Separating lungs in CT scans for improved COVID19 detection

Robert Turnbull
Melbourne Data Analytics Platform
University of Melbourne
Grattan St. Carlton VIC Australia 3052
robert.turnbull@unimelb.edu.au

Simon Mutch
Melbourne Data Analytics Platform
University of Melbourne
Grattan St. Carlton VIC Australia 3052
simon.mutch@unimelb.edu.au

Abstract

This paper outlines our submission for the 4th COV19D competition as part of the 'Domain adaptation, Explainability, Fairness in AI for Medical Image Analysis' (DEF-AI-MIA) workshop at the Computer Vision and Pattern Recognition Conference (CVPR). The competition consists of two challenges. The first is to train a classifier to detect the presence of COVID-19 from over one thousand CT scans from the COV19-CT-DB database. The second challenge is to perform domain adaptation by taking the dataset from Challenge 1 and adding a small number of scans (some annotated and other not) for a different distribution. We preprocessed the CT scans to segment the lungs, and output volumes with the lungs individually and together. We then trained 3D ResNet and Swin Transformer models on these inputs. We annotated the unlabeled CT scans using an ensemble of these models and chose the high-confidence predictions as pseudo-labels for fine-tuning. This achieved the winning macro F1 score of 94.89% for Challenge 1 of the competition. It also achieved a second-best macro F1 score of 77.21% for Challenge 2.

1. Introduction

Deep learning models are becoming an increasingly common tool used for medical image analysis. In combination with expert medical professionals, these models can aid in the accurate detection of diseases such as COVID-19 [6, 7]. Here, deep learning models have been shown to provide accurate predictions for the presence of the disease from CT scans alone.

The 4th COV19D competition is being run as part of the 'Domain adaptation, Explainability, Fairness in AI for Medical Image Analysis' (DEF-AI-MIA) workshop [12] at the Computer Vision and Pattern Recognition Conference (CVPR) in 2024. It follows on from previous competitions held as part of the IEEE ICCV 2021 [8], ECCV 2022 [9] and ICASSP 2023 [2, 10] workshops. In the 2024 competition, two challenges presented to participants. The first is to take over one thousand CT scans from the COV19-CT-DB database [1, 11], annotated as belonging to patients with or without COVID, and train a classifier. The second challenge is to perform domain adaptation. A smaller dataset with CT scans from a different distribution to Challenge 1 is provided. This also includes almost 500 scans which have not been annotated. The challenge is to use the dataset for challenge 1 and make the best classifications on data from a distribution like the additional dataset.

In our submission, we build on work for previous years [18, 19] where we trained 3D ResNet and SwinTransformer models. In our 2023 submission, we segmented the lungs and cropped the CT scans accordingly. Here we experiment with segmenting individual lungs and training additional models with each lung separately as input. We also use pseudo-labels for augmenting the annotated dataset in the domain adaptation challenge.

2. Dataset

The 2024 competition dataset is divided between the two challenges. The Challenge 1 dataset comprises a total of 3,107 scans, with 1,684 used for training and validation (table 1). We divided the training dataset into four partitions which together with the official validation set gives five partitions for cross-validation. The Challenge 2 specific dataset comprises 4,979 scans, including 912 scans to be used in training and validation. Of these, 494 scans are not labeled as to whether or not the subject is infected with COVID-19. We combined both training and validation partitions from the Challenge 2 dataset and then divided this into roughly equal partitions for five-fold cross-validation.

We also included the public STOIC dataset [15] which includes 2,000 CT scans labeled as COVID-19 positive or negative. We ignored the severity categories of COVID-19 positive scans.

	COVID	NON-COVID	Total
Training	703	655	1358
Validation	170	156	326
Test	—	—	1,413

Table 1. The Challenge 1 Dataset.

	COVID	NON-COVID	Total
Training	120	120	240
Validation	65	113	178
Unannotated	—	—	494
Test	—	—	4,055

Table 2. The Challenge 2 Dataset.

