A Closer Look at Spatial-Slice Features Learning for COVID-19 Detection

[†1]Chih-Chung Hsu, [‡1]Chia-Ming Lee, [2]Yang Fan Chiang, [1]Yi-Shiuan Chou,
[1]Chih-Yu Jiang,[1]Shen-Chieh Tai, [1]Chi-Han Tsai
[1]Institute of Data Science, National Cheng Kung University, Taiwan
[2]Department of Electrical Engineering, National Cheng Kung University, Taiwan
[†]cchsu@gs.ncku.edu.tw, [‡]zuw408421476@gmail.com

Abstract

Conventional Computed Tomography (CT) imaging recognition faces two significant challenges: ***(1) There is often considerable variability in the resolution and size of each CT scan****, necessitating strict requirements for the input size and adaptability of models.* ***(2) CT-scan contains large number of out-of-distribution (OOD) slices.*** *The crucial features may only be present in specific spatial regions and slices of the entire CT scan. How can we effectively figure out where these are located? To deal with this, we introduce an enhanced* ***S****patial-****S****lice* ***F****eature* ***L****earning* ***(SSFL++)*** *framework specifically designed for CT scan. It aims to filter out OOD data within the entire CT scan, enabling us to select crucial spatial slices for analysis by reducing 70% redundancy totally. Meanwhile, we proposed* ***K****ernel-****D****ensity-based slice* ***S****ampling* ***(KDS)*** *method to improve the stability during training and inference stage, therefore speeding up the rate of convergence and boosting performance. As a result, the experiments demonstrate the promising performance of our model using a simple EfficientNet-2D (E2D) model, even with only 1% of the training data. The efficacy of our approach has been validated on the COVID-19-CT-DB datasets provided by the DEF-AI-MIA workshop, in conjunction with CVPR 2024. Our code is available at https://github.com/ming053l/E2D.*

1. Introduction

Computed Tomography (CT) [53]has become essential in detecting and managing diseases. This technology excels at revealing abnormalities within the body, such as ground-glass opacities and bilateral patchy shadows, which are crucial for the early detection and monitoring of diseases. In diagnosing COVID-19, doctors rely on analyzing lung CT scans of patients. However, since a single patient's CT scan can include hundreds of images, manual examination becomes a time-consuming task, especially when doctors have to evaluate CT scans from dozens or hundreds of patients. This may result in false negatives when dealing with numerous scans.

With the rapid development of deep learning (DL), DL methods [17, 18, 25, 26, 48, 51, 63] have gained prominence for their ability to quickly and accurately identify COVID-19 features while efficiently handling large volumes of data. Furthermore, convolution neural networks (CNNs) have proven to be more effective than methods based on frequency-domain [49, 68] and low-level features for CT image analysis [41].

To address the terribly spreading COVID-19, Kolliaz *et al.* proposed the COVID-19-CT-DB dataset [2, 3, 34–39], which encompasses a vast amount of labeled COVID-19 and non-COVID-19 data, advancing the DL methodology and tackling the challenge faced by the huge requirement of high quality dataset for DL-based analysis. Many researchers have designed several methods to deal with COVID-detection task [11, 29, 30, 66].

Despite the effectiveness of CT imaging as a tool for detecting abnormalities, it suffers from varying resolutions and quality due to different data servers and scanning machines. The resolution and number of slices in CT images can differ based on the specific scanning machine used, potentially compelling researchers to devise more complex network architectures. Additionally, medical analysis for COVID-19, unlike typical DL-based tasks that focus solely on performance and applications, necessitates maintaining the explainability of model predictions for security and safety reasons [11, 12, 47].

Inspired by [57], Tran *et al.* presented that factorizing the 3D convolution filters (R3D) into separate spatial and temporal components (R(2+1)D) can yielding significantly gains in accuracy for action recognition. Its effectiveness have been demonstrated by several works on the fields of Video Understanding (VU) [6, 20, 42, 44] and Human Action Recognition (HAR) [58, 64]. One video may contains huge redundant information, such as noise from the audio track or each frame, and meaningless background, these factors make it difficult to train the model well [7], resulting

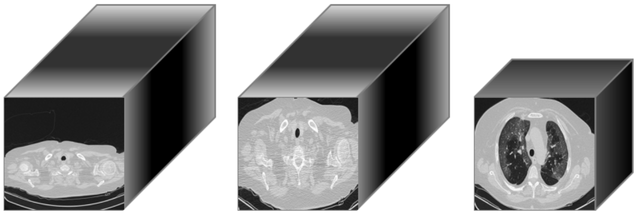

Figure 1. The brief illustration for SSFL++. It aim to reduce redundancy in spatial and slice dimension on whole CT-scan to improve model and data quality. (1)Left: original CT-scan. (2)Middle: after reduction at spatial. (3)Right: after reduction at slices.

in a significant increase in potential costs for data collecting. Likewise, CT scans can be regarded as a special case of video, it also contains various noise resulted from machine aging, and non-important spatial-slice pattern due to its imaging process [53]. Therefore, the different convolution methods on CT-scan is worthy of discussion.

In this work, we introduce a **Spatial-Slice Feature Learning (SSFL++)** method, an unsupervised approach designed to reduce computational complexity by effectively removing out-of-distribution (OOD) slices and redundant spatial information. Furthermore, previous works [11, 30] have struggled to identify the most influential slices while considering global sequence information. Based on this observation, we believe there is room for improvement. Therefore, we propose the **Kernel-Density-based Slice Sampling (KDS)** strategy, which leverages Kernel Density Estimation to simultaneously achieve both objectives. Experimental results have demonstrated our 2D model's outstanding performance, even in the face of data insufficiency.

