Specialized Image Descriptors Adaptation for Generated Images Recognition

Aleksei Samarin, Aleksei Toropov,
Egor Kotenko, Artem Nazarenko,
Elena Mikhailova, Valentin Malykh
ITMO University
St. Petersburg, Russia
avsamarin@itmo.ru, toropov.ag@hotmail.com,
kotenkoed@gmail.com, aanazarenko@itmo.ru,
e.mikhailova@itmo.ru, valentin.malykh@phystech.edu

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

***Abstract*—This study introduces an innovative method for recognizing automatically generated images by utilizing adapted descriptors specifically designed to analyze unique structural and morphological features characteristic of artificially created content. The methodology focuses on analyzing features inherent to image generation processes, ensuring the optimization of descriptors for identifying complex and subtle patterns associated with generative algorithms. The integration of these specialized descriptors not only enhances recognition accuracy but also enables the extraction of interpretable features that provide deeper insights into the key principles of artificial content creation. To validate the effectiveness of the proposed method, an annotated dataset was developed and used to test and compare performance against various classification algorithms, including deep learning-based neural networks. Experimental results demonstrated that the proposed approach achieves high efficiency. These findings underscore the significance of the proposed methodology in advancing automated systems for the analysis of generated content in areas such as authenticity verification, media security, marketing, and digital validation while maintaining high computational efficiency and interpretability of results.**

I. Introduction

In recent years, the automation of visual data analysis has become a cornerstone of numerous studies across various fields. These include medicine, industry, forensic science, environmental monitoring, and other domains where visual information plays a crucial role in decision-making processes. One of the most dynamically evolving areas within this domain is the analysis of images generated using artificial neural networks and other generative models.

With the growing popularity of generative models, such as Generative Adversarial Networks (GANs) and other content creation tools, the volume of artificially generated images has been increasing exponentially. This surge highlights the growing importance of distinguishing between generated and real content. Such solutions are applicable in various domains, including marketing and advertising, where verifying the authenticity of visual materials before publication is essential, as well as in mass media, which aims to prevent the dissemination of fake news and manipulated images. Additionally, security applications require identifying falsified documents or photographs. Verification and identification systems also demand media file authenticity checks in the context of digital control and forensic investigations. For example, these solutions can be used to detect counterfeit product images on e-commerce platforms or identify manipulated photographs in news publications.

The automation of image analysis significantly enhances the accuracy and speed of data processing. It offers several benefits, such as accelerating the processes of media file validation and verification, reducing the costs associated with manual content inspection, improving analysis accuracy through the use of specialized algorithms, and enabling scalability for large data streams in real time.

However, the task of recognizing artificially generated images presents several challenges. Modern generative models produce content with a high level of detail and realism, complicating its identification. Creating large, high-quality datasets for training algorithms remains a challenging task, as does the high computational complexity of modern deep learning models, which require substantial resources for training and inference. Artificially generated images often contain hidden defects, such as inconsistencies in lighting levels, perspective, or textures. For instance, an object placed against a shoreline background may exhibit significant differences in water level on either side of the object. Another common defect is the unnatural alignment or merging of repetitive patterns, such as distorted tiling in brick walls or uneven spacing between elements in geometric designs, which can break the visual coherence of the scene. Some examples of these defects are shown in Fig. 1).

Deep neural networks (DNNs) [1]–[4], including convolutional neural networks (CNNs) [5] and transformer-based models [6], [7], have demonstrated high accuracy in solving image classification tasks and in different appications [8], [9]. These architectures are widely used in computer vision due to their ability to efficiently identify complex patterns. However, their application has limitations, including dependence on large volumes of labeled data, high computational complexity requiring powerful hardware, and a lack of interpretability,

Fig. 1. The examples of the generated images with inconsistencies and hidden defects

which complicates result analysis in tasks where understanding object characteristics is critical.

