2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00100

Robust COVID-19 Detection in CT Images with CLIP

Li Lin[1+], Yamini Sri Krubha[1+], Zhenhuan Yang[2], Cheng Ren[3], Thuc Duy Le[4], Irene Amerini[5], Xin Wang[3], Shu Hu[1*]

[1]Purdue University, {lin1785, ykrubha, hu968}@purdue.edu
[2]Etsy, Inc, Brooklyn, New York, USA zhenhuan.yang@hotmail.com
[3]University at Albany, State University of New York {cren, xwang56}@albany.edu
[4]University of South Australia thuc.le@unisa.edu.au
[5]Sapienza University of Rome amerini@diag.uniroma1.it

***Abstract*—In the realm of medical imaging, particularly for COVID-19 detection, deep learning models face substantial challenges such as the necessity for extensive computational resources, the paucity of well-annotated datasets, and a significant amount of unlabeled data. In this work, we introduce the first lightweight detector designed to overcome these obstacles, leveraging a frozen CLIP image encoder and a trainable multilayer perception (MLP). Enhanced with Conditional Value at Risk (CVaR) for robustness and a loss landscape flattening strategy for improved generalization, our model is tailored for high efficacy in COVID-19 detection. Furthermore, we integrate a teacher-student framework to capitalize on the vast amounts of unlabeled data, enabling our model to achieve superior performance despite the inherent data limitations. Experimental results on the COV19-CT-DB dataset demonstrate the effectiveness of our approach, surpassing baseline by up to 10.6% in 'macro' F1 score in supervised learning. The code is available at https://github.com/Purdue-M2/COVID-19_Detection_M2_PURDUE.**

***Index Terms*—COVID-19, CLIP, Detection, Robust, CT Images**

I. Introduction

COVID-19 detection [2] based on 3-D chest CT scans is a diagnostic technique that uses computed tomography (CT) imaging to capture detailed images of the lungs and chest area in three dimensions. This method has been explored and utilized during the COVID-19 pandemic as a means to detect and assess the severity of infections caused by the SARS-CoV-2 virus. CT scans are particularly useful for visualizing the condition of the lungs and can help in identifying characteristic signs of COVID-19, such as ground-glass opacities and bilateral pulmonary lesions, which are not always visible on standard X-rays [24].

Deep learning has emerged as a pivotal technology in medical image analysis, demonstrating remarkable success in enhancing the detection and diagnosis of various diseases, including COVID-19 [27]. By leveraging complex neural network architectures, deep learning models can automatically learn from vast amounts of medical imaging data, such as X-rays, CT scans, and MRI images, to identify intricate patterns and anomalies that may elude human experts. In the context of COVID-19, deep learning algorithms have been particularly instrumental in analyzing 3-D chest CT scans, enabling rapid, accurate identification of viral infections and assessment of disease severity [5]. This capability has proved invaluable in managing the pandemic, as it assists healthcare professionals in making informed decisions quickly, optimizing patient management, and ultimately saving lives.

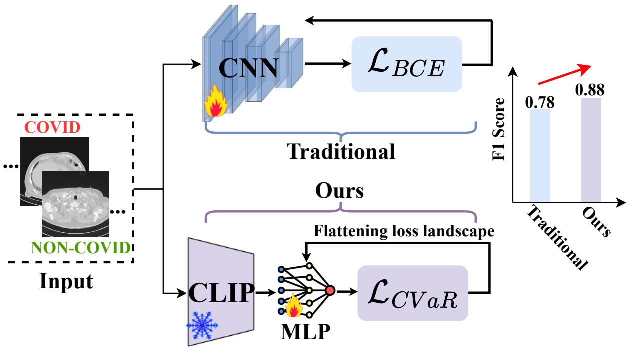

Fig. 1: *Comparison between our method with traditional method.* ***First row****: The traditional method trains a whole deep learning model* (*e.g., CNN*) *with a binary cross-entropy loss $\mathcal{L}_{BCE}$.* ***Second row****: Our method enhances COVID-19 detection by unitizing a frozen CLIP and a lightweight MLP classifier with Conditional Value at Risk (CVaR) loss $\mathcal{L}_{CVaR}$ across a flattened loss landscape.*

The integration of deep learning in medical imaging for COVID-19 detection not only exemplifies the potential of AI in healthcare but also paves the way for its broader application in diagnosing a wide range of pathologies, promising a future where medical diagnostics are more efficient, precise, and accessible [36]. However, the detection of COVID-19 using computational methods, especially through deep learning models, faces significant challenges that are critical to address for improving diagnostic accuracy and reliability. Firstly, the effectiveness of these models often hinges on the availability of extensive, high-quality datasets for training. However, real-world COVID-19 positive data is scarce, limiting the ability of these models to learn diverse manifestations of the disease [38]. This scarcity is compounded by the high variability in symptom presentation among different populations, making it difficult to develop models that are universally effective.

