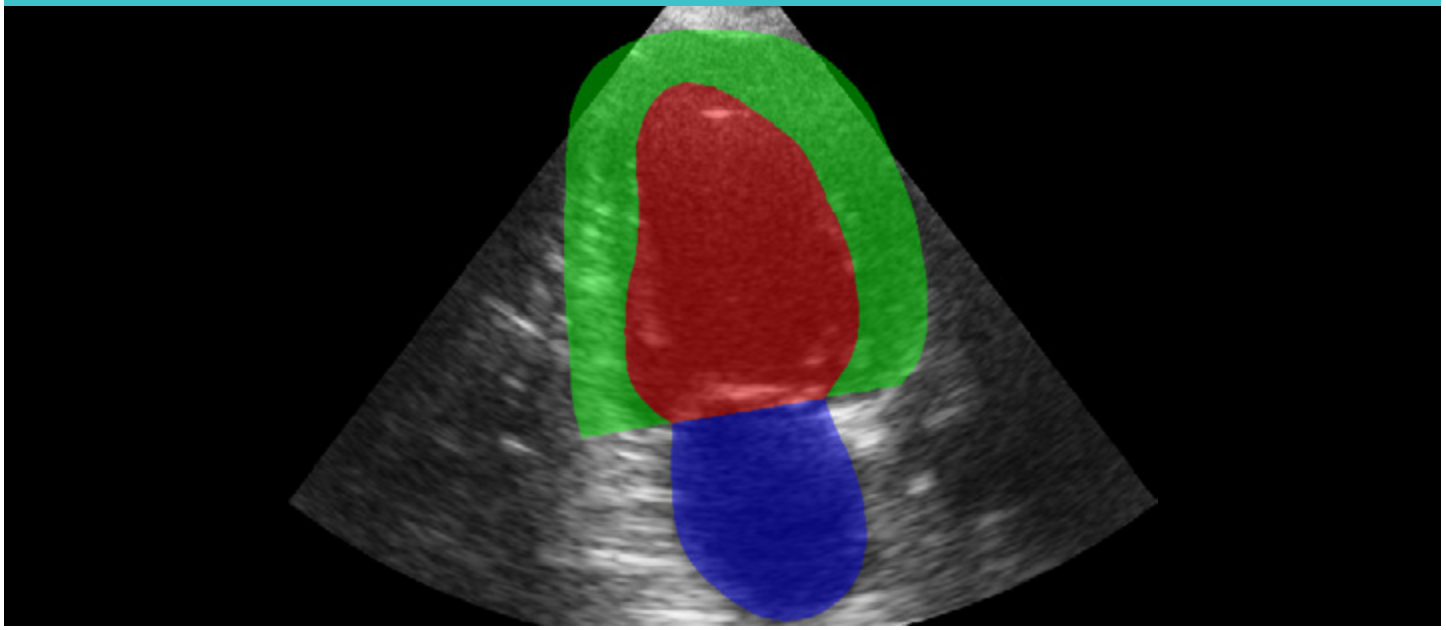
Participants will have plenty of time to network and exchange information in a relaxed atmosphere, with coffee breaks, lunches, and evening events, including cocktails and a museum visit with a banquet dinner.

Joining Pierre-Marc on the organizing committee and doing a terrific job to make the school succeed are **Christian Desrosiers, Jose Dolz, Thomas Grenier, Michaël Sdika,** and **Martin Vallières.**

All on-site participant slots are already taken, but you still have a small window of time to register to **attend virtually**. Registration closes on 4 April 2022, so get in quickly if you want to participate this year. However, do not be too disheartened if you miss out, as the plan is for the school to continue as an annual event.

