НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2026, том 36, № 4, с. 105–117 *ISSN 0868-5886*

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.932.72'1
EDN: YQPHHC

© А. С. Фатеев, Е. Л. Киселев, М. В. Терентьев, Е. В. Бокшанский, Н. В. Саушкин, Е. А. Татаринова, Д. Н. Щелкунов, С. О. Русскин, 2026

ПРЕДВАРИТЕЛЬНЫЙ ОТБОР ФРАГМЕНТОВ ГЕОПРИВЯЗАННОЙ КАРТЫ ПО ГЛОБАЛЬНЫМ ДЕСКРИПТОРАМ В СИСТЕМЕ ВИЗУАЛЬНОЙ НАВИГАЦИИ БПЛА

Рассматривается задача предварительного отбора фрагментов геопривязанной карты для визуальной навигации беспилотного летательного аппарата (БПЛА) при росте неопределенности положения после длительного отсутствия глобальной коррекции. Предлагается двухэтапная схема, в которой кандидаты сначала ранжируются по глобальным дескрипторам, а затем проверяются полной моделью сопоставления. Дескриптор формируется агрегированием промежуточной карты признаков замороженного кодировщика DINOv2, используемого в составе модели сопоставления, что позволяет отказаться от отдельной нейросетевой модели. Для поиска сравниваются точный индекс и приближенные индексы IVF и HNSW. Эксперименты на наборе UAV-VisLoc и карте Esri World Imagery показали, что приближенный поиск сохраняет полноту точного перебора при пренебрежимо малом времени запроса, а предварительный отбор существенно сокращает число запусков модели сопоставления. Дополнительно исследовано влияние размера области поиска, источника картографических данных и искажений входного изображения.

Ключевые слова: беспилотные летательные аппараты, визуальная навигация, сопоставление со спутниковыми снимками, кросс-модальное сопоставление, приближенный поиск ближайших соседей, глобальные дескрипторы изображений, предварительный выбор фрагментов карты, геопривязанные изображения

ВВЕДЕНИЕ

В современных навигационных системах беспилотных летательных аппаратов (БПЛА) положение оценивается по данным из нескольких взаимодополняющих источников. Инерциальная и другие системы локальной навигации обеспечивают высокую частоту обновления, однако их ошибки накапливаются со временем. Для ограничения этой погрешности используются периодические глобальные коррекции, получаемые от глобальной навигационной спутниковой системы (ГНСС) или путем визуального сопоставления кадров с геопривязанной картой [1–3].

При визуальной привязке кадр бортовой камеры сопоставляется со спутниковым или аэрофотосъемки изображением. По найденным соответствиям оценивается геометрическое преобразование и определяется положение наблюдаемого участка. Такое измерение непосредственно связано с картой, поэтому его ошибка не накапливается от кадра к кадру.

Однако перед сопоставлением необходимо выбрать подходящие фрагменты карты. В штатном режиме область поиска ограничивается прогнозом навигационного фильтра и содержит небольшое число фрагментов. При длительном отсутствии глобальной коррекции неопределенность положения возрастает, а область поиска может охватывать тысячи фрагментов. Поскольку ее площадь увеличивается пропорционально квадрату радиуса, полный перебор быстро становится вычислительно неприемлемым. Современные модели сопоставления используют детектируемые ключевые точки [4–6] либо строят полуплотные и плотные поля соответствий [7–9]. Обработка одной пары изображений на бортовом вычислителе может занимать от сотен миллисекунд до нескольких секунд. Поэтому последовательная проверка тысяч фрагментов не обеспечивает своевременное формирование координатного измерения. Таким образом, перед запуском полной модели сопоставления число кандидатов необходимо предварительно сократить.

В работе предлагается двухэтапный метод выбора фрагментов геопривязанной карты. Сначала кандидаты ранжируются по глобальным дескрипторам, применяемым в задачах распознавания местности [10–12], после чего лучшие из них передаются полной модели сопоставления для геометрической проверки. Поиск в базе дескрипторов

выполняется с помощью методов приближенного поиска ближайших соседей (Approximate Nearest Neighbor, ANN) [13–16].

Чтобы не использовать на борту отдельную модель глобального описания, дескриптор формируется из промежуточной карты признаков модели сопоставления. Таким образом, предварительный отбор и последующая геометрическая проверка используют один признаковый кодировщик.

В работе решены следующие задачи:

1) разработана двухэтапная процедура выбора фрагментов карты, сочетающая пространственное ограничение области поиска и ранжирование по глобальным дескрипторам;

2) исследовано построение глобального дескриптора путем агрегирования промежуточной карты признаков модели сопоставления;

3) оценено влияние размера области поиска, источника картографических данных и искажений входного изображения на полноту отбора, а также измерены вычислительные затраты метода.

Предлагаемый метод дополняет разработанный ранее алгоритм фильтрации видеопотока по информативности кадров [17]. Фильтрация уменьшает число обрабатываемых кадров, а рассматриваемый метод — число проверяемых фрагментов карты.

ПОСТАНОВКА ЗАДАЧИ

Пусть бортовая база содержит N фрагментов карты. Для каждого фрагмента заданы координаты центра c_i и заранее вычисленный дескриптор. Для поступившего кадра I_q требуется выбрать фрагменты, с которыми будет выполнено полное сопоставление.

Область поиска ограничивается по данным навигационной системы. По оценке положения $\hat{p}_t$ и радиусу неопределенности R_t формируется множество допустимых кандидатов:

$$\mathrm{C}_t = \left\{ i : \mathrm{dist}\left(c_i, \hat{p}_t\right) \le R_t \right\}, \qquad (1)$$

где $\mathrm{dist}(\cdot,\cdot)$ обозначает расстояние на местности. При отсутствии достоверной априорной оценки множество C_t совпадает со всей базой снимков.

