UDC 004.932.2:004.89
DOI: https://doi.org/10.17721/1812-5409.2026/1.28

Volodymyr KUBYTSKYI, PhD Student
ORCID ID: 0000-0002-1529-8677
e-mail: vova.kubytskyi@gmail.com
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Artem BOZHOK, PhD Student
ORCID ID: 0009-0009-9572-6501
e-mail: artembozh@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

INTERMEDIATE LAYER CONFIGURATIONS FOR IMAGE SIMILARITY: EVALUATING MULTI-LEVEL FEATURE EXTRACTION ACROSS CNN AND TRANSFORMER ARCHITECTURES

Evaluating visual similarity between images requires vector representations that preserve information across multiple levels of abstraction. Multi-level feature aggregation from intermediate CNN layers has shown good results in near-duplicate detection tasks, yet no structured analysis exists of how architectural properties of different backbones affect the quality of such multi-level representations. This study addresses the question of which structural characteristics make an architecture suitable for intermediate layer aggregation.

We formalize the multi-level aggregation pipeline as Global Average Pooling applied to intermediate activations, followed by concatenation and L2-normalization, producing a fixed-length descriptor. A compact MLP classifier is trained on the resulting pairwise representations. We evaluate six backbone architectures – VGG-16, Inception v3, DenseNet-121, MobileNetV2, ResNet-50, and ViT-B/16 – analyzing their stage structure, spatial resolution progression, intermediate feature dimensionality, and information flow properties. Quantitative evaluation is performed on the INRIA Holidays dataset using F1-score for pairwise similarity classification. ResNet-50 with layers C2+C3+C5 achieves F1 = 0.77, outperforming all other configurations. VGG-16 multi-level aggregation yields F1 = 0.72 but suffers from parameter inefficiency (138M vs. 23M). DenseNet-121 reaches F1 = 0.74 despite feature redundancy from dense connections. ViT-B/16 intermediate blocks produce F1 = 0.70, limited by homogeneous block structure without clear spatial hierarchy. MobileNetV2 and Inception v3 show F1 = 0.68 and F1 = 0.71 respectively.

Clear stage hierarchy with residual connections, progressive spatial-to-semantic transition, and moderate intermediate dimensionality are the key architectural requirements for effective multi-level feature aggregation.

K e y w o r d s : ***multi-level feature aggregation, image similarity, convolutional neural networks, vision transformers, intermediate representations, near-duplicate detection, architectural comparison, ResNet-50.***

AMS 2020 classification: 68T45, 68T07.

Introduction

Automated detection of visually similar images across large collections remains an open challenge in applied computer vision. When two photographs of the same room differ only by a watermark or a slight crop, a human recognizes them as near-duplicates instantly. Automated systems must solve the same problem at web scale, and their accuracy depends heavily on how images are encoded as vectors (Thyagharajan, & Kalaiarasi, 2021). The standard approach – extracting a fixed-length descriptor from a pre-trained deep network and comparing via a distance metric (Razavian et al., 2014) – works well in many settings, yet the choice of network architecture is rarely justified beyond convention. The same challenge appears across many visual domains – from detecting reposted content on social platforms to matching product images in e-commerce and verifying document authenticity.

A known weakness of single-layer representations from final network layers is the loss of low- and mid-level visual information. These features encode high-level semantic content but suppress textures, edges, and local structures that matter for near-duplicate detection and structural correspondence (Kordopatis-Zilos et al., 2017). This trade-off motivates multi-level feature aggregation – combining activations from several intermediate layers to preserve information across the full abstraction hierarchy (Kubytskyi, & Panchenko, 2023; Panchenko, Bozhok, & Kubytskyi, 2026).