One of the key features of generated images is the presence of artifacts related to the inconsistency of object shapes. Detecting such artifacts requires an analysis of the global features of the image. Specialized descriptors enable the identification of inconsistencies in object structure, the formation of interpretable features that help understand the nature of generation errors, reduced dependency on large datasets, and high computational efficiency. Examples of such light-weight, analytically defined descriptors [10], [11] include histograms of oriented gradients [12], [13] and local binary patterns [14], [15], which have proven effective in structural image analysis tasks.

This study builds upon previous research on specialized image descriptors [16]–[19] and adapts them to classify various images as artificially generated or created through traditional methods. Our method employs analytically defined descriptors to characterize the unique global features of artificially generated images, enabling accurate and interpretable classification. As part of our research, we have also collected and annotated a specialized dataset with synthetic images, which enables an objective evaluation of the proposed approach (for more details, see Section III.A).

Integrating specialized image descriptors into automated systems for validation, verification, and identification addresses the challenges of analyzing media samples across domains such as security, marketing, and journalism while minimizing computational resource requirements. The proposed approach combines advanced analytical methods with practical applications, fostering the development of accessible and efficient tools for analyzing diverse images and media content.

II. Proposed solution

In biomedical image analysis, previous research has underscored the importance of customizing methodologies to effectively address specific challenges [6], [7], [20]. For example, numerous investigations have highlighted the advantages of tailoring neural network approaches for tasks such as processing CT scans and identifying polyps in endoscopic imagery. These customizations not only improve computational performance and accelerate processing but also enhance the accuracy and reliability of outcomes for the targeted applications.

Furthermore, significant advancements have been achieved in the analysis of images containing embedded textual elements. This work has resulted in the development of resource-efficient techniques capable of operating on standard central processing units, thereby achieving substantial computational savings while maintaining high levels of accuracy. Such methods have proven particularly successful in multimodal tasks requiring the integration of both visual and textual information [10], [16]–[18], [21]. Additional studies have demonstrated the effectiveness of specialized descriptors for managing complex images, where the semantic understanding relies on a hierarchical structure of visual components, including descriptors. These developments highlight the potential of descriptor-driven methodologies in solving intricate multimodal challenges.

Building on this foundation, our research focuses on applying specialized descriptors to classify artificially generated images. By drawing from insights in synthetic image analysis and descriptor-based methodologies, we propose a robust framework tailored to the unique requirements of distinguishing generated content from authentic visuals. The subsequent sections elaborate on the configurations and types of specialized descriptors developed for this purpose, demonstrating their effectiveness in addressing the complexities inherent to generated image recognition.

A. Special image descriptors

We introduce a novel approach to specialized image descriptors, originally developed for analyzing images with embedded textual content. The computation process for these descriptors is inherently parallelizable, ensuring minimal demand for computational power. This design makes the method highly efficient and suitable for applications in environments with limited computational resources, such as edge devices, real-time processing systems, and mobile applications. The adaptability of the method allows it to be integrated seamlessly into a variety of imaging pipelines, enhancing its usability across multiple domains.

These descriptors are designed to extract the most meaningful information from the spatial relationships of areas exhibiting significant brightness fluctuations. By focusing on these key variations, the method ensures high sensitivity to subtle changes in image structure, which is particularly beneficial for tasks involving text recognition, document analysis, and scene understanding. Additionally, the method allows simultaneous extraction of data from multiple image regions, ensuring a more comprehensive analysis of complex patterns. This parallel extraction process enhances efficiency while preserving accuracy, making it particularly useful in large-scale image processing tasks. Based on this concept, the descriptors are formulated as paths traced by agents navigating the image from predefined starting points, following specific movement rules. These paths serve as structural representations of the

image, capturing essential spatial characteristics. Two movement strategies, which demonstrated superior performance on our dataset, are highlighted in this study, showcasing the method's adaptability to diverse imaging scenarios. The flexibility of these strategies allows them to be fine-tuned for specific applications, further improving their effectiveness in real-world use cases.

