RadioTransformer: A Cascaded Global-Focal Transformer for Visual Attention–guided Disease Classification — Supplementary Material —

Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna

Stony Brook University, Stony Brook, New York, USA
{moinak.bhattacharya,prateek.prasanna}@stonybrook.edu

In this supplementary material, we provide detailed illustration of the global-focal block (Section 1), additional information on the datasets used in this work (Section 2), the different augmentations in student-teacher network (Section 3), more quantitative (Section 4), and qualitative (Section 5) results. We also present an analogy of the global-focal block with cellular pathways (Section 6).

1 Illustration of Global-Focal block

The global-focal block in the RadioTransformer architecture is detailed in Figure 1. The global and focal blocks are cascaded in parallel. The shifting window for each block is shown with the window in red color. High contrast patterns are learned by the focal blocks, shown in the orange box in Figure 1 and low contrast patterns are learned by global blocks, shown in the blue box in Figure 1. The TWL connection averages the features between the intermediate global and focal blocks.

2 Datasets

RSNA Pneumonia Detection challenge[14], and Cell Pneumonia[8] are pneumonia classification datasets consisting of radiographs with presence and absence of pneumonia. SIIM-FISABIO-RSNA COVID-19 Detection[9] dataset categorizes radiographs as negative for pneumonia, and typical, indeterminate, or atypical for COVID-19. COVID-19 Radiography database[1,12] comprises chest radiographs with COVID-19, normal, lung opacity and viral pneumonia classes. NIH Chest X-rays[17] and VinBigData Chest X-ray Abnormalities Detection[11] datasets comprise 14 common thorax diseases. We further include the more recent large-scale RSNA-MIDRC[16,15,2] and TCIA-SBU COVID-19 datasets [13,2] that contain only COVID-19 chest radiographs.

3 Augmentation

Figure 2 illustrates the various augmentations for different blocks of *RadioTransformer*. The images in the first and second rows are the inputs to the student

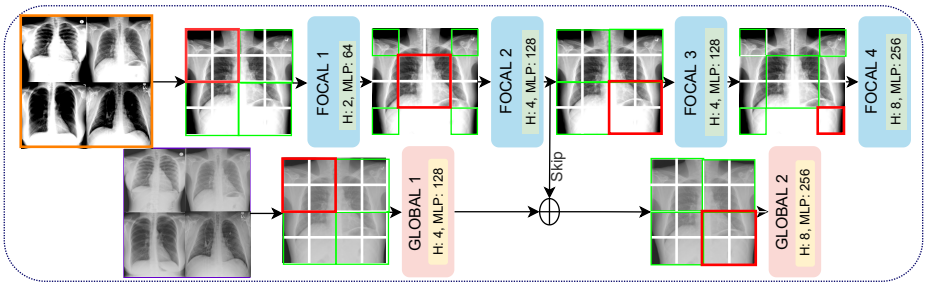

Fig. 1. Illustration of Global-Focal Network. The Focal network (top row) learns low-level representations with high-contrast images as input, as shown in the orange box. The global network (bottom row) learns high-level representations with low-contrast images as input, as shown in the violet box. The shifting windows, shown as red boxes, are implemented with incremental shift size, shown as the traversing of the red boxes diagonally. For the global network, there is a single shifting, and for the focal network, there are three incremental shifting of the windows. Both the windows shift from top-left to bottom-right. The number of Attention heads (H) and MLP heads (MLP) for different global and focal blocks are also shown.

focal and global blocks, respectively. The images in the third and fourth rows are the inputs to teacher focal and global blocks, respectively. As seen in the images, the teacher network implements hard augmentations compared to the student network. The focal block has a higher contrast value than the global block. For stateless augmentations, we use tf.image.stateless_random_contrast(.), tf.image.stateless_random_brightness(.), tf.image.stateless-_random_hue(.), and tf.image.stateless_random_saturation(.). More details on the augmentation parameters are provided in Supplementary table 1.