Геометрическая интерпретация пространственного ограничения приведена на рис. 1. В центральном круге показана область возможного положения БПЛА радиусом R_t = 8 км и ограничивающий ее квадрат со стороной $2R_t$ = 16 км, используемый для первичного пространственного запроса. Сетка с шагом 2 км нанесена только для оценки масштаба. При фактическом размере фрагмента 256 × 256 м и шаге разбиения карты 128 м каждой изображенной ячейке соответствует около 250 фрагментов.

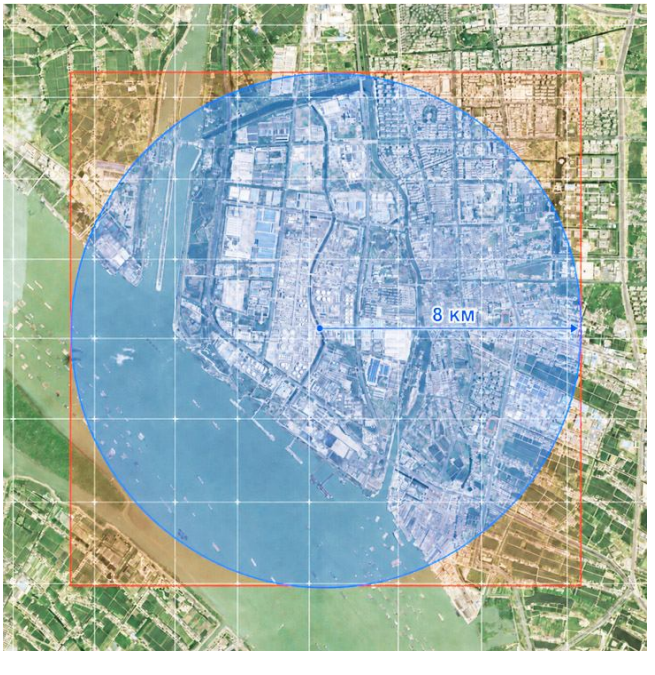

Рис. 1. Пример области поиска (ограничена кругом) при накопленной координатной неопределенности

По мере накопления координатной ошибки радиус R_t увеличивается, поэтому в область поиска попадает все больше фрагментов карты. При R_t = 8 км площадь этой области составляет около 200 км^2, что соответствует примерно 12×10^3 фрагментов.

Когда $|C_t|$ невелико, полное сопоставление можно выполнить для всех допустимых фрагментов. Если же кандидатов много, из C_t отбирается подмножество K_t меньшего размера. Насколько при этом сокращается число запусков модели сопоставления, показывает отношение

$$\rho = \frac{|C_t|}{|K_t|}, \qquad (2)$$

здесь числитель и знаменатель обозначают мощности множеств.

Решение о включении предварительного отбора опирается на текущий размер области неопределенности, а не на показания отдельного датчика. Пока глобальная привязка остается точной, система перебирает все фрагменты из C_t. Как только коррекция долго не поступает и $|C_t|$ растет, подключается поиск по дескрипторам.

От предварительного отбора требуется обеспечить попадание географически верного фрагмента в K_t при жестком ограничении на его размер и небольших вычислительных затратах.

МЕТОД

Характеристика метода

Предлагаемый метод основан на двухэтапном выборе фрагментов геопривязанной карты. На первом этапе выполняется быстрое ранжирование кандидатов по сходству глобальных дескрипторов, а на втором — полное сопоставление бортового кадра только с ограниченным числом фрагментов, получивших наибольшие оценки сходства. Такое разделение позволяет исключить заведомо неподходящие участки карты без запуска вычислительно сложной модели сопоставления для каждого кандидата.

Вычисления распределены между подготовкой на земле и работой в полете. До вылета собирается база картографических фрагментов, для каждого из которых заранее вычисляются глобальный дескриптор и географические координаты. В полете такой же дескриптор строится для текущего кадра бортовой камеры, после чего в базе отыскиваются наиболее близкие к нему элементы. Отобранные фрагменты передаются основной модели сопоставления для установления локальных соответствий и последующей геометрической проверки. Общая схема метода приведена на рис. 2.