Combining features from different depths has precedents in HyperColumns (Hariharan et al., 2015) and Feature Pyramid Networks (Lin et al., 2017), but both target dense prediction and require spatially aligned outputs. Image similarity needs a single compact descriptor per image. The pipeline in (Kubytskyi, & Panchenko, 2023) solves this by applying Global Average Pooling to each intermediate layer, concatenating the results, and L2-normalizing the combined vector. While effective on real estate image similarity tasks (Kubytskyi, & Panchenko, 2023; Panchenko, Bozhok, & Kubytskyi, 2026), the backbone architecture has not been compared against alternatives. Vision transformers (ViT) (Dosovitskiy et al., 2021) and contrastive vision-language models (CLIP) (Radford et al., 2021) offer architecturally distinct alternatives to CNNs, yet whether their intermediate features are useful for multi-level aggregation has not been studied.

The research object is the multi-level intermediate feature aggregation pipeline and its dependence on the structural properties of the backbone architecture.

The aim of this research is to evaluate the suitability of six deep learning architectures – VGG-16 (Simonyan, & Zisserman, 2014), Inception v3 (Szegedy et al., 2015), DenseNet-121 (Huang et al., 2017), MobileNetV2 (Sandler et al., 2018), ResNet-50 (He et al., 2016), and ViT-B/16 (Dosovitskiy et al., 2021) – for multi-level intermediate feature aggregation in image similarity tasks, and to identify the architectural properties that determine aggregation quality.

The structure of the paper is as follows. Section 1 describes the aggregation pipeline and evaluation protocol. Section 2 presents the architectural analysis and experimental results. The paper concludes with a discussion of findings and practical recommendations for backbone selection.

1. Multi-level aggregation pipeline and evaluation protocol

The aggregation pipeline, originally proposed in (Kubytskyi, & Panchenko, 2023), operates as follows. Given a backbone network, we select K intermediate layers at distinct depth levels. For each selected layer k, the activation tensor is reduced to a fixed-length vector via Global Average Pooling (GAP). The resulting vectors are concatenated and L2-normalized:

$$e = [GAP(F_1); GAP(F_2); \ldots; GAP(F_k)], \quad e\_norm = e / \|e\|_2, \tag{1}$$

where F_k denotes the activation tensor at layer k. The normalized descriptor (1) is architecture-agnostic and can be applied to any backbone that produces spatially-structured intermediate activations.

For pairwise similarity classification, we construct a symmetric feature vector from two image descriptors $x^{(a)}$, $x^{(b)}$ and train a compact MLP (2048→1024→512→256, ELU activation, batch normalization, dropout) with binary cross-entropy loss (Panchenko, Bozhok, & Kubytskyi, 2026). The base CNN remains frozen; only the MLP is trained.

Evaluation is performed on the INRIA Holidays dataset (Jégou, Douze, & Schmid, 2008), containing 1,491 images grouped into 500 scene classes with variations in viewpoint, scale, illumination, and partial occlusions. Positive pairs are drawn from the same class; negative pairs are randomly sampled across classes in a balanced 1:1 ratio. The MLP is trained on 100 labeled pairs. All architectures use ImageNet-pretrained weights with frozen parameters, ensuring that differences in performance reflect the quality of intermediate representations rather than training procedures.

All experiments are implemented in Python 3.10 using PyTorch 2.1 with torchvision model implementations. Feature extraction and MLP training are performed on a single NVIDIA RTX 3090 GPU. Each configuration is evaluated over 10 independent runs with different random pair splits (seed values 0–9); we report the mean F1-score and standard deviation across runs. Input images are resized to 224×224 pixels with standard ImageNet normalization.

For each architecture, we select intermediate layers that correspond to distinct processing stages, aiming to capture features at maximally different levels of abstraction. The specific layer selections are: VGG-16 – after conv2_2, conv3_3, conv5_3 (128+256+512 = 896-D); Inception v3 – outputs of mixed stages at spatial resolutions 35×35, 17×17, 8×8 (concatenated branch outputs, 1280-D); DenseNet-121 – after denseblock1, denseblock2, denseblock4 (128+256+1024= = 1664-D, skipping denseblock3 to reduce redundancy); MobileNetV2 – after bottleneck stages 2, 4, 7 (32+64+160= = 256-D); ResNet-50 – C2, C3, C5 outputs (256+512+2048 = 2816-D, excluding C4 due to overlap with neighboring stages); ViT-B/16 – outputs of attention blocks 3, 6, 9, 12 (768×4 = 3072-D).

