EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

SAM2CLIP2SAM: Vision Language Model for Segmentation of 3D CT Scans for Covid-19 Detection

Dimitrios Kollias[1,2*], Anastasios Arsenos[3], James Wingate[4]and Stefanos Kollias[3]

[1]School of Electronic Engineering & Computer Science, Queen Mary University of London, UK
[2]Digital Environment Research Institute (DERI), Queen Mary University of London, UK
[3]School of Electrical & Computer Engineering, National Technical University Athens, Greece
[4]School of Computer Science, University of Lincoln, UK

Abstract

This paper presents a new approach for effective segmentation of images that can be integrated into any model and methodology; the paradigm that we choose is classification of medical images (3-D chest CT scans) for Covid-19 detection. Our approach includes a combination of vision-language models that segment the CT scans, which are then fed to a deep neural architecture, named RACNet, for Covid-19 detection. In particular, a novel framework, named SAM2CLIP2SAM, is introduced for segmentation that leverages the strengths of both Segment Anything Model (SAM) and Contrastive Language-Image Pre-Training (CLIP) to accurately segment the right and left lungs in CT scans, subsequently feeding these segmented outputs into RACNet for classification of COVID-19 and non-COVID-19 cases. At first, SAM produces multiple part-based segmentation masks for each slice in the CT scan; then CLIP selects only the masks that are associated with the regions of interest (ROIs), i.e., the right and left lungs; finally SAM is given these ROIs as prompts and generates the final segmentation mask for the lungs. Experiments are presented across two Covid-19 annotated databases which illustrate the improved performance obtained when our method has been used for segmentation of the CT scans.

Received on 28 August 2024; accepted on 01 November 2024; published on 02 April 2025

Keywords: RACNet, SAM, CLIP, segmentation, classification, Covid-19 detection, COV19-CT-DB

doi:10.4108/eetpht.11.9010

1. Introduction

The application of Deep Learning (DL) techniques in medical image analysis has revolutionized the field, leading to substantial improvements in the accuracy and reliability of diagnoses [1–8]. In pathology and radiology, DL models have proven superior in extracting clinically relevant information from medical images compared to traditional manual assessments, which are often subjective and inconsistent. However, integrating these AI-based methods into routine clinical workflows requires significant development and rigorous validation.

This paper focuses on the diagnosis of COVID-19 using advanced medical image analysis techniques, particularly utilizing three-dimensional (3-D) chest CT scans. Advanced approaches target the segmentation and automatic detection of pneumonia regions in the lungs, followed by identifying anomalies associated with COVID-19, such as ground-glass opacities, consolidation, and interlobular septal thickening, which are typically found under the pleura. These methods necessitate large, annotated datasets for effective model training.

To meet this need, the COV19-CT-DB [9–11, 11–16] was developed, a comprehensive database of 3-D chest CT scans, which includes 7,756 scans (1,661 annotated as COVID-19 and 6,095 as non-COVID-19 cases) of around 2.5 million CT slices. In addition, we introduce RACNet, an innovative deep neural network architecture designed to address the challenges of analyzing 3-D CT scans. RACNet is engineered to: i) process volumetric CT scan data, ii) handle the variability in slice numbers across scans, and iii)

*Dimitrios Kollias. Email: d.kollias@qmul.ac.uk

deliver high diagnostic performance for COVID-19. This architecture incorporates dynamic routing and feature alignment mechanisms that selectively process relevant Recurrent Neural Network (RNN) outputs for making diagnostic decisions.

However the performance of RACNet (and of any classification model) depends on the input 3-D CT scans and whether they are segmented or not. Segmentation of chest CT scans is crucial for several reasons in the context of classification, particularly for diseases like COVID-19. Firstly, segmentation isolates specific regions of interest (ROIs), such as the lungs, which are critical for diagnosing respiratory conditions. By focusing on these relevant areas, the classification model can avoid distractions from irrelevant regions, leading to more accurate diagnoses. Secondly, segmentation allows for the extraction of detailed and localized features from the segmented regions. This detailed feature extraction can significantly improve the performance of classification models by providing more precise and relevant information. Thirdly, medical images often contain noise and artifacts from surrounding tissues and organs. Segmentation helps in reducing this noise by isolating the lung regions, thus providing cleaner input data for the classification model. Finally, segmentation ensures that the classification model consistently analyzes the same anatomical regions across different scans and patients. This consistency is vital for training robust and generalizable models.