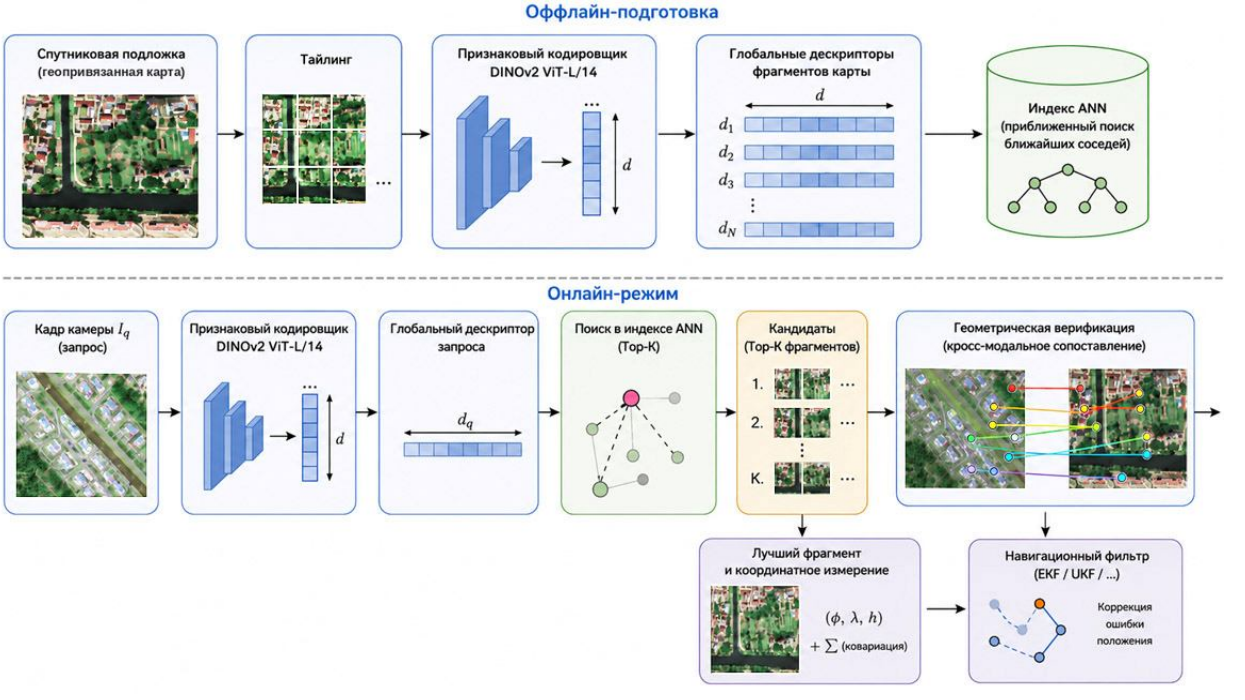

Рис. 2. Схема ускоренного выбора фрагментов карты

Подготовка базы

Исходное картографическое изображение разбивается на квадратные фрагменты одинакового размера с заданным перекрытием. Использование перекрытия уменьшает зависимость результата от положения наблюдаемого участка относительно границ сетки и повышает вероятность того, что он целиком или преимущественно попадет хотя бы в один фрагмент.

Каждый фрагмент T_i проходит через признаковый кодировщик модели сопоставления, дающий карту признаков $F(T_i)$. Пространственная размерность сворачивается обобщенным усреднением [11]

$$\mathrm{Pool}\left(F\right)_c = \left(\frac{1}{|\Omega|}\sum\nolimits_{u\in\Omega} F_c\left(u\right)^p\right)^{1/p}, \qquad (3)$$

где c — номер канала, Ω — множество пространственных позиций, а p — параметр обобщенного усреднения. Полученный вектор нормируется по евклидовой норме:

$$v_i = \frac{\mathrm{Pool}\left(F\left(T_i\right)\right)}{\left\|\mathrm{Pool}\left(F\left(T_i\right)\right)\right\|_2}. \qquad (4)$$

После L2-нормировки скалярное произведение двух дескрипторов равно косинусу угла между ними, поэтому близость фрагмента к запросу удобно измерять величиной $q^{\mathrm{T}}v_i$ и упорядочивать фрагменты по ее убыванию.

Пары $\left\{\left(v_i, c_i\right)\right\}_{i=1}^{N}$ образуют бортовую базу дескрипторов. Поиск ближайших к запросу фрагментов ведется по ней через индекс приближенного поиска ближайших соседей [14, 15]. Без априорной оценки положения искать приходится по всей базе, и здесь индекс особенно ценен. Он быстро выделяет небольшой набор вероятных кандидатов, не сравнивая запрос с каждым дескриптором. Приближенный поиск способен сдвинуть состав первых кандидатов относительно точного перебора, поэтому его влияние на полноту отбора оценивается отдельно. Соответствие фрагмента текущему кадру подтверждается лишь после полного сопоставления и геометрической проверки.

Поиск в полете

Для кадра I_q тем же кодировщиком и способом агрегирования вычисляется дескриптор запроса

$$q = \frac{\mathrm{Pool}\left(F\left(I_q\right)\right)}{\left\|\mathrm{Pool}\left(I_q\right)\right\|_2}. \qquad (5)$$

Из множества, прошедшего пространственное ограничение (1), выбираются k векторов с наибольшим значением скалярного произведения:

$$\mathrm{K}_t = \mathrm{Top\text{-}K}_{i\in C_t}\left(q^{\mathrm{T}}, v_i\right). \qquad (6)$$

Для каждого отобранного фрагмента модель сопоставления строит соответствия и оценивает преобразование между кадром и картой. Несогласованные соответствия отбраковываются при геометрической проверке. Если ни один кандидат не удовлетворяет заданным критериям, координатное измерение по текущему кадру не формируется.

Реализация метода

В экспериментах использовалась модель RoMa [9] в модально-инвариантной модификации MINIMA [18]. Для построения дескриптора выполнялся только признаковый кодировщик, без декодера сопоставления. В качестве $F(\ldots)$ использовались патч-токены DINOv2 ViT-L/14 [19] с пространственным шагом 14 пикселей и 1024 каналами.

Кодировщик DINOv2 в составе RoMa заморожен и не изменяется при дообучении матчера. Поэтому получаемые признаки совпадают с признаками исходной DINOv2. Преимущество их повторного использования состоит не в повышении качества дескриптора, а в отсутствии необходимости применять отдельную нейросетевую модель. Это также подтверждается экспериментами, приведенными ниже.

Входные изображения масштабировались до 518 × 518 пикселей и нормировались по статистикам ImageNet. Параметр p в (3) не обучался и был принят равным 3.0. Размерность дескриптора составляла d = 1024, а объем одного вектора в формате float32 — 4 КБ. Для поиска использовалась библиотека FAISS [16]. Точный перебор Index Flat IP служил эталоном. В качестве приближенных индексов рассматривались Index IVFFlat с параметрами $n_{\mathrm{list}} = \left\lfloor\sqrt{N}\right\rfloor$ и $n_{\mathrm{probe}} = 16$, а также Index HNSW Flat с параметрами M = 32, efConstruction = 200 и efSearch = 64. Поскольку все дескрипторы нормированы по L2-норме, их сходство вычислялось по скалярному произведению.

Когда априорная оценка положения отсутствует, множество C_t охватывает всю базу и поиск идет прямо по индексу. Именно этот режим — с грубым или утраченным начальным приближением — и есть основной сценарий для метода. Если же оценка положения известна, множество C_t заранее сужается пространственным запросом (1).

Индекс строится по всей базе, поэтому ранжировать внутри такого географического подмножества проще напрямую — вычислением $|C_t|$ скалярных произведений. Для нескольких тысяч кандидатов это занимает доли миллисекунды и обращения к индексу не требует.