2. Architectural analysis and experimental results

We analyze six architectures along four structural criteria that determine their suitability for multi-level feature aggregation: (i) stage clarity – whether the architecture has well-defined, separable processing stages; (ii) spatial progression – whether spatial resolution decreases monotonically across stages; (iii) feature independence – whether features at different stages capture genuinely distinct information; and (iv) residual information flow – whether architectural mechanisms preserve input information through the network depth. Table 1 summarizes the structural properties.

Table 1

Structural properties of backbone architectures for multi-level feature aggregation

Architecture	Params (M)	Stage structure	Intermediate dims	Residual flow	Stage clarity	Inf. time (ms)	Aggregation suitability
VGG-16	138	Sequential	128/256/512	No	Med	~120	Limited by params
Inception v3	24	Branched	Mixed/varied	Partial	Low	~95	Ambiguous layers
DenseNet-121	8	Dense	128/256/512/1024	Dense	Low	~110	Feature redundancy
MobileNetV2	3.4	Inv. residual	32/64/160	Yes	Med	~35	Low-capacity features
ResNet-50	23	4-stage residual	256/512/2048	Yes	High	~75	Optimal
ViT-B/16	86	Uniform blocks	768/768/768	No	Low	~130	No spatial hierarchy

Note: inference times measured on NVIDIA RTX 3090, batch size 1, 224×224 input.

VGG-16 (Simonyan, & Zisserman, 2014) has a purely sequential structure of 3×3 convolutional blocks, making intermediate layer selection straightforward (e.g., conv2_2, conv3_3, conv5_3). The spatial resolution decreases by factor 2 at each pooling layer, creating a clear hierarchy from fine-grained (128-D at 56×56) to semantic (512-D at 7×7) features. However, VGG-16 lacks residual connections, meaning that low-level information is progressively diluted through the depth – early layer details must be re-learned at each subsequent stage rather than being preserved via skip connections. The fully-connected layers alone contain 123M of the total 138M parameters, making the architecture computationally inefficient for feature extraction when used as a frozen backbone. Multi-level aggregation produces a 896-dimensional descriptor (128 + 256 + 512), achieving F1 = 0.72 – lower than ResNet-50 despite having 6× more parameters.

Inception v3 (Szegedy et al., 2015) processes each spatial resolution through parallel branches with different kernel sizes (1×1, 3×3, 5×5, pooling). This multi-branch structure creates an ambiguity problem for multi-level aggregation: at each stage, there is no single canonical feature tensor to extract. Selecting one branch discards information from others; concatenating all branches yields an inconsistent representation across stages since the number and configuration of branches varies. While the total parameter count is moderate (24M), the internal topology does not support transparent extraction of hierarchically organized features. Aggregation from mixed outputs yields F1 = 0.71, but the result is sensitive to branch selection choices, making the pipeline less reproducible.

DenseNet-121 (Huang et al., 2017) connects each layer to all subsequent layers within a dense block via concatenation. While this ensures strong gradient flow and feature reuse, it creates a problem for multi-level aggregation: features at different stages are not independent. Each dense block's output already contains information from all preceding layers within the block, and transition layers between blocks further compress this accumulated information. Concatenating features from multiple dense blocks therefore introduces substantial redundancy – the same edge and texture information may appear in multiple components of the aggregated descriptor. Despite this, DenseNet-121 achieves F1 = 0.74, benefiting from its compact architecture (8M parameters) and efficient internal feature reuse that partially compensates for the redundancy at the aggregation level.