Many state-of-the-art approaches for COVID-19 classification either do not perform segmentation at all or perform segmentation as a combination of morphological transforms and a pre-trained model, like U-Net [17–19]. More specifically, for each CT-scan, every slice first passes through the pre-trained U-Net. After all slices of the CT-scan are segmented by the U-Net model, there is a checking procedure to assure that all slices are segmented. If a slice has a mask area less than 40 % of the largest mask area of the CT-scan, then morphological transforms are used to segment this slice. However, the issue with such approaches is that the segmentation is not very accurate; for instance, the segmentation mask includes the lungs and the mediastinal mass between them, or a lung is segmented along with its surrounding area and so on. To solve this issue, in this work, we present an innovative framework that integrates Segment Anything Model (SAM) [20] and Contrastive Language-Image Pre-Training (CLIP) [21] to segment only the right and left lungs in CT scans, with these outputs subsequently being fed into RACNet for classification to detect COVID-19 and non-COVID-19 cases.

SAM and CLIP are two exemplary Vision Foundation Models (VFMs) that have showcased exceptional capabilities in segmentation and zero-shot recognition, respectively. SAM, a prompt-driven segmentation model, excels across diverse domains. SAM has been trained on an extensive dataset of over one billion masks, making it highly adaptable to a wide range of downstream tasks through interactive prompts. It can operate in two distinct modes: segment everything mode and promptable segmentation mode. In our approach, we employ both modes to achieve optimal segmentation results. SAM has shown impressive results in a broad range of tasks for natural images, but its performance has not been state-of-the-art when being directly applied to medical imaging. Conversely, CLIP's training with millions of text-image pairs has endowed it with an unprecedented ability in zero-shot visual recognition.

Despite their individual successes, their unified potential for medical image segmentation remains largely unexplored. Existing methods for adapting SAM to medical imaging often rely on tuning strategies that require extensive data or prior prompts tailored to the specific task, posing significant challenges when data samples are limited [22].

Medical imaging segmentation tasks exhibit inherent variability based on the specific clinical scenario, complicating the adaptation process. To assign semantic labels to SAM-provided masks, our method involves cropping the original image according to these masks. This set of cropped regions is then processed by CLIP, which retrieves the corresponding crop in a zero-shot manner using visually descriptive sentences related to the lungs [23], generated via GPT. The retrieved region of interest (ROI) mask is subsequently used for bounding box prompt generation, guiding SAM to deliver the final lung segmentation.

The remainder of this paper is organized as follows: Section 2 reviews the related work we have developed in medical image analysis and COVID-19 diagnosis and introduces the vision-language models. Section 3 outlines the proposed pipeline, including the vision-language models and the RACNet architecture. Section 4 presents the experimental setup and results, whilst Section 5 concludes the paper and suggests future research directions.

2. Related Work

SAM is a promptable vision foundation segmentation model designed to segment everything in an image based on different types of prompts, such as bounding boxes and point prompts. It comprises an image encoder, a prompt encoder, and a lightweight mask decoder. A pretrained Vision Transformer (ViT) [24] is used as the image encoder, transforming the input image into dense features. The prompt encoder processes prompts, which can be sparse or dense, encoding them into a format suitable for mask

generation. SAM can operate in two modes: segment everything mode and promptable segmentation mode. The former segments everything in the image using a grid of keypoints as prompts, while the latter segments specific regions based on provided prompts. Our framework utilizes both modes of SAM in conjunction with CLIP.

CLIP is a pre-trained large Vision-Language Model known for its strong generalizability and impressive zero-shot domain adaptation capabilities. It aligns image and text modalities within a shared embedding space, enabling it to perform image classification directly on the target dataset without any fine-tuning. By employing prompt engineering, CLIP can be adapted to various domains, incorporating relevant semantic details related to the specific target task. Our framework leverages CLIP's zero-shot recognition capabilities to identify and retrieve the ROI in CT scans, facilitating accurate lung segmentation.

Various 3-D CNN models have been employed for detecting COVID-19 and differentiating it from common pneumonia (CP) and normal cases using volumetric 3D CT scans [18, 19]. For instance, [25] utilized a pretrained DenseNet-201 model, which was fine-tuned on CT scan images to classify them into COVID-19 or non-COVID-19 categories. The performance of this model was compared against other pretrained and fine-tuned models (VGG16, ResNet152V2, and Inception-ResNetV2). [26] combined CNNs with RNNs to process CT scan images, successfully distinguishing between COVID-19 and non-COVID-19 cases. Similarly, [27] proposed a multi-task architecture featuring a shared encoder for 3D CT scans and three branches: a decoder for image reconstruction, a second decoder for COVID lesion segmentation, and a multi-layer perceptron for classification into COVID-19 and non-COVID-19 categories.

Additionally, a weakly supervised deep learning framework was introduced by [28] for COVID-19 classification and lesion localization using 3D CT volumes. This framework utilized a pretrained UNet to segment lung regions in each CT slice, which were then fed into a 3D DNN for classification. COVID-19 lesions were localized by combining activation regions from the DNN with connected components through an unsupervised method. Furthermore, [29] established baseline performance using 3D models such as ResNet3D101 and DenseNet3D121. They then proposed a differentiable neural architecture search (DNAS) framework to automatically identify optimal 3D DL models for CT scan classification. The study also published the training, validation, and test datasets used, providing a resource for future research.