[+]Co-first Authors
[*]Corresponding Author

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00100

Secondly, the issue of unlabeled data represents a substantial hurdle. Many datasets consist of images that have not been annotated with diagnostic outcomes, rendering them unusable for supervised learning methods without considerable effort to label them accurately. Finally, improving detection performance with limited data while simultaneously enhancing model generalization requires innovative approaches. Addressing these challenges is vital, as it directly impacts the models' ability to accurately diagnose COVID-19 across diverse global populations and varying stages of the disease, thereby playing a crucial role in managing the pandemic effectively [40].

To address the challenges in COVID-19 detection due to limited and unlabeled data, existing methods have adopted several innovative strategies, yet they also encounter inherent limitations and gaps. To tackle the issue of limited data, techniques such as transfer learning have been widely used, where models pre-trained on vast, diverse datasets are fine-tuned using the smaller available COVID-19 datasets. This approach, however, may not always capture the unique characteristics of COVID-19-related anomalies due to the potential domain shift between the original and target datasets. For handling unlabeled data, semi-supervised and self-supervised learning methods have gained popularity. These methods leverage unlabeled data to learn general features or patterns, which can then be fine-tuned with a smaller labeled dataset for specific tasks. While effective to a degree, these methods can sometimes introduce biases or inaccuracies if the unlabeled data is not representative of the wider population or the specific nuances of COVID-19 pathology [39].

To improve generalization capabilities, existing approaches often employ data augmentation techniques to artificially expand the training dataset and introduce more variability, simulating a broader range of cases. Ensemble learning methods, which combine multiple models or predictions, are also used to enhance generalization. However, these strategies may still fall short when faced with highly diverse or novel cases outside of the training dataset's scope, highlighting a gap in the ability to robustly handle unseen variations. Overall, while existing methods have made significant strides in COVID-19 detection, challenges remain in ensuring high accuracy and generalizability across varied clinical settings and populations, underscoring the need for continuous innovation and validation in diverse real-world scenarios. These challenges also lead many competitions such as the 4^{th} COV19D Competition [3], [4], [17]–[23]

In this work, we propose a novel framework as depicted in Fig.2, comprising three modules: frozen Contrastive Language-Image Pre-training (CLIP) [32] ViT as feature extractor, trainable classifier MLP, and optimization. This straightforward framework is used for both supervised and semi-supervised learning to detect COVID-19 from CT scan images. Specifically, for *Supervised Learning*, we leverage CLIP ViT-L/14 [15] image encoder to capture high-level representations of the CT images. These representations are then fed into a 3-layer MLP trained to detect COVID-19 and non-COVID-19. We enhance the model's robustness and ability to focus on the most challenging cases by incorporating Conditional Value at Risk (CVaR) into the binary cross-entropy (BEC) loss. This is complemented by the optimization module, which enhances the model's generalizability by flattening the loss landscape. For *Semi-Supervised Learning*, we design a teacher-student framework in Fig.3 to capitalize on the abundance of unlabeled data. The teacher model, after training on annotated data, assigns pseudo-labels to the unlabeled dataset. This process augments the training set, which is then used to train the student model. Through this transfer of knowledge, the student model is equipped to potentially surpass the teacher in detecting COVID-19. Our contributions are summarized as follows:

1) We propose the first lightweight detector for exposing COVID-19 based on labeled 3-D CT scans.
2) We also propose a teacher-student framework for improving the COVID-19 detection performance by integrating unlabeled data.
3) Our method outperforms state-of-the-art approaches, as demonstrated in extensive experiments on the COV19-CT-DB dataset.

II. Related Work

A. COVID-19 Detection

According to [35], AI-empowered methods are employed for the detection of COVID-19 using medical images such as X-rays and CT scans. These include the preprocessing method and segmentation. Data preprocessing [1] involves resizing to 224x224 pixels, cropping for relevant regions, and sharpening filters to enhance edges, while data augmentation includes rescaling, zooming, flipping, and shearing to increase sample diversity and training robustness. The segmentation method proposed by [28] combines a DRD U-Net for image segmentation, integrating residual modules to enhance feature extraction, with a WGAN-based DNN classifier for efficient multi-class classification of COVID-19 images to train the classifier and optimize model parameters. Additionally, methods like transfer learning, fine-tuning, and novel architectures [37] are employed in this domain.Transfer learning [31] leverages pre-trained models like VGG-16, originally trained on large datasets like ImageNet, to classify COVID-19 CT-Scan images by fine-tuning some pre-trained layers, replacing the classifier layer, and utilizing features extracted by convolutional and pooling layers for classification.