Augmentation	Contrast		Brightness	Hue	Saturation	
Parameter	lower	upper	max_delta	max_delta	lower	upper
Teacher-Global	2.0	2.2	0.8	0.8	2.0	2.5
Teacher-Focal	2.8	3.0	0.8	0.8	2.0	2.5
Student-Global	0.5	1.0	0.5	0.5	1.5	2.0
Student-Focal	1.0	1.5	0.5	0.5	1.5	2.0

Table 1. Augmentation parameters.

4 Quantitative Analysis

In addition to the AUC and F1 scores provided in the main paper, here we show the accuracy, precision, and recall values for classification tasks in the 8 datasets. In Supplementary table 2, the performance metrics for pneumonia classification

Name	Cell[8]					RSNA[14]					SIIM[9]					Rad[1,12]				
-	Acc.	AUC	F1	Pr.	Re.	Acc.	AUC	F1	Pr.	Re.	Acc.	AUC	F1	Pr.	Re.	Acc.	AUC	F1	Pr.	Re.
R50[5]	71.35	81.70	59.78	68.34	73.44	94.56	98.91	93.75	94.78	94.15	89.90	98.85	43.01	89.90	89.90	94.41	99.27	94.03	94.54	94.25
R101[5]	78.47	83.64	71.93	78.73	77.78	95.66	99.21	94.84	95.56	95.83	80.29	96.98	39.22	80.29	80.29	87.97	97.62	85.36	88.31	87.78
R152[5]	79.86	87.49	74.30	80.81	79.69	93.57	98.57	91.97	93.23	94.02	88.62	98.18	43.04	88.62	88.62	66.62	87.90	70.21	67.20	66.03
R50v2[6]	82.53	87.32	78.96	82.53	82.53	97.12	99.44	96.60	97.12	97.12	96.79	99.79	47.99	96.79	96.79	94.22	99.06	92.82	94.44	94.18
R101v2[6]	68.40	71.23	52.11	68.40	68.40	97.01	99.33	96.39	97.01	97.01	93.67	99.26	45.83	93.67	93.67	97.85	99.82	97.46	97.85	97.85
R152v2[6]	69.10	71.97	53.44	69.10	69.10	96.08	99.01	95.30	96.08	96.08	95.31	99.71	47.10	95.31	95.31	98.30	99.82	97.76	98.30	98.30
D121[7]	77.43	81.97	70.05	77.43	77.43	96.84	99.34	96.25	96.84	96.84	96.22	99.82	47.59	96.22	96.22	96.52	99.51	95.72	96.65	96.45
D169[7]	71.70	76.56	59.18	71.70	71.70	89.96	95.60	88.86	89.96	89.96	94.49	99.68	46.40	94.49	94.49	95.48	99.52	94.33	95.63	95.43
D201[7]	78.47	82.98	71.93	78.47	78.47	96.23	99.04	95.43	96.23	96.23	97.12	99.83	48.17	97.12	97.04	97.80	99.85	97.81	97.82	97.77
ViT-B16[3]	76.91	83.40	73.85	76.91	76.91	78.08	86.06	76.35	78.08	78.08	78.45	95.74	36.22	78.53	78.21	89.51	98.42	88.25	90.04	89.02
ViT-B32[3]	70.41	76.41	70.02	70.14	70.14	82.83	90.74	79.11	82.83	82.83	68.91	92.12	30.42	69.22	68.26	88.40	98.09	86.73	89.15	87.76
ViT-L16[3]	75.69	83.31	69.59	75.69	75.69	87.85	94.53	85.41	87.85	87.85	78.29	95.75	34.16	78.29	78.29	90.91	98.70	90.11	91.36	90.67
ViT-L32[3]	80.38	87.07	76.38	80.38	80.38	79.24	88.86	69.32	79.24	79.24	70.07	92.54	28.45	70.31	69.90	89.44	98.35	88.40	89.94	88.85
CCT[4]	71.18	74.59	62.10	71.18	71.18	83.84	92.04	80.60	83.84	83.84	78.12	95.33	32.63	78.12	78.12	92.19	99.11	92.52	92.33	92.09
Swin0[10]	74.83	83.74	66.04	75.13	73.96	96.87	99.57	96.27	96.79	97.11	96.38	99.66	47.63	72.19	99.92	97.94	99.92	97.53	98.31	97.54
Swin1[10]	78.65	86.91	73.74	78.25	79.34	97.17	99.58	96.65	97.14	97.22	95.72	99.56	47.30	66.48	99.67	95.48	99.64	94.94	95.71	95.17
RadT w/o (HVAT+VAL)	82.05	89.82	79.56	82.05	82.05	98.19	99.78	97.85	98.19	98.19	97.47	99.69	48.42	97.47	97.47	98.51	99.94	98.13	98.58	98.51
RadT	80.73	88.80	77.40	77.65	82.64	98.94	99.85	98.75	98.94	98.94	98.10	99.65	48.74	98.10	98.10	99.43	99.98	99.39	99.48	99.41