Выигрыш от приближенного поиска становится заметен с ростом базы. Для 3685 дескрипторов точный перебор занимает 0.650 мс — на фоне $t_{\text{кодир}} \approx 92$ мс, уходящих на построение дескриптора запроса, это пренебрежимо мало. Время точного поиска растет линейно и сравнивается с временем кодирования лишь при $N \approx 5 \cdot 10^5$. При d = = 1024 и формате float32 база такого объема занимает около 2 ГБ, а сами такие размеры возникают при переходе от карты отдельного района к региональной — там приближенный индекс и начинает по-настоящему окупаться.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Данные и метрики

Для основной серии экспериментов использовался набор UAV-VisLoc [20], регион 03 (Taizhou-1). Выборка содержала все 768 бортовых снимков региона, полученных с высоты около 466 м размером 3976 × 2652 пикселей. В качестве картографической основы использовался снимок размером 35092 × 24308 пикселей с пространственным разрешением около 0.3 м/пиксель. Карта разбивалась на квадратные фрагменты со стороной 256 м. Шаг сетки составлял 128 м, что соответствовало перекрытию 50 %. Всего было получено 3685 фрагментов размером 512 × 512 пикселей с пространственным разрешением 0.5 м/пиксель. Размер фрагмента выбирался так, чтобы приблизительно соответствовать площади местности, наблюдаемой бортовой камерой.

Для оценки влияния источника картографических данных аналогичная база была построена по снимкам Esri World Imagery [21]. Размер фрагментов, шаг сетки и пространственное разрешение (0.5 м/пиксель, зум 18) сохранялись такими же, что позволяло сравнивать источники при одинаковом масштабе картографических изображений.

На вход признакового кодировщика изображения подавались с разрешением 518 × 518 пикселей. Поэтому картографические фрагменты практически не изменялись по размеру, тогда как бортовые кадры существенно уменьшались. Влияние изменения разрешения на качество отбора отдельно исследовалось. Эталонное множество G_q формировалось по истинному положению БПЛА в момент съемки. В него включались фрагменты, центры которых находились на расстоянии не более 128 м от истинной позиции. Такой допуск учитывал положение наблюдаемого участка относительно границ сетки и гарантировал наличие хотя бы одного соответствующего фрагмента.

Рост координатной неопределенности моделировался изменением числа кандидатов $|C|$. В множество C включались ближайшие к истинному положению фрагменты, что соответствовало последовательному расширению области поиска. Для характеристики размера этой области использовалась эквивалентная площадь, равная произведению числа кандидатов на площадь одной ячейки сетки:

$$S_c = |C|\,(0.128\,\text{км})^2 \approx |C| \cdot 0.0164\,\text{км}^2. \qquad (7)$$

Так, $|C| = 50$ отвечает площади около 0.8 км^2, а $|C| = 200$ — уже 32.8 км^2.

Качество отбора измерялось долей запросов, для которых верный фрагмент попал в первые k позиций:

$$\text{recall}@k = \frac{1}{Q}\sum_{q=1}^{Q} 1\left[\exists\, j \le k :\ i_j^{(q)} \in G_q\right], \qquad (8)$$

где Q — число запросов, $i_j^{(q)}$ — кандидат с рангом j для запроса q, а $1[\cdot]$ — индикатор условия. На данном этапе оценивалась только полнота предварительного отбора. Точность координат после геометрической проверки в этот эксперимент не включалась.

При этом $\text{recall}@k$ — метрика промежуточная, которая задает лишь верхнюю границу доли успешных повторных локализаций. Попадание географически верного фрагмента в K_t необходимо для координатного измерения, но не гарантирует его — модель сопоставления может не набрать достаточного числа согласованных соответствий.

Расширение области поиска

Для проверки устойчивости отбора к росту координатной неопределенности, число допустимых кандидатов $|C|$ увеличивали от 50 до 2000 — эквивалентная площадь поиска при этом расширяется с 0.8 до 32.8 км^2. Чем больше $|C|$, тем больше визуально сходных фрагментов попадает в выборку и тем труднее верному участку удержаться в первых позициях ранжированного списка. Результаты приведены в табл. 1 и на рис. 3.

При увеличении числа кандидатов с 50 до 2000 значение recall@10 снизилось с 0.97 до 0.71.

Табл. 1. Полнота отбора на UAV-VisLoc при использовании карты из состава набора данных

Площадь, км^2	$\lvert C\rvert$	recall@ 3	recall@ 5	recall@ 8	recall@10	recall@15	recall@30
0.8	50	0.78	0.90	0.95	0.97	0.98	1.00
1.6	100	0.70	0.81	0.89	0.92	0.95	0.99
3.3	200	0.66	0.74	0.82	0.85	0.89	0.95
6.6	400	0.62	0.69	0.76	0.78	0.83	0.90
9.8	600	0.59	0.66	0.73	0.76	0.81	0.87
16.4	1000	0.57	0.64	0.70	0.74	0.78	0.84
32.8	2000	0.55	0.62	0.67	0.71	0.76	0.82