B. CLIP

Contrastive Language-Image Pre-training (CLIP) [32], a simple yet effective pre-training paradigm, successfully introduces text supervision to vision models. It has shown promising results across various tasks in medical imaging (*i.e.*, classification [25], [41], detection [7], and segmentation [30]), attributable to its generalizability [32].

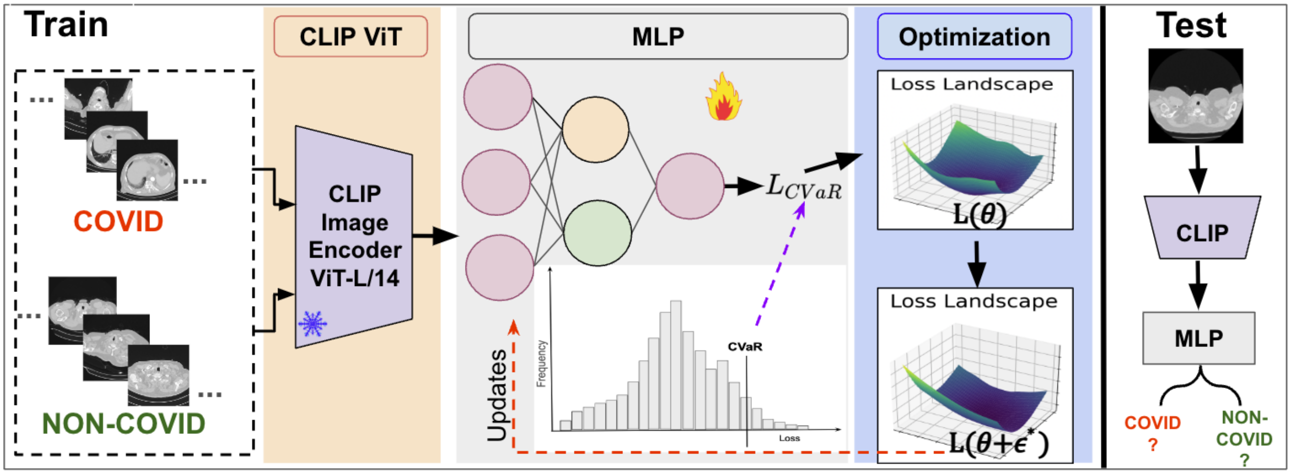

Fig. 2: *Overview of our proposed model using CLIP ViT for encoding the input images, an MLP module with robust CVaR loss, and an optimization step involving a flattened loss landscape for detecting COVID-19 cases apart from NON-COVID-19.*

III. Method

A. Supervised Learning

Feature Space Modeling. We propose a simple procedure to tell apart COVID-19 CT scan series from non-COVID-19 based on features extracted from the image encoder of CLIP ViT L/14 [15]. CLIP ViT is trained on an extraordinarily large dataset of 400M image-text pairs, so the high-level feature extracted from it is sufficient exposure to the visual world. Additionally, since ViT L/14 has a smaller starting patch size of 14 × 14 (compared to other ViT variants), we believe it can also aid modeling the low-level CT-scan slice details (*i.e.*, Ground-Glass Opacities and Consolidation) needed for COVID-19 VS non-COVID-19 classification. Given a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$ with size n, where X_i is the i-th CT scan slice and $Y_i \in \{0, 1\}$ is the i-th sample label (0 means non-COVID-19, 1 means COVID-19). Feed CLIP visual encoder with dataset $\mathcal{D}$, and use its final layer to map the training data to their feature representations (of 768 dimensions). We get the resulting feature bank $\mathcal{C} = \{(F_i, Y_i)\}_{i=1}^{n}$ and further use the feature bank, which is our training set, to design an MLP classifier.

MLP Classifier. After constructing the feature bank, we use those feature embeddings to train a binary classifier to detect COVID-19 and non-COVID-19 CT scan slices. Our classifier is a straightforward 3-layer Multilayer Perceptron (MLP), and to foster a stable learning process and enhance the model's ability to generalize, we incorporate batch normalization after each linear transformation. This is followed by a ReLU activation, allowing for the model to capture intricate data patterns effectively. To further combat the risk of overfitting, a dropout layer is included following the activation function.

