The Value of Visual Attention for COVID-19 Classification in CT Scans

Adrit Rao
Palo Alto High School
Palo Alto, CA, USA
adrit.rao@gmail.com

Jongchan Park
Lunit Inc.
Seoul, Korea
jcpark@lunit.io

Oliver Aalami
Stanford University
Stanford, CA, USA
aalami@stanford.edu

Abstract

Detecting COVID-19 in early stages is crucial in order to initiate timely treatment of disease. COVID-19 screening with chest CT scans has been utilized due to the rapidity of results and robustness. Computer vision aided medical diagnosis with deep learning models can improve accuracy and efficiency of screening. When developing models for high-risk medical classification tasks, it is important to aim to reach radiologist level interpretation in terms of cognition. When the human brain analyzes visual information, cognitive visual attention is applied in order to apply more focus onto higher frequency regions of interest. Using attention mechanisms in order to infer channel and spatial attention maps within convolutional neural networks can improve the performance in classification of COVID-19 changes. Through performing a compact study with a quantitative accuracy measure along with a qualitative visualization of activation heat-maps, we study the benefits of visual self-attention for the classification of COVID-19.

1. Introduction

The novel coronavirus disease 2019 or SARS-CoV-2 (COVID-19) has caused a severe rising global pandemic with its highly contagious and drastic effects on the human body [8, 20, 23]. The early detection of this disease can prevent the progression into severe stages of respiratory illness and help mitigate the spread of the disease through early isolation [12, 16]. One of the common and effective ways to screen for COVID-19 is through the analysis and interpretation of chest CT scan images of the lung region [1, 14]. Images are commonly derived through CT or computed tomography of the chest which produces a scan that visualizes the heart, lung and airways for diagnostic evaluation [5, 19]. Commonly reported visual features identify the coronavirus infection in effects through ground-glass opacities, vascular enlargement, bilateral abnormalities, lower lobe involvement, and posterior predilection in CT scan images [14]. The open-source crowd-sourced *UCSD COVID-CT* dataset [26] consists of single slice CT scan images with both COVID positive and negative lesions (Fig. 1). Changes consistent with the coronavirus disease can be identified through a single slice CT scan image [4, 26].

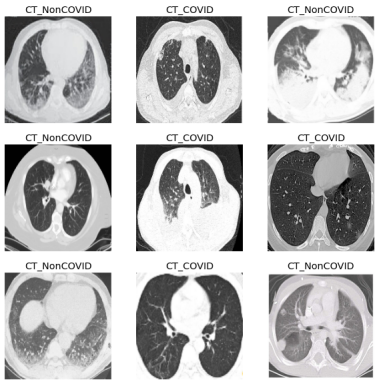

Figure 1: **COVID-19 CT dataset example images** [26]. Classes: COVID Positive, COVID Negative.

The radiologist's workload is increasing year over year. The ability to aid radiologists in COVID-19 diagnosis by triaging images can not only reduce the burden but potentially increase the efficiency of disease screening [3].

Thus, the ability to leverage deep learning based approaches for the automated and accurate diagnosis of COVID-19 is being widely studied in the computer vision community. Current approaches are achieving significantly high accuracy at the classification task [2, 9, 18]. However, when using a standard neural network based classifier to identify infection, it is crucial for the end prediction to represent a clinically relevant focal-point of the image to approach radiologist-level diagnosis. When the human brain assesses visual information, cognitive attention is utilized in order to apply higher amounts of focus onto more important regions [7, 24]. The trained radiologist can likely subconsciously apply attention onto important regions of a CT scan and judge the clinical importance of features. However, a standard neural network is potentially unable to do so and will simply extract features, clinically relevant or not.