Table 2. Quantitative Comparison:1. F1(↑) and AUC(↑) are reported for the baselines and the proposed methodology.

Name	NIH[17]					VBD[11]					MIDRC[15,16]					SBU[13,2]				
-	Acc.	AUC	F1	Pr.	Re.	Acc.	AUC	F1	Pr.	Re.	Acc.	AUC	F1	Pr.	Re.	Acc.	AUC	F1	Pr.	Re.
R50[5]	35.28	74.04	11.91	36.03	34.38	62.96	95.86	21.76	97.11	48.50	85.44	96.32	23.04	85.83	85.20	43.29	65.16	15.11	43.29	42.93
R101[5]	32.95	73.30	11.20	33.81	31.98	66.41	96.24	32.77	88.45	55.03	80.59	93.87	22.31	80.88	80.35	93.92	99.20	24.22	94.37	93.52
R152[5]	30.13	71.37	10.67	30.51	29.28	66.47	96.58	32.42	90.62	53.24	62.42	83.09	19.22	62.78	62.01	96.73	99.61	24.58	96.93	96.54
R50v2[6]	35.04	73.11	11.42	35.56	34.30	66.02	96.32	34.11	90.27	53.04	91.78	98.72	23.93	91.85	91.69	59.81	78.27	18.71	59.85	59.71
R101v2[6]	36.37	73.46	11.99	37.02	35.80	66.26	96.55	32.18	91.18	53.27	10.77	42.13	04.86	10.55	10.44	63.55	82.47	19.43	64.35	62.75
R152v2[6]	32.05	73.23	11.93	32.94	31.04	66.04	96.54	32.69	91.76	53.27	85.69	95.89	23.07	85.75	85.61	85.39	96.25	23.03	85.50	85.21
D121[7]	33.67	78.83	13.81	15.34	73.37	64.31	96.01	28.71	91.02	50.63	99.01	99.82	24.88	64.85	100.00	70.47	88.35	20.67	70.68	70.24
D169[7]	33.27	79.90	15.21	16.60	73.67	66.04	96.46	32.90	90.25	53.77	99.75	99.84	24.97	56.30	100.00	67.38	85.95	20.13	67.49	67.22
D201[7]	36.02	81.38	14.84	17.76	75.45	66.42	96.41	34.66	88.69	55.09	99.92	99.99	24.99	65.59	100.00	72.92	89.53	21.08	73.06	72.71
ViT-B16[3]	34.54	82.06	07.50	42.99	21.24	64.14	95.69	34.80	83.00	54.67	20.39	42.15	08.47	19.93	19.33	29.84	50.22	11.49	28.85	26.41
ViT-B32[3]	38.19	83.77	06.51	48.48	22.94	60.75	94.58	30.57	88.86	47.69	53.87	76.52	17.50	54.97	51.81	57.52	77.75	18.26	58.42	56.19
ViT-L16[3]	32.22	81.60	08.16	43.32	16.28	63.66	95.40	33.99	80.80	55.15	28.78	47.79	11.17	28.52	27.47	45.08	62.72	15.54	45.38	43.69
ViT-L32[3]	38.66	84.96	06.35	47.04	25.42	63.44	95.36	33.24	86.29	52.51	25.66	47.35	10.21	24.87	23.68	08.52	30.82	03.92	06.01	05.41
CCT[4]	38.69	85.37	08.08	52.10	22.06	62.02	95.12	30.25	89.69	49.51	92.19	98.53	23.98	92.93	91.94	63.57	83.21	19.43	64.20	62.84
Swin0[10]	31.65	74.62	07.90	33.21	28.14	64.81	95.08	34.30	16.57	97.32	37.91	63.07	13.74	37.96	37.34	55.12	75.47	17.77	55.42	54.43
Swin1[10]	31.17	74.18	08.30	32.71	27.34	65.03	95.13	34.27	16.48	97.83	44.82	69.00	15.47	45.10	44.24	54.50	73.68	17.64	54.77	53.99
RadT w/o (HVAT+VAL)	38.56	85.48	05.97	49.92	20.33	65.96	96.83	37.64	83.87	56.49	50.17	71.78	16.70	50.47	49.67	79.79	93.75	22.19	81.84	77.71
RadT	38.52	85.43	04.21	45.48	26.73	66.54	96.84	37.32	82.35	57.90	57.07	79.60	18.17	43.48	72.45	79.69	94.76	22.18	83.89	75.15

