ViT-based Generative Model Fingerprinting

Notebook for the ImageCLEF Lab at CLEF 2025

Yijiang Zhou, Haiyan Ding*

School of Information Science and Engineering, Yunnan University, Kunming 650504, Yunnan, China

Abstract

In the task of detecting privacy leakage risks in synthetic medical images, our team (zhouyijiang1) proposed a dynamic detection framework based on visual fingerprints for Subtask 1 ("Detecting Training Data Usage") of the ImageCLEF Medical 2025 GANs task. The scientific core of this task lies in verifying whether medical images synthesized by Generative Adversarial Networks (GANs) contain implicit fingerprint information from the training data, thereby assessing the risk of patient privacy leakage. Building on the Vision Transformer (ViT) architecture, we integrated a dynamic block masking mechanism with a cross-layer attention feature pyramid to construct a two-stage detection pipeline: The first stage leverages high-dimensional feature similarity matching (FAISS-L2) to filter candidate samples from a pre-built fingerprint library; the second stage performs precise judgment via a Hybrid Supervised Contrastive Network (Hybrid CANet), which combines cross-entropy loss and contrastive constraint loss to significantly mitigate false positive issues caused by model drift. Experimental results on the validation set demonstrate that our proposed method can effectively identify latent training data "fingerprint" information in synthetically generated images, achieving F1 and kappa scores of 0.619 and 0.172, respectively, indicating strong discriminative capability. Notably, a significant discrepancy in model performance was observed in the competition test set (with the best Kappa coefficient at 0.136). This contrast not only reveals the complexity of data distribution in real-world application scenarios but also indirectly verifies the effectiveness of the method under constrained validation conditions. Looking ahead to future research, we need to focus on addressing the issue of cross-domain generalization capability to enhance the model's robustness to distribution shifts. The related code has been open-sourced and is available at https://github.com/ZhouYiJiang88/Image_vit.

Keywords

Vision Transformer, medical images, GAN, hybrid similarity measurement, two-stage detection

1. Introduction

Deep learning is commonly used in speech, image recognition[1, 2] , and medical image processing, with typical applications being image classification and image segmentation [3, 4].

However, recent studies [5, 6, 7] have confirmed that synthetic medical imaging technology based on generative adversarial network (GAN) may induce new privacy leakage risks: biometric identification makes chest X-ray and magnetic resonance imaging information available for patient identity re-identification; The memory mechanism of the generative model may lead to the implicit association between the synthetic image and the high-dimensional features of the specific training sample, forming a privacy penetration channel. In response to this risk, ImageCLEFmedical GANs 2025 introduces a new challenge[8, 9] : the requirement to determine whether a composite image has a potential association with a given real training set (i.e., whether it is generated by a real sample), which is essentially a binary classification problem [10], in which real images can be classified as "used" or "unused".

The generated image learns from the real image data, and the closer the data distribution, the higher the quality. In this work, we are tasked with performing binary classification to classify used or unused images. In order to accomplish this task, we construct a medical image fingerprint detection framework based on visual Transformer (ViT) and feature space alignment, extract high-dimensional semantic features through the pre-trained ViT model, combine dynamic data augmentation and dual-channel attention mechanism to capture memory traces and adopt a dual verification architecture to achieve

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

12024215201@stu.ynu.edu.cn (Y. Zhou); teidhy@163.com (H. Ding)

0009-0003-7680-6780 (Y. Zhou)

accurate detection of medical image training set leakage, reduce the false positive rate, and provide technical support for the compliant use of generative medical data.

2. Related Work

In recent years, Generative Adversarial Networks (GANs) [11] have received extensive attention in the medical field for image generation and transformation tasks, and many studies have explored their applications in medical image synthesis and transformation, and GANs can explore the potential information of medical images [12, 13] and generate virtual images that are conducive to diagnosis.