Objective Function. To obtain a robust model, we apply a distributionally robust optimization (DRO) technique called *Conditional Value-at-Risk* (CVaR) [8], [10]–[14], [16], [26], [29], [33], [34]. By integrating CVaR into the binary cross-entropy (BEC) loss, the model is encouraged to pay more attention to the riskiest predictions. In the context of COVID-19 detection, these could be cases where the model is most uncertain and where misclassification could lead to the worst outcomes, such as failing to detect COVID-19 in patients with the disease. To this end, in what follows, we assume that $\mathcal{C} = \{(F_i, Y_i)\}_{i=1}^{n}$ consists of i.i.d. samples from a joint distribution $\mathbb{P}$, F_i is the i-th data point's feature, and Y_i is the i-th point's label. Given some variant of minibatch gradient descent, in the COVID-19 detection task, we are minimizing the empirical risk of the loss $\mathcal{L}_{avg}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, F_i, Y_i)$ for $\theta \in \Theta$, instead of minimizing the true unknown risk $\mathcal{R}_{avg}(\theta) = \mathbb{E}_{(F,Y)\sim\mathbb{P}}[\ell(\theta; F, Y)]$, where ℓ is the individual loss function (*e.g.*, binary cross-entropy) of the COVID-19 detection model, which has parameters θ for MLP.

However, the average loss is not robust to the imbalanced data, which is a common charismatic of the existing COVID-19 datasets. In addition, the training data distribution is usually not consistent with the testing data distribution, which is called the domain shift problem that widely exists in real-world COVID-19 detection scenarios. For example, the training set may be from one hospital but the test data may be from a different hospital. Therefore, we explore a DRO technique CVaR for handling imbalanced data, which can be formulated as: $\text{CVaR}_\alpha(\theta) = \inf_{\lambda\in\mathbb{R}} \left\{\lambda + \frac{1}{\alpha}\mathbb{E}_{(F,Y)\sim P}\left[(\ell(\theta; F, Y) - \lambda)_+\right]\right\}$, where $[a]_+ = \max\{0, a\}$ is the hinge function, the conditional value at risk at level $\alpha \in (0, 1)$. As $\alpha \to 0$, we are concerned about minimizing the risk of 'hard' samples (cases that are difficult to diagnose). In contrast, as $\alpha \to 1$, it becomes minimizing the $\mathcal{R}_{avg}(\theta)$. Inspired by [42], we can minimize a loss function that aims to minimize an upper bound on the worst-case risk by employing the CVaR. In practice, we minimize an empirical version of $\text{CVaR}_\alpha(\theta)$. This gives us the following optimization problem:

$$\mathcal{L}_{CVaR}(\theta) = \min_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\alpha n}\sum_{i=1}^{n}[\ell(\theta; F_i, Y_i) - \lambda]_+ . \quad (1)$$

Suppose for a moment that we have obtained the optimal

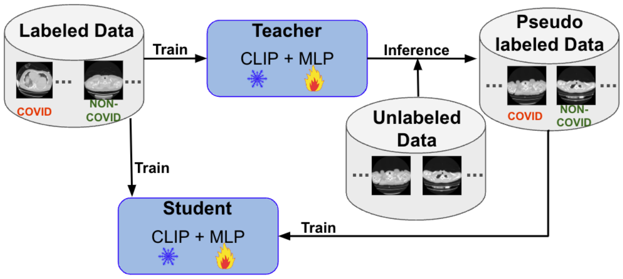

Fig. 3: *Diagrammatic representation of our robust model with teacher-student framework by leveraging unlabeled data for enhancing detection performance.*

value of λ^* in (1), then the only training points that contribute to the loss are the 'hard' ones with a loss value greater than λ^*, whereas the 'easy' training points with low loss smaller than λ^* are ignored. To this end, a robust model is obtained. These procedures are demonstrated in Fig.2.

B. Semi-Supervised Learning

In addressing the challenge of limited labeled data for COVID-19 diagnosis from imaging, we propose a semi-supervised learning framework [9] that leverages knowledge distillation to utilize unlabeled data effectively. Our methodology comprises two primary phases: the teacher model training and the student model training. Note the model we use here is the same architecture as in supervised learning, both are fixed CLIP with a trainable 3-layer MLP. The whole training framework is illustrated in Fig.3. We explain each module as follows.

Teacher Model Training. The teacher model is first trained on a small set of labeled data. The aim of this phase is to develop a robust initial model that captures the high-level features and complexities associated with the diagnosis from the labeled dataset.

Pseudo-Labeling with Teacher Model. Once the teacher model is trained, it is employed to perform inference on a larger corpus of unlabeled data. The model's predictions are used to assign pseudo-labels to the unlabeled CT scan slices, creating a new set of training data that, while not verified by human experts, carries the inferred knowledge of the teacher model. These pseudo-labels are subject to uncertainty but offer a valuable starting point for expanding the training dataset beyond the limits of the labeled data.