Our novelties and contributions can be briefly divided into two parts as following mentions:

- **Improved spatial-slice feature learning module:** SSFL++ is a morphology-based approach for CT scans that removes redundant areas in both spatial and slice dimensions. This significantly reduces computational complexity and efficiently identifies the Regions of Interest (RoI) without the need for complicated designs or configurations. Remarkably, we were able to eliminate 70% of the area in the COVID-19-CT-DB datasets without any degradation in performance.
- **The comparison between 2D, (2+1)D, and 3D for CT-scan is discussed:** To facilitate the development of related research, we conducted a thorough discussion on the use of 2D, 2+1D, and 3D convolutions for CT scan data in COVID-19 detection. Based on experimental results, we believe that the 2D convolutional architecture holds more potential for future applications compared to 3D and 2+1D convolutions.
- **Density-aware slice sampling method:** Coupled with SSFL++'s ability to adaptively remove redundant spatial areas and slices, KDS further adaptively samples the most crucial slices while preserving global sequence information. This approach enhances data efficiency and strengthens the model's few-shot capabilities. Experimental results have shown that our E2D model maintains strong and robust performance under scenarios with few CT scans and slices.

2. Related Work

In this section, we introduce the related works on COVID-19 recognition in recent years, along with traditional spatial-temporal feature learning for Video Understanding (VU) and Human Action Recognition (HAR). The philosophy behind these approaches is important for analyzing CT-Scans.

2.1. Region of Interests for Computed Tomography

Background. CT [53] harnesses X-rays, which encircle a specific plane of the human body, while detectors on the opposite side capture the resultant signals. This technique exploits the differential attenuation of X-rays by various tissues, combined with signals obtained from multiple irradiation angles traversing the body, to compile a sinogram. This sinogram facilitates the reconstruction of cross-sectional imagery [4, 5]. Nonetheless, the CT imaging paradigm, necessitating multi-angular signal acquisition for reconstruction, engenders scans replete with extraneous data, potentially escalating labor costs.

Although this technology has been around for a long time, designing a robust and reliable Region of Interest (RoI) selection algorithm for CT scans remains an open problem. Noise and redundancy harm model performance. In recent years, most researchers have still focused on enhancing the feature extraction pipeline [45], or improving the quality of image reconstruction [27], to address the aforementioned challenges. Cobo *et al.* [14] suggested that standardizing medical imaging workflows could improve the performance of radiomics and deep learning systems. Jensen *et al.* [32] proposed enhancing the stability of CT radiomics with parametric feature maps. Gaidel *et al.* [23] introduced a greedy forward selection-based method for lung CT images, but its development was limited due to a lack of robustness against data shifting and noise.

2.2. COVID-19 Recognition

In recent years, substantial progress has been achieved in developing methods for COVID-19 recognition. Kollias *et al.* [34] have contributed to this field by analyzing the prediction results of deep learning models based on latent representations. Chen *et al.* [11] integrated maximum likelihood estimation with the Wilcoxon test, adopting a statistical learning perspective to adaptively select slices and design models with explainability.

Furthermore, Hou *et al.* proposed a method based on contrastive learning to enhance feature representation. Turnbull *et al.* applied a 3D ResNet [28] for COVID-19 severity classification. Hsu *et al.* [29] introduced a two-step model that combines 2D feature extraction with an LSTM [19] and Vision Transformer [16]. They presented a 2D and (2+1)D approach [30], achieving outstanding results in the AI-MIA 2023 COVID-19 detection competition.

2.3. Spatiotemporal Feature Learning for Video

Video analysis is crucial for computer vision, as videos contain far more information than single images. This analysis focuses on extracting spatiotemporal features, with traditional methods relying on optical flow [8, 55] and trajectory analysis [50, 67]. With the advent of deep learning (DL), a strategy employing 2D Convolution Neural Networks (CNNs) was proposed [22, 65]. This strategy includes temporal feature pooling to aggregate features from different frames for classification. Subsequently, approaches combining CNNs with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [1] were introduced, aiming to capture the long-range dependencies across various frames. 3D convolution kernels (C3D [56], I3D [9]) are used in video understanding, capturing channel interactions and local interactions simultaneously. However, they lead to a computational burden and have been regarded as an inefficient approach.

Subsequently, strategies offering greater efficiency were introduced, such as the Non-local network [59], S3D [62], CoST [43], SlowFast [21], and CSN [58]. These methods more efficiently learn the spatiotemporal features of videos by either reducing the number of sampled frames or replacing the use of 3D convolution with (2+1)D convolution. The prevailing consensus has moved away from the necessity of utilizing a large number of video frames or 3D convolution as the optimal approach for learning spatiotemporal features. Similarly, considering the resemblance between CT scans and videos, it is plausible to learn the feature representation of CT scans using only a small number of slices, without relying on 3D CNNs.

3. 2D, (2+1)D, 3D Convolutions for CT Scan

In this section, we discuss the three types of convolutions within framework of COVID-19 detection. The detailed architecture is described in Section 5-1.

2D: 2D Convolution over the sampled slices. The use of 2D convolution networks for extracting spatio-temporal features from 3D-cube data faces certain limitations, such as the requirement for strong spectral band or temporal continuity. Without these prerequisites, 2D convolutions may struggle to perform effectively due to their focus on spatial features and a lack of comprehensive sequence modeling. In applications involving CT scans, 2D convolutions are generally considered less effective compared to architectures like 2+1D convolutions, CNN-LSTM, or CNN-RNN, which are capable of capturing spatiotemporal features more efficiently. However, previous 2D CNN approaches often involve pre-processing, where crucial slices are selected and sampled to serve as inputs for the network. This sampling process tends to be overly simplistic, for instance, by manually selecting slices with the least artifacts or best quality, or randomly selecting a few slices to train a 2D CNN model. This limits the network's potential for global sequential modeling.