3. The Proposed Approach

The complete framework of the proposed approach is depicted in Figure 1. It consists of the segmentation framework (which has the goal of segmenting only the right and left lungs in CT scans) and the COVID-19 classification framework (RACNet). Let us stress that the proposed segmentation framework can be integrated in any classification model for any similar task (not just RACNet, or not for COVID-19 detection). More details about each framework follow.

3.1. The SAM2CLIP2SAM Segmentation Framework

Our proposed framework leverages the combined strengths of the Segment Anything Model (SAM) and Contrastive Language-Image Pre-Training (CLIP) for the task of lung segmentation in CT scans, followed by classification using RACNet. This proposed methodology, inspired by [30], can be delineated into three primary components: i) part-based segmentation using SAM; ii) region of interest (ROI) extraction using CLIP; and iii) final segmentation using bounding box prompts from SAM.

Part-based Segmentation using SAM. In the first phase, SAM is employed to generate part-based segmentation masks from the input CT scan image. SAM operates in its segment-everything mode to produce an exhaustive set of masks. Given an input image $\mathrm{I} \in \mathfrak{Re}^{H\times W\times 3}$ (with H and W denoting the Height and Width of the image) and a grid of keypoints $\mathcal{G}$, SAM generates a set of part-based masks $\mathcal{A}$:

$$\mathcal{A} = SAM_{seg-everything}(I, \mathcal{G}) \tag{1}$$

where $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ and each a_i is a mask corresponding to a specific region in I.

ROI Extraction using CLIP. Once the part-based masks are generated, the next step involves extracting the mask corresponding to the Region Of Interest (ROI) using CLIP. The masks from $\mathcal{A}$ are used to crop the input image I, resulting in a set of cropped images $\mathcal{C}$:

$$\mathcal{C} = \{I \cdot a_i \mid a_i \in \mathcal{A},\ \mathrm{area}(a_i) > \tau\} \tag{2}$$

where $\cdot$ denotes element-wise multiplication and τ is the area threshold used to filter out background masks.

Visually descriptive textual prompts for lung anatomy are generated using GPT. These prompts are passed through CLIP's text encoder to obtain a textual embedding $\mathcal{W}$:

$$\mathcal{W} = CLIP_{text-encoder}(VDT) \tag{3}$$

where VDT represents the visually descriptive text.

Each cropped image $c \in \mathcal{C}$ is then passed through CLIP's vision encoder to obtain a set of vision embeddings $\mathcal{V}$:

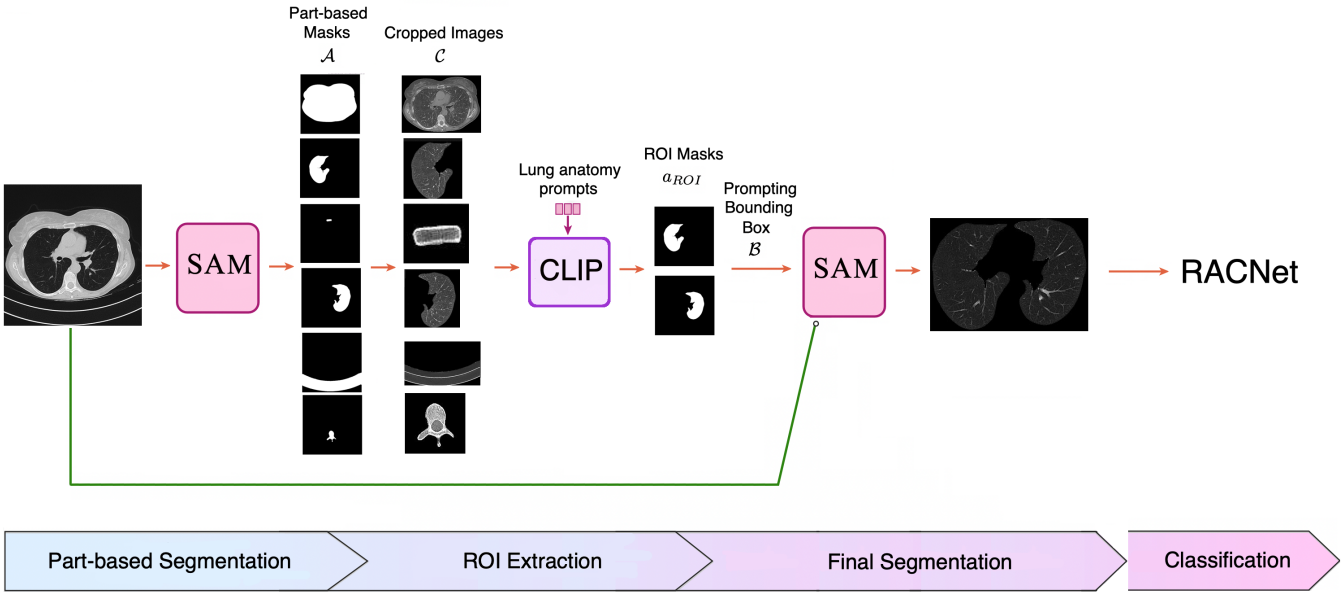

Figure 1. Our whole proposed pipeline that includes segmentation and classification tasks

$$\mathcal{V} = \{v_i \mid v_i = CLIP_{vision-encoder}(c_i), c_i \in \mathcal{C}\} \quad (4)$$