Student Model Training. Subsequently, a student model is trained on a combination of the original labeled data and the newly created pseudo-labeled data. This process allows the student model to learn from both the ground truth in the labeled data and the nuanced patterns inferred by the teacher in the unlabeled data. The student model, through this extended training, is expected to outperform the teacher model by generalizing better to unseen data, thanks to the larger and more diverse training set it has been exposed to.

```
Algorithm 1: Optimization (CVaR+SAM)
Input: A training dataset $\mathcal{C}$ with size $n$, $\alpha$, $\gamma$
       max_iterations, num_batch, learning rate $\beta$
Output: A robust COVID-19 detection model
Initialization: $\theta_0$, $l = 0$
for $e = 1$ to max_iterations do
    for $b = 1$ to num_batch do
        Sample a mini-batch $\mathcal{C}_b$ from $\mathcal{C}$
        Compute $\ell(\theta_l; F_i, X_i)$, $\forall (F_i, Y_i) \in \mathcal{C}_b$
        Use binary search to find $\lambda$ that minimizes (1)
         on $\mathcal{C}_b$
        Compute $\epsilon^*$ based on Eq. (2)
        Compute gradient approximation for (3)
        Update $\theta$: $\theta_{l+1} \leftarrow \theta_l - \beta \nabla_\theta \mathcal{L}_{CVaR}\big|_{\theta_l + \epsilon^*}$
        $l \leftarrow l + 1$
    end
end
return $\theta_l$
```

C. Optimization

Last, to further improve the detector's generalization capability, we optimize the COVID-19 detection model by utilizing the sharpness-aware minimization (SAM) method [6] to flatten the loss landscape. Note that this optimization module can be used in both supervised learning and semi-supervised learning. As shown in Fig.2, by utilizing such a technique, the model yields a more flattened loss landscape indicating a stronger generalization capability [29]. As a reminder, the model's parameters are denoted as θ, flattening is attained by determining an optimal ϵ^* for perturbing θ to maximize the loss, formulated as:

$$\begin{aligned} \epsilon^* &= \arg \max_{\|\epsilon\|_2 \le \gamma} \mathcal{L}_{CVaR}(\theta + \epsilon) \\ &\approx \arg \max_{\|\epsilon\|_2 \le \gamma} \epsilon^\top \nabla_\theta \mathcal{L}_{CVaR} = \gamma \mathtt{sign}(\nabla_\theta \mathcal{L}_{CVaR}), \end{aligned} \tag{2}$$

where γ is a hyperparameter that controls the perturbation magnitude. The approximation is obtained using first-order Taylor expansion with the assumption that ϵ is small. The final equation is obtained by solving a dual norm problem, where $\mathtt{sign}$ represents a sign function and $\nabla_\theta \mathcal{L}_{CVaR}$ being the gradient of $\mathcal{L}_{CVaR}$ with respect to θ. As a result, the model parameters are updated by solving the following problem:

$$\min_\theta \mathcal{L}_{CVaR}(\theta + \epsilon^*). \tag{3}$$

Perturbation along the gradient norm direction increases the loss value significantly and then makes the model more generalizable while detecting COVID-19.

End-to-end Training. In practice, we first initialize the model parameters θ and then randomly select a mini-batch set C_b from C, performing the following steps for each iteration on C_b (see Algorithm 1):

- Fix θ and use binary search to find the global optimum of λ since (1) is convex w.r.t. λ.
- Fix λ, compute ϵ^* based on Eq. (2).
- Update θ based on the gradient approximation for (3): $\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{CVaR}\big|_{\theta+\epsilon^*}$, where β is a learning rate.

IV. Experiments

A. Experimental Settings

1) Datasets: The COV19-CT-DB dataset, as referenced in [19], forms the basis of our study, comprising 3-D chest CT scans. This collection encompasses 7,756 3-D CT scans, with 1,661 being COVID-19 positive and 6,095 being negative for COVID-19. It aggregates to approximately 2,500,000 images, of which 724,273 images are categorized under the COVID-19 class and 1,775,727 images under the non-COVID-19 class. Our analysis employs slices, meaning 2-D images derived from these scans, with each scan series containing 50 to 700 slices. Each slice has a resolution of 512×512. This dataset also features in the 4^{th} COV19D Competition as documented in [20]. Based on the guidelines from [20], our approach involves: (a) using 703 3-D COVID-19 and 655 3-D non-COVID-19 CT scans from the dataset for supervised learning training, while reserving 170 3-D COVID-19 and 156 3-D non-COVID-19 CT scans for testing; (b) employing a semi-supervised learning framework, which includes CT scans obtained from various hospitals and medical facilities to ensure data diversity. In addition to the labeled scans used for supervised learning, our training set is augmented with 239 additional annotated 3-D CT scans (120 COVID-19 and 119 non-COVID-19) and 494 3-D CT scans without annotations, with the test set comprising 178 3-D CT scans (65 COVID-19 and 113 non-COVID-19). Further dataset details are elaborated in [20].

