

Specialized Non-local Blocks for Recognizing Tumors on Computed Tomography Snapshots of Human Lungs

Aleksei Samarin, Aleksei Toropov,
Alina Dzestelova, Artem Nazarenko,
Egor Kotenko, Elena Mikhailova
ITMO University
St. Petersburg, Russia
avsamarin@itmo.ru, toropov.ag@hotmail.com,
aldzestelova@gmail.com, aanazarenko@itmo.ru,
kotenkoed@gmail.com, e.mikhailova@itmo.ru

Alexander Savelev, Alexander Motyko
St. Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
algsavelev@gmail.com, aamotyko@etu.ru

Abstract—This research endeavor is dedicated to the integration of specialized attentional mechanisms within the intricate web of deep neural network architectures aimed at discerning indications of lung carcinoma from monochromatic snapshots derived from computerized axial tomography. Within this exploration, we propose a myriad of adaptations to the traditional non-local blocks, infusing them with bespoke attentional nuances to resonate with the idiosyncrasies of medical imaging data. These bespoke adaptations ushered in discernible ameliorations in the performance metrics of the fundamental deep neural network model. Our solution facilitated a reduction in the model parameter count without compromising classification efficiency significantly. Additionally, it enabled a streamlined approach to feature extraction, contributing to enhanced interpretability and efficiency in the recognition process. These advancements were meticulously validated across test subsets meticulously curated from the Open Joint Monochrome Lungs Computer Tomography dataset, the Lung Image Database Consortium and Image Database Resource Initiative dataset, the Iraq-Oncology Teaching Hospital / National Center for Cancer Diseases dataset, Radiology Moscow and The Cancer Imaging Archive and from several others.

I. INTRODUCTION

In the realm of modern medicine, lung cancer stands out as a particularly lethal adversary, contributing significantly to the global toll of 1.76 million annual deaths [1]. Advanced diagnostic techniques heavily rely on computer tomography (CT) scans, offering detailed cross-sectional views of internal organs.

In parallel, the rise of online medical consultation platforms [2]–[5] highlights a growing demand for user-friendly interfaces capable of processing single-channel monochrome CT images. This shift underscores the need for simplified yet effective image analysis methods accessible to everyday users.

While existing methodologies for processing CT data exist [6]–[8], their suitability for rapid online diagnostics is hampered by resource-intensive hardware requirements and sluggish processing speeds [9], [10]. As a result, alternative

approaches for detecting lung neoplasms have gained traction, with some showcasing impressive results, such as the two-stage self-attention-based neural model [11]. However, the sequential nature of such approaches often leads to prolonged inference times, prompting investigations into streamlined architectures [12]–[14].

Our research focuses on refining neoplasm recognition frameworks to improve classification and segmentation accuracy. By introducing of specialized self-attention blocks and integrating them into deep neural network (DNN) structure, we successfully reduce inference times while maintaining high classification performance.

II. RELATED WORK

Biomedical imaging holds paramount significance in the detection of neoplasms within the pulmonary domain, necessitating robust methodologies for image classification and instance segmentation within automated diagnostic systems. An array of solutions has been devised to address the intricate task of neoplasm segmentation across both general and biomedical imaging domains [11]–[20].

Transformer-based architectures have risen to prominence as leading contenders in numerous benchmarks for biomedical image segmentation and classification [21]–[23]. However, it is imperative to acknowledge that the attention mechanism inherent within transformers was not originally tailored for biomedical imaging modalities [24]–[26]. The utilization of multi-headed attention [27], [28] within transformer architectures often incurs a substantial parameter count, computational resource demands, and relatively sluggish inference speeds. Moreover, extant solutions typically neglect to adapt the attention mechanism to the idiosyncrasies of biomedical image processing or capitalize on distinctive visual cues characteristic of such data [29], [30].

In our current study, we pivot our focus towards the refinement of the attention mechanism for the precise iden-

tification of neoplasms within lung CT images, leveraging single-channel monochromatic representations derived from CT scans. These high-resolution images afford detailed cross-sectional insights into pulmonary structures, facilitating the discernment of anomalies such as neoplastic growths. Central to our investigation is the adaptation of the attention mechanism encapsulated within non-local blocks [31], [32], which serve to bolster self-attentive capabilities within deep neural network architectures tailored for biomedical image analysis.

