Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (SD TIB PAN)

SUBJECT:

Generalizable Methods for Detecting Synthetic Materials Generated by Artificial Intelligence Models in Multi-Domain Environments

SUPERVISORS:

Mariusz Kamola PhD DSc Eng. (main supervisor)
mariusz.kamola@nask.pl
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa

Ewelina Bartuzi-Trokielewicz PhD (auxiliary supervisor)
ewelina.bartuzi@nask.pl
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa

DESCRIPTION:

The dynamic development of generative artificial intelligence models has led to a rapid increase in both the number and quality of fully synthetic materials. Contemporary generative systems make it possible to generate content with a high level of realism, correct semantics, and increasingly greater structural coherence. As a result, distinguishing synthetic materials from real ones is becoming increasingly difficult, while trust in digital information may be weakened [1]. This problem is further intensified by the rapid evolution of successive generations of models, which may lead to the partial obsolescence of existing detection methods and limit their effectiveness under real-world conditions [2].

A particularly important research gap remains the effective detection of synthetic deepfakes while maintaining the ability to generalize to data that differ from the training set. Many current methods achieve high effectiveness on known or similar data; however, their performance may significantly deteriorate when the generator [3], visual domain [2], compression level, resolution, or the method of material processing changes [4]. This means that effectiveness obtained under controlled benchmark conditions does not always translate into real-world usage scenarios, in which synthetic content originates from new models and various distribution sources, and is subject to additional transformations [5]. For this reason, it becomes crucial to study deepfake detection methods not only in terms of classifying known examples, but above all in terms of robustness to data distribution shifts.

The dissertation should concern the development and evaluation of methods for detecting synthetic materials generated by artificial intelligence models. The scope of the research should be framed broadly, so as not to limit the proposed approach to a single type of data, a specific generator, or a selected application domain. Particular attention should be devoted to the ability to generalize, understood as maintaining effectiveness under conditions that differ from the data used during training.

As part of the dissertation, it would be necessary to investigate which representations and regions of the material contain the information most useful for distinguishing real content from synthetic content. It would be important to determine which properties of generated materials are most relevant from the perspective of detection and remain most stable when the generator, domain, and data quality change [2]. Such an approach would make it possible to distinguish features related to the synthetic generation process itself from incidental properties of a particular dataset or compression method [4].

One possible research direction is the analysis of frequency-domain features and low-level signal properties. Generative models may leave subtle traces in the frequency distribution, in the relationships between local structures, or in residuals [4]. In the case of older GAN-based models, artifacts of this type were often more regular and easier to detect unambiguously. However, contemporary diffusion models generate materials of significantly higher quality, which means that frequency-domain traces may be weaker and less explicit than in the case of older GAN-based models [6]. At the same time, it should be taken into account that compression and further processing of the material may weaken some of the information present in the spectrum, which is important when assessing the persistence of such features under realistic operating conditions for detectors [4].

A second possible direction is the use of pre-trained models such as CLIP [7] or DINOv2 [8]. Models of this type learn representations from large and diverse datasets, and therefore may provide more general information about the semantics, structure, texture, and relationships present in the material. In the context of synthetic content detection, it would be important to verify whether representations derived from such models make it possible to separate features related to the general characteristics of the material from artifacts specific to particular generators. This direction may be especially valuable in multi-domain environments, where the detector should maintain its effectiveness despite changes in the data source [2].

In the case of materials with a temporal dimension, the analysis of temporal information may also be important. Real data are usually characterized by the natural continuity of changes, motion coherence, and the stability of relationships between scene elements. Materials generated by artificial intelligence, on the other hand, may contain subtle inconsistencies in dynamics, structural instability, texture changes, or distortions that become visible only when successive fragments of a sequence are analyzed [9]. For this reason, temporal features may constitute an important source of information, particularly when individual frames appear locally realistic.

Moreover, it is worth considering different representations of synthetic materials, since it is currently difficult to identify a single type of feature that would ensure stable detection when the generator, visual domain, and data quality change [2]. Therefore, relationships between objects and the coherence of scene geometry [10], as well as features related to texture, semantics, and the context of the material [11], [12], may be important. The analysis of such relationships could make it possible to detect synthetic content even when local artifacts are weakly visible or have been partially removed by compression and post-processing [4].

