

Візуальна локалізація на основі зображень: методи, виклики та напрями розвитку (2019–2025)

Іван Ільницький¹, Петро Пукач²

¹аспірант, Національний університет «Львівська політехніка»,
вул. Ст. Бандери, 12, 79013, Львів, email: ivan.v.ilnytskyi@lpnu.ua

²д. т. н., професор, Національний університет «Львівська політехніка»,
вул. Ст. Бандери, 12, 79013, Львів, email: petro.v.pukach@lpnu.ua

У статті представлено огляд домену систем локалізації на основі зображень, присвячений ключовим методам, викликам і сучасним напрямом розвитку у період 2019–2025 років. Розглянуто класифікацію підходів до візуальної локалізації, включно з традиційними підходами на основі локальних ознак і 3D-структур, підходами на основі пошуку зображень (візуальне розпізнавання місцевості), а також новітніми методами глибокого навчання. Проаналізовано набори даних, основні виклики галузі та перспективні напрями подальших досліджень.

Ключові слова: візуальна локалізація, локалізація на основі зображень, 6DoF позиції камери, гібридні методи, *imageretrieval*, локальні ознаки, регресія пози, нейронні радіальні поля, SLAM, глибоке навчання.

Вступ. Візуальна локалізація — це задача визначення позиції та орієнтації камери (шість ступенів вільності) за допомогою аналізу зображень сцени. Цей підхід широко застосовується в автономній навігації, доповненій (AR) та віртуальній реальності (VR), а також мобільних картографічних сервісах завдяки доступності та гнучкості сучасних камер порівняно з GPS або іншими засобами локалізації, які мають обмежене використання (наприклад, тільки для зовнішніх середовищ).

Формально задача абсолютної візуальної локалізації передбачає визначення 6DoF-пози камери на основі вхідних зображень відносно заздалегідь побудованої моделі сцени або карти. Залежно від наявності моделі, задача може бути локалізацією у відомому середовищі, коли доступна 3D-модель, або одночасною локалізацією та побудовою карти (SLAM/SfM), коли така модель будується динамічно (або в режимі реального часу). У цьому огляді основну увагу приділено першому випадку.

Класичні підходи базуються на геометричному співставленні локальних ознак (наприклад, SIFT, SURF) з 3D-моделлю [25]. Останнім часом популярними стали методи глобального пошуку зображень (*imageretrieval*) та методи глибокого навчання, що безпосередньо прогнозують координати чи позу камери (наприклад, PoseNet) [9]. Проте дослідження показують, що методи глибокого навчання наразі поступаються класичним підходам у точності [19]. Це мотивує

подальший розвиток гібридних підходів, що поєднують переваги нейромереж і традиційних геометричних методів.

Ця стаття є оглядом, присвяченим сучасним досягненням і викликам у сфері візуальної локалізації за період з 2019 по 2025 рік. У ній узагальнено ключові методи, проаналізовано популярні набори даних для їх оцінювання та визначено перспективні напрями подальших досліджень.

1. Класифікація підходів.

Системи візуальної локалізації можна класифікувати за характером використання даних і знань про середовище. Основна відмінність — локалізація у відомому середовищі (із попередньо побудованою 3D-моделлю або базою зображень) проти локалізації у невідомому середовищі (SLAM / SfM) [19, 9]. У цьому огляді розглядаються методи для відомого середовища. Їх можна поділити на чотири основні категорії:

Методи на основі локальних ознак і 3D-моделі (structure-based). Використовують відповідності між 2D-точками зображення та 3D-точками сцени для розв’язання PnP-задачі. Надійність таких методів забезпечується геометричними властивостями сцени та застосуванням RANSAC для відсіювання аномалій.

Методи на основі пошуку зображень (globalimageretrieval). Локалізація здійснюється через пошук схожих зображень у великій базі еталонних фото за допомогою глобальних дескрипторів (VLAD, NetVLAD [1]). Отримана позиція оцінюється на основі знайдених орієнтирів.

Методи глибокого навчання (learning-based).

(а) Пряма регресія пози: нейронна мережа (наприклад, PoseNet [9, 19]) безпосередньо прогнозує параметри пози камери.

