УДК 004.932.2:004.89
DOI https://doi.org/10.24144/2616-7700.2026.49(2).197-204

В. О. Кубицький[1], А. В. Божок[2]

[1] Київський національний університет імені Тараса Шевченка,
аспірант кафедри теорії та технології програмування
vova.kubytskyi@gmail.com
ORCID: https://orcid.org/0000-0002-1529-8677

[2] Київський національний університет імені Тараса Шевченка,
аспірант кафедри теорії та технології програмування
artembozh@knu.ua
ORCID: https://orcid.org/0009-0009-9572-6501

МІЖДОМЕННЕ УЗАГАЛЬНЕННЯ БАГАТОРІВНЕВИХ CNN-ПРЕДСТАВЛЕНЬ ЗОБРАЖЕНЬ ДЛЯ ЗАДАЧ ОЦІНКИ ПОДІБНОСТІ

Запропоновано та досліджено багаторівневе векторне представлення зображень, яке агрегує ознаки з проміжних шарів C2, C3 та C5 згорткової нейронної мережі ResNet-50 за допомогою глобального усереднення, конкатенації та L2-нормалізації, формуючи єдиний 2816-вимірний дескриптор. Принциповою особливістю підходу є архітектурне відокремлення побудови універсального представлення від залежного від задачі механізму прийняття рішень, реалізованого як компактний багатошаровий перцептрон. Проведено оцінку міждоменної переносимості цього представлення на незалежному наборі даних INRIA Holidays, який суттєво відрізняється від домену первинної апробації методу: багаторівневий дескриптор перевершив одношарове CNN-представлення на 9 відсоткових пунктів за F1 при адаптації лише компактного MLP на 100 розмічених прикладах. Опубліковані результати на фіксованому наборі даних підтвердили обґрунтованість вибору ResNet-50 як базової архітектури: R-MAC на основі ResNet-50 перевершує R-MAC на основі VGG-19 на 4,9 в.п. mAP при шестиразово меншій кількості параметрів, а також є конкурентоспроможним порівняно з ViT-дескрипторами, які мають у 4–13 разів більший обсяг моделі.

Ключові слова: подібність зображень, згорткові нейронні мережі, багаторівневі представлення, ResNet-50, міждоменне узагальнення, виявлення майже дублікатів, перенесення навчання, INRIA Holidays.

1. Вступ. Веб-платформи та інформаційні системи з користувацьким контентом постійно накопичують колекції зображень, значна частина яких є майже ідентичними внаслідок републікацій, рекомпресії та графічних накладок [1]. Існуючі підходи до оцінки подібності — порівняння на рівні пікселів, локальні дескриптори [2], перцептивне хешування [5] — демонструють обмежену стійкість при складних перетвореннях. CNN-представлення [3] суттєво підвищили якість, однак одношарові ознаки з глибоких шарів пригнічують низько- та середньорівневу інформацію, критичну для виявлення майже дублікатів [4].

У роботах [5, 6] запропоновано метод побудови багаторівневих описових векторів зображень шляхом агрегації ознак із шарів C2, C3 та C5 архітектури ResNet-50 із застосуванням GAP, конкатенації та L2-нормалізації. Метод продемонстрував високу ефективність на задачах подібності у предметній області нерухомості (F1 0,96/0,87/0,77 для трьох підзадач) [5, 6]. Було також запропоновано принцип архітектурного відокремлення побудови представлення від залежного від задачі механізму оцінки подібності [6].

Водночас залишаються відкритими питання міждоменної переносимості отриманих представлень, порівняльного аналізу архітектур CNN для багаторівневої агрегації та зіставлення з альтернативними парадигмами — ViT [7] і CLIP [8]. Метою роботи є дослідження здатності запропонованого представлення до міждоменного узагальнення на незалежному наборі INRIA Holidays [9] та порівняльний аналіз з альтернативними архітектурами.

2. Аналіз існуючих методів представлення зображень для задач подібності. Визначення подібності зображень базується на перетворенні візуальних даних у представлення, яке дозволяє здійснювати змістовне порівняння. Формально, представлення зображення визначається як відображення:

$$f : \mathcal{I} \to \mathbb{R}^d,$$

де $\mathcal{I}$ — простір зображень, $\mathbb{R}^d$ — простір ознак фіксованої розмірності d. Подібність двох зображень I_1, I_2 обчислюється як:

$$s(I_1, I_2) = g\left(f(I_1), f(I_2)\right), \tag{1}$$

де g — функція подібності або прийняття рішення. Ефективність оцінки подібності визначається в першу чергу властивостями функції представлення f, тоді як функція g розкриває свою ефективність лише за умови інформативного простору ознак [5].