In summary, the dissertation should make a scientific contribution by determining which types of information and representations support stable deepfake detection, and by developing methods that use these properties under realistic conditions. The results obtained in this way could support the development of systems for analyzing the authenticity of digital content, which are important for cybersecurity, content moderation, and the media.

REQUIREMENTS:

- MSc degree in computer science, mathematics or a related field
- Programming skills
- Experience with knowledge engineering
- Knowledge about current research in computer vision, particularly synthetic image and video detection methods.
- Advanced Level in English (speaking and writing)

BIBLIOGRAPHY:

[1] Ma, L., Xue, Z., Wang, Y., Yan, Z., Xu, J., Jiang, X., Yu, H., Liao, Y., & Bi, Z. (2026). Your One-Stop Solution for AI-Generated Video Detection. arXiv. Available: https://arxiv.org/pdf/2601.11035 [Accessed: 19 May 2026].

[2] Amanzadi, A., Dehghanian, Z., Beigy, H., & Rabiee, H. R. (2025). Redefining Generalization in Visual Domains: A Two-Axis Framework for Fake Image Detection with FusionDetect. arXiv. Available: https://arxiv.org/pdf/2510.05740 [Accessed: 19 May 2026].

[3] Ma, L., Zhang, J., Deng, H., Zhang, N., Liao, Y., & Yu, H. (2024). DeCoF: Generated Video Detection via Frame Consistency. arXiv. Available: https://arxiv.org/pdf/2402.02085v1 [Accessed: 19 May 2026].

[4] Corvi, R., Cozzolino, D., Prashnani, E., De Mello, S., Nagano, K., & Verdoliva, L. (2025). Seeing What Matters: Generalizable AI-generated Video Detection with Forensic-Oriented Augmentation. arXiv. Available: https://arxiv.org/pdf/2506.16802 [Accessed: 19 May 2026].

[5] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. arXiv. Available: https://arxiv.org/pdf/2503.02857 [Accessed: 19 May 2026].

[6] Ricker, J., Damm, S., Holz, T., & Fischer, A. (2024). Towards the Detection of Diffusion Model Deepfakes. arXiv. Available: https://arxiv.org/pdf/2210.14571 [Accessed: 19 May 2026].

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv. Available: https://arxiv.org/pdf/2103.00020 [Accessed: 19 May 2026].

[8] Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., & Bojanowski, P. (2024). DINOv2: Learning Robust Visual

Features without Supervision. arXiv. Available: https://arxiv.org/pdf/2304.07193 [Accessed: 19 May 2026].

[9] Kundu, R., Mohanty, V., Xiong, H., Jia, S., Balachandran, A., & Roy-Chowdhury, A. K. (2026). SAGA: Source Attribution of Generative AI Videos. arXiv. Available: https://arxiv.org/pdf/2511.12834 [Accessed: 19 May 2026].

[10] Chang, C., Liu, J., Liu, Z., Lyu, X., Huang, Y.-H., Tao, X., Wan, P., Zhang, D., & Qi, X. (2025). How Far are AI-generated Videos from Simulating the 3D Visual World: A Learned 3D Evaluation Approach. arXiv. Available: https://arxiv.org/pdf/2406.19568 [Accessed: 19 May 2026].

[11] Li, Y., Zhen, W., Zhang, Y., Sun, R., Zheng, Y., Chen, L., Zhou, J., & Lu, J. (2026). Skyra: AI-Generated Video Detection via Grounded Artifact Reasoning. arXiv. Available: https://arxiv.org/pdf/2512.15693 [Accessed: 19 May 2026].

[12] Liu, X., Song, X., Zheng, H., Bai, L., Liu, X., & Zhai, G. (2025). Consolidating Diffusion-Generated Video Detection with Unified Multimodal Forgery Learning. arXiv. Available: https://arxiv.org/pdf/2511.18104 [Accessed: 19 May 2026].