3. Methods

3.1. Preprocessing

As in [19], we first segment the lungs using an adapted version of the methodology from [16] and crop the full 3D volumes to the resulting bounding box. In this work, we then additionally identify and crop each lung separately.

This is achieved by taking each transverse slice in turn, applying a binary threshold using Otsu's method, identifying all contours in the resulting image, removing contours with enclosed areas below a 500 pixels^2 and which are clearly not associated with lungs (e.g. span the entire width of the slice), and finally taking the two largest contours which overlap by less than 20% of their horizontal axis extent. An example is presented in the top panel of fig. 1 for a single slice. Once we have identified the lungs in each slice, we determine their axis-aligned bounding boxes. The left (/right) lung is cropped to be from the left (/right) side of the volume, to the left (/right) edge of the largest bounding box surrounding the right (/left) lung. In this manner, we find the maximum crop that guarantees each lung will be fully contained and contamination from the opposite lung is minimized. An example of the resulting crops for each lung are shown in fig. 1 in both the transverse and coronol planes (upper and lower panels, respectively). A volume containing each lung is stored and these are used as input to the model.

The volumes are interpolated to a single size. The cropped volumes including both lungs are interpolated to $256 \times 256 \times 176$ (in axes normal to the axial, sagittal and frontal planes respectively). The individual lungs are interpolated to a size of $320 \times 160 \times 224$.

3.2. Models

Heterogeneity in the number of scan slices is a common issue for deep-learning models in an medical imaging context. A number of methods have been proposed to deal with this (see e.g. [11] and citations therein), including the use of 3D-CNN architectures with fixed input lengths, achieved by interpolation or duplicating/subtracting slices as necessary. In this work, we note that the high-resolution scans which comprise our dataset closely resemble motion videos, where deep-learning models face similar issues with the heterogeneity of input lengths. In both cases we have two spatial dimensions with a fixed extent, plus another with a varying extent (the scan slice in our case and the frame in the case of videos). The information in consecutive slices/frames are also typically highly correlated. Noting this similarity, we opt to trial two neural network architectures, both pre-trained on the Kinetics 400 video classification dataset [4]. The first is a 3D ResNet [3, 17] with adaptations discussed in [19]. The second is a 3D Swin Transformer of size 'Tiny' [14]. These architectures were chosen to have models which processed the 3D volumes as input as a whole and

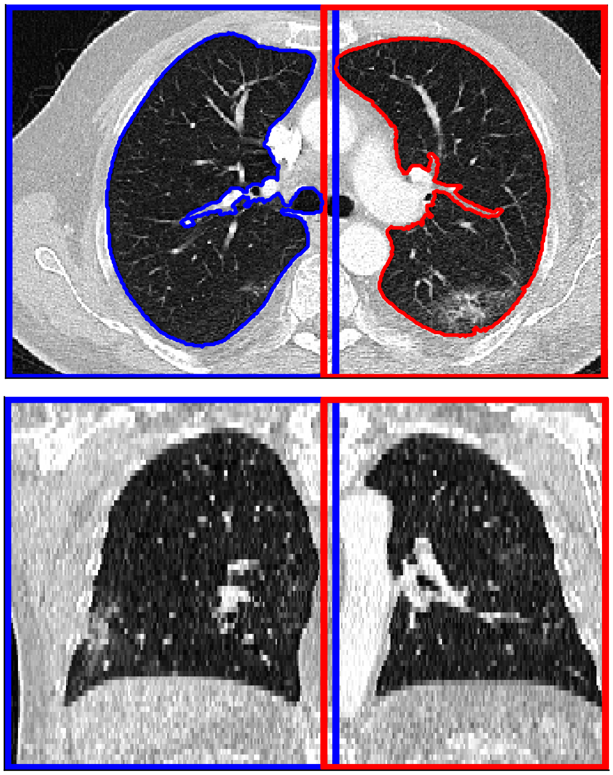

Figure 1. Example of segmenting and cropping individual lungs. The scan is first cropped to a bounding box which fully contains both lungs, as shown in the transverse and coronol planes by the top and bottom panels, respectively. In each transverse slice, the individual lungs are identified (blue and red contours in top panel) and a bounding box found for each lung which guarantees the lung is fully enclosed (red and blue boxes in both panels). Note that we allow small overlaps in the bounding boxes for each lung.

still be able to train on a single GPU.