(б) Прогноз координат сцени (SceneCoordinateRegression): нейромережа передбачає для кожного пікселя 3D-координати сцени, після чого поза обчислюється із застосуванням геометричного алгоритму (DSAC [2]).

Гібридні та ієрархічні підходи. Поєднують кілька стратегій задля покращення точності та масштабованості. Приклад — ієрархічна локалізація, де глобальний пошук слугує для грубої оцінки, а точна поза визначається локальним зіставленням ознак [14,15,18]. Також застосовуються ансамблі методів або каскадні архітектури [4].

У наступному розділі наводиться детальний огляд методів кожної з цих категорій, їх еволюції та практичної ефективності.

2. Огляд методів.

2.1 Локальні ознаки та 3D-структура: класичний підхід.

Традиційний підхід до image-based локалізації базується на побудові та використанні явної 3D-моделі сцени (набору 3D-точок із дескрипторами ознак). Для нового зображення локалізація відбувається шляхом знаходження

відповідностей між 2D-точками на зображенні та 3D-точками моделі, з подальшим розв'язанням задачі PnP.

Формулювання задачі PnP. У методах structure-based задача локалізації формулюється як пошук таких параметрів положення і орієнтації камери R, t , що для відомих 3D-точок сцени X_i справджується рівняння:

$$x_i = K(RX_i + t),$$

де:

- x_i — координати проєкції точки на зображенні (в однорідних координатах),
- X_i — 3D-точка у глобальній системі координат,
- K — матриця внутрішніх параметрів камери,
- R, t — шукані обертання та зсув (поза камери).

Розв'язання цієї задачі зазвичай здійснюється методами PnP у поєднанні з відсіюванням аномальних відповідностей за допомогою RANSAC. Ключовими компонентами є алгоритми виявлення та опису локальних ознак у зображенні, а також ефективний пошук відповідностей між 2D-дескрипторами і дескрипторами 3D-точок. Ранні локальні ознаки (SIFT, ORB) та методи побудови моделей за допомогою SfM зробили можливими перші реалізації таких систем [25].

Прикладні алгоритми. Одним із перших ефективних методів став алгоритм ActiveSearch [17], що оптимізував пошук відповідностей за допомогою попередньої індексації дескрипторів, пришвидшуючи локалізацію. Ці методи заклали основу структурної локалізації. Як показано в роботі [18], класичні feature-based методи залишаються дуже робастними до змін умов зйомки.

Обмеження. Попри високу точність, підходи на основі 2D–3D зіставлення мають обмеження: висока обчислювальна складність при великих сценах, вимоги до пам'яті для зберігання густої 3D-моделі, чутливість до екстремальних змін умов. Як зазначено у [14], пряме зіставлення 2D–3D є або робастним, але обчислювально важким, або прискореним, але менш надійним.

Покращені локальні ознаки. В останні роки було впроваджено навчальні дескриптори: SuperPoint [7], R2D2, а також навчання процесу співставлення, як у методі SuperGlue [15]. SuperGlue, побудована на графових нейронних мережах із механізмом уваги, значно підвищує точність співставлення навіть у складних умовах, що дозволило демонструвати рекордні результати на бенчмарках CVPR 2020.

Ієрархічна локалізація. Для вирішення проблеми масштабованості було запропоновано ієрархічні підходи: спершу виконується пошук схожих зображень у великій базі (наприклад, NetVLAD [1]), далі — локальне співставлення ознак лише в межах top-N кандидатів. Система HF-Net [14]) поєднує глобальний дескриптор і локальні ознаки в одній нейромережі, що дозволяє ефективно будувати ієрархічний pipeline. Такі методи стали стандартом для масштабованої локалізації в міських сценах.

2.2 Пошук по зображеннях (ImageRetrieval) і візуальне місцевизначення.

Методи *imageretrieval* не будують явної 3D-моделі сцени, а покладаються на колекцію еталонних зображень з відомими позами або геотегами. Завдання локалізації зводиться до пошуку в базі знімка, зробленого найближче до положення камери запиту.