Moreover, the final architecture resembles a two-headed neural network, where one branch is responsible for the calculation of descriptors, and the second serves as a general neural network encoder. The computation results from both branches are projected into a joint space, creating a unified representation that effectively combines structural and learned features. Within this space, a fully connected classification layer is employed to distinguish patterns and make predictions with high accuracy. This hybrid architecture provides a robust foundation for various computer vision tasks, enabling precise and efficient image analysis while leveraging both handcrafted and learned representations.

Image descriptor type A employs a specific strategy for guiding agent movement. At each step, the agent determines its movement direction based on the following rule:

$$m^v_{i+1} = \underset{m \in M_p}{}(|R^i_1 - R^i_2| + c_p * 1_{\{p\}}(m)),$$

$$(R^i_1, R^i_2) = \begin{cases} (R^i_{up}, R^i_{down}), \textit{ if} \\ |R^i_{up} - R^i_{down}| > |R^i_{left} - R^i_{right}| \\ (R^i_{left}, R^i_{right}), \textit{ otherwise} \end{cases},$$

$$R^i_{up} = I[x_i - s/2 : x_i + s/2; y_i - s : y_i],$$

$$R^i_{down} = I[x_i - s/2 : x_i + s/2; y_i : y_i + s],$$

$$R^i_{left} = I[x_i - s : x_i; y_i - s/2 : y_i + s/2],$$

$$R^i_{right} = I[x_i : x_i + s; y_i - s/2 : y_i + s/2],$$

where m^v_i stands for a movement direction, i stands for a step number, I denotes an input image, and (x, y) denotes position of a pixel on the input image, p stands for priority movement direction, M_p is a subset of $\{up, down, left, right\}$ that denotes allowed movements according to priority direction p, c_p denotes bonus for movement along the priority direction and s stands for a step size in pixels.

The path traced by an agent with an assigned priority direction can be expressed as follows:

$$T^p(x_0, y_0) = (m_1(x_0, y_0), ..., m_N(x_0, y_0)),$$

where $T^p(x_0, y_0)$ – representing the trajectory of an agent starting at with priority movement direction p and initial position (x_0, y_0), $m_i(x_0, y_0)$ stands for chosen agent movement at step i with predefined initial position and priority direction and N denotes the length of each trajectory (if an edge of the image achieved before making N steps then trace is padded with a special value). Consequently, the trajectories for this descriptor type are typically aligned with the contours of the input image (Fig. 2).

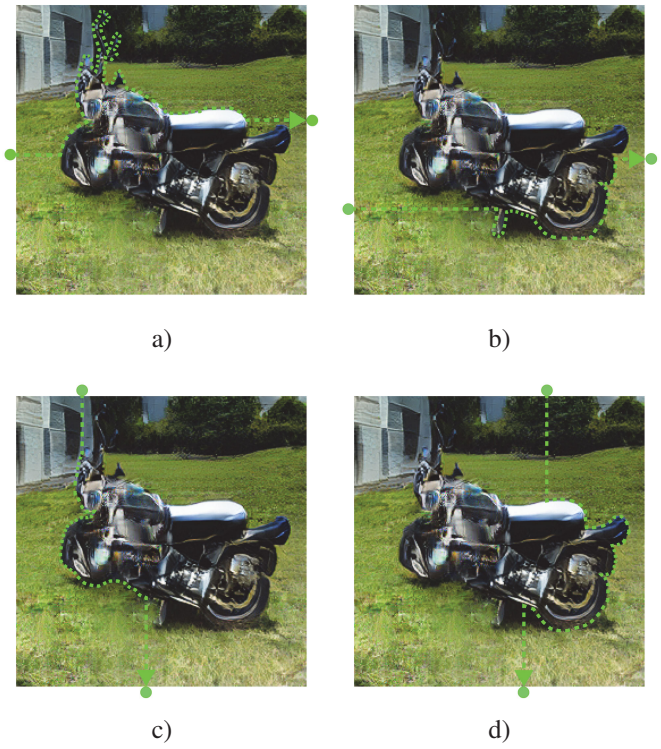

Fig. 2. Type A descriptors trace illustration: a,b) horizontally oriented trajectories; c,d) vertically oriented trajectories.