3.3. Training Procedure

The models were trained for 30 epochs with a batch size of 2 using cross-entropy loss with the Adam optimizer [5]. Each volume was included in the training and validation datasets twice with the second one reflected through the sagittal plane. The brightness and contrast for each scan was randomly adjusted during the training according to the scheme discussed in [19].

3.4. Pseudo-Labels

For Challenge 2, we train an ensemble of models on the annotated scans and then make predictions on the unannotated scans. These predictions can be used as pseudo-labels [13]. To mitigate against training with too many scans with incorrect pseudo-labels, we only include predictions with higher confidence, meaning that only include predictions with a probability of the 0.7 or greater. These scans with their pseudo-labels are then included in the training dataset for fine-tuning the models for an additional ten epochs.

4. Cross-Validation Results

4.1. Challenge 1

Three models were trained for Challenge 1: a ResNet model, a Swin Transformer model and a ResNet model trained on the individual left and right lungs (ResNet-LR). The best performing model was the ResNet which achieved a mean F1 score of 92.55% across the five cross-validation partitions (fig. 2a). The ResNet-LR and Swin-LR models had a lower mean F1 score than the same architecture with both lungs at the same time but they had a lower variance across the five cross-validation partitions. Averaging the ResNet and the Swin Transformer results gave the highest F1 score overall at 93.5%, although including the ResNet-LR model results in the ensemble produced a slightly lower F1 score of 93.4% but with a smaller variance.

4.2. Challenge 2

A ResNet, Swin Transformer and ResNet-LR model were trained to predict the pseudo-labels. The best single model was the Swin Transformer with a mean F1 score of 90.73. As with Challenge 1, the ResNet-LR model had a lower mean F1 score than the ResNet model for both lungs but the individual lung model had a lower variance. An ensemble of the ResNet and Swin Transformer models achieved an F1 score of 91.2% (fig. 2b). If we filter the validation datasets for only high-confidence scans with a probability of being with or without COVID-19 above 0.7, then the F1 score increases to 95.8%. Using this ensemble, predictions were made on the 494 unannotated scans for Challenge 2. Of these, 414 predictions were above the threshold of 0.7 and these were assigned as pseudo-labels. This improved the F1 score for the Swin Transformer to 91.22% but the result for the ResNet decreased a small amount (fig. 2c). The ResNet-LR model improved from 88.6% to 89.85%. An ensemble of both Swin Transformer models (with and without pseudo-labels) together with the ResNet-LR trained with pseudo-labels achieved the highest F1 score of 92.15%.

5. Competitions Results

Each challenge was allowed five submissions for the competition. For both challenges, we chose models which performed best on the five cross-validation partitions to be used for our submissions to the competition. All submissions average results across models trained on the five cross-validation partitions.

The five competition submissions for Challenge 1 were:

1. ResNet
2. Swin Tranformer
3. Ensemble of ResNet and ResNet-LR
4. Ensemble of ResNet and Swin Transformer
5. Ensemble of ResNet, Swin Transformer and ResNet-LR

The five competition submissions for Challenge 2 were:

1. ResNet
2. Swin Tranformer
3. Swin Tranformer with pseudo-labels
4. Ensemble of Swin Transformer and Swin Transformer with pseudo-labels
5. Ensemble of Swin Transformer and Swin Transformer with pseudo-labels and the ResNet model with individual lungs and pseudo-labels.