Table 3. Quantitative Comparison:2. F1(↑) and AUC(↑) are reported for the baselines and the proposed methodology.

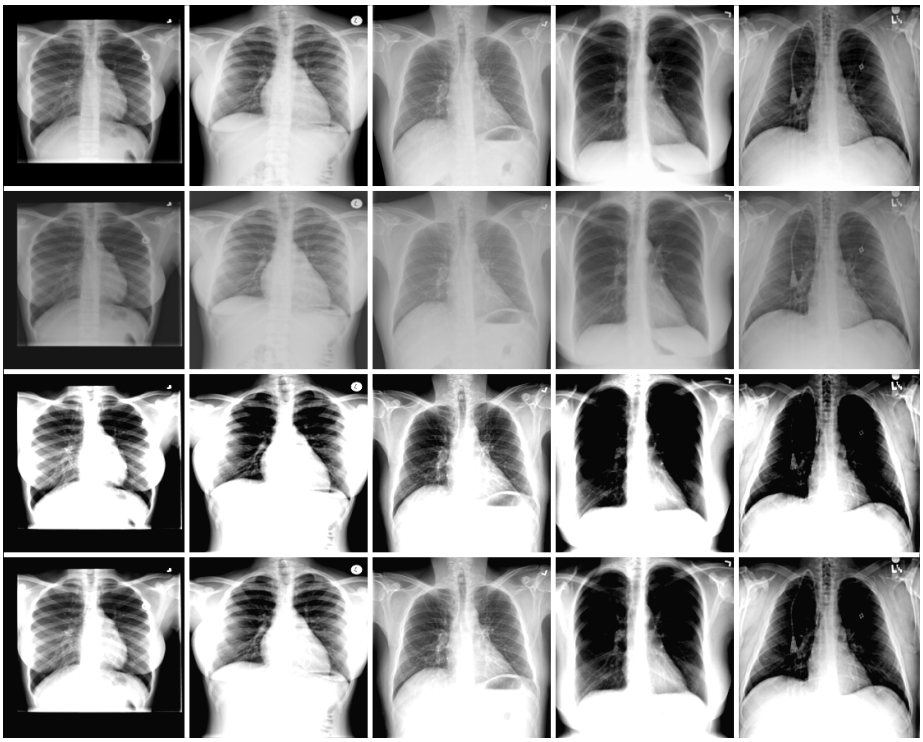

Fig. 2. Augmentations: The different augmentations to input images of student global-focal, and teacher global-focal blocks are shown.

datasets such as Cell, and RSNA Pneumonia Challenge dataset, and COVID-19 classification datasets such as SIIM-RSNA-FISABIO COVID-19 challenge, and Radiography dataset are shown. In Supplementary table 3, we show the performance metrics for 14 thoracic diseases classification tasks (in the NIH, and VinBigData datasets), and the COVID-19 classification task (in MIDRC and TCIA-SBU datasets).