2.1. Класичні підходи. Порівняння на рівні пікселів є непридатним через чутливість до геометричних та фотометричних перетворень [1]. Локальні дескриптори (SIFT [2], SURF, ORB) забезпечують стійкість лише до обмеженого класу трансформацій і погано масштабуються до великих колекцій. Перцептивне хешування (pHash на основі DCT) [5] створює компактні бінарні представлення для швидкої фільтрації за відстанню Хеммінга, однак не може слугувати самостійним засобом оцінки подібності через обмежену стійкість до накладень та структурних варіацій [5].

2.2. Векторні представлення на основі CNN. CNN навчаються ієрархічним ознакам: ранні шари фіксують краї та текстури, глибокі — семантичні концепції [3, 10]. Ідея агрегації ознак з декількох шарів розвинена в HyperColumn [16] та Feature Pyramid Networks [17]. Поширена практика вилучення представлення з останнього шару CNN пригнічує низько- і середньорівневу інформацію, критичну для розрізнення майже дублікатів [4]. Агрегування ознак з декількох шарів зберігає доповнювальну інформацію на різних рівнях абстракції [4] — саме цей принцип покладено в основу методу [5, 6].

2.3. Альтернативні архітектури CNN. Вибір базової архітектури CNN суттєво впливає на якість вилучених ознак. VGG [11] має прозору послідовну структуру, але характеризується надлишковою кількістю параметрів ($\sim$ 138 млн). Inception [12] підвищує ефективність через мультигілкові модулі, однак розгалужена структура ускладнює вилучення проміжних ознак. ResNet [13] запровадила залишкові зв'язки, що стабілізують навчання та забезпечують збереження інформації по всій глибині. ResNet-50 має чітку поетапну структуру з послідовним зменшенням просторової роздільної здатності — саме ця властивість робить її найбільш придатною для багаторівневого вилучення ознак [5].

2.4. Візуальні трансформери та візуально-мовні моделі. Окремою парадигмою є візуальні трансформери (ViT) [7], які використовують механізм

self-attention для фіксації глобальних залежностей між частинами зображення. В оригінальній постановці ViT потребує значно більших обсягів навчальних даних порівняно з CNN [7], хоча сучасні методи self-supervised навчання частково знімають це обмеження. Візуально-мовні моделі, зокрема CLIP [8], забезпечують семантичні представлення високого рівня узагальнення, здатні до zero-shot перенесення, однак можуть пригнічувати дрібні візуальні деталі, критичні для виявлення майже дублікатів [5].

3. Метод побудови багаторівневого векторного представлення зображення.

3.1. Побудова представлення. Метод, детально описаний у [5, 6], використовує ResNet-50 [13], попередньо навчену на ImageNet, як фіксований екстрактор ознак (~ 23 млн замороженних параметрів). Ознаки вилучаються з шарів C2 (256 каналів, низькорівневі текстури), C3 (512 каналів, структурні патерни) та C5 (2048 каналів, семантика); C4 виключено через перекриття з сусідніми шарами [5]. До кожного тензора активації застосовується GAP, отримані вектори об'єднуються та нормалізуються:

$$\mathrm{e} = [\mathrm{v}_{C2}; \mathrm{v}_{C3}; \mathrm{v}_{C5}] \in \mathbb{R}^{2816}, \quad \mathrm{e}_{norm} = \mathrm{e}/\|\mathrm{e}\|_2. \tag{2}$$

3.2. Модель прийняття рішень. Для пари представлень $x^{(a)}$, $x^{(b)}$ будується симетричне парне представлення:

$$z = \mathrm{contact}\left(\left|x^{(a)} - x^{(b)}\right|, \left(x^{(a)} - x^{(b)}\right)^2\right). \tag{3}$$

Вектор (3) подається в MLP ($2048 \to 1024 \to 512 \to 256$, ELU [14], пакетна нормалізація, відсів), який навчається з BCE-втратами. Базова CNN залишається замороженою; для адаптації до нової задачі навчається лише MLP [6]. На задачах нерухомості метод досяг F1 0,96/0,87/0,77 для трьох підзадач [5].

4. Експериментальне дослідження.