Глобальні дескриптори. Ключовим компонентом підходів цього класу є функція, що перетворює зображення у компактний вектор ознак. Ранні роботи використовували VLAD, а згодом — NetVLAD [1], нейромережну модифікацію VLAD, що навчалася для задачі *placerecognition*. Сучасні глобальні дескриптори (NetVLAD, AP-GeM, DELF) стали стандартом для задач великомасштабної геолокації [13].

Локалізація за знайденими зображеннями. Простий підхід — брати позу найбільш схожого зображення як оцінку пози запиту. Для більш високої точності застосовують комбінацію з геометрією: після пошуку кандидата виконується локальне співставлення (локальні ознаки + PnP). Такий підхід використовувався, наприклад, у датасеті *AachenDay-Night* [18].

Переваги і недоліки. Сильна сторона підходів *imageretrieval* — масштабованість і ефективність пам'яті. Вони добре працюють для великих сцен (наприклад, ціле місто), а глобальні дескриптори демонструють стійкість до змін освітлення та сезону [23]. Основний недолік — обмежена точність у випадку відсутності схожих зображень у базі або при наявності періодичних структур.

Розвиток (2019–2025). У цей період було показано, що навіть при наявності 3D-моделі використання *globalimageretrieval* у *pipeline* локалізації дає великі вигоди за швидкодією без втрати точності [23]. З'явилися нові дескриптори з підвищеною інваріантністю до ракурсу (наприклад, *EigenPlaces*). Для задач автономного водіння розроблено спеціалізовані схеми локалізації за картами зображень (*MapillaryResearch*, 2021 (*MapillaryResearch* 2021)).

2.3 Нейронні мережі: пряма регресія пози.

Методи цієї групи прагнуть відмовитися від ручної побудови відповістей чи моделей і покладаються виключно на навчання нейронної мережі, що безпосередньо прогнозує положення камери. Основоположною роботою стала *PoseNet* [9], де CNN навчалася видавати 7-вимірний вектор (кватерніон орієнтації + 3D-координати) для вхідного зображення.

Loss-функція для навчання. У методах прямої регресії пози (наприклад, *PoseNet*) використовують *loss-функцію*, що суміщує похибку по позиції та орієнтації:

$$L = ||\hat{t} - t||_2 + \beta \cdot D(\hat{q}, q),$$

де:

- $\hat{t}, \hat{q}$ — передбачена позиція (вектор зсуву) та орієнтація (кватерніон),
- t, q — істинні (еталонні) значення,
- $D(\hat{q}, q)$ — функція різниці між кватерніонами (наприклад, кутова відстань),
- β — ваговий коефіцієнт для балансування впливу компонент позиції та орієнтації.

Оптимальний вибір β є важливим для забезпечення збалансованого навчання.

Переваги. Головна перевага таких методів — швидкість і простота використання. Після навчання модель може прогнозувати позу майже миттєво, що зручно для мобільних застосунків AR. Відсутність явної 3D-моделі зменшує обсяг пам'яті: знання про сцену зберігаються у вагових параметрах нейромережі.

Обмеження і точність. Попри активні дослідження, прямі нейронні методи поки поступаються альтернативам за точністю, особливо у великих сценах. Було показано [19], що такі моделі погано екстраполюють позу за межами області, охопленої тренувальними даними. Крім того, на практиці вони працюють подібно до *imageretrieval*: мережа фактично відшуковує найбільш схожий приклад із навчальної вибірки.

Новітні покращення. У 2019–2021 роках з'явилися підходи для підвищення надійності нейронних регресорів пози [5].

2.4 Прогнозування координат сцени та диференційований RANSAC.

Альтернативним шляхом використання глибокого навчання у локалізації є не пряма видача пози, а прогнозування проміжної інформації — координат сцени. Ця задача відома як *SceneCoordinateRegression* (SCR): для пікселів зображення прогножуються їхні координати у глобальній системі сцени.

Розвиток DSAC. Першим проривом у цій сфері став метод Shottonetal. (CVPR 2013, [20]), де було показано ефективність SCR для індор-локалізації. Надалі було запропоновано диференційовану версію RANSAC — DSAC [2], яка поєднує нейромережу для прогнозування координат з геометричним модулем. Подальша версія DSAC++ [3] підвищила точність і ефективність методу.