Prior research endeavors have explored diverse methodologies for neoplasm recognition within lung CT imagery, with certain approaches demonstrating considerable promise [11]–[14], which exhibited commendable performance on select datasets, underscoring the potential for specialized approaches within this domain. Nevertheless, there persists a notable lacuna in tailored methodologies that optimize non-local block construction specifically for neoplasm identification and segmentation within lung CT images.

Our ongoing research endeavors encompass the iterative development and meticulous evaluation of modified non-local blocks [11], [31], [32], meticulously considering the structural intricacies of neoplasms and the overarching visual characteristics inherent within lung CT imagery. By integrating these refined non-local blocks within established architectural frameworks such as U-Net [15], our overarching objective is to augment the efficiency and efficacy of neoplasm segmentation and classification within lung CT images, thereby catalyzing advancements in diagnostic precision and clinical decision-making.

III. PROPOSED SOLUTION

A. Model structure overview

Our approach draws upon the methodological framework delineated in the cited papers [12]–[14], a selection made due to its direct relevance to our research objective. The model expounded therein was meticulously crafted to optimize the two-stage methodology [11] tailored for the analysis of biomedical images, with a particular focus on lung neoplasm detection employing single-channel monochrome CT images. The authors proposed a solution adept at concurrent scene classification and neoplasm segmentation, exhibiting promising performance metrics within the domain of lung image analysis. Consequently, the foundational version we opted to scrutinize closely mirrors the conceptual framework articulated in the cited literature.

Central to our baseline solution is a U-Net [15] architecture augmented with attention blocks strategically incorporated into intermediate representations across diverse scales, as visually represented in Fig. 1.

In an earnest endeavor to streamline the architectural design, we embarked on a systematic exploration of the potentiality of eliminating non-local attention blocks. Regrettably, such an endeavor yielded a discernible decline in performance metrics, as meticulously documented in the Experimental Results section.

Moreover, our experimental findings shed light on the inherent constraints in the baseline architecture’s capacity for generalization, a phenomenon elucidated in detail within the corresponding section of the referenced literature.

Thus, the architectural framework under consideration serves as an exemplary foundational scaffold upon which to iteratively refine and tailor non-local blocks specifically for the task of lung neoplasm segmentation, leveraging single-channel monochrome CT images.

B. Basic non-local block (Basic NLB) construction

As discussed above, we adopted self-attention as the fundamental implementation of the non-local block, utilizing grid tokenization techniques (depicted in Fig. 2).

The operational framework of this attention mechanism, as depicted in Fig. 2, is widely recognized. Initially, the input tensor undergoes grid-based tokenization, partitioning it into rectangular segments. Each token (grid element) is subsequently transformed into an isomorphic tensor x_i , where $|x_i| = N = n * n$, and $dim(x_i) = d$. Subsequently, three sets of tensors Q, K, V are derived through trainable projection operations ($f(x), g(x), h(x)$) applied to the original tensor set x_i , resulting in

$$Q = f(x_i), K = g(x_i), V = h(x_i).$$

Following this, similarity coefficients between each element $q_i \in Q$ and other tensors are computed by evaluating dot products $q_i * k_j \forall j \in [1..n]$. Normalization, such as *SoftMax* normalization, is then applied to these dot products, often preceded by division by the square root of the dimension $dim(k_i) = d_k$, yielding

$$\alpha_i = SoftMax(q_i k_1 / d_k, q_i k_2 / d_k, \dots, q_i * k_n / d_k).$$

Subsequently, an attention map is constructed by linearly combining these coefficients α_i with the respective tensors v_i and applying a trainable projection operation $v(x)$ to the resultant expression:

$$o_i = v\left(\sum_{j \in [1..n]} (\alpha_i * v_j)\right).$$

Alternatively, this sequence of operations can be represented in matrix form, where Q, K, V are denoted in matrix format as

$$\begin{aligned} Q &= [f(x_1), \dots, f(x_n)], \\ K &= [g(x_1), \dots, g(x_n)], \\ V &= [h(x_1), \dots, h(x_n)]. \end{aligned}$$

Consequently, the resultant attention map can be computed as:

$$O = v\left(SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)\right).$$

The resulting attention map is employed to reweight the elements of the input tensor $X = [x_1, \dots, x_n]$ in conjunction with the residual connections concept, yielding: $Y = WO + X$.