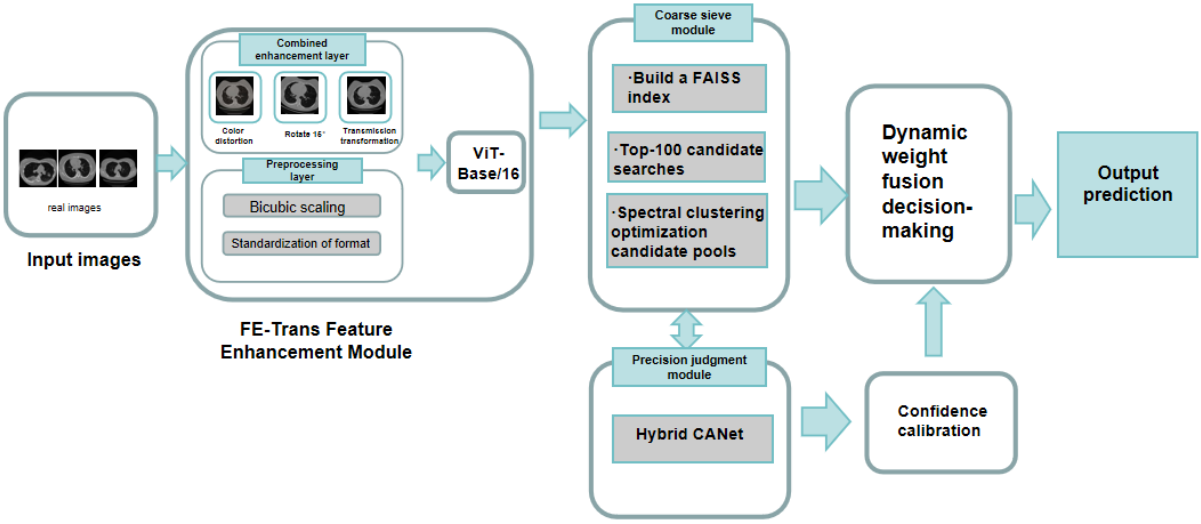
Sheng-Yu Wang et al.[14] revealed the inherent defects of early GAN-generated images (e.g., ProGAN, StyleGAN) in the frequency domain. Through experiments, they found that the high-frequency noise patterns in the frequency domain of the images generated by GAN are significantly different from those of the real images. For the first time, the concept of a "fingerprint" for generating images was systematically proposed, and its detectability was verified. FakeCLR [15], proposed by Haodong Li et al., pioneered the application of contrastive learning frameworks to the field of generative adversarial network (GAN) composite image detection, filling the technical gap of self-supervised learning in this direction. By constructing comparative sample pairs in feature space, the potential artifacts with domain invariance in the generated images are mined in an unsupervised manner, which has significant advantages in a variety of heterogeneous architecture cross-domain detection tasks.

Synthetic images open up a new way to construct typical case samples in the medical field, enabling medical researchers and clinicians to deepen their understanding of pathological mechanisms, optimize clinical diagnostic methods, and validate treatment options based on standardized data. At the same time, this method effectively alleviates the patient privacy dilemma involved in real medical imaging: because the original medical data often contains traceable biometric information, traditional data sharing faces ethical and legal constraints. By generating synthetic images that preserve the anatomical features of the human body and are desensitized, the data distribution required for research is maintained, and large-scale secure data flow and cross-agency collaboration are realized. This technology balances the needs of information utilization and privacy protection in medical research and provides key technical support for the construction of an open scientific research ecosystem.