Переваги та недоліки. Методи SCR демонструють високу точність у локальних сценах і дозволяють явно враховувати геометрію. Вони особливо ефективні в індор-сценах з низькою текстурованістю. Проте такі підходи мають обмеження: складність навчання (потрібна густо анотована сцена), погана масштабованість до великих сцен, потреба у великому обсязі даних і пам'яті.

Нові напрямки SCR. Останнім часом з'явилися методи, які поєднують SCR із сучасними нейронними представленнями сцени. Наприклад, використання *NeuralRadianceFields* (NeRF, [10]) для локалізації [26]. Метод PNeRFLoc [26] комбінує точкове представлення NeRF з традиційним 2D–3D співставленням для підвищення точності локалізації за рахунок використання повної інформації про вигляд сцени.

3. Аналіз датасетів.

Прогрес у галузі багато в чому стимулюється наявністю публічних наборів даних та бенчмарків, на яких порівнюються алгоритми. Нижче наведено основні датасети, що активно використовувалися у 2019–2025 роках, та їх особливості.

7-Scenes. 7-Scenes (Microsoft, 2013) — класичний індор-датасет із 7 різними кімнатами, знятих камерою Kinect (RGB-D). Датасет став базовим для багатьох робіт, зокрема DSAC та PoseNet. Проблеми: повторювані текстури та динамічні об'єкти.

CambridgeLandmarks. CambridgeLandmarks [9] — набір із 5 міських сцен у Кембриджі, зображення зняті смартфоном. Використовувався для тестування PoseNet. Оригінальна похибка — 1–2 м; сучасні методи (AtLoc) [24] знизили її до десятків сантиметрів.

AachenDay-Night. AachenDay-Night [18] — міський датасет для оцінки локалізації в умовах дня і ночі. Виявилося, що нічна локалізація особливо складна; сучасні методи (SuperGlue, NightAug) підвищили якість локалізації до понад 80% із похибкою <5 м.

OxfordRobotCar / RobotCarSeasons. OxfordRobotCar / RobotCarSeasons [11] — великий набір, зібраний на автомобілі протягом року у різних погодних та сезонних умовах. Головний виклик — сезонні зміни (літо-зима), що значно ускладнюють класичне зіставлення ознак. Кращі підходи поєднували *imageretrieval* з навчанням відповідностей [15, 21].

InLoc. InLoc [22] — індор-датасет для локалізації у великих будівлях. Інтегрує 3D-модель (лазерний скан) та фотографії. Pipeline включає *imageretrieval*, локальний *matching* та рендеринг глибини. Датасет стимулював розвиток локалізації в складних індор-сценах.

Бенчмарки та змагання. Крім окремих датасетів, важливу роль відіграють спільні бенчмарки: сайт *visuallocalization.net* агрегує різні набори та підтримує єдину лідерборд [18]. Серія змагань Long-Term Visual Localization Challenge (2019–2021) [16] стимулювала розробку універсальних методів.

Підсумок. У 2019–2025 роках дослідники мали у розпорядженні широкий спектр тестових середовищ — від маленьких кімнат до цілих міст, від ідеальних умов до нічних зйомок. Загальна тенденція: покращення результатів завдяки комбінації *retrieval* та *matching*. Однак задачі локалізації в динамічних або слабо структурованих середовищах залишаються відкритими.

4. Основні виклики.

Розглянемо ключові проблеми та виклики, з якими стикаються системи локалізації на основі зображень.

Зміни умов освітлення та пори доби. Один з найскладніших викликів — це нічні умови та сильні тіні/яскраве сонце. Алгоритми, що покладаються на кольорові чи текстурні ознаки, можуть зазнавати невдачі при зміні освітлення. Нічні зображення містять менше деталізації, багато шуму, домінують локальні джерела світла, що спричиняє появу нових візуальних елементів. Інваріантні описи (наприклад, навчання дескрипторів на парах день-ніч) частково

допомагають, але повного рішення ще немає. Використання інфрачервоних камер або активного освітлення виходить за рамки чисто vision-підходів.

Сезонні зміни та погодні умови. Сцена може виглядати дуже по-різному влітку та взимку. Листя на деревах може закривати будівлі, сніг — покривати текстури землі. Методи, які "запам'ятали" літній вигляд, погано впізнають локацію взимку. Один із підходів — використання семантичної абстракції (наприклад, контурів будівель, доріг), що дозволяє ігнорувати змінені частини зображення.