All of the trajectories are grouped by priority directions and initial positions:

$$T^{up} = (T^{up}(x_0, H), ..., T^{up}(x_A, H)),$$

$$T^{down} = (T^{down}(x_0, 0), ..., T^{down}(x_A, 0)),$$

$$T^{left} = (T^{left}(W, y_0), ..., T^{left}(W, y_B)),$$

$$T^{right} = (T^{right}(0, y_0), ..., T^{right}(0, y_B)),$$

where A and B stands for horizontal-oriented and vertical-oriented agents number, W denotes an input image width and H denotes image height. Finally, we merge groups of trajectories for each direction into the complete image descriptor that can be described as follows:

$$T = (T^{up}, T^{down}, T^{left}, T^{right}).$$

Based on this formulation, it is straightforward to design a procedure for calculating the descriptor in several steps proportional to the pixel count of the input image. Additionally, the computation process is inherently parallelizable and can be implemented with minimal complexity.

Image descriptor type B differs from type A only with the movement direction rule:

$$m^v_{i+1} = \underset{m \in M_p}{}(|R^i_1 - R^i_2| + \alpha * c_p * 1_{\{p\}}(m)),$$

where equals 1 if movement direction was changed at least once and 0 otherwise. That modification allows the agent to

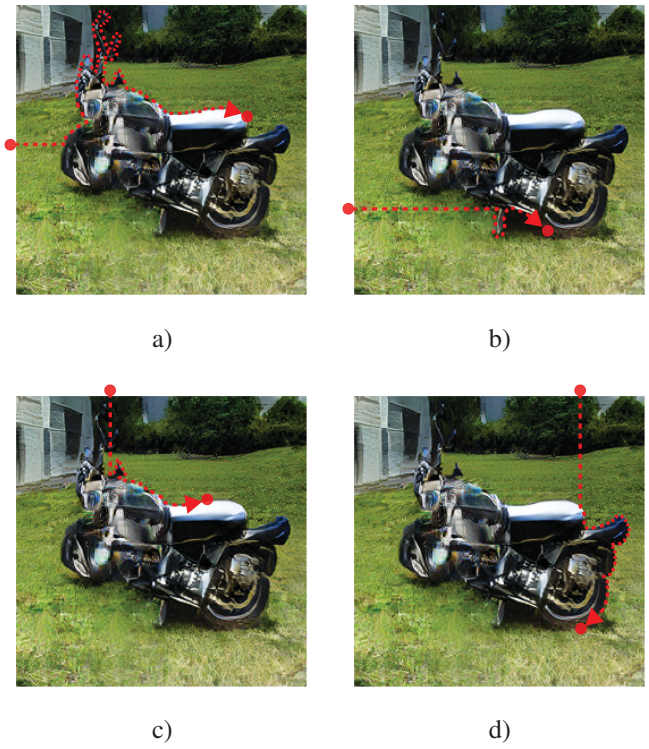

Fig. 3. Type B descriptors trace illustration: a,b) horizontally oriented trajectories; c,d) vertically oriented trajectories.

avoid priority direction influence if the significant element was found during the agent's movement (Fig. 3).

Image descriptor type C (presented in [21]–[23]) is also based on the agent's trace concept. However, each movement type is selected according to the following expression:

$$m^v_{i+1} = \underset{m \in M}{}(\mathrm{Var}[I[x,y]] + c(m,v)),$$

where

$$(x,y) \in P((x_i,y_i),s),\ v \in \{up, down, left, right\}.$$

Image descriptor type D represents a method developed as part of the combined framework discussed in this work. Following the methodology outlined in [19], [23], we implement a dual-path feature extraction mechanism to serve as the image descriptor. This method employs two variants of EfficientNet concurrently to generate image embeddings. These embeddings are subsequently concatenated and transformed to align with the classifier's input requirements. Unlike agent-based approaches, this method leverages well-established components, resulting in a simpler and more streamlined architecture.