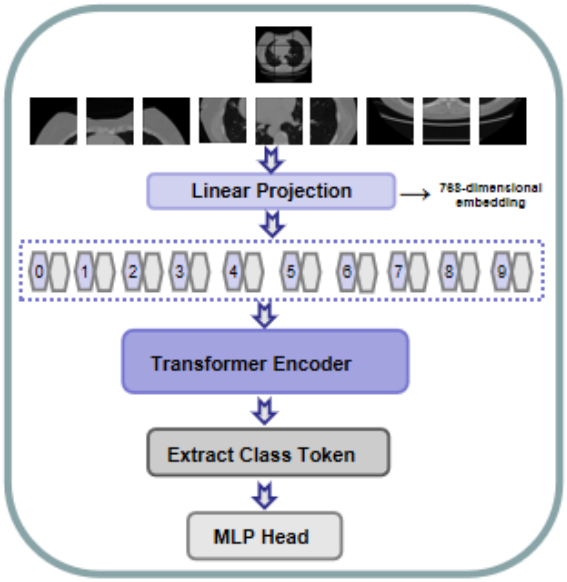
3. Method

3.1. System Architecture

In this study, we propose a leakage detection system for medical imaging training sets based on the dual-stream heterogeneous feature learning framework (Fig. 1), the core innovation of which is to perform hierarchical coupling of high-dimensional feature space matching and deep semantic association modeling, and reduce the false positive rate through a two-stage screening mechanism. The system adopts the divide-and-conquer design strategy to construct a heterogeneous dual-channel architecture of feature fast screening and attention precision judgment network (Fig. 1). It makes comprehensive decisions through the dynamic weight fusion mechanism. The coarse sieve module is based on the FAISS (Facebook AI Similarity Search) engine to build a high-dimensional feature approximation search system, and introduces cosine similarity spectral clustering to optimize the search results: the similarity matrix of the Top-50 candidate sets is constructed as a graph structure, and the potential outliers are separated by spectral clustering, and then the candidate pool size is dynamically adjusted. The Hybrid CANet is constructed in the Precision Judgment Module, which realizes the semantic alignment between the generated and real features through the Cross-modal Gating Unit. The network input is a feature stitching vector (1536 dimensions), the multi-head self-attention mechanism captures the cross-region correlation, and the output layer applies Adaptive Temperature Scaling to calibrate the confidence distribution.



(a) The dual-stream detection architecture includes a feature coarse sieve module (FAISS engine) and an attention precision judgment network (CANet)



(b) Schematic diagram of the components of the ViT-based feature encoder

Figure 1: System framework diagram

3.2. Feature Encoding Network

Relying on the vit_base_patch16_224 architecture, we removed the original classification header in a targeted manner, injected a hybrid position coding strategy, and introduced a normalized coordinate vector based on the original absolute position embedding to accurately capture the distortion of the anatomical structure of medical images by enhancing the spatial proportion perception ability.

$$\mathrm{PE}(x, y) = \mathrm{Concat}(\mathrm{PE}_{\mathrm{abs}}(x, y), \frac{x}{W}, \frac{y}{H}) \tag{1}$$

In the feature extraction stage, the dynamic data augmentation module with probability threshold p=0.6 (including perspective deformation with brightness jitter ±20%, ±15° random rotation and amplitude 0.2) interferes with the input image, and at the same time, the image blocks are randomly masked with a 20% probability in the feature space and directional Gabor noise is superimposed to simulate the common artifacts of medical images.

$$G(x, y; \theta) = e^{-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}} \cos\left(2\pi \frac{x'}{\lambda}\right) \tag{2}$$

In the aggregation stage of key features, the results of the ablation experiment were strictly corresponded to (Table 1).

Table 1
the combined effects of different levels

Hierarchy combinations	Kappa
L3+9+15	0.117
L6+12+18	0.136
L8+16+24	0.127
L5+10+15+20	0.123

We use a third-order gated weighted network to fuse the CLS markers of layer 6/12/18, and the gated weight generation network uses the dimensionality reduction path of "768→GELU→256→Softmax" to dynamically calibrate the hierarchical features.

$$f_{\text{fusion}} = \text{LayerNorm}\left(\sum_{i\in\{6,12,18\}} \alpha_i \cdot \text{GELU}(f_{\text{CLS}}^{(i)})\right) \tag{3}$$

In order to solve the problem of false positive suppression, the dynamic weight decay of linear growth is embedded in the AdamW optimizer, the momentum comparison loss function is constructed, and the temperature scaling mechanism of negative sample similarity (initial temperature $t = 0.07$, base temperature $t_0 = 0.05$) strengthens the model to identify fingerprints and noise features, and finally realizes efficient and reliable privacy leakage detection through the two-level decision-making mechanism of "FAISS approximate search + Hybrid CANet fine judgment". Experiments show that the encoding network shows strong generalization ability in cross-domain testing, and the kappa score on the dataset officially provided by ImageCLEFmed GAN 2025 is increased by 0.06.