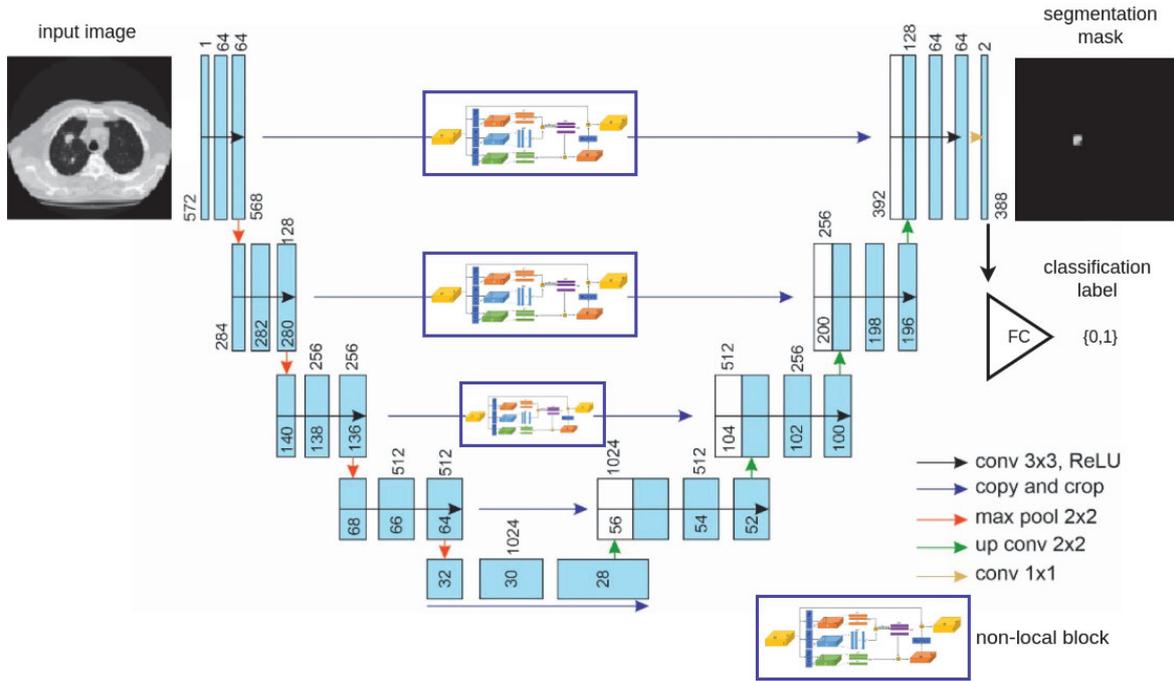


Fig. 1. U-Net based DNN model with attention modules for neoplasms presence recognition.

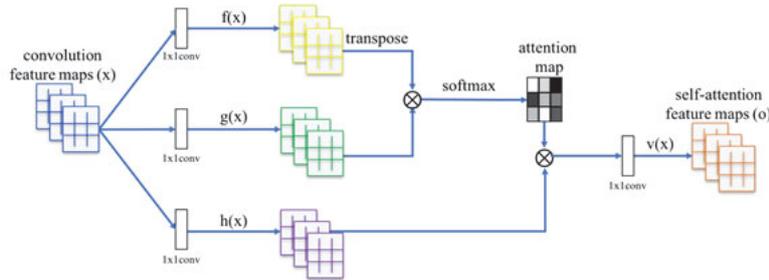


Fig. 2. Original non-local block scheme [12].