5 Qualitative Analysis

We supplement our qualitative results (in Section 5.2 of the main paper) with additional class activation maps for both the datasets i.e., RSNA, and Radiography. In Figure 3, the RadT w/o (HVAT+VAL) and RadT class activation maps are shown for Normal and Pneumonia cases. Similarly, in Figure 4, the RadT w/o (HVAT+VAL), and RadT class activation maps are shown for Normal and COVID-19 cases. For both the datasets, the maps of RadT w/o (HVAT+VAL) show discrete patterns, and those of RadT show comparatively continuous patterns. In addition to all the previous discussions, we discuss another interesting

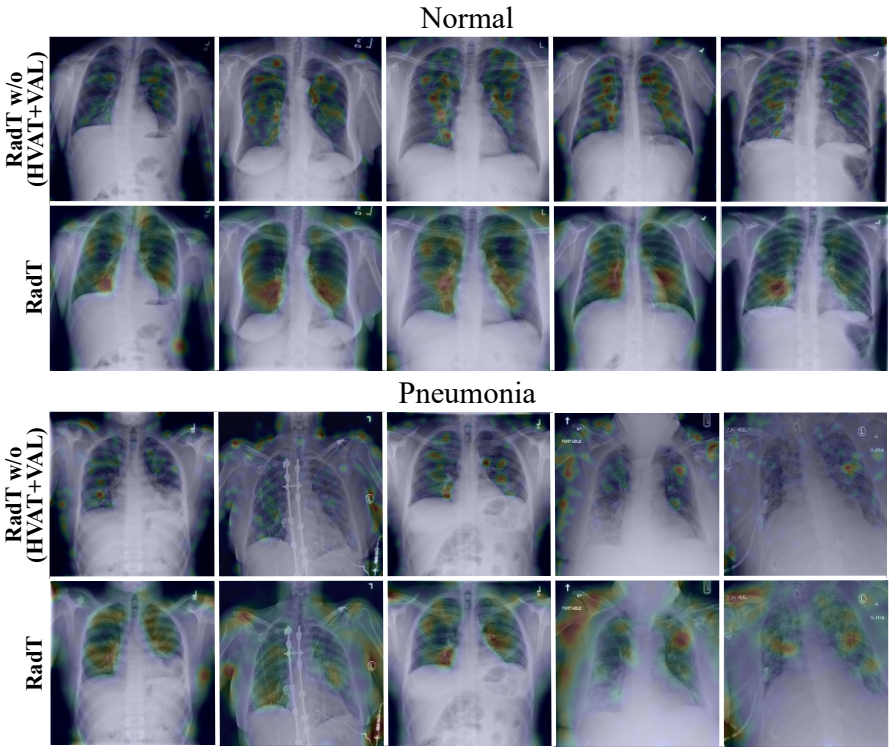

Fig. 3. Qualitative Comparison on RSNA dataset: Comparison of class activation maps from RadioTransformer w/o (HVAT+VAL) and RadioTransformer.

finding. In the fourth row of Figure 3, we observe that apart from clear attention on the white/fluid regions, there are some extraneous attention regions in the shoulders. Again, this phenomenon is not observed in the fourth row of Figure 4. This is clearly explainable from the ablation study in the main paper. For the RSNA dataset, the global block is showing better performance, hence the global block is activated in this case. The global block focuses on high-level features and in this case, it hypothesizes to identify features from non-relevant regions(like shoulder, etc) in addition to the white/fluid regions in the lungs. Whereas in the Radiography dataset, the focal block is activated and the attention regions perfectly intersect with the white/fluid regions.