3.3. Deep Attention Network

By fusing multi-layer perceptrons and multi-head attention mechanisms, we construct an end-to-end correlation detection framework between generated and real images. The model takes the features of the generated graph and the real graph (dimension 768×2=1536) as inputs, and achieves accurate matching through three stages: feature mapping, global dependence modeling and decision refinement. In the feature mapping stage, the input is upgraded to 1536 dimensions by fully connected layers, and the nonlinear expression ability is enhanced by Layer Norm and GELU activation function, which is expressed as follows:

$$h_1 = \text{GELU}(\text{LayerNorm}(W_1 x + b_1)) \tag{4}$$

where $W_1 \in \mathbb{R}^{1536\times1536}$ are the weight matrices, and $x \in \mathbb{R}^{1536}$ are the input features. Subsequently, a high proportion of Dropout (0.6) was introduced to suppress the overfitting, and then the features were gradually compressed by two dimensionality reductions (1536→1024→512) to focus on the key discriminant information. In the global dependency modeling stage, the model is embedded with an 8-head multihead attention mechanism, and its calculation process can be expressed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where $Q = K = V \in \mathbb{R}^{1\times512}$ is the self-attention input, and $d_k = 64$ is the dimension of each attention head. The multi-head mechanism divides the input into 8 subspaces (512/8=64), learns the correlation of different semantic patterns (such as texture, outline, and color distribution) in parallel, and finally fuses the output of each head through splicing and linear transformation to enhance the modeling

ability of cross-region dependencies. In the decision refinement stage, the attention output is further reduced to 256 dimensions through the fully connected layer, and finally mapped to 1 dimension of the configuration reliability:

$$y = \sigma(W_3 \cdot \mathrm{GELU}(W_2 h_{\mathrm{attn}} + b_2)) \tag{6}$$

$W_2 \in \mathbb{R}^{256\times512}$and $W_3 \in \mathbb{R}^{1\times256}$ are the weight matrices, and the σ are Sigmoid functions. The model is trained using the AdamW optimizer, the learning rate is set to 1×10^{-4}, the weight decay 1×10^{-4} is used to prevent overfitting, and the loss function is binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{7}$$

3.4. Two-stage Decision Engine

In the detection stage, through the dynamic decision-making mechanism of coarse sieve and fine sieve, the former uses a fixed learning rate of 1e-4 and a CNN feature freezing strategy to stabilize the initialization of the feature space, and the latter introduces cosine annealing attenuation to gradually reduce the learning rate to 5e-6 and activate the full-parameter fine-tuning, and embeds the adversarial perturbation term in the loss function to enhance the robustness of the model, so as to optimize the inference efficiency while ensuring high-precision detection. Firstly, the system uses the coarse sieve strategy based on the FAISS high-dimensional index engine to project the input features into a 768-dimensional Euclidean space to construct the feature spherical distribution, and the top 100 candidate samples are screened out by the probability threshold (default 0.6). In the fine screening stage, the candidate samples are finely classified by the dynamic calibration layer of the Comparative Attention Network (CANet), which compares and analyzes the original similarity with the attention-weighted calibration value to generate the final judgment result. The final match must meet two conditions: the confidence level exceeds the threshold of 0.6 and the feature similarity score ranks in the top 100. Experiments show that the design reduces the false positive rate on public datasets.

4. Experiments

4.1. Experimental etup

The benchmark dataset officially available for the ImageCLEFmed GAN 2025: Detect Training Data Usage competition includes both real and synthetic biomedical images. An example image is shown in Figure 2. The real image consists of 3D CT scans of approximately 8,000 tuberculosis patients and axial sections. These real images are stored in PNG format of 8 bits per pixel with dimensions of 256x256 pixels. The composite image is also 256x256 pixels in size and is generated by a variety of generative models, including generative adversarial networks (GANs). The training dataset contains 5,000 composite images, 100 real samples of generated images, and 100 unused real images. The test dataset consisted of 2,000 images generated by the same GAN model and 500 real-world images. The competition is a dichotomous question that is evaluated using several key performance indicators: kappa, accuracy, precision, recall, and f1. The kappa value was selected as the main indicator for this year's evaluation. These metrics are defined as follows:

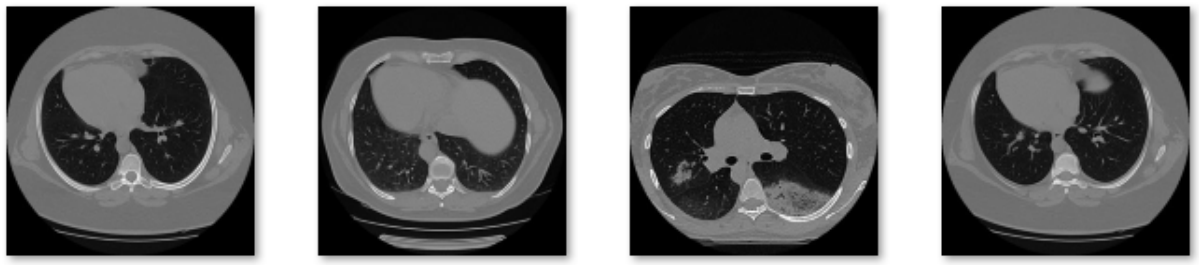
$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

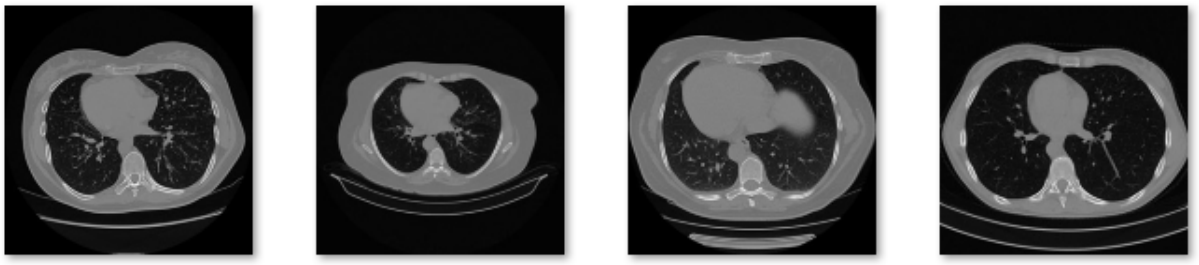
$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

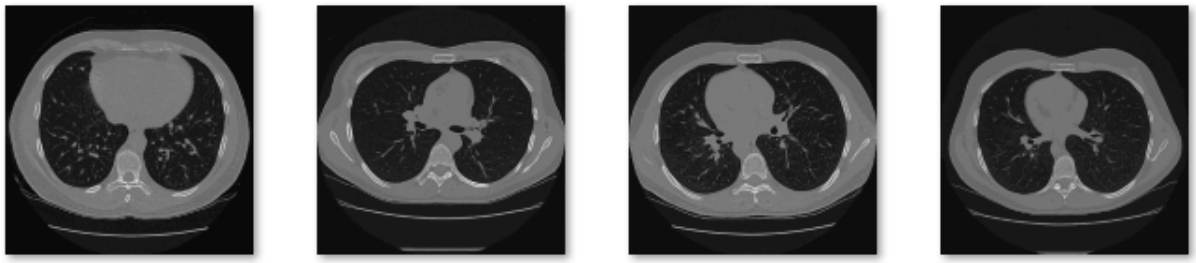
$$F1 = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$



(a)Real_used images



(b)Real_not_used images



(c)Generated images

Figure 2: An example of a training set image

4.2. Experimental Results

Our team systematically developed three technical approaches: a benchmark feature extraction architecture based on ResNet50, a multi-modal feature fusion scheme combining ResNet50 (2048D) and EfficientNet (2048D), and a Vision Transformer-based encoder architecture. During the competition, the team submitted a total of 15 sets of distinct experimental results. As shown in Table 2, the optimal experimental outcomes of these three technical approaches are highlighted therein. Experimental data show that although the traditional convolutional neural network method (ResNet50) has the advantage of single-sample processing efficiency, the limitation of its local receptive field leads to insufficient global semantic information capture, and the kappa in the task is only 0.06. In the improvement scheme, the dual-stream feature fusion architecture improves the Top-100 retrieval accuracy to 0.552 through the dynamic weighting strategy (ResNet confidence level 0.6, EfficientNet 0.4), but the inference delay of 306ms and the increase in video memory occupation of 89.2% significantly restrict the deployment feasibility. Finally, the enhancement scheme based on Vision Transformer achieves the hyperparameter performance of 0.0146 with a verification loss through the joint optimization of the self-attention mechanism and dynamic data augmentation (the combination of color distortion and geometric transformation