The ability to mimic the cognitive capability of visual attention has been a popular research topic in the computer vision community. In the computer vision community, the *attention mechanism* has been an important research topic [6, 10, 11, 21, 22, 25]. The simplest operation of this mechanism is to generate attention masks for intermediate convolutional feature maps and multiply them together. Attention mechanisms implicitly learn to keep important values in the feature maps and scale down less important values. A notable implementation of this is Convolutional Block Attention Module (CBAM) [25]. To reduce the heavy computational overhead of 3D (channel, height, and width) attention map generation, CBAM decomposes 3D attention maps into 2D spatial attention maps and 1D channel attention maps. CBAM can be easily integrated into any CNN architecture with minimal overhead to achieve significant performance improvements. CBAM is explained in detail in Section 3. In the high-risk medical domain, especially the COVID-19 classification problem, it can be important for deep learning models to employ such capabilities. With the aim of improving COVID-19 classification performance and starting to approach radiologist-level interpretation of CT scan images, we study the application of integrated visual attention mechanisms within COVID-19 detection pipelines and observe changes in both quantitative (statistical F1 score analysis) and qualitative results (attention heat-map visualization). Our final goal is to understand and analyze the effects of attention for this medical task.

2. Related Work

COVID-19 Classification. Classification of the coronavirus infection in chest CT scans using deep learning methods has been widely studied. The ability to accurately do so can prove to be highly useful within a clinical setting to reduce screening time and act as a quality assurance (QA) tool. Methods vary between both 2D (single slice) and 3D (volume slices) convolutional neural network (CNN) approaches. The 3D CNN approach can prove to be computationally expensive and 2D approaches may be more feasible. 2D CNN approaches study the use of different state-of-the-art CNN architectures (ResNet, Inception, EfficientNet, etc.) for classification of single slice images. The COVID-CT dataset [26] is an open-source crowd-sourced repository collected by a team at the UCSD. The dataset contains COVID-19 positive and negative images collected from coronavirus related research papers via medRxiv, bioRxiv, NEJM, JAMA, Lancet, etc. Using this dataset, multiple studies have been carried out around the use of different state-of-the-art neural networks and transfer learning approaches with the goal of identifying the highest performing system [9]. As previously discussed in Section 1, the ability to use attention can prove to bring significant value to these systems. With state-of-the-art attention mechanisms, deep learning models are able to infer the importance of visual features in a data driven manner. With this capability, we can enable the ability for COVID-19 classifiers to focus on features of higher clinical importance and reduce computation across all features present with the goal of enabling radiologist-level interpretation. The goal of our paper is to understand the value of attention within the standard CNN for the COVID-19 classification task through a compact quantitative and qualitative analysis study.

Attention Mechanism. One of the pioneering approaches using attention mechanisms is the Residual Attention Network (RAN) [21]. The RAN uses a separate network to generate the attention mask, using the same size of the intermediate feature. This direct computation is simple and intuitive and improves baseline performance, yet the computational cost is quite high. Squeeze-and-Excitation (SE) [10] is also a prominent approach that focuses on channel attention. For each given intermediate feature map, an SE module generates a per-channel attention value from the global-average-pooled features. SE has been shown to improve performance with minimal overhead [10]. The Style-based Recalibration Module (SRM) [15] is a simple yet powerful channel attention module that accounts for channel statistics (mean and standard deviation) when scaling the channel values. The Convolutional Block Attention Module (CBAM) [25] is a computationally efficient method that decomposes the heavy attention generation into separate dimensions. Specifically, while RAN directly generates full-sized attention maps, CBAM generates 2D spatial attention maps and 1D channel attention maps. CBAM has been shown to improve performance in various tasks consistently [25]. Due to the simplicity of the method and accuracy received, we choose CBAM as a proxy for attention mechanisms and analyze the effect of it in the COVID-19 CT classification task with standard 2D CNN architectures.

3. Convolutional Block Attention Module

The basic idea of the *attention mechanism* is to focus on important values in the intermediate features or tensors. While there can be various interpretations on what *attention* is, in this paper, we will define *attention* as scaling the values according to their importance. If we use 2D image inputs with a 2D CNN architecture, the shape of the intermediate feature (or tensor) is 3D (channel, height, and width). So, for the given 3D tensor, the attention mechanism computes a 3D mask to increase important values, and decrease less important values. When the attention tensor is computed with the 3D feature itself, it is called 'self-attention'.