4.1. Набір даних та протокол оцінки. Для перевірки міждоменної здатності до узагальнення запропонований метод оцінюється на наборі даних INRIA Holidays [9] — широко використовуваному еталонному наборі даних для задач пошуку подібних зображень. Набір містить 1491 зображення, згруповане у 500 класів сцен, де кожен клас представляє окрему сцену або об'єкт. Колекція охоплює різноманітні типи сцен — природні ландшафти, архітектурні об'єкти, предмети повсякденного вжитку — з варіаціями ракурсу, масштабу, освітлення та часткових перекриттів.

Цей набір суттєво відрізняється від домену первинної апробації методу [5, 6], де варіативність є переважно систематичною. Протокол оцінки відповідає [5]: зображення пропускаються через заморожену ResNet-50, ознаки з шарів C2, C3 та C5 об'єднуються та нормалізуються за L2, MLP навчається на 100 випадково вибраних розмічених парах.

Позитивні пари утворюються зі зображень одного класу (500 класів, 2–3 зображення на клас), негативні — випадковим вибором з різних класів у співвідношенні 1:1. Усі базові методи оцінено за тим самим протоколом та розбиттями. Оскільки задача є попарною класифікацією, а не ранжуванням (retrieval), якість вимірюється за F1 [18], а не за mAP.

Таблиця 1.

Порівняння методів на наборі даних INRIA Holidays

Метод	**F1**
Багаторівневе C2+C3+C5 + MLP (запропонований)	**0,77 ± 0,02**
Одношарове представлення ResNet-50	0,68 ± 0,03
Ознаки проміжного шару pool5 CNN [15]	0,65 ± 0,03
Перцептивний хеш DCT	0,63 ± 0,03

Запропоноване багаторівневе представлення перевершує найближчий базовий метод (одношарове представлення ResNet-50) на 9 відсоткових пунктів за F1. Порівняно з підходом Mousavian та Kosecka [15], який використовує ознаки лише одного проміжного шару (pool5), покращення становить 12 відсоткових пунктів.

4.2. Роль багаторівневої агрегації на різних доменах. У роботі [5] було проведено абляційне дослідження різних комбінацій шарів ResNet-50 на трьох задачах подібності зображень у предметній області нерухомості: задача A — виявлення майже дублікатів фотографій, які відрізняються лише графічними накладеннями (водяні знаки, логотипи, рамки); задача B — ідентифікація зображень одного приміщення, знятих із різних ракурсів; задача C — зіставлення схематичних планів поверхів у різному графічному оформленні. Результати абляції наведено в таблиці 2.

Таблиця 2.

Абляція комбінацій шарів ResNet-50 на задачах з предметної області нерухомості (F1) [5]

Комбінація шарів	**Задача A**	**Задача B**	**Задача C**
Тільки C2	0,77 ± 0,03	0,68 ± 0,03	0,63 ± 0,04
Тільки C3	0,80 ± 0,03	0,70 ± 0,02	0,67 ± 0,03
Тільки C5	0,87 ± 0,02	0,80 ± 0,02	0,69 ± 0,02
C2 + C3	0,82 ± 0,02	0,73 ± 0,03	0,69 ± 0,03
C2 + C5	0,89 ± 0,02	0,83 ± 0,02	0,71 ± 0,02
C3 + C5	0,92 ± 0,02	0,85 ± 0,02	0,74 ± 0,03
C2 + C3 + C5	**0,96 ± 0,01**	**0,87 ± 0,02**	**0,77 ± 0,02**

Конфігурація {C2, C3, C5}, оптимальність якої було встановлено абляційним дослідженням у [5], застосована без змін до набору INRIA Holidays. Повна конфігурація C2+C3+C5 досягає F1 = 0,77, тоді як одношарове представлення з останнього шару ResNet-50 — F1 = 0,68. Перевага багаторівневого представлення становить 9 в.п. — співставний результат з покращенням, спостережуваним на задачах нерухомості (від 7 до 9 в.п. при переході від C5 до C2+C3+C5).

Подібна величина покращення на двох різних доменах свідчить про те, що внесок низько- та середньорівневих ознак не є специфічним для одного домену. Водночас абсолютні значення F1 на INRIA Holidays (0,77) є нижчими, ніж на задачі A нерухомості (0,96), що пояснюється значно більшою візуальною різноманітністю набору INRIA Holidays. Остаточне підтвердження оптимальності конфігурації {C2, C3, C5} на інших доменах потребує повної абляції на кожному з них, що є напрямком подальших досліджень.