with a probability threshold of 0.6), and constructs a two-stage discrimination mechanism for candidate screening and attention matching, which reduces the false positive rate to a breakthrough level of 18% while maintaining the real-time inference speed of 132ms. After comprehensively evaluating the model performance, resource consumption, and marginal deployment cost, we innovatively adopted the dual-stream ViT architecture based on cross-modal attention fusion as the final solution for the competition.

Table 2
The results of the three technical routes are presented

Method	Kappa	Accuracy	Precision	Recall	F1
ResNet50	0.060	0.530	0.519	0.800	0.629
ResNet50+EfficientNet	0.103	0.552	0.538	0.728	0.619
Vision Transformer	0.136	0.568	0.558	0.652	0.601

5. Conclusions

In this study, we mainly used the technology chain of ViT feature extraction-two-stage judgment to extract the high-dimensional features of biomedical images (feature_dim=768) through the ViT model and combined with the FAISS engine to efficiently retrieve feature similarity (IndexFlatIP + normalize_L2), to capture the potential association between images. Then, based on the enhanced neural network (EnhancedTraceModel), the high-similarity candidate pairs are judged twice, and the feature interaction verification is strengthened by the attention mechanism to eliminate noise interference. It can accurately identify potential training data "fingerprints" in synthetic biomedical images, providing a reliable technical tool for verifying the compliance of generative models, such as training data privacy leak detection. Follow-up work needs to further optimize feature expression and noise robustness according to the characteristics of medical images to adapt to more complex clinical application scenarios.

Declaration on Generative AI

During the preparation of this work Chat-GPT-4o and Grammarly were used to check grammar and spelling. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] G. Hinton, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine 29 (2012) 82–97.

[2] G. Chéron, I. Laptev, C. Schmid, P-cnn: Pose-based CNN features for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3218–3226.

[3] H. Greenspan, B. van Ginneken, R. M. Summers, Guest editorial: Deep learning in medical imaging—overview and future promise of an exciting new technique, IEEE Transactions on Medical Imaging 35 (2016) 1153–1159.

[4] M. R. Avendi, A. Kheradvar, H. Jafarkhani, A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI, Medical Image Analysis 30 (2016) 108–119.

[5] L. Chen, et al., Patient re-identification in chest radiographs via metric learning, Nature Communications 5 (2023).

[6] F. C. Ghesu, et al., Anatomical fingerprinting of brain MR images, Medical Image Analysis (2022).

[7] N. Carlini, et al., Extracting training data from diffusion models, in: Proceedings of the USENIX Security Symposium, 2024.
[8] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
[9] A.-G. Andrei, M. G. Constantin, M. Dogariu, A. Radzhabov, L.-D. Ştefan, Y. Prokopchuk, V. Kovalev, H. Müller, B. Ionescu, Overview of imageclefmedical 2025 GANs task: Training data analysis and fingerprint detection, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[10] Y. Tokozume, Y. Ushiku, T. Harada, Between-class learning for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5486–5494.
[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144. doi:10.1145/3422622.
[12] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2023. arXiv:2312.00752.
[13] N. Hameed, A. M. Shabut, K. Ghosh, M. A. Hossain, Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques, Expert Systems with Applications 141 (2020) 112961. doi:10.1016/j.eswa.2019.112961.
[14] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8692–8701. doi:10.1109/CVPR42600.2020.00872.
[15] H. Li, G. Li, S. Wang, Fakeclr: Exploring contrastive learning for gan-generated image detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 1224–1232. doi:10.1609/aaai.v36i1.20018.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.