2) Evaluation Metrics: In line with the evaluation protocol established in [20], we adopt the macro F1 score as our primary metric for assessing the performance of all methods. This metric is essentially the unweighted mean of the F1 scores across different classes or labels, such as averaging the F1 scores for both the COVID-19 and non-COVID-19 categories. Given that our models are designed to classify individual images or slices rather than entire CT scans, we employ a majority voting strategy to aggregate the slice-level predictions into a singular diagnostic outcome for each CT scan. This approach allows us to compile the discrete predictions across all slices from a specific CT scan to arrive at a consolidated final diagnosis, thereby aligning our methodology with the comprehensive evaluation framework referenced in [20].

3) Baseline Methods: The baseline methods for our study are sourced from [20], which outlines two primary approaches: (a) Within the supervised learning framework, we process input 3-D CT scans by applying padding to standardize their dimensions, which then proceed to a Convolutional Neural Network (CNN) segment for initial analysis. This CNN component is tailored to extract pivotal features, predominantly from lung areas, on an individual 2D slice basis. Following this feature extraction phase, a Recurrent Neural Network (RNN) sequentially processes the CNN-derived features, directing them through a Fully Connected (FC) layer and culminating in a softmax activation-based output layer dedicated to COVID-19 classification. A Dropout layer is integrated within this architecture to mitigate the risk of overfitting. This model is denoted as **CNN+RNN**. (b) The semi-supervised learning strategy incorporates Monte Carlo Dropout to gauge uncertainty levels during the training phase with labeled data. This uncertainty measurement guides the annotation of unlabeled data, especially highlighting COVID-19 instances where the model demonstrates substantial confidence in its predictions. This technique is referred to as **Dropout**.

	Method	'macro' F1 Score
Supervised	CNN+RNN [20]	0.780
	Ours	**0.886**
Semi-Supervised	Dropout [20]	0.730
	Ours	**0.734**

TABLE I: *Comparison with the baseline method in terms of 'macro' F1 score under supervised and semi-supervised learning, respectively. The best results are shown in **Bold**.*

4) Implementation Details: We adopt the CLIP framework, incorporating the Vision Transformer (ViT) as the image processing unit, specifically configured to the L/14 scale. This setup is augmented with three Multi-Layer Perceptron (MLP) layers, each consisting of 768 neurons, serving as our primary computational model. Our training protocol is executed with a batch size of 32, ensuring a precise management of samples during each training iteration.

The optimization process is facilitated by the Adam optimizer, which kicks off with an initial learning rate set at $\beta = 1e-3$. Additionally, we employ a Cosine Annealing Learning Rate Scheduler to modulate the learning rate adaptively across the training duration, aiming to bolster the model's path to convergence. The tuning of hyperparameters involves adjusting α within the range of $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ as delineated in Eq. (1), alongside setting γ to 0.05 as specified in Eq. (2). These experiments are conducted using the PyTorch framework and leverage the computational prowess of an NVIDIA RTX A6000 GPU for training purposes.

B. Results

Table I shows our results compared with the baseline method CNN+RNN [20] for two primary approaches: *supervised learning* and *semi-supervised learning*. It is clear that, in supervised learning, our method has superior COVID-19 detection ability compared to CNN+RNN [20]. It enhances the 'macro' F1 score by **10.6%**. When applied to semi-supervised learning, our method slightly outperforms the baseline. This is because our method simplifies the semi-supervised learning process by employing a direct pseudo-labeling approach without the added complexity of Monte Carlo Dropout for uncertainty estimation. Despite this simplification, our method achieves a higher 'macro' F1 score. This indicates that the quality of pseudo-labels generated by our model is high, and our model is particularly effective at identifying and learning from the most informative unlabeled instances.

C. Ablation Study

Visualization of Loss Landscape. Fig.4 visually illustrates the impact of incorporating SAM optimization in our proposed method. The loss landscape, without SAM, shows a more

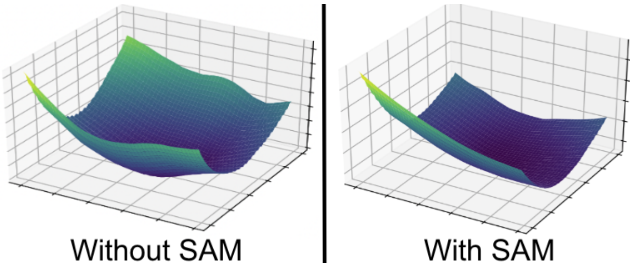

Fig. 4: *The loss landscape visualization of our proposed method without (left) and with (right) using the sharpness-aware minimization (SAM) method. The axis's scales are the same for both figures.*