4.3. Порівняння архітектур CNN на основі опублікованих результатів. Для обґрунтування вибору ResNet-50 як базової архітектури запропонованого методу використано опубліковані результати на наборі даних INRIA Holidays (mAP). Слід зазначити, що retrieval протокол (mAP) відрізняється від попарної класифікації, що використовується в даній роботі, і ранжування архітектур за mAP не обов'язково зберігається при зміні протоколу. Тому наведені дані використовуються як допоміжний аргумент на користь вибору архітектури, а не як пряме порівняння методів. Результати наведено в таблиці 3.

Таблиця 3.

Результати пошуку зображень на INRIA Holidays (mAP, %) за опублікованими даними

Метод / Архітектура	**Параметри**	**INRIA mAP**	**Джерело**
MAC / VGG-19	∼144 млн	76,3	[19]
R-MAC / VGG-19	∼144 млн	87,7	[19]
R-MAC / ResNet-50	**∼23 млн**	**92,6**	[19]
ViT-B/16 (без навч.)	∼86 млн	∼90,5	[20]
ViT-L/32 (без навч.)	∼307 млн	∼87,1	[20]

Дані таблиці 3 підтверджують обґрунтованість вибору ResNet-50 у запропонованому методі. При однаковій стратегії агрегації (R-MAC) ResNet-50 перевершує VGG-19 на 4,9 в.п. mAP (92,6% проти 87,7%) при шестиразово меншій кількості параметрів, що свідчить про те, що залишкові зв'язки забезпечують більш інформативні проміжні ознаки для багаторівневої агрегації.

4.4. Порівняння базової архітектури із сучасними парадигмами. Опубліковані результати ViT на INRIA Holidays (таблиця 3) дозволяють оцінити обґрунтованість вибору ResNet-50 як базової архітектури порівняно з трансформерними моделями. За даними [19, 20], R-MAC на основі ResNet-50 досягає 92,6% mAP, тоді як ViT-B/16 без додаткового навчання — ∼90,5% mAP, а ViT-L/32 — ∼87,1% mAP. Слід підкреслити, що R-MAC є окремим методом агрегації [19], а не запропонованим підходом; ці дані використовуються для підтвердження того, що ResNet-50 формує більш дискримінаційні проміжні ознаки, ніж трансформерні архітектури порівнянного або більшого розміру (86–307 млн параметрів у ViT проти 23 млн у ResNet-50).

CLIP [8] забезпечує zero-shot перенесення, проте пригнічує дрібні візуальні деталі при значно більшому обсязі моделі (∼300 млн параметрів). Запропонований CNN-підхід орієнтований на збереження дрібних деталей при обчислювальній ефективності (∼75 мс/зобр. [5]) та мінімальних вимогах до адаптації (100 прикладів). Пряме порівняння з ViT та CLIP у задачі попарної класифікації є напрямком подальших досліджень.

5. Висновки та перспективи подальших досліджень. У роботі досліджено міждоменну переносимість запропонованого багаторівневого CNN-представлення зображень та проведено порівняльний аналіз з альтернативними архітектурами екстракторів ознак. Отримано такі основні результати.

1. Експериментально підтверджено міждоменну переносимість запропонованого багаторівневого представлення. На незалежному наборі даних INRIA Holidays, який суттєво відрізняється від домену первинної апробації, багаторівневий дескриптор на основі шарів C2, C3 та C5 досяг F1 = 0,77,

перевершивши одношарове CNN-представлення на 9 в.п. та ознаки pool5 на 12 в.п. за ідентичним протоколом оцінки. Адаптація до нового домену потребувала навчання лише компактного MLP на 100 розмічених прикладах без модифікації базового представлення, що підтверджує ефективність принципу архітектурного відокремлення екстрактора ознак від залежного від задачі механізму прийняття рішень.

2. Встановлено, що ефект багаторівневої агрегації відтворюється на різних доменах. Перевага конфігурації C2+C3+C5 над одношаровим C5 становить 9 в.п. на INRIA Holidays та 7–9 в.п. на задачах нерухомості [5] — кількісно співставний результат на двох принципово різних доменах. Це свідчить про загальний, а не доменоспецифічний характер внеску низько- та середньорівневих ознак у дискримінаційну здатність представлення. Подальша верифікація на додаткових наборах даних дозволить підтвердити цю закономірність з більшою статистичною достовірністю.
3. Обґрунтовано вибір ResNet-50 як оптимальної базової архітектури для багаторівневої агрегації ознак. За опублікованими даними [19], R-MAC на основі ResNet-50 перевершує R-MAC на основі VGG-19 на 4,9 в.п. mAP (92,6% проти 87,7%) на INRIA Holidays при шестиразово меншій кількості параметрів (23 млн проти 144 млн). Поетапна ієрархічна структура із залишковими зв'язками забезпечує найбільш придатну основу для прозорого та ефективного вилучення проміжних ознак.