Варіації ракурсу та масштабу. Якщо камера знімає сцену під іншим кутом або з іншої висоти, класичні алгоритми можуть не знайти відповідностей. Частково це вирішується за рахунок багаторакурсних моделей або використанням дескрипторів із підвищеною інваріантністю до трансформацій. Також застосовуються трансформери та тривимірні згорткові мережі [21, 6].

Повторювані та схожі структури. У міських сценах багато періодичних елементів (вікна, однакові будівлі). Це призводить до помилкових відповідностей. Для боротьби з цим застосовують покращені фільтри (наприклад, SuperGlue) та вимоги геометричної узгодженості. Проте в сильно однорідних сценах доводиться залучати додаткові сенсори (IMU, магнетометр).

Масштабування і продуктивність. Для мобільних пристроїв важлива швидкість роботи та здатність обробляти великі сцени. Сучасні ієрархічні методи наближаються до реального часу навіть на смартфонах, однак для великих баз потрібні серверні обчислення, що піднімає питання приватності (privacy-preserving localization).

Динамічні сцени та об'єкти. Рухомі об'єкти (автомобілі, люди, тварини) можуть створювати хибні відповідності. Основний підхід — маскування динамічних об'єктів за допомогою попереднього розпізнавання. Інтеграція сегментації та локалізації дозволяє підвищити стійкість алгоритмів у динамічному середовищі [6].

5. Останні напрями досліджень.

У період 2019–2025 рр. у галузі візуальної локалізації спостерігалось кілька чітких тенденцій, що визначають вектор розвитку досліджень.

Інтеграція глибокого навчання з класичними методами. Якщо раніше існувало протиставлення "традиційні vs. нейронні" підходи, то останні методи об'єднують їх сильні сторони. Наприклад, ієрархічні методи (гібрид global+local), де нейронні мережі генерують ознаки, а геометричні алгоритми забезпечують точну позу. Методи типу NG-RANSAC (2019) [4] інтегрують нейронний модуль у класичний фільтр RANSAC, підвищуючи його ефективність. Багато сучасних рішень включають каскадні підходи, де мережа (наприклад, PoseNet) забезпечує початкову оцінку, що потім уточнюється за допомогою локального оптимізатора (PnP).

Використання трансформерів і attention-механізмів. Трансформери, які досягли успіху в обробці природної мови, тепер активно застосовуються у комп'ютерному зорі. Архітектури, такі як LoFTR [21] (CVPR 2021), використовують multi-head attention для знаходження цільної карти відповідностей між зображеннями. Це дозволяє визначати відповідності навіть на текстурно слабких ділянках. Крім того, трансформери сприяють поєднанню даних з різних модальностей, наприклад, зображень і хмар точок LiDAR (TransformerFusion, [6]).

Семантично обізнані методи. Додавання семантичної інформації допомагає підвищити стійкість локалізації. Прикладом є SemanticVLAD [8], який агрегує ознаки для різних семантичних класів (дороги, будівлі, рослинність), роблячи дескриптори менш чутливими до тимчасових об'єктів. Семантика також використовується в методах SCR для покращення узагальнення на нові сцени.

NeuralSceneRepresentations (NeRF та інші). Представлення середовища за допомогою нейронних радіальних полів (NeRF) дозволяє компактно зберігати сцену та відтворювати її з будь-якого кута. Методи iNeRF (2021) оптимізують параметри поз для знаходження відповідностей між реальними зображеннями та NeRF-сценами. Інші підходи, такі як CrossLoc, генерують синтетичні вигляди для тренування моделей локалізації.

Узагальнення та динамічна адаптація (або оперативне перенавчання). Створюються моделі, здатні адаптуватися до нових сцен без повного перенавчання. Прикладом є Learning to localize in minutes, де модель DSAC швидко переналаштовується під нові дані. Інший підхід — zero-shotgeneralization, коли великі моделі VisionTransformer, натреновані на різних сценах, демонструють здатність до узагальнення без додаткового тренування.