MobileNetV2 (Sandler et al., 2018) uses depthwise separable convolutions and inverted residual blocks to minimize computational cost and parameter count (3.4M). While the architecture does include residual connections and has a stage-like structure, the resulting intermediate features have very low channel dimensionality (32/64/160 at the selected stages). This low capacity limits the representational richness of each component in the aggregated descriptor. The multi-level descriptor is only 256-dimensional – an order of magnitude smaller than ResNet-50's 2816-D – and the F1 of 0.68 confirms that extreme parameter efficiency comes at the cost of feature richness for similarity tasks requiring fine-grained discrimination.

ResNet-50 (He et al., 2016) provides the most suitable structure for multi-level aggregation due to three key properties. First, its four clearly defined stages (C2, C3, C4, C5) with progressively decreasing spatial resolution (56×56 → 28×28 → 14×14 → 7×7) and increasing channel dimensionality (256 → 512 → 1024 → 2048) enable transparent layer selection at distinct levels of abstraction. Second, residual connections within each stage preserve low-level information through the entire depth, ensuring that early-stage features retain discriminative details that would otherwise be lost in a feedforward architecture. Third, the moderate total parameter count (23M) and intermediate dimensionality (256+512+2048 = 2816-D with C4 excluded) balance expressiveness and computational efficiency. The C2+C3+C5 configuration achieves F1 = 0.77, consistently outperforming all other architectures. C4 is excluded due to significant feature overlap with neighboring stages (Kubytskyi, & Panchenko, 2023).

ViT-B/16 (Dosovitskiy et al., 2021) consists of 12 identical transformer blocks, each producing a 768-dimensional output over a 14×14 grid of patch tokens. Unlike CNNs, ViT lacks a spatial hierarchy: all blocks operate at the same patch resolution, and the self-attention mechanism allows each block to aggregate information globally from the first layer. Intermediate features from blocks 3, 6, 9, 12 have identical dimensionality and similar structural properties, limiting the complementarity that drives multi-level aggregation gains in CNNs. The concatenated 3072-dimensional descriptor achieves F1 = 0.70 – only 2 percentage points above the single-block baseline, compared to the 9 p.p. gain observed with ResNet-50. ViT-B/16 also has 86M parameters – 3.7× more than ResNet-50 – and longer inference time.

CLIP ViT-B/32 (Radford et al., 2021) shares the architectural limitations of standard ViT but adds a further constraint: contrastive text-image training suppresses fine-grained visual details in favor of semantic alignment, making it poorly suited for near-duplicate detection tasks (Gkelios, Boutalis, & Chatzichristofis, 2021).

To quantify the practical impact of architectural properties, we evaluate each backbone in the multi-level aggregation pipeline described in Section 1. For each architecture, we report the F1-score achieved by the multi-level configuration alongside the single-layer baseline (final convolutional or attention block only) to measure the aggregation gain – the improvement attributable to intermediate layer inclusion. Results are presented in Tab. 2.

Table 2

Multi-level aggregation performance on INRIA Holidays (F1-score)

Architecture	Multi-level config	Dim.	Multi-level F1	Single-layer F1	Gain (p.p.)	Key limitation
ResNet-50	C2+C3+C5	2816	0.77 ± 0.02	0.68 ± 0.03	+9	–
DenseNet-121	dense1+2+4	1664	0.74 ± 0.02	0.68 ± 0.03	+6	Feature redundancy
VGG-16	conv2+3+5	896	0.72 ± 0.03	0.66 ± 0.03	+6	No residual flow
Inception v3	mixed stages	1280	0.71 ± 0.03	0.67 ± 0.03	+4	Ambiguous layers
ViT-B/16	blocks 3+6+9+12	3072	0.70 ± 0.03	0.68 ± 0.03	+2	No spatial hierarchy
MobileNetV2	stages 2+4+7	256	0.68 ± 0.03	0.63 ± 0.03	+5	Low capacity

The results in Tab. 2 reveal two key patterns. First, multi-level aggregation consistently improves over single-layer baselines across all architectures, confirming that intermediate features carry complementary information regardless of backbone design. Second, the magnitude of improvement varies widely and correlates with the structural criteria defined above. ResNet-50 shows the largest gain (+9 p.p.), followed by DenseNet-121 and VGG-16 (both +6 p.p.), MobileNetV2 (+5 p.p.), Inception v3 (+4 p.p.), and ViT-B/16 (+2 p.p.).