The structure of our feature extraction module is illustrated in Fig. 4. For this implementation, we utilize EfficientNet B2 and EfficientNet B3 as the two branches of the descriptor, as this pairing demonstrated optimal performance in our feature extraction experiments.

Image descriptor type E relies on a learnable procedure for extracting image descriptors, bearing similarities to *Image descriptor type D*). However, rather than training a convolutional neural network, this method employs gradient descent optimization to fine-tune the parameters governing movement strategies. The descriptor construction follows the same agent-based trace approach as used in *image descriptors types A, B* and *C* an adapted trainable rule for selecting step directions during agent movement:

$$m^v_{i+1} = \underset{m \in M_p}{}(SoftMax(1_{M_p}(up) * \\ * Patch(I,(x_i,y_i),d) \circledast Kernel_{up},$$

$$1_{M_p}(down) * Patch(I,(x_i,y_i),d) \circledast Kernel_{down},$$

$$1_{M_p}(left) * Patch(I,(x_i,y_i),d) \circledast Kernel_{left},$$

$$1_{M_p}(right) * Patch(I,(x_i,y_i),d) \circledast Kernel_{right})),$$

where

$$Patch(I,(x_i,y_i),d) = \\ = I[x_i - d/2 : x_i + d/2; y_i - d/2 : y_i + d/2].$$

$Kernel_{up}, Kernel_{down}, Kernel_{left}, Kernel_{right}$ — a set of trainable kernels of size d. Convolutions with presented kernels stand for the usefulness of movement along the corresponding direction (Fig. 5). It should be noted that kernel values are trained as a part of the resulting NN-based classifier using gradient descent weights optimization.

All the described descriptors are concatenated into a single tensor, passing through a projection layer and a fully connected classification network.

III. Evaluation

A. Dataset description

To develop and evaluate the performance of our classifier, as well as to benchmark it against other models, we created and publicly released the Synthetic Images Dataset (SID) [24]. This dataset comprises a balanced collection of approximately 8000 images, equally divided between two classes: artificially generated images and real images. The generated images originate from a variety of generative models and exhibit diverse artifacts, such as inconsistent contours, compression defects, unnatural textures, uneven lighting, distorted facial features, unrealistic object boundaries, and irregular noise patterns (Fig.6).

The dataset is organized into three subsets designated for training, testing, and hyperparameter tuning, consisting of approximately 6000, 1000, and 1000 images, respectively. Each subset maintains class balance to ensure robust and unbiased evaluation. The SID dataset is publicly available to facilitate further research in the field of synthetic image detection.