C. Localized non-local block (Localized NLB) modification

It is worth noting a significant aspect of the self-attention mechanism elucidated above, namely its non-local nature. Specifically, the process of weighting the components of O , as detailed earlier, involves computing the following linear combination:

$$o_i = v \left(\sum_{j \in [1..n]} (\alpha_i * v_j) \right).$$

As such, each tensor's attention map relies on similarity coefficients

$$\alpha_i = \text{SoftMax}(q_i k_1 / d_k, q_i k_2 / d_k, \dots, q_i * k_n / d_k)$$

that remain invariant regardless of the spatial arrangement of the corresponding tensors q_i and k_j . However, the specific context of our investigation, particularly the distinct characteristics and spatial organization of lung neoplasms, prompted us to explore methods to enhance the significance of neighboring regions relative to those more distant from the region of interest.

To implement this concept technically, we introduce an additional coefficient to scale the dot products of vectors from corresponding regions. This coefficient diminishes as the regions move away from the focal point. Consequently, we employ modified coefficients calculated using the formula:

$$\alpha_i^* = \text{SoftMax}(\text{dist}(i, 1)q_i k_1 / d_k, \text{dist}(i, 2)q_i k_2 / d_k, \dots, \text{dist}(i, n)q_i k_n / d_k),$$

where

$$\text{dist}(i, k) = 1 / \min(|\text{column}_i - \text{column}_k|, |\text{row}_i - \text{row}_k|),$$

with column_x and row_x representing the column and row numbers of the grid position of region x , respectively. The overall operational schema of the proposed modification to the non-local block is depicted in Fig. 3.

D. Symmetry-based non-local block (S-B NLB) modification

In our exploration of modifying the attention mechanism for the task at hand, we investigated an alternative heuristic

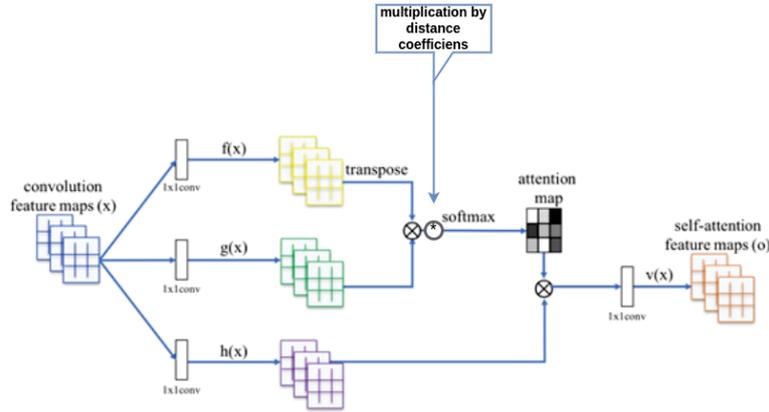


Fig. 3. Localized non-local block scheme.

rooted in the concept of symmetry within vectorized tokens. This approach was motivated by observations pertaining to the spatial distribution and morphology of neoplasms within lung CT images. Unlike nodules, which manifest as irregular protrusions in lung CT images, neoplasms within lung CT scans often exhibit characteristic shapes and occupy considerable spatial extents.

To accommodate these distinct features observed in lung CT images, we adjusted the weighting scheme of tokens within the intermediate attention map based on dot product vectors derived from the partitioning of original token vectors. Specifically, we introduced modified self-attention feature maps defined as

$$o_i^* = v(\theta_i * \sum_{j \in [1..n]} (\alpha_i * v_j)),$$

where

$$\theta_i = (x_i[1, \dots, \lfloor (n+1)/2 \rfloor] * x_i[\lfloor (n+1)/2 \rfloor + 1, n]) / |x_i[1, \dots, \lfloor (n+1)/2 \rfloor] * x_i[\lfloor (n+1)/2 \rfloor + 1, n]|.$$

This modification, illustrated in Fig. 4, aims to enhance the attention mechanism's sensitivity to the distinctive features of neoplasms in lung CT images, thereby optimizing their recognition within the analytical framework under consideration.