The Convolutional Block Attention Module (CBAM) is a self-attention mechanism designed for standard CNN architectures. The direct computation of a 3D attention tensor is

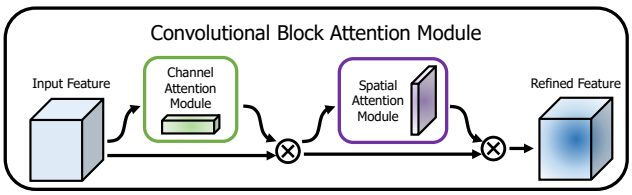

Figure 2: **A conceptual diagram for CBAM.** Image taken from original CBAM paper [25].

quite heavy [21], roughly doubling the overall computation. CBAM decomposes the 3D attention tensor into 2D spatial attention and 1D channel attention and applies them sequentially into the input feature to reduce the overhead of the attention mechanism. The design is illustrated in Fig. 2. To further reduce the overhead, CBAM extracts mean and max statistics into both channel and spatial dimensions to compute the attention tensors for each dimension. The channel and spatial attentions are sequentially applied to the input feature map F. The mathematical notation of CBAM is:

$$F' = M_c(F) * F \tag{1}$$

$$F'' = M_s(F') * F' \tag{2}$$

Eq. 1 denotes the channel attention sub-module and Eq. 2 denotes the spatial attention sub-module which follows after the channel attention sub-module. F represents the original 3D input feature and F'' represents the output from CBAM. F and $F'' \in \mathbb{R}^{C\times H\times W}$, where C is the channel size, H is the height and W is the width of the feature. Also, M_c is the 1D channel attention tensor $M_c \in \mathbb{R}^{C\times 1\times 1}$, and M_s is the 2D spatial attention tensor $M_s \in \mathbb{R}^{1\times H\times W}$.

Channel attention module The structure of the channel attention module is illustrated in Fig. 3 (top). The attention tensor for the channel dimension is a 1D tensor. To efficiently calculate the 1D tensor, the global average and the global maximum values along each channel are pooled. Then, the 1D features are fed into a 2-layer MLP with a sigmoid normalization layer at the end. The mathematical notation of the channel attention computation is:

$$M_c(F) = \sigma(MLP(F_{avg\_ch}) + MLP(F_{max\_ch})) \tag{3}$$

where σ denotes the Sigmoid function, MLP is the 2-layer multi-layered perceptron, F_{avg_ch} and F_{max_ch} are global average pooled / global max pooled features along the channel dimension, where F_{avg_ch} and $F_{max\_ch} \in \mathbb{R}^{C\times 1\times 1}$. The final output of the channel attention module is the original 3D CNN feature multiplied by the 1D attention tensor with broadcasting along the spatial dimension.

Spatial attention module The structure of the spatial attention module is illustrated in Fig. 3 (bottom). The architecture follows the same structure as the channel attention module the only difference being the fact that the spatial attention module focuses on the spatial dimension. The mathematical notation for the spatial attention computation is:

$$M_s(F) = \sigma(Conv_{7x7}(F_{avg\_sp}) + Conv_{7x7}(F_{max\_sp})) \tag{4}$$

As written in Eq. 4, the spatially max/avg pooled feature F_{max_sp} $F_{avg\_sp} \in \mathbb{R}^{1\times H\times W}$ are fed into a convolutional layer to compute the spatial attention tensor $M_s \in \mathbb{R}^{1\times H\times W}$.