The similarity between each vision embedding v_i and the text embedding $\mathcal{W}$ is computed using cosine similarity:

$$sim(v_i, \mathcal{W}) = \frac{v_i \cdot \mathcal{W}}{||v_i| \cdot ||\mathcal{W}||} \quad (5)$$

The mask corresponding to the ROI is identified as the mask with the highest similarity score:

$$a_{ROI} = a_{\arg\max_i sim(v_i, \mathcal{W})} \quad (6)$$

Final Segmentation using Bounding Box Prompts. The final segmentation step involves using the bounding box of the ROI mask to generate prompts for SAM. The bounding box $\mathcal{B}$ of the ROI mask a_{ROI} is calculated as:

$$\mathcal{B} = \big(min_{(x,y)\in a_{ROI}} x,\ min_{(x,y)\in a_{ROI}} y,\ max_{(x,y)\in a_{ROI}} x,\ max_{(x,y)\in a_{ROI}} y\big) \quad (7)$$

SAM is then prompted with $\mathcal{B}$ to generate the final segmentation mask for the lungs:

$$a_{final} = SAM_{promptable}(I, \mathcal{B}) \quad (8)$$

The segmented lung regions a_{final} are subsequently used to train and classify images in RACNet for detecting COVID-19 and non-COVID-19 cases.

3.2. The RACNet COVID-19 Classification Framework

An overview of the RACNet COVID-19 Classification Framework can be seen in Figure 2. Each segmented 3-D scan, consisting of t slices, is analyzed using a CNN-RNN based model. A routing component equipped with alignment and masking operations effectively handles the variability in t across different CT scans. The final diagnosis is produced through a dense layer followed by an output layer.

The input data is fed into the 3D Analysis component of RACNet, which comprises a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). The CNN component performs 2-D slice analysis, extracting lung features from each slice. Our objective is to replicate the diagnostic process of medical professionals by utilizing all CT slices for COVID-19 detection. The RNN component processes the sequential features extracted by the CNN, analyzing the CT scan slices in sequence. The features extracted by the RNN are subsequently fed into the RACNet Routing Component. These features are concatenated and then processed by the Mask Layer.

During the training of RACNet, the routing component dynamically selects a number of RNN outputs equal to the length l of the input, zeroing out the remaining RNN outputs. The selected outputs are then fed into the dense layer. This selection process can be implemented in two ways: by either selecting the first l RNN outputs, or through an 'alignment' approach, where l RNN outputs are positioned at equidistant intervals within the $[0, t-1]$ range, with the remaining outputs placed between them.

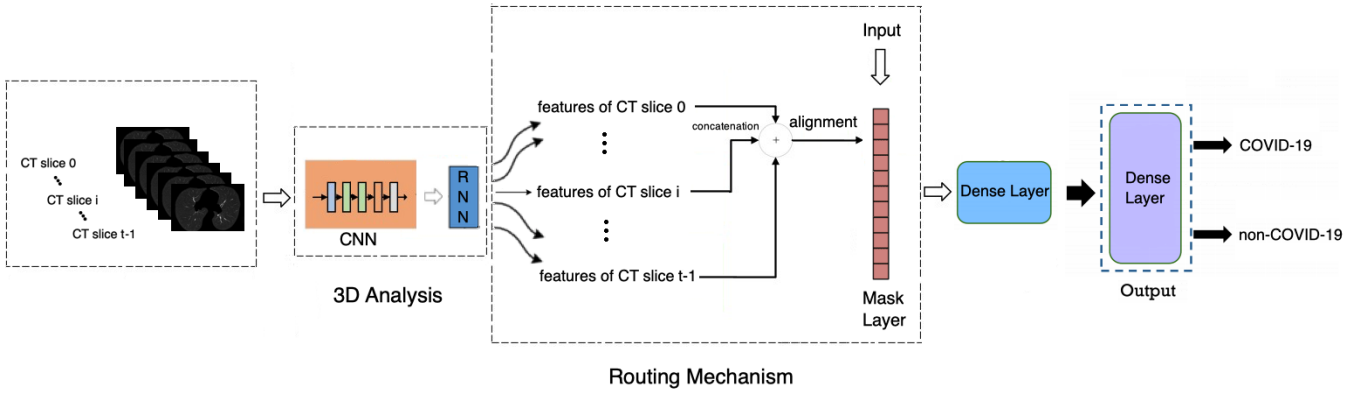

Figure 2. The RACNet model for COVID-19 Classification

The concatenated features are then processed by the RACNet Classification Component, which includes a dense layer and an output layer utilizing a softmax activation function to predict the presence or absence of COVID-19.