Приватність та розподілені обчислення. Впроваджуються федеративні методи локалізації для захисту даних користувачів. Прикладами є передача лише дескрипторів, використання шифрування запитів або гомоморфного шифрування для зіставлення на сервері без розкриття даних. Це напрям важливий для практичних застосувань (наприклад, Google VPS, AppleARKit) і активно розвивається.

В цілому, останні напрями демонструють прагнення зробити локалізацію точнішою, робастною та універсальною через поєднання різних джерел даних і сильних сторін класичних та глибоких методів.

Висновки. Системи локалізації на основі зображень демонструють стрімкий прогрес, поєднуючи алгоритми комп'ютерного зору з методами глибокого навчання. У цьому огляді висвітлено основні підходи — структурні методи з локальними ознаками і 3D-моделями, imageretrieval, прямі нейронні регресори пози та гібридні рішення.

Кожен клас методів демонструє унікальний баланс характеристик: структурні методи вирізняються точністю, підходи image retrieval — масштабованістю та ефективністю пам'яті, а методи глибокого навчання — швидкістю та адаптивністю. Водночас залишаються значні виклики:

робастність до змін оточення (освітлення, сезон, ракурс), динамічні сцени та потреба в ефективній адаптації моделей. Аналіз датасетів і бенчмарків свідчить, що інтеграція гібридних підходів (наприклад, SuperGlue, LoFTR, ієрархічні локалізатори) значно підвищила якість локалізації.

Перспективними напрямками є подальша інтеграція семантики, використання attention-механізмів і нейронних 3D-представлень (NeRF), а також розробка універсальних моделей з високою узагальнюючою здатністю.

Подальші дослідження мають бути спрямовані на забезпечення масштабованої, довготривалої та робастної локалізації в реальних, змінних умовах. Це відкриває нові можливості для застосування — від автономної навігації до доповненої реальності у повсякденному житті.

Література

- [1] Arandjelović, R., P. Gronat, A. Torii, T. Pajdla, and J. Sivic. 2016. "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition." Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5297–307. <https://arxiv.org/abs/1511.07247>
- [2] Brachmann, E., A. Krull, S. Nowozin, et al. 2017. "DSAC – Differentiable RANSAC for Camera Localization." Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 668–76. <https://arxiv.org/abs/1611.05705>
- [3] Brachmann, E., and C. Rother. 2019a. "Expert Sample Consensus Applied to Camera Re-Localization." Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 7525–34. <https://arxiv.org/abs/1908.02484>
- [4] Brachmann, E., and C. Rother. 2019b. "Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses." Proc. IEEE/CVF International Conference on Computer Vision (ICCV). <https://arxiv.org/abs/1905.04132>
- [5] Chen, C., B. Wang, C. X. Lu, N. Trigoni, and A. Markham. 2024. "Deep Learning for Visual Localization and Mapping: A Survey." IEEE Transactions on Neural Networks and Learning Systems, ahead of print. <https://arxiv.org/abs/2308.14039>
- [6] Chen, Y. et al. 2022. Transformer Fusion: Point Cloud Guided Global Feature Fusion for Image-Based Localization. <https://arxiv.org/abs/2203.10722>
- [7] DeTone, D., T. Malisiewicz, and A. Rabinovich. 2018. "SuperPoint: Self-Supervised Interest Point Detection and Description." Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). <https://arxiv.org/abs/1712.07629>
- [8] Garg, S. et al. 2019. Semantic-geometric visual place recognition: a new perspective for reconciling opposing views. https://www.researchgate.net/publication/332280264_Semantic-geometric_visual_place_recognition_a_new_perspective_for_reconciling_opposing_views
- [9] Kendall, A., M. Grimes, and R. Cipolla. 2015. "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization." Proc. IEEE International Conference on Computer Vision (ICCV). <https://arxiv.org/abs/1505.07427>
- [10] Lin, Y.-C. et al. 2021. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. <https://arxiv.org/abs/2012.05877>
- [11] Maddern, W., G. Pascoe, C. Linegar, and P. Newman. 2017. "1 Year, 1000 Km: The Oxford Robot Car Dataset." International Journal of Robotics Research (IJRR). <https://ora.ox.ac.uk/objects/uuid%3Ac5266bad-e0f8-49f1-918e-1602ef935990>
- [12] Mapillary Research. 2021. Visual Place Recognition for Autonomous Driving. <https://research.mapillary.com>
- [13] Noh, H., A. Araujo, J. Sim, T. Weyand, and B. Han. 2017. "Large-Scale Image Retrieval with Attentive Deep Local Features." Proc. IEEE International Conference on Computer Vision (ICCV). <https://arxiv.org/abs/1612.06321>
- [14] Sarlin, P.-E., C. Cadena, R. Siegwart, and M. Dymczyk. 2019. "From Coarse to Fine: Robust Hierarchical Localization at Large Scale." Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12716–25. <https://arxiv.org/abs/1812.03506>
- [15] Sarlin, P.-E., D. DeTone, T. Malisiewicz, and A. Rabinovich. 2020. "Super Glue: Learning Feature Matching with Graph Neural Networks." Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4937–46. <https://arxiv.org/abs/1911.11763>