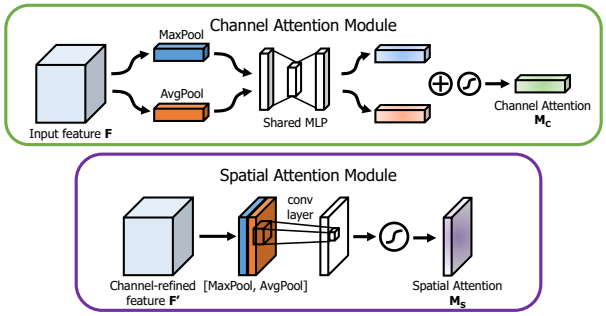

Figure 3: **Spatial and Channel Attention Sub-modules.** Image taken from original CBAM [25].

4. Methods

4.1. Experiment Details

The aim of our experiment was to compare the performance of the baseline CNN with an attention augmented CNN for the COVID-19 classification task and understand the benefits of attention. As previously discussed in Section 1, the ability to use artificial visual attention mechanisms in neural networks can simulate cognitive attention which the trained radiologist uses naturally in order to determine the importance of clinical features in CT scans. With the addition of attention, we notice major changes in a qualitative heat-map visualization over more clinically relevant features along with significant increases in quantitative accuracy metrics. As also previously mentioned, we use CBAM as our attention method for experimentation purposes. We used the COVID-CT dataset [26] for training and validation which consists of 349 COVID-19 positive CT scans and 463 COVID-19 negative CT scans (total 812 CT scan images).

4.2. Model Architectures

Baseline CNN The baseline standard CNN architecture, illustrated in Fig. 4 (left), was taken from the default Tensorflow CNN documentation. The model is a very simple form

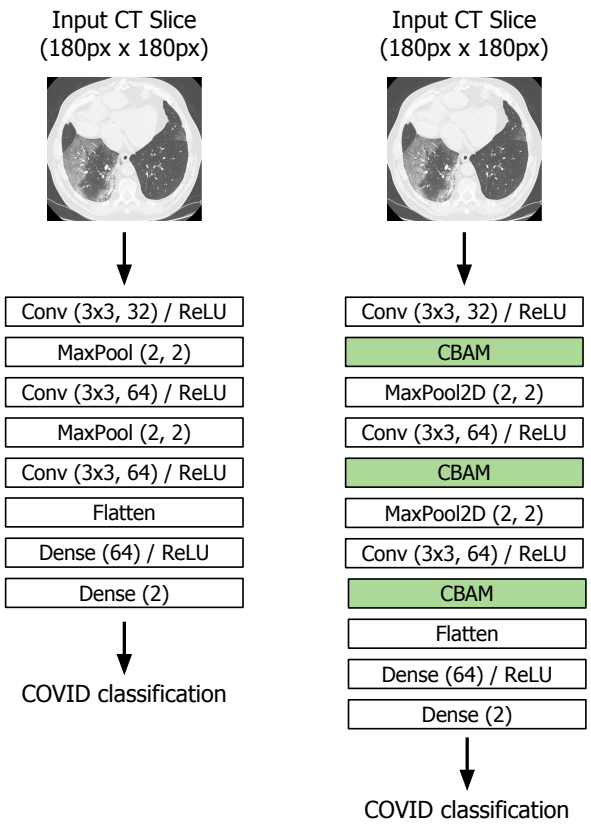

Figure 4: The *baseline CNN* (8 layers) and the *CBAM-augmented CNN* (11 layers) architectures.

of the standard CNN and consists of convolution (ReLU) paired with max pooling (2x) followed by a final convolution and a flatten with dense layer for class label prediction.

Attention CNN CBAM is a self-contained module where the input is a 3D tensor, and the output is an attention-augmented 3D tensor. Because of the self-contained characteristic, CBAM can be easily integrated in any part of the CNN architectures and augment output feature maps with attention. Placement of CBAM blocks followed standard procedure discussed in the original paper [25]. In the original paper [25], the authors suggest adding CBAM after every convolutional block. In the baseline CNN, there are 3 convolutional layers as illustrated in the left part of Fig. 4. The architecture is very simple and CBAM is placed subsequently after the convolutional layers with ReLU. As a result of this modification, the CBAM-augmented CNN is illustrated in the right side of Fig. 4 (right) varied only by the addition of attention in comparison to the baseline CNN.