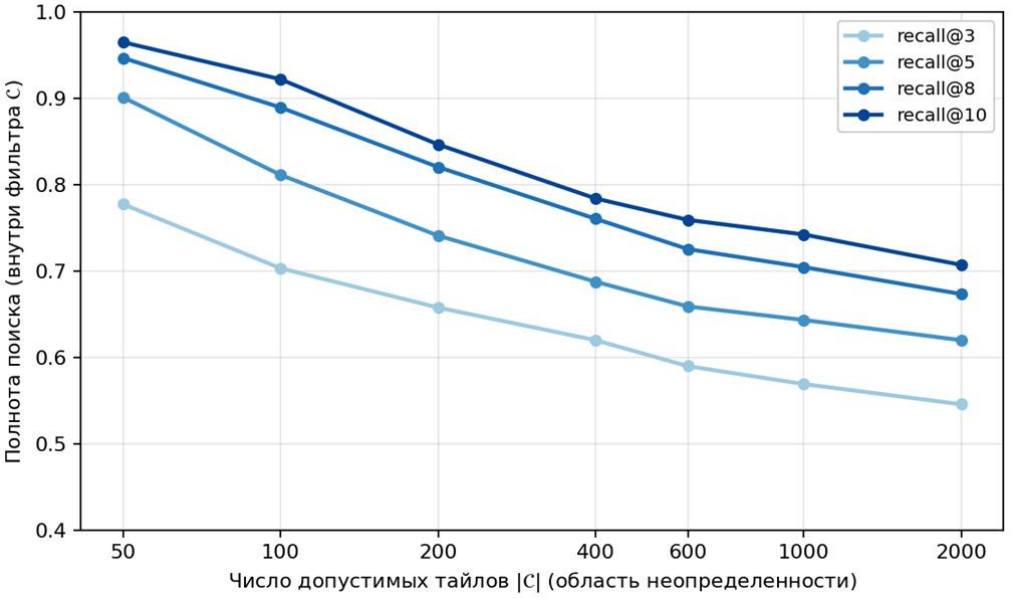

Рис. 3. Зависимость полноты отбора от размера области поиска

Увеличение числа отбираемых фрагментов частично компенсирует это снижение: при $\lvert C\rvert \geq 200$ переход от $k = 10$ к $k = 30$ повышает полноту на 0.10–0.12. При этом для $k = 10$ число запусков полной модели сопоставления сокращается в 5 раз при $\lvert C\rvert = 50$ и в 200 раз при $\lvert C\rvert = 2000$.

Характерные примеры ранжирования показаны на рис. 4. Для каждого бортового кадра приведены три первых кандидата и расстояния от их центров до истинного положения БПЛА.

При наличии выраженных линейных объектов, таких как дороги и каналы, верный фрагмент обычно занимал первую позицию. Ошибки ранжирования чаще возникали в районах однотипной жилой застройки, где соседние кварталы имеют сходную планировку. Там верный фрагмент мог оказаться на второй или третьей позиции — это и оправдывает передачу модели сопоставления не одного, а нескольких первых кандидатов.

Влияние источника карты

Для оценки влияние картографического источника базу заново построили по изображениям Esri World Imagery на ту же территорию, снятым в другие даты, нежели бортовая съемка. Результаты приведены в табл. 2.

На карте Esri World Imagery полнота оказалась ниже на всем исследованном диапазоне: разность по recall@10 составила 0.10 при $\lvert C\rvert = 50$, 0.12 при $\lvert C\rvert = 400$ и 0.18 при $\lvert C\rvert = 2000$. Такое снижение объяснимо: снимки сделаны в другой сезон, поэтому растительность и поля выглядят иначе, местами изменилась застройка, отличается и цветопередача.

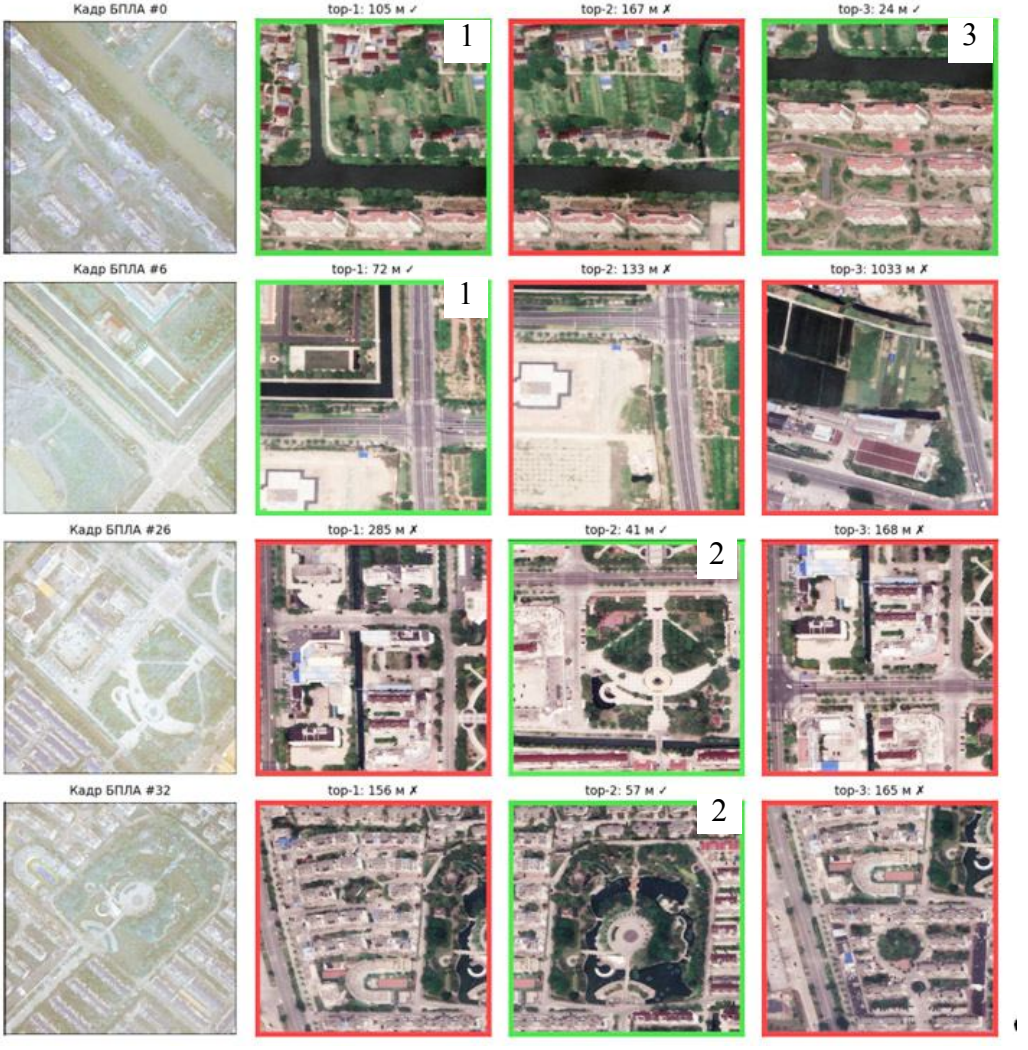

Рис. 4. Отбор фрагментов карты. Слева — кадр, справа — три кандидата: помеченные номером — верные, остальные — ошибочные