6 Analogy with cellular pathways

Parvo, Magno, and Konio cells are ganglion cells that transfer information generated by the photoreceptors in the retina to the visual cortex in the brain. Structurally, Magno cells are larger, and have thick axons with more myelin while Parvo cells are smaller, and have less myelin and thinner axons. Func-

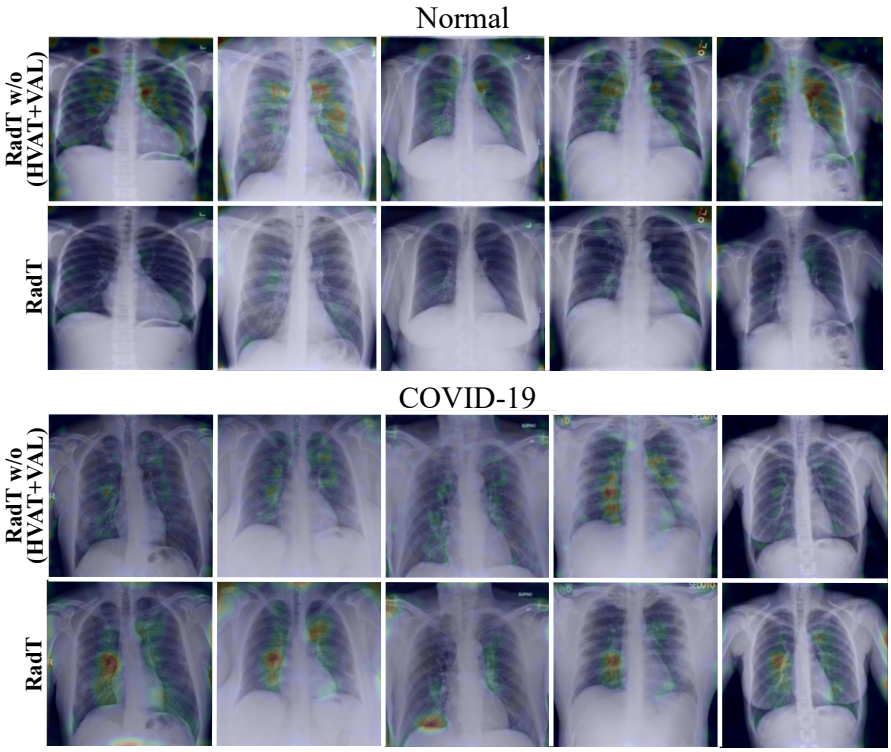

Fig. 4. Qualitative Comparison on Radiography dataset: Comparison of class activation maps from RadioTransformer w/o (HVAT+VAL) and RadioTransformer.

tionally, the Magno cells have a large receptive field; they respond rapidly to changing stimuli and detect robust/global details like luminance, motion, stereopsis, and depth. Parvo cells, on the other hand, have a smaller receptive field, respond slowly to stimuli, and detect finer/local details like chromatic modulation and the form of an object. The Global-Focal blocks in *RadioTransformer* are inspired by these cellular pathways.

References

1. Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Emadi, N.A., Reaz, M.B.I., Islam, M.T.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access **8**, 132665–132676 (2020). https://doi.org/10.1109/ACCESS.2020.3010287
2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging **26**(6), 1045–1057 (2013)

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H.: Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704 (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
8. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131 (2018)
9. Lakhani, P., Mongan, J., Singhal, C., Zhou, Q., Andriole, K.P., Auffermann, W.F., Prasanna, P., Pham, T., Peterson, M., Bergquist, P.J., et al.: The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. (2021)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
11. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. arXiv preprint arXiv:2012.15029 (2020)
12. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)
13. Saltz, J., et al.: Stony brook university covid-19 positive cases [data set] (2021)
14. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
15. Tsai, E.B., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: Data from medical imaging data resource center (midrc) - rsna international covid radiology database (ricord) release 1c - chest x-ray, covid+ (midrc-ricord-1c). The Cancer Imaging Archive (2021)
16. Tsai, E.B., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: The rsna international covid-19 open radiology database (ricord). Radiology **299**(1), E204–E213 (2021)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)