Training specifics Model training was done in the Google Colab high-ram IDE with the NVIDIA Tesla V100 GPU (tensor core). Both models were trained across 30 epochs using cross-entropy loss with the Adam optimizer (LR 0.001) in Tensorflow with Keras. Batch size was 34. Image input size was 180x180 pixels. For this preliminary study, we randomly split the dataset into training set (80%) and validation set (20%) without cross-validation. The baseline and CBAM-augmented CNN follow the same experimental settings. Starting from the randomly initialized weights, CBAM is trained jointly end-to-end with other CNN layers.

5. Results

Quantitative To measure the performance of both models in a quantitative statistical manner, we calculate the F1 score metric. The F1 score is used to evaluate the balance between precision and recall and is arguably a more robust metric in comparison to accuracy for evaluating an image classification based CNN model. Table 1 depicts the F1 score calculated at each 10 epoch iteration (total 30 epochs).

Epoch	No Attention	Attention
10	62.58%	73.58%
20	61.36%	74.62%
30	57.36%	73.89%

Table 1: F1 score comparison at epoch iterations.

The model with integrated visual attention (CBAM) received a 16% higher F1 score at the end of training compared to the model without attention (standard CNN). Attention added significant value to the predictive capabilities of the standard CNN model in our sample. Through this statistical model validation, we demonstrate the potential value of visual attention within COVID-19 classifiers for increasing statistical performance of the standard 2D CNN model.

Qualitative Radiologists are trained over many years to filter and focus on clinically relevant visual information. This enables them to spend more time on evaluating pertinent visual features. When presenting deep learning based techniques for the automated classification of CT scans, it would seem beneficial to follow a similar attention-based principle in order to start to approach radiologist level interpretation. To identify the value of visual attention within a COVID-19 detection pipeline visually, we use both models to visualize and compare qualitative heat-map activation results from the final convolutional feature extraction layer.

The Grad-CAM (gradient-weighted class activation mapping) algorithm [17] was used in order to visualize activation heat-maps from both models in order to compare

points of focus. Fig. 5 shows six images used for testing, the COVID-19 diagnosis class and heat-map for each model.

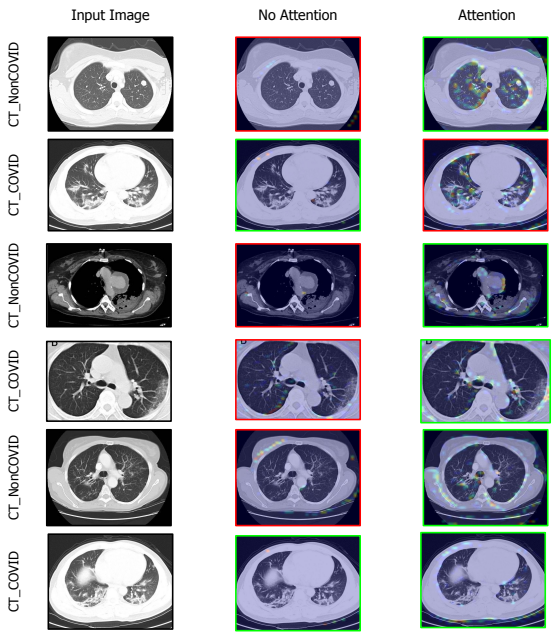

Figure 5: **Grad-CAM visualization results**. Comparison of the standard CNN and CBAM-augmented CNN activation heat-maps on CT scan images. **Red border**: incorrect prediction, **Green border**: correct prediction.

We observed major differences in activation heat-maps through this comparison. As seen, the standard model fails to cover visually important (lung) regions and is potentially extracting features from unimportant areas (border etc.). The attention augmented model infers spatial and channel feature maps and covers more significant and potentially higher frequency regions closer to the center of the CT scan.