(2+1)D: 1D Convolution over the extracted features on different dimension. The 2+1D model is widely regarded as the greatest solution for CT analysis due to its exceptional performance and lower computational costs compared to 3D models. Typically, the 2+1D model performs best as it first extracts features on the spatial scale before modeling the sequences of these extracted features, effectively achieving both. However, according to our experiments, it tends to be less robust in situations with limited samples. This is because CT scans vary greatly in terms of resolution or the number of slices, making the 2+1D model more sensitive to the quantity of training data compared to 2D models. Additionally, we believe a potential concern with the 2+1D model is its difficulty in augmentation since spatial features are encoded into the latent space, the implicit learning approach limits its scalability and interpretability in clinical applications.

3D: 3D Convolution over the contiguous slices. *Compared with 2D and 2+1D, 3D is a heavy computational resource burden for COVID-19 detetion.* The differences between CT scans and conventional videos lie in several key aspects. Firstly, videos typically contain a significantly larger number of frames compared to the number of slices in a CT scan. Secondly, videos enhance their spatio-temporal coherence through frame rates (FPS), whereas the spatial relationships between slices in CT scans are relatively weaker. Lastly, slices in CT scans often exhibit redundancy at the beginning and end, which does not substantially contribute to analysis.

In conclusion, the advantages and weaknesses of these three methods can be itemized as follows:

- **2D:** Training and testing pipeline are simple. The model is robust no matter when few-scan or few-slice. Easy to augment. There are multiple methods which provide an explaniability for 2D model's prediction, such as Grad-CAM++ [10], SHAP [46]. Uneasy to capture sequential information unless dedicated design.
- **(2+1)D:** The performance is optimal when there is enough training data and the length of the CT slice sequence is sufficiently large, allowing it to capture sequential information. However, it becomes unstable with only a few scans or slices; the pipeline is complicated. It is

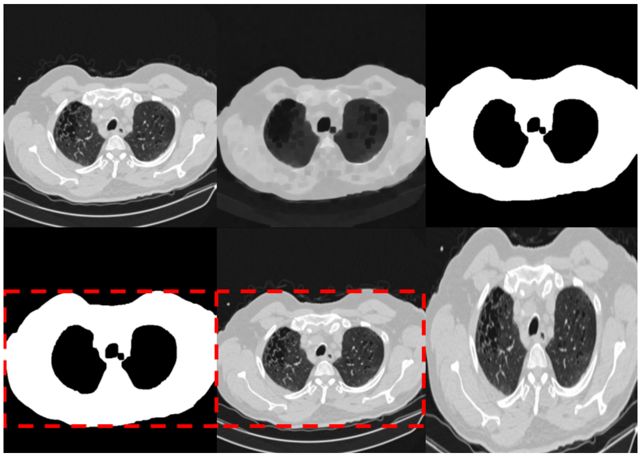

Figure 2. The illustration of spatial steps in proposed SSFL++.

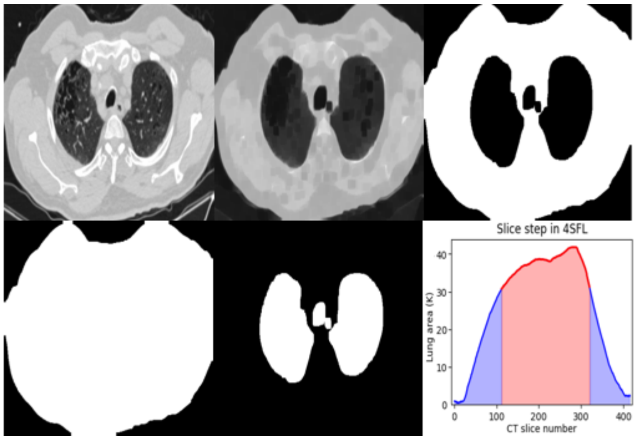

Figure 3. The illustration of slice steps in proposed SSFL++. The line graph in the bottom right corner represents the area of each slice in a CT scan. The blue area denotes OOD data that have been removed, while the red area represents the CT slices that have been selected.

also difficult to explain and augment.

- **3D:** Training and testing pipeline are simple. Can capture sequential information. Worst performance. Highest computational complexity. Unstable when few-scan and few-slice. Hard to explain and augment.

We believe 2D-CNNs have the potential to become mainstream for COVID-19 detection tasks. To enhance the ability of 2D-CNNs to learn sequence information from CT scans, we have designed the KDS method. This approach helps overcome the limitations of 2D-CNNs in this regard, with details to be introduced in Section 4.2.

4. Methodology

4.1. Spatial-Slice Feature Learning

In this section, we introduces our proposed SSFL++, which aim to figure out the RoIs in spatial and slice dimension, mainly based on the simple but effective computed morphology method and formulation of optimization problem.

Spatial Steps. The most importance concern is that CT-scan alway exists large black area between every single CT slice's background, and it will distort the RoI area when resizing to fixed shape to neural network, leading to feature vanish. In order to deal with this, a low-pass filter with a window size of $k \times k$ is applied to all CT slices $\mathbf{Z}$ to eliminate a noises. The low-pass filtering operator can be defined as:

$$\mathbf{Z}_{\text{filtered}}(i,j) = \frac{\sum_{p=-k}^{k}\sum_{q=-k}^{k} w(p,q) \times \mathbf{Z}(i+p, j+q)}{\sum_{p=-k}^{k}\sum_{q=-k}^{k} w(p,q)} \tag{1}$$

where $w(p,q)$ represents the weight at position (p,q) in the filter kernel. The above formula can determine the segmentation $\mathbf{Mask}$ of the filtered slices by a threshold t:

$$\mathbf{Mask}[i,j] = \begin{cases} 0, \text{ if } \mathbf{Z}_{\text{filter}}[i,j] < t \\ 1, \text{ if } \mathbf{Z}_{\text{filter}}[i,j] >= t \end{cases} \tag{2}$$

where i, j denote as an pixel for every single CT slice $\mathbf{Z}^c$, which resolution is $x \times y$. A Cropped region $\mathbf{Z}^c_{\text{crop}}$ can be calculated by:

$$\begin{aligned} \min(\mathbf{Z}^c_{\text{crop}}(x)) &= \min\{i \mid \mathbf{Mask}[i,j] = 1, \forall i\} \\ \max(\mathbf{Z}^c_{\text{crop}}(x)) &= \max\{i \mid \mathbf{Mask}[i,j] = 1, \forall i\} \\ \min(\mathbf{Z}^c_{\text{crop}}(y)) &= \min\{j \mid \mathbf{Mask}[i,j] = 1, \forall j\} \\ \max(\mathbf{Z}^c_{\text{crop}}(y)) &= \max\{j \mid \mathbf{Mask}[i,j] = 1, \forall j\} \end{aligned} \tag{3}$$

$\mathbf{Z}^c_{crop}$ is yielded accordingly, we can further resize the resolution of $\mathbf{Z}^c_{\text{crop}}$ to $H \times W$ for the slice steps and as an input of neural network. Spatial Steps in proposed 4SFL effectively filter out non-lung tissue regions (also known as RoIs in spatial dimension), and reduce computational complexity, as the Figure 2 illustrated.

Slice Steps. To find the lung tissue region in the CT scan, we used the binary dilation algorithm [61] to obtain the filled result $\mathbf{Mask}_{\text{filled}}$. The difference between the $\mathbf{Mask}$ and filled mask $\mathbf{Mask}_{\text{filled}}$ represents the lung tissue region. The above method can be summarized as the following formula:

$$Area(\mathbf{Z}) = \sum_i \sum_j \mathbf{Mask}_{\text{filled}}(i,j) - \mathbf{Mask}(i,j). \tag{4}$$

After the above technique, we can finally obtain a range where s and e denote the starting and ending indexes, respectively, and n_c is the constraint of the number of slices

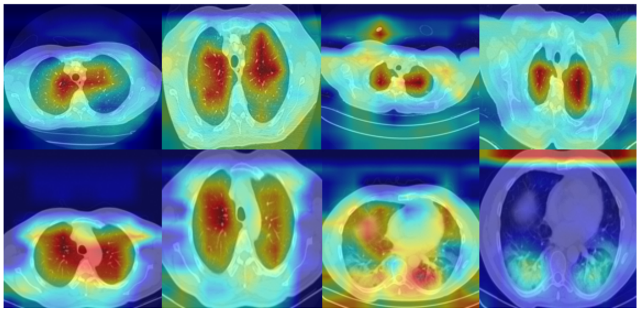

Figure 4. The GradCAM++ [10] visualization before and after proposed SSFL++. By reducing redundancy on the spatial scale, we can implicitly enhance the visual effectiveness of Explainable AI, thereby facilitating clinical applications.

for a single CT scan to select most importance RoIs in slice dimension with proportion α. The optimization problem can be formulated as following:

$$\begin{aligned} \underset{s,\,e}{\text{maximize}} \quad & \sum_{i=s}^{e} Area(\mathbf{Z}_i), \\ \text{subject to} \quad & e - s \leq n_c, \\ & \frac{\sum_{i=s}^{e} Area(\mathbf{Z}_i)}{\sum_{i=1}^{n_c} Area(\mathbf{Z}_i)} \geq \alpha. \end{aligned} \tag{5}$$

It is worth noting that we sort all CT slices according to their slice numbers n_c, as illustrated in the bottom-right corner of Figure 3.

The spatial and slice steps of proposed SSFL++ follow unsupervised learning manner, which only follow the prior knowledge of lung-CT-scan. It can be generalize to other organs or body parts CT-scan. However, it may require parameter adjustments based on their specific characteristics. Additionally, with the SSFL++, the visual explanation method can also look RoI more concentrated, as shown in Figure 4.

4.2. Density-aware Slice Sampling

Background. The SSFL proposed by Hsu *et al.* [30] employs a random sampling method to select slices, which were used for the detection of COVID-19 using 2D and 2+1D CNNs. However, random sampling may potentially introduce bias and instability when training and inference, and it does not efficiently identify the most representative CT slices, as shown in Figure 5.

In order to address this, we propose a Kernel-Density-based Slice Sampling (KDS). It performs kernel density estimation (KDE) on the selected slices-set [$\mathbf{Z}_e$,$\mathbf{Z}_s$], adaptively and wisely sampling the most crucial CT-slices. Meanwhile, it also keeps the sequence information globally and alleviates the instability during training and inference stage.

Definition. KDE is a classic method to estimate the probability density function (PDF) of a random variable in a non-parametric manner. It can be defined as:

$$\widehat{f}_h(x) = \frac{1}{s}\sum_{i=1}^{s} K_h(x - x_i) = \frac{1}{sh}\sum_{i=1}^{s} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \tag{7}$$

where h is the bandwidth constant, calculated by Scott-rule [52], K is a Gaussian kernel, s is a smooth factor of estimated density function, (the higher the smoother, we set it to 100). For a given KDE, we can create several subintervals by calculating its Cumulative Distribution Function (CDF), where the length of each sub-interval adaptively changes with its p-percentile. The CDF of KDE and its p-percentile can be calculated as following formulas:

$$F(x) = \int_{-\infty}^{x} \hat{f}_h(t)dt, F(q_p) = p \tag{8}$$

In the proposed KDS method, we determine the probability of slices being selected in each interval based on the density from KDE, while also ensuring that each subinterval has at least one sample selected. This method captures the global sequential information and increases the probability of selecting the most crucial CT slices.