The competition test set results for the submissions to Challenge 1 are shown in table 3 and for Challenge 2 in table 4. These results are plotted against the cross-validation results in fig. 3. There was a strong correlation between the performance in cross-validation and on the competition test set. The results for Challenge 1 achieved a slightly higher macro F1 score on the test set. The results for Challenge 2 were substantially lower, due to poorer F1 scores for the positive prediction of COVID19 even though the F1 for predicting non-COVID19 was high. The number of predictions of COVID19 in this submission was 364 out of 4,055 and, given the F1 scores for both cases, we can infer that the proportion of actual COVID19 cases in the test set was smaller than this. Thus the distribution of COVID19 in the test dataset was quite different to the training and validation data for Challenge 2 (table 2).

Twelve teams submitted results to Challenge 1 of the competition. Of these, six teams achieved a macro F1 score above the baseline (fig. 4a). Our's was the winning submission but the highest result in the next three teams were within 0.65%.

Ten teams submitted results to Challenge 2 and four of

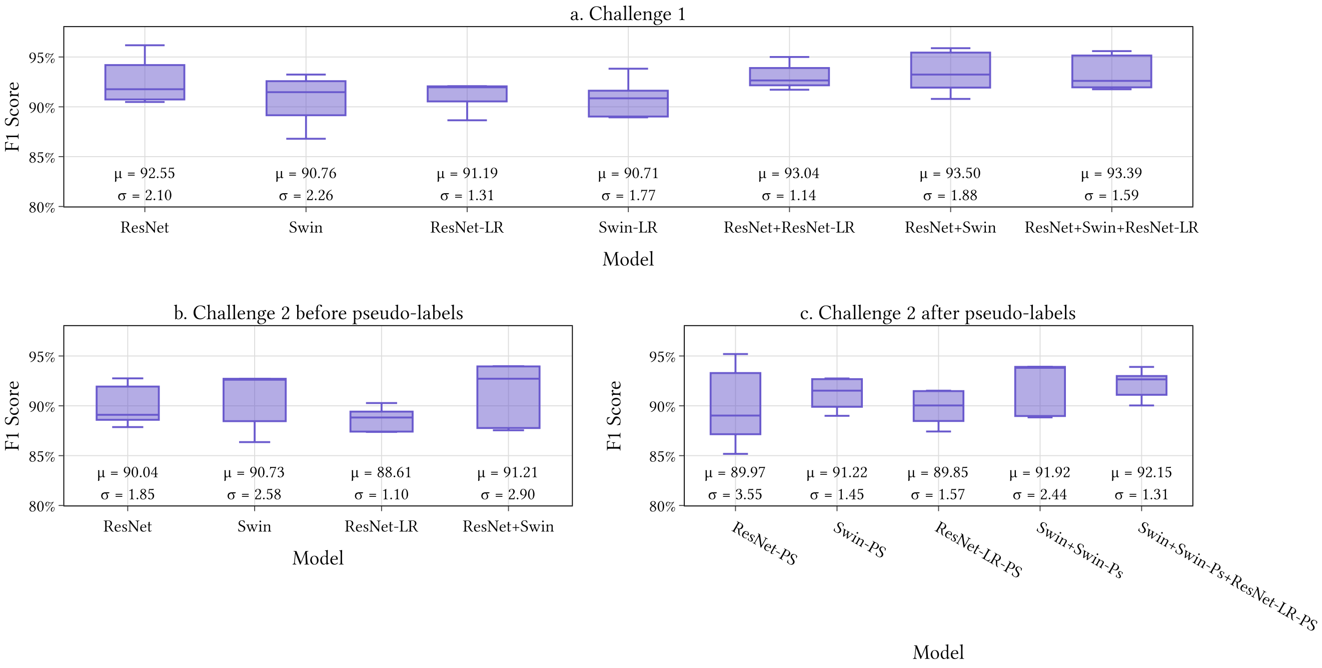

Figure 2. a. The cross-validation results for challenge 1. b. The cross-validation results for challenge 2 before adding in pseudo-labels. c. The cross-validation results for challenge 2 after adding in pseudo-labels. Model names joined with a '+' are ensembles with prediction probabilities averaged.