The dense layer is responsible for the final extraction of semantic information from the RNN outputs. During training, weight updating is guided by the routing mechanism and the Mask Layer. Only the weights corresponding to the selected RNN outputs are updated, while the others remain constant until they are selected for another scan input. We keep these weights constant during training and ignore the links of RNN outputs which are not selected by the routing mechanism [31].

4. Experimental study

4.1. Databases

The COV19-CT-DB database comprises 7,756 3-D chest CT scans collected from various medical institutions. Each CT scan contains between 50 and 700 2-D CT slices. The dataset includes 1,661 scans from COVID-19 positive patients and 6,095 from non-COVID-19 cases, totaling approximately 2.5 million images. All images have been anonymized, with 724,273 slices labeled as COVID-19 and 1,775,727 slices labeled as non-COVID-19. There is big variability in scan lengths due to factors such as required resolution and the specific features of the imaging equipment used [32]. To extend and disseminate this research, four COV19D Competitions have been organized in conjunction with workshops at ICCV 2021 [33], ECCV 2022 [14], ICASSP 2023 [34] and CVPR 2024 [10]. These competitions featured challenges on COVID-19 detection, severity assessment, and domain adaptation [35, 36], all leveraging the COV19-CT-DB database. In the presented results thereafter, we utilized the part of COV19-CT-DB [9, 33] used in the ECCV 2022 Competition [14, 37, 38]. This includes 1550 COVID-19 and 5044 non-COVID-19 3-D CT scans. The training part includes 882 COVID-19 and 1110 non-COVID-19 samples and its validation set 215 COVID-19 and 289 non-COVID-19 samples.

We also used the MosMedData database [39], which includes 856 COVID-19 CT-scans and 254 non-COVID samples (601 COVID-19 and 178 non-COVID-19 in the training set; 256 COVID-19 and 76 non-COVID-19 in the test set).

4.2. Performance Metric

The performance measure used to evaluate the models' performance in detecting Covid-19 is the average F1 Score (i.e., macro F1 Score) across all 2 categories (Covid-19 and non-Covid-19) :

$$\mathcal{P} = \frac{F_1^{Covid-19} + F_1^{non-Covid-19}}{2} \tag{9}$$

The F_1 score is a weighted average of the recall (i.e., the ability of the classifier to find all the positive samples) and precision (i.e., the ability of the classifier not to label as positive a sample that is negative). The F_1 score takes values in the range $[0, 1]$; high values are desired. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{10}$$

4.3. Implementation Details

We employed ViT-H, a variant of SAM, and ViT-L/14 trained in CLIP by OpenAI. The visually descriptive textual sentences were generated using GPT-3.5. All was implemented in PyTorch. All experiments were conducted on Tesla V100 32GB GPU.

Regarding implementation of RACNet: i) we used EfficientNetB0 as CNN model, stacking a global average pooling layer on top, a batch normalization layer and

Table 1. Performance comparison on the test set of COV19D-CT-DB (of the ECCV 2022 Competition) between RACNet and the state-of-the-art, when images are segmented with the conventional approach and when images are segmented with our proposed SAM2CLIP2SAM framework. F1 Score is given in %

COV19D-CT-DB (ECCV 2022)	F1		
	Macro	COVID	non-COVID
MDAP [40]	87.87	78.80	96.95
MDAP with SAM2CLIP2SAM	89.87	81.50	97.25
FDVTS [41]	89.11	80.92	97.31
FDVTS with SAM2CLIP2SAM	90.61	82.22	97.51
ACVLab [42]	89.11	80.78	97.45
ACVLab with SAM2CLIP2SAM	90.61	82.08	97.65
RACNet without segmentation	93.06	92.18	93.95
RACNet with conventional segmentation	95.06	94.18	95.95
RACNet with SAM2CLIP2SAM	**96.81**	**95.68**	**97.95**

dropout (with keep probability 0.8) [5]; ii) we used a single one-directional GRU layer consisting of 128 units as RNN model; iii) the first dense layer consisted of 128 hidden units. Regarding implementation details of RACNet training, batch size was equal to 5 (i.e, at each iteration our model processed 5 CT scans) and the input length 't' (see Figure 2) was 700 (the maximum number of slices found across all CT scans). Our model was fed with 3-D CT scans composed of CT slices; each slice was resized from its original size of 512×512 to 256×256. Loss function was cross entropy. Adam optimizer was used with learning rate 10^{-4}.

4.4. Experimental Results

In the following we provide an extensive experimental study comparing various networks and segmentation methods on the two above described databases.