The ordering of aggregation gains is not random. Architectures with clear stage hierarchy and residual connections (ResNet-50) benefit most because their intermediate features are both diverse and well-preserved. Architectures where intermediate features are structurally similar (ViT-B/16) or already entangled through dense connections (DenseNet-121) benefit less, because concatenation adds redundant rather than complementary information.

The descriptor dimensionality does not predict performance: ViT-B/16 produces the largest descriptor (3072-D) yet achieves lower F1 than ResNet-50 (2816-D) and DenseNet-121 (1664-D). This confirms that the structure of intermediate features matters more than their quantity. From a practical standpoint, ResNet-50 achieves the best accuracy with moderate computational cost (23M parameters, ~75 ms/image), making it the most efficient backbone for production similarity systems.

Discussion and conclusions

This study identifies three architectural requirements for effective multi-level feature aggregation in image similarity tasks. First, clear stage hierarchy with distinct spatial resolutions ensures that features from different layers capture complementary information – from fine-grained edges and textures at early stages to semantic concepts at deep stages. Second, residual connections preserve low-level details through the network depth, preventing the information loss that degrades aggregation quality in architectures like VGG. Third, moderate intermediate dimensionality provides sufficient representational capacity without the excessive redundancy observed in DenseNet or the insufficient capacity of MobileNet.

ResNet-50 uniquely satisfies all three requirements, which explains its best performance across all tested configurations. Its four-stage structure provides a natural progression from spatial to semantic features, while skip connections ensure that each stage retains access to input-level information.

ViT-B/16 shows the smallest aggregation gain (+2 p.p.) despite having the largest descriptor. The uniform block structure – all 12 blocks produce 768-dimensional outputs at the same spatial resolution – means that intermediate features differ primarily in degree of abstraction rather than in spatial resolution or feature type. Hierarchical vision transformers such as Swin Transformer (Liu et al., 2021) may address this limitation by introducing a stage-based structure more analogous to CNNs. CLIP adds a further constraint: contrastive pre-training suppresses fine-grained details, making it poorly suited for near-duplicate detection.

For practitioners selecting a backbone for multi-level similarity systems, these results translate into concrete guidelines. If the task requires fine-grained discrimination, ResNet-50 with C2+C3+C5 is the recommended choice. If inference speed is the primary constraint, MobileNetV2 offers 2× faster processing at lower accuracy. DenseNet-121 provides a middle ground when memory is limited. VGG-16, Inception v3, and ViT-B/16 are not recommended for this pipeline due to the structural limitations identified above.

Limitations of this study include the use of a single benchmark dataset and frozen pre-trained weights. Future work should extend the evaluation to additional visual domains, investigate fine-tuning of intermediate layers, and test hierarchical vision transformers that combine the stage-based structure of CNNs with the self-attention mechanism. Another direction is learned aggregation weights, where the contribution of each layer is optimized jointly with the similarity classifier.

Authors' contribution: Volodymyr Kubytskyi – conceptualization, methodology, software, formal analysis, writing – original draft. Artem Bozhok – data validation, formal analysis, writing – review and editing.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of ICLR 2021*. arXiv:2010.11929

Gkelios, S., Boutalis, Y., & Chatzichristofis, S. A. (2021). Investigating the vision transformer model for image retrieval tasks. In *Proceedings of IEEE MMSP 2021*. https://doi.org/10.1109/MMSP53017.2021.9733553

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2015). Hypercolumns for object localization and fine-grained localization. In *Proceedings of CVPR 2015* (pp. 447–456).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR 2016* (pp. 770–778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of CVPR 2017* (pp. 4700–4708).