MICCAI ENDORSED EVENT

Courtesy Olivier Bernard - Ultrasound Cardiac



INTERNATIONAL NEURO IMAGING SUMMER SCHOOL

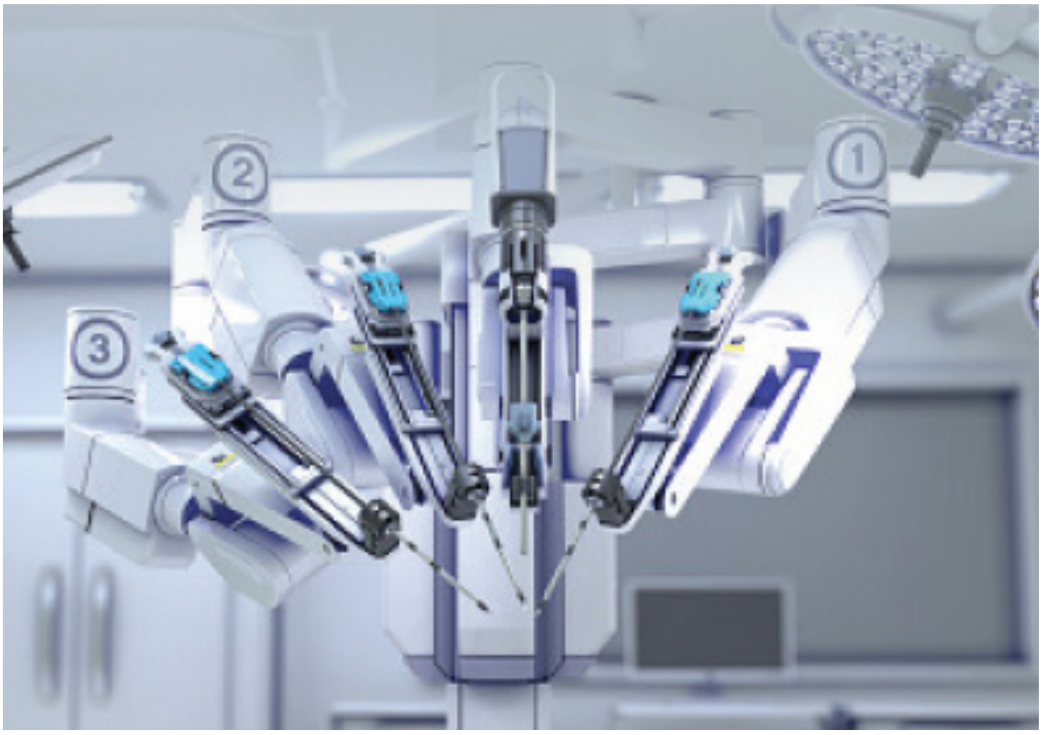
4TH-9TH JULY 2022

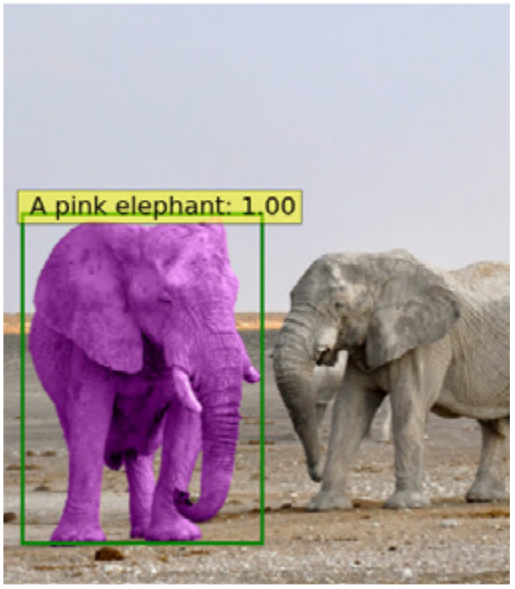
IT'S TIME TO PUT OUR BRAINS TOGETHER

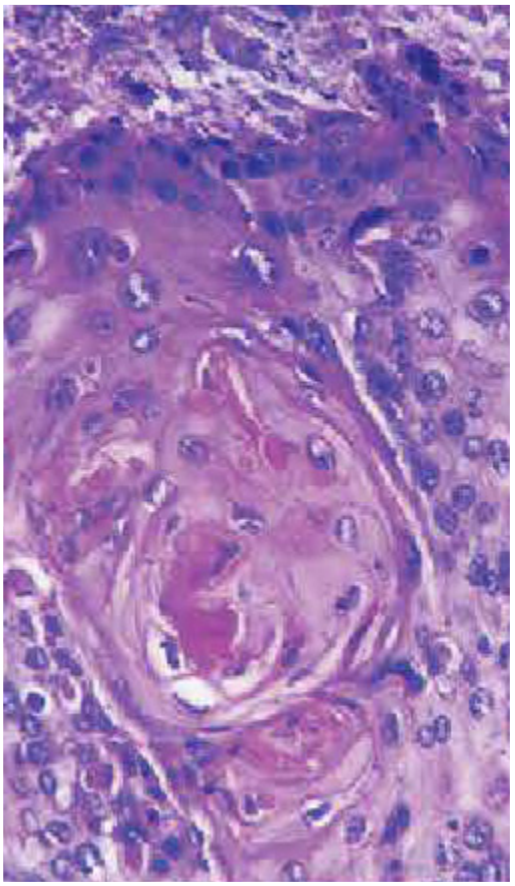
sano

www.neurosummerschool.org

N3S2







RSIP VISION