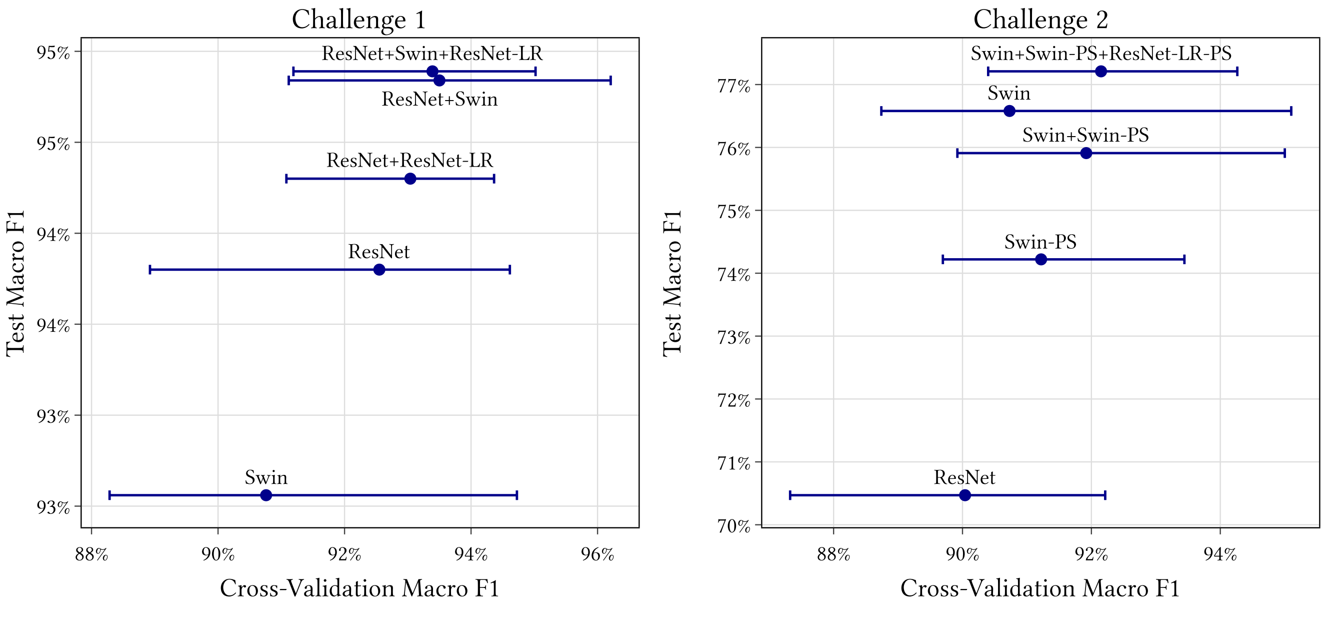

Figure 3. The test set results against the cross-validation results. Cross validation results shown with a circle at the mean and the error bars showing the minimum and maximum values across the five cross-validation partitions.

these achieved a macro F1 higher than the baseline (fig. 4b). Our best result was 0.34% below the winning submission, giving us the place of 'Runner-Up'. All teams which improved upon the baseline had much lower F1 scores for predicting COVID19 than for predicting non-COVID19. This tendency may have arisen since the proportion of COVID19 scans in the test dataset being low relative to the training and validation datasets, leading to over-prediction of COVID19.