Табл. 2. Полнота отбора на UAV-VisLoc, карта Esri World Imagery

Площадь, км^2	$\|C\|$	recall@3	recall@5	recall@8	recall@10	recall@15	recall@30
0.8	50	0.63	0.76	0.84	0.87	0.92	0.97
1.6	100	0.58	0.66	0.76	0.78	0.84	0.91
3.3	200	0.51	0.59	0.68	0.72	0.77	0.84
6.6	400	0.47	0.54	0.63	0.66	0.71	0.79
9.8	600	0.44	0.52	0.59	0.62	0.68	0.76
16.4	1000	0.41	0.48	0.54	0.58	0.64	0.72
32.8	2000	0.37	0.45	0.50	0.53	0.59	0.69

При самой большой области поиска верный фрагмент попал в первую десятку кандидатов лишь для 53 % запросов.

Выбор способа агрегирования и сравнение с исходной моделью

Для оценки влияния способа агрегирования карты признаков на качество предварительного отбора сравнивались несколько вариантов построения глобального дескриптора. Во всех экспериментах сохранялись одинаковыми разрешение входного изображения 518 × 518 пикселей, размерность дескриптора d = 1024, L2-нормировка и протокол оценки.

Рассматривалось обобщенное усреднение (3) при p = 1, 2, 3 и 4, а также агрегирование по максимальному значению в каждой компоненте карты признаков. При p = 1 обобщенное усреднение соответствует обычному усреднению по пространственным позициям.

Дополнительно сравнивались дескрипторы, построенные по признакам кодировщика DINOv2 ViT-L/14 в составе матчера и по признакам исходной модели DINOv2. Поскольку кодировщик в составе матчера заморожен, такое сравнение позволяет проверить, изменяются ли его выходные признаки при использовании в основном конвейере сопоставления.

Для одного изображения все варианты агрегирования вычислялись по одной и той же карте признаков, поэтому повторный запуск кодировщика для каждого значения p не требовался. Результаты приведены в табл. 3.

Обобщенное усреднение при $p = 3$ и $p = 4$ обеспечило более высокую полноту отбора, чем обычное усреднение и максимум. Наилучшие результаты получены при $p = 4$, однако его преимущество над $p = 3$ не превышало 1.7 процентного пункта. Дескрипторы, построенные по признакам кодировщика в составе матчера и исходной модели DINOv2, дали практически одинаковые результаты. Следовательно, использование замороженного кодировщика не повышает качество отбора, но позволяет отказаться от отдельной модели построения глобальных дескрипторов.

Устойчивость дескриптора к искажениям

Устойчивость дескриптора оценивалась с помощью синтетических искажений картографических фрагментов. К исходным изображениям применялись уменьшение разрешения, перевод в градации серого, поворот, гамма-коррекция, размытие, JPEG-сжатие и кадрирование с последующим восстановлением исходного размера. Эксперимент проводился на карте района летных испытаний вертолета-носителя. База включала 1172 фрагмента, из которых 83 относились к району испытаний, а остальные 1089 использовались как отвлекающие кандидаты. Всего было сформировано 150 запросов. Правильными считались исходный фрагмент и его соседи, расположенные в радиусе 80 м. Результаты приведены в табл. 4.

Табл. 3. Значения recall@10 при различных способах агрегирования карты признаков

Способ агрегирования	Карта набора данных		Esri World Imagery	
	$\|C\| = 200$	$\|C\| = 2000$	$\|C\| = 200$	$\|C\| = 2000$
Среднее, p = 1	0.835	0.652	0.688	0.495
GeM, p = 2	0.850	0.686	0.704	0.514
GeM, p = 3	0.845	0.707	0.717	0.526
GeM, p = 4	0.850	0.719	0.720	0.543

Табл. 4. Влияние синтетических искажений на полноту отбора

Преобразование	recall@1	recall@5	recall@10
Без преобразований	1.000	1.000	1.000
Размытие, 7 × 7	1.000	1.000	1.000
Гамма-коррекция	1.000	1.000	1.000
Поворот на 15°	0.947	0.987	1.000
Градации серого	0.933	0.980	0.987
JPEG-сжатие	0.927	0.980	0.993
Уменьшение в 4 раза	0.920	0.980	1.000
Кадрирование, 0.7	0.807	0.940	0.953
Уменьшение в 8 раз	0.693	0.853	0.893

Такой эксперимент позволяет сравнить влияние отдельных искажений, но не воспроизводит условия сопоставления бортовых и картографических изображений. Запрос и база получены из одного источника, поэтому различия, связанные с сезоном, условиями съемки и характеристиками сенсоров, отсутствуют. Размытие и гамма-коррекция не повлияли на результат. При повороте, JPEG-сжатии, переводе в градации серого и четырехкратном уменьшении разрешения значение recall@10 осталось не ниже 0.987. Наиболее заметное снижение наблюдалось при сильном изменении разрешения. После уменьшения изображения в восемь раз значения recall@1 и recall@10 составили соответственно 0.693 и 0.893. Это показывает, что при подготовке картографической базы необходимо учитывать пространственное разрешение бортовых изображений.

Вычислительные затраты

Среднее время построения дескриптора одного изображения составило около 92 мс. Измерения проводились на видеокарте NVIDIA RTX 3090 для изображений размером 518 × 518 пикселей при пакетной обработке по 8 изображений. Время относится только к признаковому кодировщику DINOv2 ViT-L/14 без декодера сопоставления.