Method	BCE	BCE+SAM	CVaR	Ours(CVaR+SAM)
'macro' F1 score	0.868	0.874	0.877	**0.886**

TABLE II: *An ablation study of our key components. 'BCE' denotes cross-entropy loss, 'BCE+SAM' represents BCE loss with the sharpness-aware minimization (SAM) optimization, 'CVaR' and 'Ours' represent Conditional Value at Risk without SAM and with SAM, respectively.*

rugged and uneven loss surface. This unevenness can make the optimization process challenging, as it may lead to inconsistent generalization. In contrast, the right side reveals a much smoother loss landscape when SAM is applied. The smoother surface indicates a more robust model with parameters that generalize better to new data. The consistency in the loss surface with SAM also suggests that the optimization process is more straightforward, leading to improved learning during training. This visualization underscores the significance of the optimization module in our method for enhancing the detector's generalization.

Effects of CVaR and SAM. The results in Table II reveal the effects of CVaR technique and SAM optimization we applied in our proposed method. Compared with 'BCE', 'BCE+SAM' improves the 'macro' F1 score by 0.6%, 'CVaR' enhances performance by 0.9%, indicating the effectiveness of SAM optimization and CVaR loss, respectively. When we incorporate CVaR with SAM, it achieves the best performance surpassing 'BCE' by 1.8%. Overall, our method with both CVaR and SAM optimization yields the most substantial gains in 'macro' F1 score.

D. Sensitive Analysis

Fig.5 shows the 'macro' F1 score to different α values in Eq. (1). The F1 scores are significantly lower when α is 0.1, 0.2, and 0.3. This is because the model focuses on the worst-case outcomes (extreme risks), it is either not predicting the positive class at all or predicting the negative class (F1 positive is 0 or F1 negative is 0). It shows a steep increase in the F1 Score at an α value of 0.4, indicating that the model's performance improves drastically when it moves away from the most extreme risk assessments to slightly more moderate ones. At this point, the balance between precision and recall that the F1 Score represents is much more favorable.

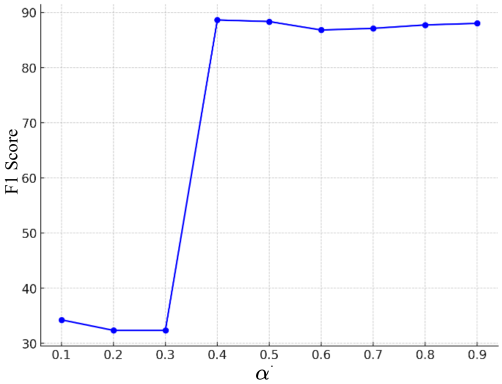

Fig. 5: *'Macro' F1 score to different α values.*

V. Conclusion

The current landscape of COVID-19 detection through deep learning models is faced with various challenges: requires large computational resources, the scarcity of high-quality, labeled datasets, and the abundance of unlabeled data. Addressing these issues, we introduce a streamlined detector that employs a frozen CLIP image encoder and a trainable MLP, augmented with CVaR and a loss landscape flattening strategy. The CVaR integration bolsters our model's robustness, while the loss flattening strategy enhances generalization. Moreover, our teacher-student framework adeptly leverages unlabeled data, ensuring effective model training even with limited labeled data. Experimental results showcase the superiority of our method compared with the baseline.

Limitation. A notable limitation of our study is that our proposed methodologies overlook the inherent correlations among CT images derived from the same 3-D CT scan. This oversight potentially results in the omission of valuable information that could enhance the diagnostic accuracy of our models.

Future Work. We plan to apply the CLIP text encoder by utilizing the medical diagnosis report data with the COVID-19 CT images to improve the detector's performance further.

Acknowledgments. This work is supported by the U.S. National Science Foundation (NSF) under grant IIS-2348419 and the National Artificial Intelligence Research Resource (NAIRR) Pilot and TACC Lonestar6. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of NSF and NAIRR Pilot.

References

[1] Khabir Uddin Ahamed, Manowarul Islam, Ashraf Uddin, Arnisha Akhter, Bikash Kumar Paul, Mohammad Abu Yousuf, Shahadat Uddin, Julian MW Quinn, and Mohammad Ali Moni. A deep learning approach using effective preprocessing techniques to detect covid-19 from chest ct-scan and x-ray images. *Computers in biology and medicine*, 139:105014, 2021.

[2] Moutaz Alazab, Albara Awajan, Abdelwadood Mesleh, and Salah Alhyari. Covid-19 prediction and detection using deep learning. *International Journal of Computer Information Systems and Industrial Management Applications*, 12:14–14, 2020.

[3] 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.

[4] 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.