Jégou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In ECCV 2008 (pp. 304–317). Springer.

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). Near-duplicate video retrieval by aggregating intermediate CNN layers. In *Multimedia Modeling*. Springer.

Kubytskyi, V., & Panchenko, T. (2023). Enriched image embeddings as a combined outputs from different layers of CNN for various image similarity problems. In Lecture Notes on Data Engineering and Communications Technologies (Vol. 180, pp. 321–333). Springer. https://doi.org/10.1007/978-3-031-36115-9_30

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of CVPR 2017* (pp. 2117–2125).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of ICCV 2021 (pp. 10012–10022).

Panchenko, T., Bozhok, A., & Kubytskyi, V. (2026). Multi-level CNN feature fusion from ResNet50 for near-duplicate image detection in real estate imagery. Informatica, 50(9). https://doi.org/10.31449/inf.v50i9.12111

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. In *Proceedings of ICML 2021* (Vol. 139, pp. 8748–8763). PMLR.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of CVPRW 2014* (pp. 806–813).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of CVPR 2018* (pp. 4510–4520).

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of CVPR 2015* (pp. 1–9).

Thyagharajan, K. K., & Kalaiarasi, G. A. (2021). A review on near-duplicate detection of images using computer vision techniques. *Archives of Computational Methods in Engineering*, 28(3), 897–916. https://doi.org/10.1007/s11831-020-09400-w

Отримано редакцією журналу / Received: 01.12.25
Прорецензовано / Revised: 26.03.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Володимир КУБИЦЬКИЙ, асп.
ORCID ID: 0000-0002-1529-8677
e-mail: vova.kubytskyi@gmail.com
Київський національний університет імені Тараса Шевченка, Київ, Україна

Артем БОЖОК, асп.
ORCID ID: 0009-0009-9572-6501
e-mail: artembozh@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

КОНФІГУРАЦІЇ ПРОМІЖНИХ ШАРІВ ДЛЯ ЗАДАЧ ПОДІБНОСТІ ЗОБРАЖЕНЬ: ОЦІНКА БАГАТОРІВНЕВОГО ВЕКТОРНОГО ПРЕДСТАВЛЕННЯ, ОТРИМАНОГО З АРХІТЕКТУР CNN І ТРАНСФОРМЕРІВ

Оцінка візуальної подібності зображень потребує векторних представлень, що зберігають інформацію на різних рівнях абстракції. Багаторівнева агрегація ознак із проміжних шарів CNN демонструє високу ефективність у задачах виявлення майже дублікатів, однак систематичний аналіз впливу архітектурних властивостей різних backbone-мереж на якість таких представлень досі не проводився. У пропонованій роботі формалізовано конвеєр багаторівневої агрегації: GAP, конкатенація та L2-нормалізація проміжних активацій з подальшою класифікацією компактним MLP. Проаналізовано шість архітектур – VGG-16, Inception v3, DenseNet-121, MobileNetV2, ResNet-50 та ViT-B/16 – за критеріями чіткості стадій, просторової прогресії, незалежності ознак і залишкового потоку інформації. Кількісну оцінку виконано на

наборі даних INRIA Holidays за F1. ResNet-50 із шарами C2 + C3 + C5 досягає F1 = 0,77, перевершуючи всі інші конфігурації. VGG-16 дає F1 = 0,72 за надлишкової кількості параметрів. DenseNet-121 досягає F1 = 0,74 попри надмірність ознак. ViT-B/16 показує F1 = 0,70, обмежений однорідною структурою блоків. Чітка ієрархія стадій із залишковими зв'язками, поступовий перехід від просторових до семантичних ознак і помірна проміжна розмірність є ключовими архітектурними вимогами для ефективної багаторівневої агрегації.

К л ю ч о в і с л о в а : ***багаторівнева агрегація ознак, подібність зображень, згорткові нейронні мережі, візуальні трансформери, проміжні представлення, виявлення майже дублікатів, порівняння архітектур, ResNet-50.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.