Table 1 shows the performance comparison (in terms of F_1 Score) of RACNet (trained and tested on COV19D-CT-DB of the ECCV 2022 Competition) when its input is: i) unsegmented images; ii) segmented images with the conventional approach [43]; iii) segmented images with our proposed SAM2CLIP2SAM framework. It is evident that when segmentation is performed with our SAM2CLIP2SAM framework, RACNet achieves 3.75% and 1.75% superior performance to the cases when no segmentation is performed, or segmentation is performed with the conventional approach, respectively. This validates our notion that our segmentation approach enhances the classification model's feature extraction and assists it into focusing on only the important ROIs which are the left and right lungs. This also validates our observation that conventional segmentation approaches contain some noise and artifacts, whereas our method removes them.

Figure 3 illustrates this further. It presents three cases of unsegmented slices of a CT scan (left column), along with their cases when they are segmented with conventional approaches (middle column) and with our proposed framework SAM2CLIP2SAM (right column). It is evident that the segmentation result with our approach is more accurate and error-prune. In the first case (top row), the mediastinal mass between the left and right lungs is kept when the slices are segmented with conventional approaches, whereas it is not kept (i.e., it is black) when the slices are segmented with our SAM2CLIP2SAM framework. In all cases, one can also note that a bit of the pleural space (e.g. on the peripheral of the lungs) is also kept and is not masked when the slices are segmented with conventional approaches; this is not the case when the slices are segmented with our SAM2CLIP2SAM framework.

Additionally, Table 1 presents a comparison between the performance of RACNet to that of the state-of-the-art in the COV19D-CT-DB of the ECCV 2022 Competition. One can see that RACNet, especially when trained with CT scans segmented with our proposed SAM2CLIP2SAM framework, outperforms (in terms of F1 Score) all state-of-the-art methods by large margins (between 7.69% and 8.94%). Finally, Figure 1 shows that our proposed SAM2CLIP2SAM framework can be applied to the state-of-the-art methods as well, enhancing their performance by between 1.5% and 2% (most improvement in performance is seen for Covid-19 class, which is the most important).

Table 2 shows the performance comparison (in terms of F_1 Score) of RACNet (trained and tested on MosMedData) when its input is: i) unsegmented images; ii) segmented images with the conventional approach [43]; iii) segmented images with our proposed SAM2CLIP2SAM framework. It is evident that when segmentation is performed with our SAM2CLIP2SAM framework, RACNet achieves 3.43% and 1.43% superior performance to the cases when no segmentation is performed, or segmentation is performed with the conventional approach,

Original-Unsegmented Conventionally segmented SAM2CLIP2SAM segmented

Figure 3. Illustration of the improvement in segmentation quality when the CT scan slices are segmented with our proposed approach, the SAM2CLIP2SAM framework (right column) vs when they are segmented with conventional approaches (middle column).

Table 2. Performance comparison on the test set of MosMedData between RACNet and the state-of-the-art, when images are segmented with the conventional approach and when images are segmented with our proposed SAM2CLIP2SAM framework. F1 Score, Precision and Sensitivity are given in %

MosMedData	Precision	Sensitivity	F1
MC3_18 [44]	79.43	98.43	87.92
MC3_18 with SAM2CLIP2SAM	80.93	98.73	88.95
Densenet3D121 [45]	84.23	92.16	88.01
Densenet3D121 with SAM2CLIP2SAM	85.73	93.66	89.51
CovidNet3D [29]	79.50	98.82	88.11
CovidNet3D with SAM2CLIP2SAM	81.50	98.92	89.37
EMARS-APS [46]	93.52	90.59	92.03
EMARS-APS with SAM2CLIP2SAM	95.02	92.09	93.53
RACNet without segmentation	92.69	90.85	91.76
RACNet with conventional segmentation	94.69	92.85	93.76
RACNet with SAM2CLIP2SAM	**96.12**	**94.86**	**95.49**

respectively. Additionally, Figure 2 presents a comparison between the performance of RACNet to that of the state-of-the-art on MosMedData. One can see that RACNet, especially when trained with CT scans segmented with our proposed SAM2CLIP2SAM framework, outperforms (in terms of F1 Score) all state-of-the-art methods by large margins (between 16.69% and 2.6%). Finally, Figure 1 shows that our proposed SAM2CLIP2SAM framework can be applied to the state-of-the-art methods as well, enhancing their performance by between 1.03% and 2%.

5. Conclusions and Future Work

In this paper, we introduced a novel approach for COVID-19 diagnosis that leverages the RACNet deep neural architecture combined with vision-language models. Our method utilizes the SAM and CLIP models to perform precise segmentation of the right and left

lungs in CT scans. The segmented outputs are then fed into RACNet, which demonstrates enhanced accuracy in detecting COVID-19 and non-COVID-19 cases.