- [16] Sattler, T. et al. 2019. “Long-Term Visual Localization Challenge.” Proc. CVPR Workshops. <https://www.visuallocalization.net/>
- [17] Sattler, T., B. Leibe, and L. Kobbelt. 2017. “Efficient&Effective Prioritized Matching for Large-Scale Image-Based Localization.” https://www.cvlb.net/projects/autonomous_vision_survey/literature/Sattler2016PAMI.pdf
- [18] Sattler, T., W. Maddern, C. Toft, et al. 2018. “Benchmarking 6DOF Out door Visual Localization in Changing Conditions.” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8601–10. <https://arxiv.org/abs/1707.09092>
- [19] Sattler, T., Q. Zhou, M. Pollefeys, and L. Leal-Taixé. 2019. “Understanding the Limitations of CNN-Based Absolute Camera Pose Regression.” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3297–307. <https://arxiv.org/abs/1903.07504>
- [20] Shotton, J. et al. 2013. “Scene Coordinate Regression Forests for Camera Relocalization in RGB-d Images.” Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). <https://www.microsoft.com/en-us/research/publication/scene-coordinate-regression-forests-for-camera-relocalization-in-rgb-d-images/>
- [21] Sun, J., Z. Shen, T. Lin, L. Huang, H. Bao, and X. Zhou. 2021. “LoFTR: Detector-Free Local Feature Matching with Transformers.” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8922–31. <https://arxiv.org/abs/2104.00680>
- [22] Taira, H., M. Okutomi, T. Sattler, A. Torii, and T. Pajdla. 2018. “InLoc: In door Visual Localization with Dense Matching and View Synthesis.” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). <https://arxiv.org/abs/1803.10368>
- [23] Masone, Carlo & Caputo, Barbara. (2021). A Survey on Deep Visual Place Recognition. IEEE Access. 9. 19516–19547. 10.1109/ACCESS.2021.3054937.
- [24] Wang, B., C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham. 2020. “AtLoc: Attention Guided Camera Localization.” Proc. AAAI Conference on Artificial Intelligence (AAAI). <https://ojs.aaai.org/index.php/AAAI/article/view/6608>
- [25] Wu, Y., F. Tang, and H. Li. 2018. “Image-Based Camera Localization: An Overview.” Visua l Computing for Industry, Biomedicine, and Art 1 (8). <https://arxiv.org/pdf/1610.03660>
- [26] Zhao, B., L. Yang, M. Mao, H. Bao, and Z. Cui. 2023. PNeRFLoc: Visual Localization with Point-Based Neural Radiance Fields. <https://arxiv.org/abs/2312.10649>

Visual localization based on images: methods, challenges, and trends (2019–2025)

Ivan Illytskyi, Petro Pukach

The article provides an overview of the domain of image-based localization systems, focusing on key methods, challenges, and current trends in development for the period 2019–2025. It considers the classification of approaches to visual localization, including traditional approaches based on local features and 3D structures, approaches based on image search (visual recognition of terrain), as well as the latest deep learning methods. It analyzes datasets, key industry challenges, and promising areas for further research.

Keywords: visual localization, image-based localization, 6DoF camera position, hybrid methods, image retrieval, local features, position regression, neural radial fields, SLAM, deep learning.

Отримано 21.10. 2025