5. Experiment

Dataset description. In our experiments, we used a total of 1,684 COVID-19-CT-DB data, provided by Kollias *et al.* [40]. The dataset information have shown in Table 2. Our loss function is binary cross-entropy. In order to ensure stability and fairly check performance during the experiments, group-5-fold-cross-validation is used. Data augmentation and hyperparameters are kept consistent in all experiments.

Hyperparameter settings. The Adam [33] optimizer was used with a learning rate of $1e-4$ and a weight decay of $5e-4$. The batch-size is set to 16.

Data Augmentation. In our experiments, we utilized common augmentation strategy like HorizontalFlip, RandomScaleShifting to prevent overfitting and enlarge feature space. Additionally, we find that HueSaturationValue, RandomBrightnessContrast and CoarseDropout [15] are also used.

Evaluation Metric. We mainly used F1-score in the experiments for model evaluation. F1-score is a metric used to determine the accuracy of a binary classification model. It combines the harmonic mean of Precision and Recall.

$$\text{f1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{9}$$

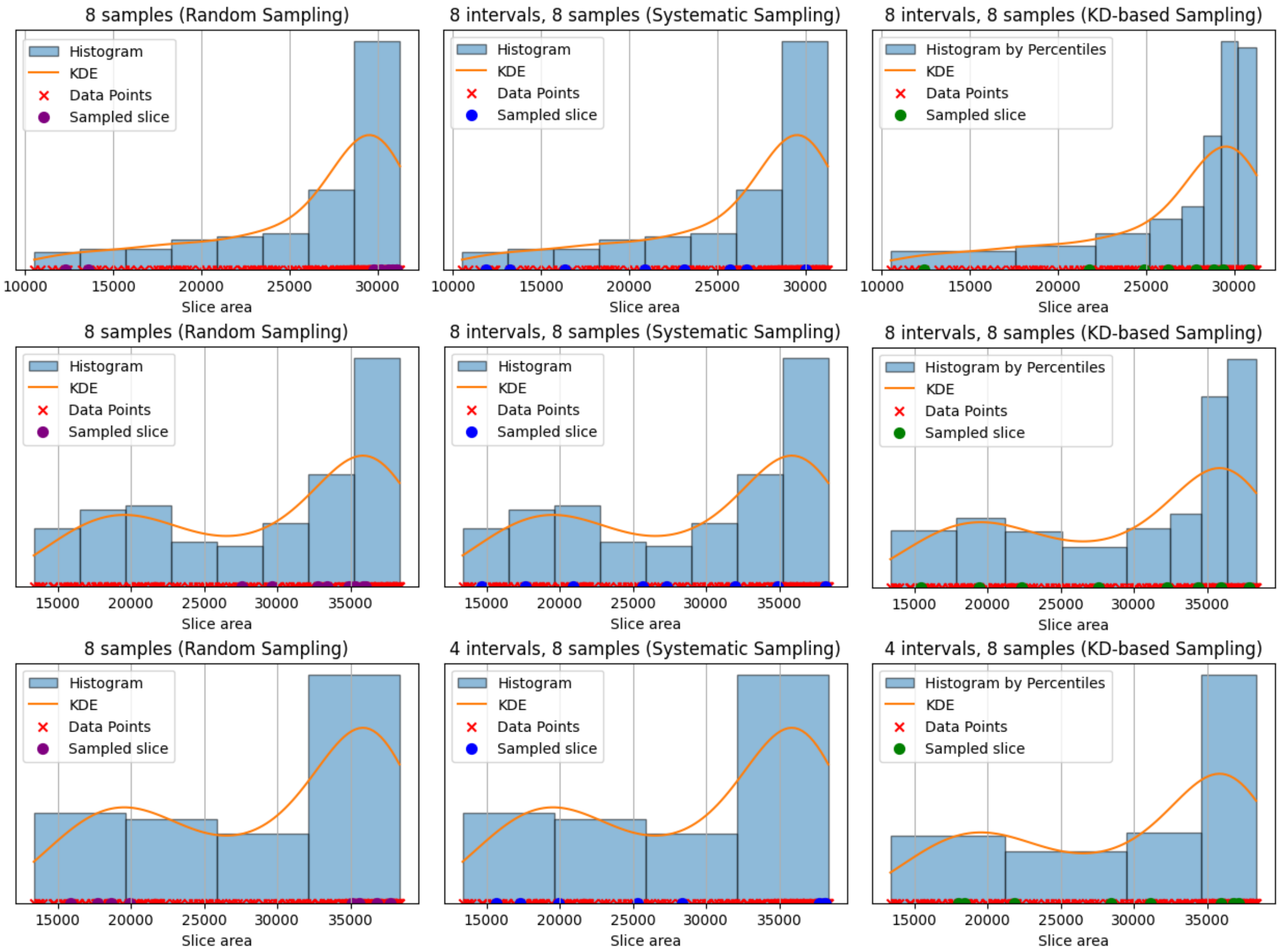

Figure 5. The comparison between random sampling, systematic sampling, and the proposed KDS method is noteworthy. As illustrated, random sampling fails to uniformly sample CT slices of varying area sizes, tending to select larger areas while neglecting global information. This results in greater bias and randomness during training and inference. On the other hand, systematic sampling divides the area into equally lengthened sub-intervals before randomly selecting samples from them. Although this approach can capture global information, it is ineffective at sampling the most crucial CT slices. Our proposed KDS method combines the advantages of both methods without their drawbacks, achieving a better balance. KDS can implicitly improve data efficiency, thereby enhancing the model's few-shot capability.