E. Symmetry-localized non-local block (S-L NLB) modification

Upon confirming the efficacy of the aforementioned adaptations within our research framework, detailed in Section IV, we made the decision to integrate these concepts. This amalgamation resulted in the refinement of the attention mechanism's architecture (refer to Fig. 5).

Consequently, to implement this amalgamated modification, we introduced localized adjustments while computing similarity coefficients

$$\alpha_i = \text{SoftMax}(\text{dist}(i, 1)q_i k_1 / d_k, \text{dist}(i, 2)q_i k_2 / d_k, \dots, \text{dist}(i, n)q_i k_n / d_k),$$

as elucidated earlier. Additionally, we incorporated posterior modifications to the attention map based on the symmetric heuristic

$$o_i^* = v(\theta_i * \sum_{j \in [1..n]} (\alpha_i * v_j)),$$

as discussed previously.

IV. EVALUATION

A. Dataset description

In this study, we have collected publicly available lung CT-snapshots datasets: LIDC-IDRI, IQ-OTH/NCCD, Lung-PET-CT-Dx and Radiology Moscow and The Cancer Imaging Archive.

1) *LIDC-IDRI*: The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset contains diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions. It encompasses over 1,000 cases, each with associated radiologist annotations of nodules, including their locations and characteristics.

2) *IQ-OTH/NCCD*: Iraq-Oncology Teaching Hospital / National Center for Cancer Diseases (IQOTH/NCCD) encompasses CT scans from both healthy individuals and lung cancer patients at various disease stages. Expert oncologists and radiologists from these centers annotated the slides within the dataset, which comprises a total of 1,190 images derived from CT scans across 110 cases.

3) *Lung-PET-CT-Dx*: The Lung-PET-CT-Dx dataset contains paired PET and CT scans of patients with lung cancer. These images are annotated with information regarding the location, size, and type of lung tumors. The dataset contains more than 200,000 images from 355 patients.

4) *Radiology Moscow and The Cancer Imaging Archive*: The dataset employed for training, validation, and testing purposes encompasses a total of 10,052 single-channel monochrome images. These images, sourced from combined datasets available from Radiology Moscow and The Cancer Imaging Archive, are meticulously categorized based on the presence or absence of neoplasms. Notably, the dataset is evenly divided into two distinct classes, facilitating balanced

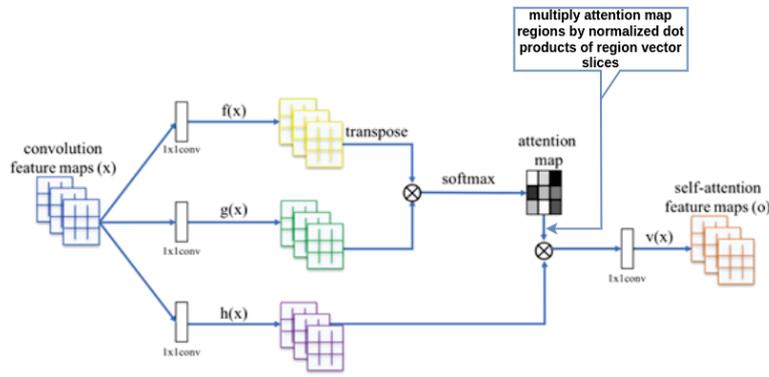


Fig. 4. Symmetry-based non-local block scheme.

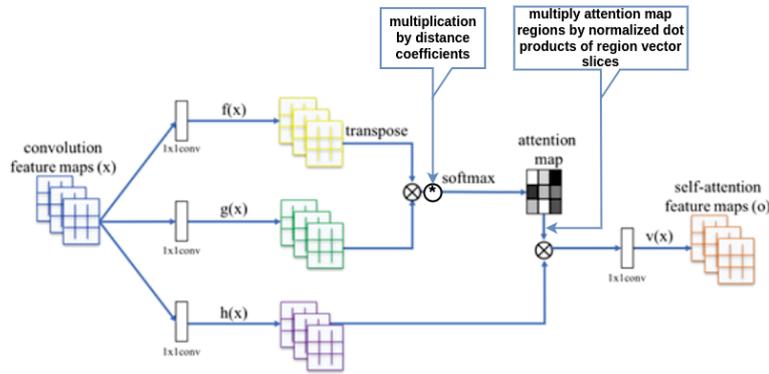


Fig. 5. Symmetry-localized non-local block scheme.

representation during training, testing, and validation phases, adhering to a partitioning ratio of 7,196 for training, 1,428 for validation, and 1,428 for testing.