Our experimental results highlight the significant benefits of integrating vision-language models with deep learning architectures for medical image analysis. The proposed method shows promising potential for managing the variability of medical images collected from different healthcare institutions. Furthermore, this research defines a critical direction for future studies, including the exploration of uncertainty estimation [47] to further improve diagnostic performance.

The findings of this study suggest that the combination of advanced segmentation techniques and robust classification models can substantially enhance the accuracy and reliability of COVID-19 detection. Future work will focus on refining these techniques and exploring their applicability to other medical imaging tasks, with the aim of creating more efficient and scalable diagnostic frameworks.

References

[1] Tagaris, A., Kollias, D. and Stafylopatis, A. (2017) Assessment of parkinson's disease based on deep neural networks. In *Engineering Applications of Neural Networks: 18th International Conference, EANN 2017, Athens, Greece, August 25–27, 2017, Proceedings* (Springer): 391–403.

[2] Tagaris, A., Kollias, D., Stafylopatis, A., Tagaris, G. and Kollias, S. (2018) Machine learning for neurodegenerative disorder diagnosis—survey of practices and launch of benchmark dataset. *International Journal on Artificial Intelligence Tools* **27**(03): 1850011.

[3] Kollias, D., Vendal, K., Gadhavi, P. and Russom, S. (2023) Btdnet: A multi-modal approach for brain tumor radiogenomic classification. *Applied Sciences* **13**(21): 11984.

[4] Chowdhury, D., Das, A., Dey, A., Banerjee, S., Golec, M., Kollias, D., Kumar, M. *et al.* (2023) Covidetector: A transfer learning-based semi supervised approach to detect covid-19 using cxr images. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* **3**(2): 100119.

[5] Kollias, D., Tagaris, A., Stafylopatis, A., Kollias, S. and Tagaris, G. (2018) Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems* **4**(2): 119–131.

[6] Kollias, D., Vlaxos, Y., Seferis, M., Kollia, I., Sukissian, L., Wingate, J. and Kollias, S.D. (2020) Transparent adaptation in deep medical image diagnosis. In *TAILOR*: 251–267.

[7] Kollias, D., Bouas, N., Vlaxos, Y., Brillakis, V., Seferis, M., Kollia, I., Sukissian, L. *et al.* (2020) Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044* .

[8] Salpea, N., Tzouveli, P. and Kollias, D. (2022) Medical image segmentation: A review of modern architectures. In *European Conference on Computer Vision* (Springer): 691–708.

[9] Arsenos, A., Kollias, D. and Kollias, S. (2022) 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)* (IEEE): 1–5.

[10] Kollias, D., Arsenos, A. and Kollias, S. (2024) 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* .

[11] Kollias, D., Arsenos, A. and Kollias, S. (2023) 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)* (IEEE): 1–5.

[12] Arsenos, A., Davidhi, A., Kollias, D., Prassopoulos, P. and Kollias, S. (2023) Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* (IEEE): 1–5.

[13] Kollias, D., Arsenos, A. and Kollias, S. (2023) A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing* **542**: 126244.

[14] Kollias, D., Arsenos, A. and Kollias, S. (2023) Ai-mia: Covid-19 detection and severity analysis through medical imaging. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII* (Springer): 677–690.

[15] Gerogiannis, D., Arsenos, A., Kollias, D., Nikitopoulos, D. and Kollias, S. (2024) Covid-19 computer-aided diagnosis through ai-assisted ct imaging analysis: Deploying a medical ai system. *arXiv preprint arXiv:2403.06242* .

[16] Kollias, D., Arsenos, A., Soukissian, L. and Kollias, S. (2021) Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 537–544.

[17] Azad, R., Asadi-Aghbolaghi, M., Fathy, M. and Escalera, S. (2019) Bi-directional convlstm u-net with densley connected convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[18] Morani, K., Ayana, E.K., Kollias, D. and Unay, D. (2024) Covid-19 detection from computed tomography images using slice processing techniques and a modified xception classifier. *International Journal of Biomedical Imaging* **2024**(1): 9962839.

[19] Morani, K., Ayana, E.K., Kollias, D. and Unay, D. (2024) Detecting covid-19 in computed tomography images: A novel approach utilizing segmentation with unet architecture, lung extraction, and cnn classifier. In *Science and Information Conference* (Springer): 450–465.

[20] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T. *et al.* (2023) Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* : 3992–4003URL https://api.semanticscholar.org/CorpusID:257952310.

[21] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G. *et al.* (2021)

Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. [eds.] *Proceedings of the 38th International Conference on Machine Learning* (PMLR), *Proceedings of Machine Learning Research* **139**: 8748–8763. URL https://proceedings.mlr.press/v139/radford21a.html.

[22] Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N. and Zhang, Y. (2023) Segment anything model for medical image analysis: an experimental study. *Medical image analysis* **89**: 102918. URL https://api.semanticscholar.org/CorpusID:258236547.