	Spatial Area (K)			Slice Length			Spatial × Slice (M)		Total
	Before	After	Δ (%)	Before	After	Δ (%)	Before	After	Δ (%)
Training	267.25	155.53	0.4184	285.32	142.91	0.4983	76.25	22.22	0.7085
Positive	266.42	157.69	0.4088	295.90	148.18	0.4985	78.83	23.36	0.7036
Negative	268.21	153.03	0.4296	273.97	137.26	0.4981	73.48	21.00	0.7141
Validation	265.62	155.23	0.4172	281.95	141.23	0.4984	74.89	21.92	0.7072
Positive	268.94	160.48	0.4061	280.53	140.55	0.4984	75.45	22.55	0.7010
Negative	262.12	149.69	0.4288	283.49	141.97	0.4984	74.30	21.25	0.7139
(T+V) Positive	267.25	155.53	0.4184	292.96	146.72	0.4985	78.29	22.81	0.7085
(T+V) Negative	267.01	152.37	0.4294	275.78	138.16	0.4982	73.64	21.05	0.7141
Total	266.94	155.47	0.4182	284.68	142.59	0.4983	75.99	22.16	0.7082
Testing	279.55	153.41	0.4520	309.39	154.67	0.5003	86.48	23.72	0.7256

Table 1. The reduction in redundant data achieved by the SSFL++ module is evaluated across three dimensions: spatial, slice, and overall. This approach quantifies the efficiency of the SSFL++ module in reducing unnecessary information in CT scans, enabling more focused analysis and processing. By minimizing data redundancy, the module enhances computational efficiency and potentially improves the accuracy of subsequent analyses or models applied to the CT data.

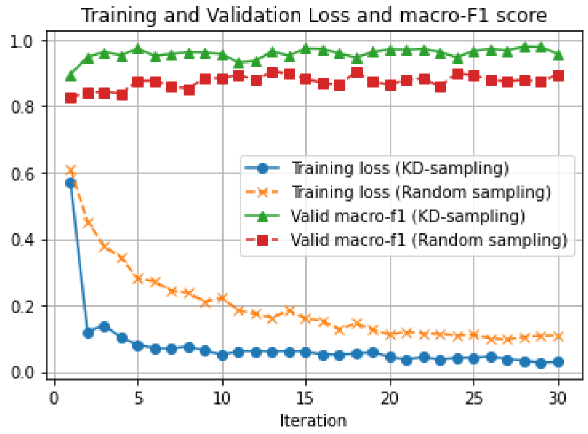

Figure 6. In terms of optimizing procedure, our proposed KDS approach, compared to the random sampling used by Hsu *et al.* [30], is more capable of learning the global information of CT scans, thereby accelerating the convergence rate and enhancing the model performance.

Type	Positive Scan	Negative Scan	Total Scan
Training	703	655	1358
Valid	170	156	326
Total	873	811	1684
Testing	-	-	1413
Type	Positive Slice	Negative Slice	Total Slice
Training	206608	178722	385330
Valid	46042	43679	89721
Total	252650	222401	475051
Testing	-	-	437185

Table 2. The number of data samples at the scan and slice level.

where precision and recall are computed for COVID and non-COVID. The macro f1-score is the average of the f1-scores for all classes:

$$\text{macro f1-score} = \frac{1}{N}\sum_{i=1}^{N} \text{f1-score}_i \tag{10}$$

where N is the number of classes, and f1-score$_i$ is the f1-score for the i-th class. These metrics provide a balanced evaluation of the model's ability to classify each class accurately and its overall performance across all classes.

5.1. Model Details and Performance Comparison

To provide a more comprehensive comparison and improve future research, we designed simple E2D, E2+1D, E3D in our experiments. The backbones are all based on EfficientNet-b3 [54, 60]. The baseline method and detailed pipeline are as follows:

Baseline: The baseline method is presented in [40], Kollias *et al.* adopted CNN-RNN to extract feature within all CT-slice. First, all CT-slices are resized to 224 × 224 to extract feature, then RNN (GRU [13] with 128 neurons) analyzed the 2D-CNN (ResNet-50 [28]) features. The output of the RNN element is then forwarded to a fully connected layer. In addition, this also includes a dropout layer (the dropout rate is set to 0.8) before the fully connected layer.

E2D: From the CT-scans processed by SSFL++, subsequently, we use our proposed KDS. These sampled slices are resized to 384 × 384 and extracted to high-representation features.

E2+1D: Similar to E2D, firstly, the CT scans processed by SSFL++ are resized to 384 × 384. And 100 slices are selected to be encoded. therefore, we used 2D encoder to get an encoded vectors. By doing so, the CT scans will be encoded into latent feature queue, which size is 224 × 100. Subsequently, we randomly sampled 50 features from latent feature queue, and utilized a simple 1D convolution with kernel size 1 × 1 in e or l dimensions to capture sequential information.

E3D: We first utilized SSFL++ to remove OOD slices and redundant spatial information, and then sample a certain number of CT slices for modeling.

The experimental results, as presented in Table 3, highlight the E2D model's exceptional performance when paired with KDS on the COVID-19 database 2024 validation set. It also showcases remarkable robustness in few-scan scenarios, delivering results that instill confidence. Comparatively, the E2D model utilizing KDS achieves a significant improvement in scan-level f1-score compared to its counterpart that employs random sampling. This underscores the capability of 2D convolutions to implicitly capture global sequence information through an appropriate sampling method. In contrast, the E3D model demands a large sample size, resulting in limited performance and higher computational requirements.

5.2. Ablation Study

To further analyze the impact of SSFL++ and KDS on the COVID-19 detection task, the ablation study were conducted, with results presented in Table 4. All experiments are based on the E2D model, with all experimental hyperparameters kept constant. The results demonstrate that the proposed SSFL++ significantly enhances performance, implying the importance of spatial redundancy in CT scans and efficient slice selection. On the other hand, KDS further improves the model's prediction ability at the slice-level and makes significant progress at the scan-level, achieving convincing performance. KDS effectively addresses the lack of global sequential modeling capability in 2D-CNN when analyzing CT images.