In terms of interpretability, the ability to allow practitioners to understand how a deployed deep learning model is making decisions or where it is focusing is vital in making clinical decisions. With attention, we can aim to enhance interpretability by enabling neural networks to employ the ability to filter features similar to the trained radiologists.

6. Conclusion

We have presented a compact study which aims to empirically demonstrate the efficacy of visual attention within COVID-19 classification models. A high performing and fast deep learning model can aid radiologists with high work loads and in low resourced areas to interpret CT scans of the lung during the COVID-19 pandemic. This technology could be implemented to help in triaging of work lists and also run in the background to support quality assurance (QA). Attention shows promise as an efficient way to improve the performance of standard CNNs with minimal overhead [25]. In this comparison between a standard CNN and an attention-augmented CNN, the attention mechanism demonstrates higher performance through both enhanced qualitative visualization of activation heat-maps and higher F1 measures, in the COVID-19 single-slice CT classification task. This compact informational study was performed and submitted to the ICCV MIA-COV19D workshop [13].

References

[1] R. Alizadehsani, Z. Alizadeh Sani, M. Behjati, Z. Roshanzamir, S. Hussain, N. Abedini, F. Hasanzadeh, A. Khosravi, A. Shoeibi, M. Roshanzamir, et al. Risk factors prediction, clinical outcomes, and mortality in covid-19 patients. *Journal of medical virology*, 93(4):2307–2320, 2021.

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

[3] H. X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, T. M. L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei, et al. Performance of radiologists in differentiating covid-19 from non-covid-19 viral pneumonia at chest ct. *Radiology*, 296(2):E46–E54, 2020.

[4] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, et al. Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. *Radiology*, page 200463, 2020.

[5] T. M. Buzug. Computed tomography. In *Springer handbook of medical technology*, pages 311–342. Springer, 2011.

[6] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[7] M. M. Chun and J. M. Wolfe. Visual attention. *Blackwell handbook of perception*, 272310, 2001.

[8] A. S. Fauci, H. C. Lane, and R. R. Redfield. Covid-19—navigating the uncharted, 2020.

[9] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv*, 2020.

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[11] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[12] M. Karimi-Galougahi, N. Raad, and N. Mikaniki. Anosmia and the need for covid-19 screening during the pandemic. *Otolaryngology–head and Neck Surgery*, 163(1):96–97, 2020.

[13] D. Kollias, A. Arsenos, L. Soukissian, and S. Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. *arXiv preprint arXiv:2106.07524*, 2021.

[14] T. C. Kwee and R. M. Kwee. Chest ct in covid-19: what the radiologist needs to know. *RadioGraphics*, 40(7):1848–1865, 2020.

[15] H. Lee, H.-E. Kim, and H. Nam. Srm: A style-based recalibration module for convolutional neural networks. In *ICCV*, 2019.

[16] C. Pozzessere, D. C. Rotzinger, B. Ghaye, F. Lamoth, and C. Beigelman-Aubry. Incidentally discovered covid-19 pneumonia: the role of diagnostic imaging. *European radiology*, 30(9):5211–5213, 2020.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[18] P. Silva, E. Luz, G. Silva, G. Moreira, R. Silva, D. Lucio, and D. Menotti. Covid-19 detection in ct images with deep learning: A voting-based scheme and cross-datasets analysis. *Informatics in medicine unlocked*, 20:100427, 2020.

[19] E. D. Tenda, M. Yulianti, M. Asaf, R. Yunus, W. Septiyanti, V. Wulani, et al. The importance of chest ct scan in covid-19: A case series. *Acta med indones*, 52(1):68–73, 2020.

[20] T. P. Velavan and C. G. Meyer. The covid-19 epidemic. *Tropical medicine & international health*, 25(3):278, 2020.

[21] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[22] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[23] N. Wilson, S. Corbett, and E. Tovey. Airborne transmission of covid-19. *bmj*, 370, 2020.

[24] J. M. Wolfe. Visual attention. *Seeing*, pages 335–386, 2000.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[26] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020.