It is noteworthy that for some of the listed datasets we needed to transform the image format, converting the detection markup into a segmentation markup (by applying the watershed method to the region of interest). The markup for binary classification was obtained naturally from the meta-information presented in the listed data sets.

B. Experimental results

Based on the aforementioned criteria, the our combined dataset is partitioned into two distinct classes. Consequently, we rely on conventional metrics utilized for assessing the efficacy of binary classification, including Precision, Recall, and F1-score. And mIoU and mDice for segmentation task.

Furthermore, alongside evaluating the classification performance, we conduct a comparative analysis of the inference time across various neural network architectures investigated, considering our hardware configuration and the number of model parameters.

The findings are presented in Table I for the segmentation task and in Table II for the classification task.

The outcomes depicted illustrate that proposed specialized attention blocks usage allows to increase the performance of baseline DNN architectures.

TABLE I. RESULTS OF THE CLASSIFICATION TASK

Method	classification metrics		
	precision	recall	F1
L-Net	0.782	0.692	0.734
EfficientNet-b4	0.822	0.791	0.806
ResNet50	0.813	0.774	0.793
Unet encoder with self attention	0.799	0.769	0.784
MultiAttention over ResNet50	0.837	0.798	0.817
MultiAttention X 1 over L-Net	0.846	0.799	0.822
Unet + S-B NLB	0.878	0.836	0.856
Unet S-L NLB	0.884	0.843	0.863

TABLE II. RESULTS FOR SEGMENTATION TASK

Method	seg metrics	
	IoU	Dice
L-Net	0.801	0.890
Unet++	0.823	0.903
DeepLabV3+	0.815	0.898
Unet with self attention	0.835	0.910
Unet + localized NLB	0.846	0.917
Unet + S-B NLB	0.851	0.920
Unet + S-L NLB	0.878	0.935

V. CONCLUSION

In this study, we present an investigation into the application of specialized attention blocks integration into DNN structure

for recognizing neoplasms in lung CT images. Our efforts have culminated in the development of a special types of non-local blocks that integration achieves better performance on the joint dataset for both of the classification problem statement and the segmentation problem statement, surpassing baseline models without specialized attention integration. We also conduct a comprehensive performance comparison with alternative approaches explored in our research.

In addition to detailing the neural network architectures utilized in our study, we provide supplementary segmentation labeling for the datasets under consideration. Future endeavors may focus on extending the application of our developed methodologies to address various challenges in biomedical image processing beyond neoplasm recognition in lung CT snapshots.