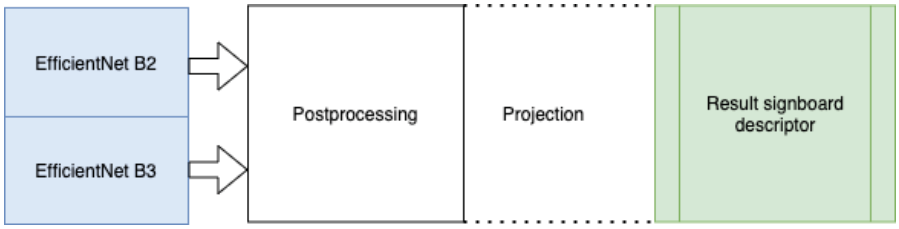

Fig. 4. CNN-based image descriptor generation scheme

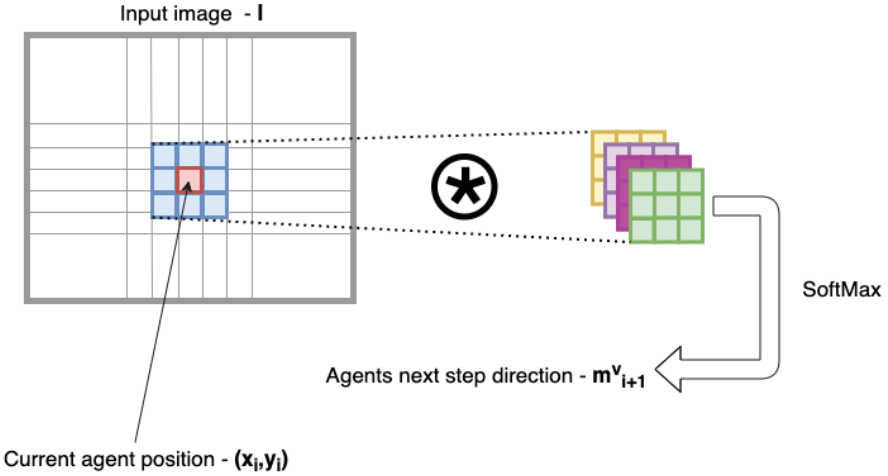

Fig. 5. Image descriptor Type E principles illustration

Fig. 6. Examples of artificially generated images with shape inconsistency artifacts from the Synthetic Images Dataset [24].

B. Experimental results

We are using the following metrics to validate our results for the classification task:

$$Precision = \frac{TP}{TP + FP}, \; Recall = \frac{TP}{TP + FN},$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

The classification outcomes are summarized in Table I. The experiments reveal that certain descriptors exhibit comparable performance in distinguishing artificially generated images. However, not all evaluated descriptor types achieved the anticipated accuracy in this context. This shortfall can likely be attributed to their limited capacity to capture subtle inconsistencies and structural artifacts, which are crucial for this task and markedly different from their effectiveness in domains such as textual feature analysis.

Furthermore, extensive hyperparameter optimization yielded no significant improvements, suggesting that some descriptor types may possess inherent constraints for this application. Nonetheless, it is important to emphasize that a subset of descriptors successfully identified distinctive characteristics of generated content with minimal parameter tuning, highlighting their robustness and potential for further refinement and application in this domain.

IV. Conclusion

This study highlights the development of an effective approach for classifying images as either artificially generated or real by leveraging specialized image descriptors. These descriptors are designed to capture unique global features and address specific challenges associated with inconsistencies and artifacts in generated content, such as irregular object shapes and lighting discrepancies. By integrating these descriptors into automated frameworks for validation, verification, and

TABLE I. LEADERBOARD TABLE

Method	classification metrics		
	precision	recall	F1
ResNet-50 [25]	0.837	0.845	0.841
MobileNet-v2 [26]	0.841	0.848	0.844
ResNet-101 [25]	0.857	0.861	0.859
InceptionResNet-v2 [27]	0.864	0.866	0.865
ResNet-152 [25]	0.868	0.870	0.869
ResNet-101 + descriptor type D [18]	0.875	0.877	0.876
EfficientNet B3 [5]	0.879	0.881	0.880
ResNext [28]	0.884	0.886	0.885
CLIP [4]	0.890	0.897	0.893
EfficientNet B4 [5]	0.899	0.903	0.901
CLIP + descriptor type C [18]	0.900	0.911	0.906
EfficientNet B6 [5]	0.904	0.915	0.909
EfficientNet B6 + descriptor type C [18]	0.913	0.916	0.915
CoAtNet [29]	0.915	0.925	0.920
EfficientNet B6 + descriptor type E [30]	0.923	0.926	0.924

identification, we provide a computationally efficient and interpretable alternative to resource-intensive deep neural networks.