Подготовка базы из 3685 картографических фрагментов заняла около шести минут. Эта операция выполняется однократно до вылета. Один дескриптор (d = 1024, формат float32) весит 4 КБ, так что вся база из 3685 векторов умещается примерно в 14.4 МБ — без учета служебных структур индекса.

Время поиска и влияние приближенного индексирования на полноту измерялись на полной базе из N = 3685 дескрипторов — это худший случай, когда априорной оценки положения нет и множество C_t охватывает всю базу. Поиск шел через FAISS в один поток, время усреднялось по 768 запросам. Результаты приведены в табл. 5.

С выбранными параметрами оба приближенных индекса не потеряли в полноте: значения recall@3, recall@5 и recall@10 у IVF и HNSW совпали с точным перебором. По времени запроса IVF оказался быстрее в 4.3 раза, а HNSW — в 9.3 раза. Правда, на базе такого размера сам поиск и без того дешев сравнительно с построением дескриптора: 0.650 мс на точный перебор против $t_{\text{кодир}} \approx 92$ мс на кодировщик. Поэтому на общем времени обработки запроса выбор индекса почти не сказывается; ощутимым выигрыш от приближенного поиска станет лишь на заметно большей базе, а этот режим в работе экспериментально не изучался.

Оценка времени обработки

Пусть $t_{\text{М}}$ обозначает время сопоставления бортового кадра с одним фрагментом карты. Тогда время полного перебора оценивается как

$$T_{\text{полн}} = |\text{C}|\, t_{\text{М}}. \tag{9}$$

Для предлагаемой двухэтапной схемы время обработки составляет

$$T_{\text{двух}} = t_{\text{кодир}} + t_{\text{поиск}} + k \cdot t_{\text{м}}, \tag{10}$$

где $t_{\text{кодир}}$ — время построения дескриптора запроса, $t_{\text{поиск}}$ — время поиска по базе, а k — число фрагментов, передаваемых модели сопоставления.

Расчет выполнен для видеокарты NVIDIA RTX 3090. Время сопоставления одной пары изображений размером 512 × 512 пикселей принято равным $t_{\text{м}}$ = 95 мс. По результатам предыдущих измерений, $t_{\text{кодир}} = 92$ мс, а $t_{\text{поиск}} \le 0.1$ мс. Для каждого значения $|\text{C}|$ выбиралось минимальное из исследованных значений k, при котором $\text{recall@}k \ge 80\%$. При $|\text{C}| = 2000$ этому условию соответствует k = 30 и recall@30 = 0.82 (табл. 1). Полный перебор в этом случае занимает $2000 \cdot 0.095 = 190$ с, а предлагаемая двухэтапная схема — $0.092 + 0.0001 + 30 \cdot 0.095 \approx 2.9$ с.

Табл. 5. Полнота и время поиска по базе из 3685 дескрипторов

Способ поиска	recall@3	recall@5	recall@10	Время, мс
Точный (IndexFlatIP)	0.530	0.594	0.686	0.650
IVF (IndexIVFFlat)	0.530	0.594	0.686	0.150
HNSW (IndexHNSWFlat)	0.530	0.594	0.686	0.070

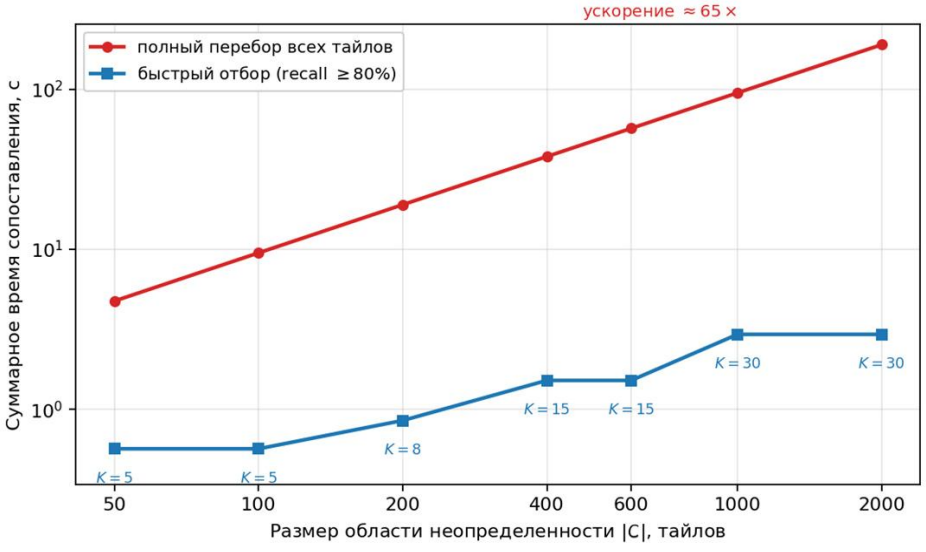

Рис. 5. Расчетное время полного перебора и двухэтапного выбора фрагментов карты

Расчетное ускорение составляет около 65 раз. Зависимость времени обработки от числа кандидатов приведена на рис. 5.

Отношение времени полного перебора ко времени двухэтапной обработки равно

$$\frac{T_{\text{полн}}}{T_{\text{двух}}} = \frac{|\mathrm{C}|\,t_{\mathrm{M}}}{t_{\text{кодир}} + t_{\text{поиск}} + k \cdot t_{\mathrm{M}}} = \frac{|\mathrm{C}|\,t_{\mathrm{M}}}{t_{\text{кодир}} / t_{\mathrm{M}} + t_{\text{поиск}} / t_{\mathrm{M}} + k}. \qquad (11)$$

Приведенная оценка относится к карте из состава UAV-VisLoc. Для Esri World Imagery при том же $k = 30$ расчетное время остается равным 2.9 с, однако полнота снижается до recall@30 = = 0.69 (табл. 2). Следовательно, значение k должно выбираться с учетом картографического источника и требуемой полноты отбора.