Подальші дослідження доцільно спрямувати на розширення експериментальної бази, зокрема повна абляція конфігурацій шарів на додаткових наборах даних дозволить перевірити стійкість встановленої закономірності. Окремої уваги потребує дослідження методів зменшення розмірності 2816-вимірного представлення (PCA, квантування) для масштабування до колекцій із сотень мільйонів зображень, а також аналіз внеску окремих шарів у розрізнення конкретних типів візуальних перетворень.

Конфлікт інтересів

Автори заявляють, що не мають конфлікту інтересів щодо даного дослідження.

Фінансування

Дослідження було проведено без фінансової підтримки.

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автори підтверджують, що при створенні даної роботи вони не використо-

вували технології штучного інтелекту.

Внесок авторів

В. О. Кубицький: концептуалізація, методологія, формальний аналіз, експерименти, написання — оригінальний проєкт. А. В. Божок: формальний аналіз, курація даних, написання — рецензування та редагування.

Список використаної літератури

1. 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
2. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision,* 60(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
3. 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). https://doi.org/10.48550/arXiv.1403.6382
4. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). *Near-duplicate video retrieval by aggregating intermediate CNN layers.* In Multimedia Modeling. Springer. https://doi.org/10.1007/978-3-319-51811-4_21
5. 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
6. 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
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., & et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR 2021.* https://doi.org/10.48550/arXiv.2010.11929
8. Radford, A., Kim, J. W., Hallacy, C., & et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of ICML 2021.* https://doi.org/10.48550/arXiv.2103.00020
9. 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. https://doi.org/10.1007/978-3-540-88682-2_24
10. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *ECCV 2014.* (pp. 818–833). https://doi.org/10.48550/arXiv.1311.2901
11. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556.* https://doi.org/10.48550/arXiv.1409.1556
12. Szegedy, C., Liu, W., Jia, Y., & et al. (2015). Going deeper with convolutions. In *Proceedings of CVPR 2015.* https://doi.org/10.48550/arXiv.1409.4842
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR 2016.* (pp. 770–778). https://doi.org/10.48550/arXiv.1512.03385
14. Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In *Proceedings of ICLR 2016.* https://doi.org/10.48550/arXiv.1511.07289

15. Mousavian, A., & Kosecka, J. (2015). Deep convolutional features for image based retrieval and scene categorization. *arXiv:1509.06033.* https://doi.org/10.48550/arXiv.1509.06033
16. 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). https://doi.org/10.48550/arXiv.1411.5752
17. Lin, T.-Y., Dollár, P., Girshick, R., & et al. (2017). Feature pyramid networks for object detection. In *Proceedings of CVPR 2017.* (pp. 2117–2125). https://doi.org/10.48550/arXiv.1612.03144
18. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
19. Cataldo, A., Bianco, S., Napoletano, P., & Schettini, R. (2018). An accurate retrieval through R-MAC+ descriptors for landmark recognition. In *Proceedings of ICIAP 2018.*
20. 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

Kubytskyi V. O., Bozhok A. V. Cross-domain generalization of multi-level CNN image representations for similarity tasks.

A multi-level image vector representation is proposed and investigated, which aggregates features from intermediate layers C2, C3, and C5 of the ResNet-50 convolutional neural network using global average pooling, concatenation, and L2-normalization, producing a single 2816-dimensional descriptor. A key feature of the approach is the architectural separation of universal representation construction from the task-dependent decision mechanism, implemented as a compact multilayer perceptron. Cross-domain transferability of this representation is evaluated on the independent INRIA Holidays dataset, which substantially differs from the domain of the method's primary validation: the multi-level descriptor outperformed single-layer CNN representations by 9 percentage points in F1, with adaptation requiring only a compact MLP trained on 100 labeled examples. Published results on the same dataset confirmed the validity of ResNet-50 as the base architecture: R-MAC based on ResNet-50 outperforms R-MAC based on VGG-19 by 4.9 p.p. mAP with six times fewer parameters, and is competitive with ViT descriptors that have 4 to 13 times larger model size.

Keywords: image similarity, convolutional neural networks, multi-level representations, ResNet-50, cross-domain generalization, near-duplicate detection, transfer learning, INRIA Holidays.

Отримано: 30.03.2026 Прийнято: 16.04.2026 Опубліковано: 30.04.2026