[5] Zhenghan Fang, Junjie Bai, Xinyu Guo, Xin Wang, Feng Gao, Hao-Yu Yang, Bin Kong, Ying Hou, Kunlin Cao, Qi Song, et al. Annotation-efficient covid-19 pneumonia lesion segmentation using error-aware unified semisupervised and active learning. *IEEE Transactions on Artificial Intelligence*, 4(2):255–267, 2022.

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

[7] Miaotian Guo, Huahui Yi, Ziyuan Qin, Haiying Wang, Aidong Men, and Qicheng Lao. Multiple prompt fusion for zero-shot lesion detection using vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–292. Springer, 2023.

[8] Shu Hu and George H Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, pages 62–87. PMLR, 2022.

[9] Shu Hu et al. Pseudoprop: Robust pseudo-label generation for semi-supervised object detection in autonomous driving systems. In *CVPR Workshop*, pages 4390–4398, 2022.

[10] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu. Tkml-ap: Adversarial attacks to top-k multi-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7649–7657, 2021.

[11] Shu Hu, Xin Wang, and Siwei Lyu. Rank-based decomposable losses in machine learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] Shu Hu, Zhenhuan Yang, Xin Wang, Yiming Ying, and Siwei Lyu. Outlier robust adversarial training. In *Asian Conference on Machine Learning*, pages 454–469. PMLR, 2024.

[13] Shu Hu, Yiming Ying, Siwei Lyu, et al. Learning by minimizing the sum of ranked range. *Advances in Neural Information Processing Systems*, 33:21013–21023, 2020.

[14] Shu Hu, Yiming Ying, Xin Wang, and Siwei Lyu. Sum of ranked range loss for supervised learning. *Journal of Machine Learning Research*, 23(112):1–44, 2022.

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open clip. https://github.com/mlfoundations/open_clip, 2021.

[16] Yan Ju, Shu Hu, Shan Jia, George H Chen, and Siwei Lyu. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4655–4665, 2024.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Ameer Sardar Kwekha-Rashid, Heamn N Abduljabbar, and Bilal Alhayani. Coronavirus disease (covid-19) cases analysis using machine-learning applications. *Applied Nanoscience*, 13(3):2013–2025, 2023.

[25] Yiming Lei, Zilong Li, Yan Shen, Junping Zhang, and Hongming Shan. Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–412. Springer, 2023.

[26] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[27] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: evaluation of the diagnostic accuracy. *Radiology*, 296(2):E65–E71, 2020.

[28] Luoyu Lian, Xin Luo, Canyu Pan, Jinlong Huang, Wenshan Hong, and Zhendong Xu. Lung image segmentation based on drd u-net and combined wgan with deep neural network. *Computer Methods and Programs in Biomedicine*, 226:107097, 2022.

[29] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. *CVPR*, 2024.

[30] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.

[31] Nurdina Gita Pratiwi, Yumna Nabila, Rian Fiqraini, and Agung W Setiawan. Effect of ct-scan image resizing, enhancement and normalization on accuracy of covid-19 detection. In *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 17–22. IEEE, 2021.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[34] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[35] Faisal Muhammad Shah, Sajib Kumar Saha Joy, Farzad Ahmed, Tonmoy Hossain, Mayeesha Humaira, Amit Saha Ami, Shimul Paul, Md Abidur Rahman Khan Jim, and Sifat Ahmed. A comprehensive survey of covid-19 detection using medical images. *SN Computer Science*, 2(6):434, 2021.

[36] Feng Shi, Jun Wang, Jun Shi, Ziyan Wu, Qian Wang, Zhenyu Tang, Kelei He, Yinghuan Shi, and Dinggang Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19. *IEEE reviews in biomedical engineering*, 14:4–15, 2020.

[37] Nandhini Subramanian, Omar Elharrouss, Somaya Al-Maadeed, and Muhammed Chowdhury. A review of deep learning-based detection methods for covid-19. *Computers in Biology and Medicine*, 143:105233, 2022.

[38] Xin Wang, YIN Youbing, KONG Bin, Yi Lu, Junjie Bai, Zhenghan Fang, and Qi Song. Method and system for diagnosis of covid-19 using artificial intelligence, August 3 2021. US Patent 11,076,824.

[39] Xin Wang, YIN Youbing, KONG Bin, Yi Lu, Junjie Bai, Zhenghan Fang, and Qi Song. Method and system for diagnosis of covid-19 disease progression using artificial intelligence, June 14 2022. US Patent 11,361,440.

[40] Xin Wang, YIN Youbing, KONG Bin, Yi Lu, Junjie Bai, Zhenghan Fang, and Qi Song. Method and system for diagnosis of covid-19 using artificial intelligence, June 14 2022. US Patent 11,357,464.

[41] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022.

[42] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355. PMLR, 2021.