The proposed method demonstrates significant potential in addressing the growing demands of various industries, including security, marketing, journalism, and digital forensics. It ensures precise and scalable analysis, facilitating the rapid detection and classification of manipulated or synthetic images in real-world applications. Additionally, the interpretability of the descriptors offers dual utility, enabling automated detection while providing insights for manual review in specialized contexts.

This research underscores the viability of descriptor-based methodologies as a complementary or alternative solution to deep learning models, particularly in environments with limited computational resources. Future work could explore the adaptation of these descriptors to a wider range of generative models and content types, further refining their capabilities and expanding their applications in domains requiring robust image classification and analysis tools.

ACKNOWLEDGMENT

The research was carried out with the financial support of the ITMO University Research Projects in AI Initiative (project No. 640113).

REFERENCES

[1] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[3] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," *CoRR*, 2022.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[5] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[6] S. Teh, S. Sivakumar, and F. Motalebi, "Vision transformers for biomedical applications *," in *2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, 2024, pp. 195–201.

[7] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523000634

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[10] V. Malykh and A. Samarin, "Combined advertising sign classifier," in *Analysis of Images, Social Networks and Texts*. Cham: Springer International Publishing, 2019, pp. 179–185.

[11] A. Samarin, V. Malykh, and S. Muravyov, *Specialized Image Descriptors for Signboard Photographs Classification*, 08 2020, pp. 122–129.

[12] C. Huang and J. Huang, "A fast hog descriptor using lookup table and integral image," 03 2017.

[13] T. Dittimi and C. Suen, "Modified hog descriptor-based banknote recognition system," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, 10 2018.

[14] J. Sun, Z. Shisong, and W. Xiaosheng, "Image retrieval based on an improved cs-lbp descriptor," 05 2010, pp. 115 – 117.

[15] A. Bachchan, A. Gorai, and P. Gupta, "Automatic license plate recognition using local binary pattern and histogram matching," 07 2017, pp. 22–34.

[16] 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.

[17] A. Samarin, A. Savelev, and Toropov, "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, 07 2023.

[18] 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 (S.A.I.ence)*, 2020, pp. 50–53.

[19] A. Samarin and V. Malykh, "Ensemble-based commercial buildings facades photographs classifier," in *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 257–265. [Online]. Available: https://doi.org/10.1007/978-3-030-72610-2_19

[20] Y. Wei, M. Yang, L. Xu, M. Liu, F. Zhang, T. Xie, X. Cheng, X. Wang, F. Che, Q. Li, Q. Xu, Z. Huang, and M. Liu, "Novel computed-tomography-based transformer models for the noninvasive prediction of pd-1 in pre-operative settings," *Cancers*, vol. 15, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2072-6694/15/3/658

[21] A. Samarin, A. Savelev, A. Toropov, A. Dzestelova, V. Malykh, E. Mikhailova, and A. Motyko, "The complete study of the movement strategies of trained agents for visual descriptors of advertising signs," in *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, J.-J. Rousseau and B. Kapralos, Eds. Cham: Springer Nature Switzerland, 2023, pp. 571–585.

[22] A. Samarin and V. Malykh, "Worm-like image descriptor for signboard classification," in *Proceedings of The Fifth Conference on Software Engineering and Information Management (SEIM-2020)*, 2020.

[23] A. Samarin, A. Savelev, A. Toropov, A. Dzestelova, V. Malykh, E. Mikhailova, and A. Motyko, "Trainable agents movement strategies for advertising sign visual descriptors," *Pattern Recognit. Image Anal.*, vol. 32, no. 3, p. 651–657, Sep. 2022. [Online]. Available: https://doi.org/10.1134/S1054661822030373

[24] "Sid: Synthetic images dataset." [Online]. Available: https://goo.su/2qUYoz

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692

[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[27] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[28] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *CoRR*, vol. abs/1611.05431, 2016. [Online]. Available: http://arxiv.org/abs/1611.05431

[29] Z. Dai, H. Liu, Q. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *CoRR*, 2021.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, 2020.