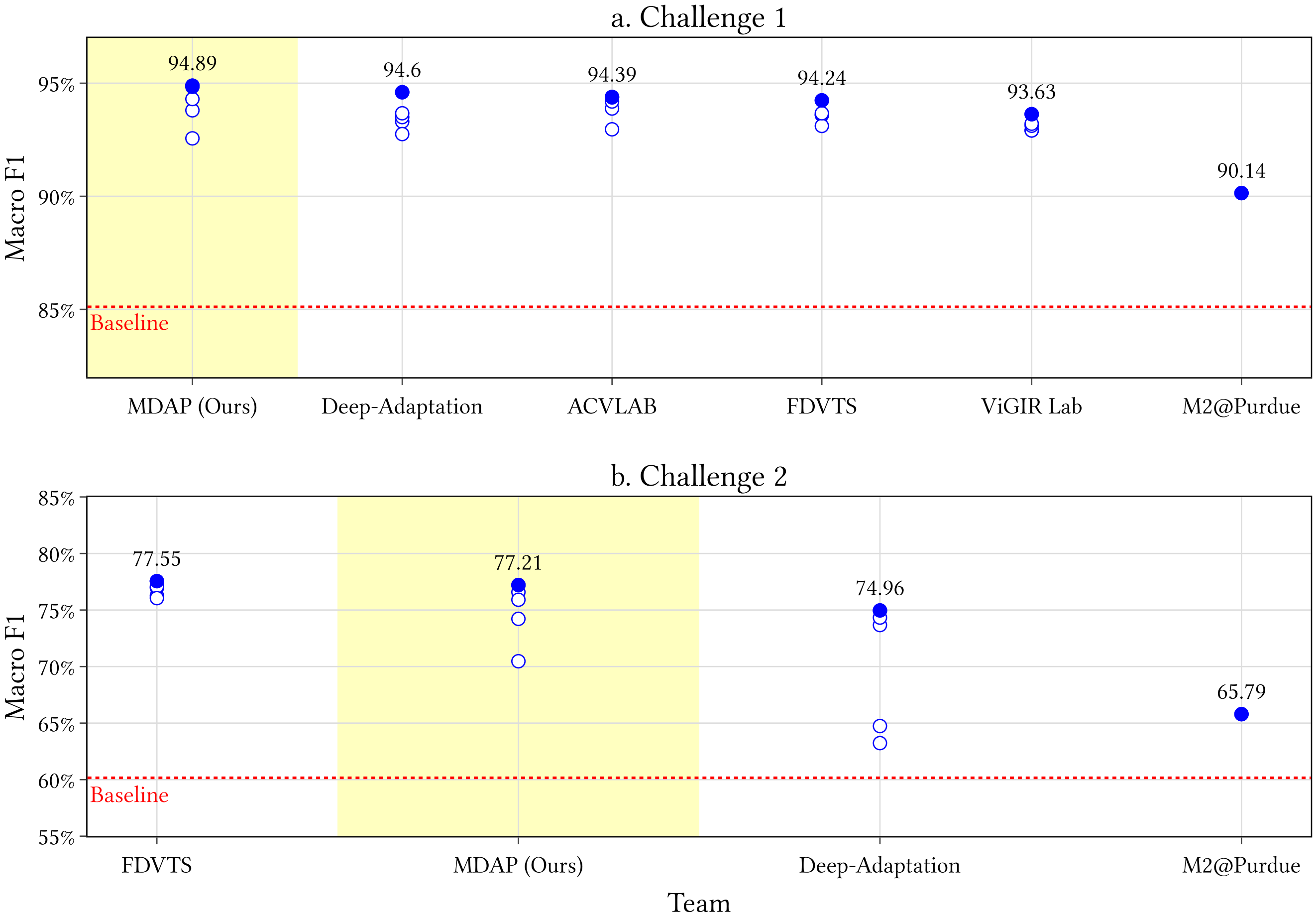

Figure 4. The competition results for all teams that surpassed the baseline. Our result, under then team name 'MDAP', is highlighted in yellow. The highest result for each team is marked with a solid blue dot and the value written above.

Submission	Macro F1	Non-COVID19 F1	COVID19 F1
ResNet	93.80	95.06	92.53
Swin	92.56	94.21	90.91
ResNet+ResNet-LR	94.30	95.50	93.09
ResNet+Swin	94.84	95.88	93.79
ResNet+Swin+ResNet-LR	**94.89**	**95.97**	**93.81**

Table 3. The Challenge 1 test set results.