[23] Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K. and O'Connor, N.E. (2023), Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. URL https://arxiv.org/abs/2307.11661. 2307.11661.

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M. *et al.* (2021), An image is worth 16x16 words: Transformers for image recognition at scale. URL https://arxiv.org/abs/2010.11929. 2010.11929.

[25] Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V. and Kaur, M. (2020) Classification of the covid-19 infected patients using densenet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics* : 1–8.

[26] Khadidos, A., Khadidos, A.O., Kannan, S., Natarajan, Y., Mohanty, S.N. and Tsaramirsis, G. (2020) Analysis of covid-19 infections on a ct image using deepsense model. *Frontiers in Public Health* **8**.

[27] Amyar, A., Modzelewski, R., Li, H. and Ruan, S. (2020) Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation. *Computers in Biology and Medicine* **126**: 104037.

[28] Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W. *et al.* (2020) A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. *IEEE transactions on medical imaging* **39**(8): 2615–2625.

[29] He, X., Wang, S., Chu, X., Shi, S., Tang, J., Liu, X., Yan, C. *et al.* (2021) Automated model design and benchmarking of deep learning models for covid-19 detection with chest ct scans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**: 4821–4829.

[30] Aleem, S., Wang, F., Maniparambil, M., Arazo, E., Dietlmeier, J., Curran, K., Connor, N.E. *et al.* (2024) Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 5184–5193.

[31] Kollias, D., Arsenos, A. and Kollias, S. (2023) A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing* **542**: 126244.

[32] Arsenos, A., Davidhi, A., Kollias, D., Prassopoulos, P. and Kollias, S. (2023) Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*: 1–5. doi:10.1109/ICASSPW59220.2023.10193437.

[33] Kollias, D., Arsenos, A., Soukissian, L. and Kollias, S. (2021) Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 537–544.

[34] Kollias, D., Arsenos, A. and Kollias, S. (2023) 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)*: 1–5. doi:10.1109/ICASSPW59220.2023.10193422.

[35] Kollias, D., Yu, M., Tagaris, A., Leontidis, G., Stafylopatis, A. and Kollias, S. Adaptation and contextualization of deep neural network models. In *2017 IEEE symposium series on computational intelligence (SSCI)* (IEEE): 1–8.

[36] Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W. and Qin, T. (2021) Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097* .

[37] Kollias, D., Bouas, N., Vlaxos, Y., Brillakis, V., Seferis, M., Kollia, I., Sukissian, L. *et al.* (2020) Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044* .

[38] Kollias, D., Vlaxos, Y., Seferis, M., Kollia, I., Sukissian, L., Wingate, J. and Kollias, S.D. (2020) Transparent adaptation in deep medical image diagnosis. In *TAILOR*: 251–267.

[39] Morozov, S.P., Andreychenko, A.E., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.P., Nikolaev, A.E., Pavlov, N.A. *et al.* (2020) Mosmeddata: data set of 1110 chest ct scans performed during the covid-19 epidemic. *Digital Diagnostics* **1**(1): 49–59.

[40] Turnbull, R. (2022) Cov3d: Detection of the presence and severity of COVID-19 from CT scans using 3D ResNets [Preliminary Preprint] URL https://doi.org/10.48550/arXiv.2207.12218.

[41] Hou, J., Xu, J., Feng, R. and Zhang, Y. (2022), Fdvts's solution for 2nd cov19d competition on covid-19 detection and severity analysis. doi:10.48550/ARXIV.2207.01758, URL https://arxiv.org/abs/2207.01758.

[42] Hsu, C.C., Tsai, C.H., Chen, G.L., Ma, S.D. and Tai, S.C. (2022), Spatiotemporal feature learning based on two-step lstm and transformer for ct scans. doi:10.48550/ARXIV.2207.01579, URL https://arxiv.org/abs/2207.01579.

[43] Salpea, N., Tzouveli, P. and Kollias, D. (2023) Medical image segmentation: A review of modern architectures. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII* (Springer): 691–708.

[44] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M. (2017), A closer look at spatiotemporal convolutions for action recognition. doi:10.48550/ARXIV.1711.11248, URL https://arxiv.org/abs/1711.11248.

[45] Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R. and Van Gool, L. (2017), Temporal 3d convnets: New architecture and transfer learning for video classification. doi:10.48550/ARXIV.1711.08200, URL https://arxiv.org/abs/1711.08200.

[46] He, X., Ying, G., Zhang, J. and Chu, X. (2022) Evolutionary multi-objective architecture search framework: Application to covid-19 3d ct classification. In Wang, L., Dou, Q., Fletcher, P.T., Speidel, S. and Li, S. [eds.] *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (Cham: Springer Nature Switzerland): 560–570.

[47] Arsenos, A., Kollias, D., Petrongonas, E., Skliros, C. and Kollias, S. (2024) Uncertainty-guided contrastive learning for single source domain generalisation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE): 6935–6939.