Model type	Scans	Sampled slice	macro f1-score (slice-level)	f1-score (scan-level)
baseline [40]	100%	-	-	78.00
E3D	1%	33(random)	-	32.55
	50%	33(random)	-	78.54
	100%	33(random)	-	86.76
	100%	50(random)	-	87.05
	100%	80(random)	-	90.24
	100%	120(random)	-	91.05
E(2+1)D	1%	8(random)	73.46	-
	50%	8(random)	87.64	-
	100%	8(random)	91.39	-
	100%	16(random)	92.31	93.69
E2D	1%	8(random)	88.94	92.11
	50%	8(random)	91.52	92.42
	100%	8(random)	92.44	93.18
	100%	16(random)	92.68	93.37
	1%	4(KDS)	91.42	96.42
	1%	8(KDS)	91.88	99.80
	100%	8(KDS)	93.46	**100.00**
	100%	16(KDS)	**94.11**	**100.00**

Table 3. Performance comparison between baseline provided by Kollias *et al.* [40], and proposed E2D, E2+1D, E3D on COVID-19-CT-DB validation set.

Spatial step	Slice step	KDS	marco f1-score (slice level)	f1-score (scan level)
			80.41	81.26
✓			88.01	88.04
	✓		90.32	90.48
✓	✓		92.68	93.37
✓	✓	✓	**94.11**	**100.00**

Table 4. The ablation study of proposed SSFL++ and KDS on COVID-19-CT-DB validation set.

	macro-F1	F1(NONCOVID)	F1(COVID)
baseline [40]	85.11	87.48	82.74
E2D (Ours) [31]	**94.39**	**95.52**	**93.26**

Table 5. The results on COVID-19-CT-DB testing set.

6. Generalizability

Our proposed SSFL++ not only excels in performance on the COVID-19-CT-DB [40] but also demonstrates commendable efficacy on CT scans from various views and body parts. We showcased the versatility of SSFL++ by selecting four distinct types of data, with the results depicted in Figure 7. From top to bottom, the figures represent the different views or body parts before and after SSFL++. Specifically, (a) (c) (d) are lung CT scans from the COVID-19-CT-DB dataset, featuring the axial, sagittal, and coronal views. Meanwhile, (b) involves a dataset provided by [24], aimed at identifying acute appendicitis from CT scans of acute abdomen cases.

Additionally, it is important that when using SSFL++ on CT slices of different body parts or from different views, its hyperparameters may need specific adjustments. For instance, in the case of (b), the original settings might select OOD slices rather than the RoI slices.

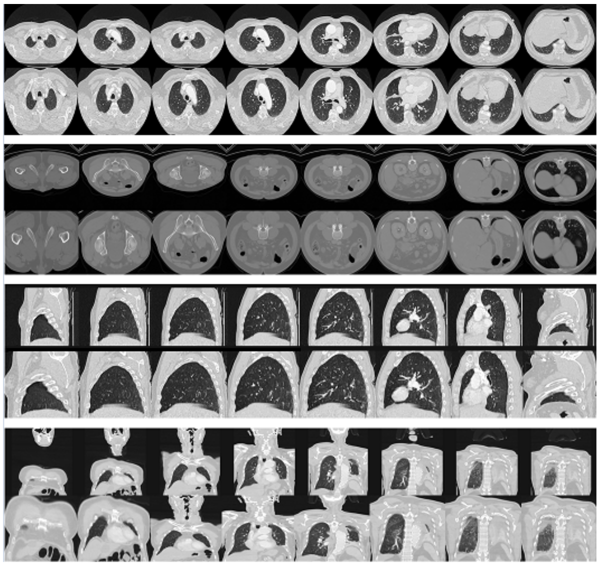

Figure 7. CT slices from different views and body parts, as well as the results after processing through the spatial step in our proposed SSFL++, are presented. From left to right, the sequence represents the process of CT imaging, where OOD data tend to concentrate at the beginning and the end. The middle section represents the RoI area. As shown in the figure, SSFL++ performs well under various conditions.

7. Conclusion

We conducted a comprehensive analysis of the COVID-19 detection task, noting that CT scans often contain a large amount of redundant information, which limits the performance of models. To address this issue, we introduced a simple morphology-based method for CT images, named Spatial-Slice Feature Learning (SSFL++), designed to efficiently and adaptively locate the Region of Interest (RoI). This method effectively reduces redundancy across both spatial and slice dimensions. Furthermore, to inspire future research, we analyzed the advantages and disadvantages of 2D, 2+1D, and 3D convolutions on CT data. After extensive experimentation, we believe that 2D-CNNs hold the greatest potential in the wild.

To overcome the limitations previously encountered by 2D-CNN in research, we combined SSFL++ with the further designed KDS, thereby addressing the instability brought about by random sampling during the training and inference. Moreover, through the global sequence modeling, we activated the potential of 2D-CNNs. Finally, our method demonstrated promising results on the validation and testing sets provided by the DEF-AI-MIA workshop.