REFERENCES

- [1] A. A. Thai, B. J. Solomon, L. V. Sequist, J. F. Gainor, and H. R. S., "Lung cancer," *The Lancet*, vol. 398, no. 10299, pp. 535–554, 2021.
- [2] "Amwell." [Online]. Available: <https://amwell.com/cm/>
- [3] "Betterhelp." [Online]. Available: <https://www.betterhelp.com/>
- [4] "Yandex health." [Online]. Available: <https://health.yandex.ru/>
- [5] "Sber med ai." [Online]. Available: <https://amwell.com/cm/>
- [6] A. Shimazaki, D. Ueda, A. Choppin, A. Yamamoto, T. Honjo, Y. Shimahara, and Y. Miki, "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," 2022.
- [7] K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek, "Explainable machine learning for lung cancer screening models," *Applied Sciences*, vol. 12, p. 1926, 02 2022.
- [8] G. Kasinathan and S. Jayakumar, "Cloud-based lung tumor detection and stage classification using deep learning techniques," *BioMed Research International*, vol. 2022, 2022.
- [9] H. Yang, L. Chen, Z. Cheng, M. Yang, J. Wang, C. Lin, Y. Wang, L. Huang, Y. Chen, g. S. Pen, Z. Ke, and W. Li, "Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study," *BMC Medicine*, 2021.
- [10] M. A. Heuvelmans, P. M. van Ooijen, S. Ather, C. F. Silva, D. Han, C. P. Heussel, W. Hickey, H.-U. Kauczor, P. Novotny, H. Peschl, M. Rook, R. Rubtsov, O. von Stackelberg, M. T. Tsakok, C. Arteta, J. Declerck, T. Kadir, L. Pickup, F. Gleeson, and M. Oudkerk, "Lung cancer prediction by deep learning to identify benign lung nodules," *Lung Cancer*, vol. 154, pp. 1–4, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169500221000453>
- [11] A. Samarin, A. Savelev, and V. Malykh, "Two-staged self-attention based neural model for lung cancer recognition," in *2020 Science and Artificial Intelligence conference (SAI ence)*. IEEE, 2020, pp. 50–53.
- [12] A. Samarin, A. Savelev, A. Toropov, A. Dzestelova, V. Malykh, E. Mikhailova, and A. A. Motyko, "One-stage classifiers based on u-net and autoencoder with attention for recognition of neoplasms from single-channel monochrome computed tomography images," *Pattern Recognition and Image Analysis*, vol. 33, pp. 132–138, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259310593>
- [13] A. Samarin, A. Savelev, A. Toropov, A. Dzestelova, V. Malykh, E. Mikhailova, and A. Motyko, *Prior Segmentation and Attention Based Approach to Neoplasms Recognition by Single-Channel Monochrome Computer Tomography Snapshots*, 08 2023, pp. 561–570.
- [14] A. Samarin, A. Savelev, A. Toropov, A. Dzestelova, V. Malykh, E. Mikhailova, and A. Motyko, "One-staged attention-based neoplasms recognition method for single-channel monochrome computer tomography snapshots," *Pattern Recognition and Image Analysis*, vol. 32, pp. 645–650, 10 2022.
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018. [Online]. Available: <http://arxiv.org/abs/1807.10165>
- [17] F. Isensee, J. Petersen, S. A. A. Kohl, P. F. Jäger, and K. H. Maier-Hein, "nnu-net: Breaking the spell on successful medical image segmentation," *CoRR*, vol. abs/1904.08128, 2019. [Online]. Available: <http://arxiv.org/abs/1904.08128>
- [18] C. Lahiff and J. E. East, "Endoscopic approach to polyp recognition," *Frontline Gastroenterology*, vol. 8, no. 2, pp. 98–103, 2017.
- [19] N. A. Obukhova, A. A. Motyko, U. Kang, S.-J. Bae, and D.-S. Lee, "Automated image analysis in multispectral system for cervical cancer diagnostic," in *2017 20th conference of open innovations association (FRUCT)*. IEEE, 2017, pp. 345–351.
- [20] N. Obukhova, A. Motyko, B. Timofeev, and A. Pozdeev, "Method of endoscopic images analysis for automatic bleeding detection and segmentation," in *2019 24th Conference of Open Innovations Association (FRUCT)*. IEEE, 2019, pp. 285–290.
- [21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
- [22] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.
- [23] H. Tao, K. Mao, and Y. Zhao, "Dbt-unetr: Double branch transformer with cross fusion for 3d medical image segmentation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1213–1218.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [25] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.
- [26] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.
- [27] G. Park, C. Han, W. Yoon, and D. Kim, "Mhsan: multi-head self-attention network for visual semantic embedding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1518–1526.
- [28] B. Ghogh and A. Ghodsi, "Attention mechanism, transformers, bert, and gpt: tutorial and survey," 2020.
- [29] D. Jha, N. K. Tomar, V. Sharma, and U. Bagci, "Transnetr: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing," *arXiv preprint arXiv:2303.07428*, 2023.
- [30] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 151–160.
- [31] 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*, 2018, pp. 7794–7803.
- [32] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 191–207.