6. Conclusion

The approach used in this paper achieved high cross-validation F1 scores for both challenges. The best result for Challenge 1 was an ensemble of the ResNet and Swin Tranformer models with an average F1 score of 93.5%. This ensemble achieved the winning result for all teams in Challenge 1 of the competition with a macro F1 score on the test set of 94.89%. The best single model for Challenge 2 was the Swin Transformer at an F1 score of 90.73%. This improved to 91.22% when pseudo-labels with high-confidence were added to the training set. An ensemble achieved even better results with an F1 score of 92.15%. This ensemble was our highest scoring submission on the test set of Challenge 2, and achieved 'Runner-Up' place in the competition with a macro F1 score of 77.21%. As deep learning methods continue to develop for analysis of medical imaging, especially when adapting to new domains, we expected improved outcomes for diagnosis and patient care.

Submission	Macro F1	Non-COVID19 F1	COVID19 F1
ResNet	70.47	94.76	46.17
Swin	76.58	96.78	56.37
Swin-PS	74.22	96.09	52.36
Swin+Swin-PS	75.91	96.56	55.27
Swin+Swin-PS+ResNet-LR-PS	**77.21**	**96.82**	**57.60**

Table 4. The Challenge 2 test set results.

7. Acknowledgments

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus. We acknowledge the help of Evelyn Mannix.

References

[1] Anastasios Arsenos, Dimitrios Kollias, and Stefanos Kollias. A large imaging database and novel deep neural architecture for covid-19 diagnosis. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, page 1–5. IEEE, 2022. 1

[2] Anastasios Arsenos, Andjoli Davidhi, Dimitrios Kollias, Panos Prassopoulos, and Stefanos Kollias. Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 2

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 3

[6] Dimitrios Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044*, 2020. 1

[7] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In *TAILOR*, page 251–267, 2020. 1

[8] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 537–544, 2021. 1

[9] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-mia: Covid-19 detection and severity analysis through medical imaging. In *European Conference on Computer Vision*, page 677–690. Springer, 2022. 1

[10] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-enabled analysis of 3-d ct scans for diagnosis of covid-19 & its severity. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023. 1

[11] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing*, 542:126244, 2023. 1, 2

[12] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Domain adaptation, explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. *arXiv preprint arXiv:2403.02192*, 2024. 1

[13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 3

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. 2

[15] Marie-Pierre Revel, Samia Boussouar, Constance de Margerie-Mellon, Inès Saab, Thibaut Lapotre, Dominique Mompoint, Guillaume Chassagnon, Audrey Milon, Mathieu Lederlin, Souhail Bennani, Sébastien Molière, Marie-Pierre Debray, Florian Bompard, Severine Dangeard, Chahinez Hani, Mickaël Ohana, Sébastien Bommart, Carole Jalaber, Mostafa El Hajjam, Isabelle Petit, Laure Fournier, Antoine Khalil, Pierre-Yves Brillet, Marie-France Bellin, Alban Redheuil, Laurence Rocher, Valérie Bousson, Pascal Rousset, Jules Grégory, Jean-François Deux, Elisabeth Dion, Dominique Valeyre, Raphael Porcher, Léa Jilet, and Hendy Abdoul. Study of thoracic ct in covid-19: The stoic project. *Radiology*, 301(1):E361–E370, 2021. PMID: 34184935. 1

[16] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L. Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 2

[17] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[18] Robert Turnbull. Using a 3D ResNet for Detecting the Presence and Severity of COVID-19 from CT Scans. In *Com-*

puter Vision – ECCV 2022 Workshops, number 7, pages 663–676, Cham, 2023. Springer Nature. 1

[19] Robert Turnbull. Lung segmentation enhances covid-19 detection. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023. 1, 2, 3