References

[1] A. R. Abdali and R. F. Al-Tuma. Robust real-time violence detection in video using cnn and lstm. In *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, pages 104–108, 2019. 3

[2] 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

[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. 1

[4] Harrison H. Barrett. Iii the radon transform and its applications. pages 217–286. Elsevier, 1984. 2

[5] J. A. Beatty. The radon transform and the mathematics of medical imaging. 2012. 2

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1

[7] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M. Khapra. Efficient video classification using fewer frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[8] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision - ECCV 2004*, pages 25–36, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. 3

[9] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 3

[10] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 3, 5

[11] Guan-Lin Chen, Chih-Chung Hsu, and Mei-Hsuan Wu. Adaptive distribution learning with statistical hypothesis testing for covid-19 ct scan classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 471–479, 2021. 1, 2

[12] M. et al. Chetoui. Explainable covid-19 detection based on chest x-rays using an end-to-end regnet architecture. *Viruses*, 2023. 1

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 7

[14] M. et al Cobo. Enhancing radiomics and deep learning systems through the standardization of medical imaging workflows. *Scientific Data*, 2023. 2

[15] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] Bin Saeedan et al. Thyroid computed tomography imaging: pictorial review of variable pathologies. *Insights Imaging*, 2016. 1

[18] Chilamkurthy S et al. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *Lancet*, 2018. 1

[19] Hochreiter et al. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[20] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 1

[21] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 3

[22] Basura Fernando and Stephen Gould. Learning end-to-end video classification with rank-pooling. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1187–1196, New York, New York, USA, 2016. PMLR. 3

[23] Andrey Gaidel. Method of automatic roi selection on lung ct images. *Procedia Engineering*, 201:258–264, 2017. 3rd International Conference "Information Technology and Nanotechnology", ITNT-2017, 25-27 April 2017, Samara, Russia. 2

[24] Wen-Jeng Lee Goman, Taiwan Radiological Society (TRS). Aocr2024 ai challenge, 2023. 8

[25] Monika Grewal, Muktabh Mayank Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 281–284, 2017. 1

[26] Bajaj V. Gupta K. Deep learning models-based ct-scan image classification for automated screening of covid-19. *Biomed Signal Process Control*, 2023. 1

[27] Hongchao et al. He. Computed tomography-based radiomics prediction of ctla4 expression and prognosis in clear cell renal cell carcinoma. *Cancer medicine*, 2023. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7

[29] Chih-Chung Hsu, Chi-Han Tsai, Guan-Lin Chen, Sin-Di Ma, and Shen-Chieh Tai. Spatiotemporal feature learning based on two-step lstm and transformer for ct scans. *arXiv preprint arXiv:2207.01579*, 2022. 1, 3

[30] Chih-Chung Hsu, Chih-Yu Jian, Chia-Ming Lee, Chi-Han Tsai, and Shen-Chieh Tai. Bag of tricks of hybrid network for covid-19 detection of ct scans. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–4. IEEE, 2023. 1, 2, 3, 5, 7

[31] Chih-Chung Hsu, Chia-Ming Lee, Yang Fan Chiang, Yi-Shiuan Chou, Chih-Yu Jiang, Shen-Chieh Tai, and Chi-Han Tsai. Simple 2d convolutional neural network-based approach for covid-19 detection, 2024. 8

[32] Laura J et al. Jensen. Enhancing the stability of ct radiomics across different volume of interest sizes using parametric feature maps: a phantom study. *European radiology experimental*, 2022. 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[34] 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, 2

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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

[40] Dimitris 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. 5, 7, 8

[41] Dong-Hoon Lee, Do-Wan Lee, and Bongsoo Han. Possibility study of scale invariant feature transform (sift) algorithm application to spine magnetic resonance imaging. *PloS one*, 11:e0153043, 2016. 1

[42] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 1

[43] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Collaborative spatiotemporal feature learning for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7872–7881, 2019. 3

[44] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[45] Lin et al. Lu. Uncontrolled confounders may lead to false or overvalued radiomics signature: A proof of concept using survival analysis in a multicenter cohort of kidney cancer. 2021. 2

[46] Scott et al. Lundberg. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. 3

[47] F. et al. Mercaldo. Coronavirus covid-19 detection by means of explainable deep learning. *Scientific Reports*, 2023. 1

[48] K et al. Moulaei. Comparing machine learning algorithms for predicting covid-19 mortality. *BMC Med Inform Decis Mak*, 2022. 1

[49] Kiran Parmar, Dr. Rahul Kher, and Falgun Thakkar. Analysis of ct and mri image fusion using wavelet transform. pages 124–127, 2012. 1

[50] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *Computer Vision – ECCV 2014*, pages 581–595. Springer International Publishing, 2014. 3

[51] André et al. Ramon. Role of dual-energy ct in the diagnosis and follow-up of gout: systematic analysis of the literature. *Clinical Rheumatology*, 2018. 1

[52] D.W. Scott. Multivariate density estimation: Theory, practice and visualization. 1992. 5

[53] Steven W. Smith. Computed tomography, 1999. 1, 2

[54] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International conference on machine learning (ICML)*, pages 6105–6114, 2019. 7

[55] Yongyi Tang, Lin Ma, and Lianqiang Zhou. Hallucinating optical flow features for video classification. In *IJCAI*, 2019. 3

[56] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[57] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018. 1

[58] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[59] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[60] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 7

[61] Wikipedia contributors. Mathematical morphology — Wikipedia, the free encyclopedia, 2022. [Online; accessed 2-July-2022]. 4

[62] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[63] Q. et al. Xu. Ai-based analysis of ct images for rapid triage of covid-19 patients. *npj digital medicine*, 2020. 1

[64] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[65] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[66] Shenghan Zhang, Binyi Zou, Binquan Xu, Jionglong Su, and Huafeng Hu. An efficient deep learning framework of covid-19 ct scans using contrastive learning and ensemble strategy. In *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 388–396. IEEE, 2021. 1

[67] Zhengwu Zhang, Jingyong Su, Eric Klassen, Huiling Le, and Anuj Srivastava. Rate-invariant analysis of covariance trajectories. *Journal of Mathematical Imaging and Vision*, 60, 2018. 3

[68] Sun Y. Zhang Y, Zhang L. Rigid motion artifact reduction in ct using frequency domain analysis. *J Xray Sci Technol*, 2017. 1