ЗАКЛЮЧЕНИЕ

Предложен метод предварительного отбора фрагментов геопривязанной карты для визуальной навигации БПЛА при значительной неопределенности положения. Кандидаты сначала ранжируются по глобальным дескрипторам, после чего ограниченное число лучших фрагментов передается полной модели сопоставления для геометрической проверки.

Глобальный дескриптор строится по промежуточным признакам замороженного кодировщика DINOv2 ViT-L/14, входящего в состав модели MINIMA. Поэтому предварительный отбор не требует применения отдельной нейросетевой модели. Сравнение с исходной DINOv2 показало практически одинаковую полноту отбора. Наилучшие результаты среди исследованных способов агрегирования получены при использовании обобщенного среднего с p, равным от 3 до 4.

На наборе UAV-VisLoc при увеличении эквивалентной площади поиска до 32.8 км^2 значение recall@10 составило 0.71, а recall@30 — 0.82. При $|\mathrm{C}| = 2000$ и $k = 30$ расчетное время обработки сократилось со 190 до 2.9 с, что соответствует ускорению примерно в 65 раз по сравнению с полным перебором. На полной базе из 3685 дескрипторов оба приближенных индекса сохранили полноту точного поиска, сократив время запроса с 0.650 мс до 0.150 мс (IVF) и 0.070 мс (HNSW). На таком объеме поиск и так намного дешевле построения дескриптора, поэтому по-настоящему приближенный индекс окупится лишь на существенно большей карте.

Переход на карту Esri World Imagery снизил полноту от 10 до 18% относительно карты из состава набора данных: при $|\mathrm{C}| = 2000$ и $k = 30$ значение recall@30 упало до 0.69. Значит, число отбираемых кандидатов приходится подбирать под конкретный источник карты и нужный уровень полноты.

В диагностическом эксперименте с синтетическими искажениями сильнее всего качество падало при заметном снижении разрешения кадра. Чаще ошибался отбор и в районах однотипной застройки, где соседние фрагменты похожи по пространственной структуре.

В целом метод дает высокую полноту отбора и при этом резко сокращает вычислительные затраты, что важно для систем визуальной навигации БПЛА, работающих с периодической коррекцией положения.

В дальнейшем метод предстоит проверить на базе регионального масштаба, провести сквозной эксперимент с геометрической проверкой и оценкой координатной ошибки, а также измерить время обработки на бортовом вычислителе.

СПИСОК ИСТОЧНИКОВ

1. *Gyagenda N. et al.* A review of GNSS-independent UAV navigation techniques // Robotics and Autonomous Systems. 2022. Vol. 152. Id. 104069. DOI: 10.1016/j.robot.2022.104069
2. *Choudhary G. et al.* Internet of drones (iod): Threats, vulnerability, and security perspectives // arXiv, preprint arXiv:1808.00203. 2018. P. 1–13. URL: https://arxiv.org/abs/1808.00203
3. *Murhij Y. et al.* Visual localization system for GPS-blind environments in unmanned aerial vehicles / // Proceedings of the 2024 8th International Conference on Information, Control, and Communication Technologies (ICCT). 2024. P. 431–435. DOI: 10.1109/ICCT62929.2024.10875019
4. *De Tone D., Malisiewicz T., Rabinovich A.* SuperPoint: Self-supervised interest point detection and description // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018. P. 224–236. DOI: 10.1109/CVPRW.2018.00060
5. *Sarlin P.-E. et al.* SuperGlue: Learning feature matching with graph neural networks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 4938–4947. DOI: 10.1109/CVPR42600.2020.00499
6. *Lindenberger P., Sarlin P.-E., Pollefeys M.* LightGlue: Local feature matching at light speed // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 17627–17638. DOI: 10.1109/ICCV51070.2023.01616
7. *Sun J. et al.* LoFTR: Detector-free local feature matching with transformers // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 8922–8931. DOI: 10.1109/CVPR46437.2021.00881
8. *Wang Y. et al.* Efficient LoFTR: Semi-dense local feature matching with sparse-like speed // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 21666–21675. URL: https://ieeexplore.ieee.org/abstract/document/10655054
9. *Edstedt J. et al.* RoMa: Robust dense feature matching // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 19790–19800. DOI: 10.1109/CVPR52733.2024.01871
10. *Arandjelović R. et al.* NetVLAD: CNN architecture for weakly supervised place recognition // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 5297–5307. DOI: 10.1109/CVPR.2016.572
11. *Radenović F., Tolias G., Chum O.* Fine-tuning CNN image retrieval with no human annotation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. Vol. 41, no. 7. P. 1655–1668. DOI: 10.1109/TPAMI.2018.2846566
12. *Berton G., Masone C., Caputo B.* Rethinking visual geo-localization for large-scale applications // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 4878–4888. DOI: 10.1109/CVPR52688.2022.00483
13. *Jégou H., Douze M., Schmid C.* Product quantization for nearest neighbor search // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011. Vol. 33, no. 1. P. 117–128. DOI: 10.1109/TPAMI.2010.57
14. *Muja M., Lowe D.G.* Scalable nearest neighbor algorithms for high dimensional data // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014. Vol. 36, no. 11. P. 2227–2240. DOI: 10.1109/TPAMI.2014.2321376
15. *Malkov Yu.A., Yashunin D.A.* Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020. Vol. 42, no. 4. P. 824–836. DOI: 10.1109/TPAMI.2018.2889473
16. *Johnson J., Douze M., Jégou H.* Billion-scale similarity search with GPUs // IEEE Transactions on Big Data. 2021. Vol. 7, no. 3. P. 535–547. DOI: 10.1109/TBDATA.2019.2921572
17. *Фатеев А.С. и др.* Структурно-ориентированный предварительный отбор видеокадров беспилотного летательного аппарата для задач кросс-модального сопоставления со спутниковыми изображениями // Научное приборостроение. 2026. Т. 36, № 2. С. 103–120. URL: http://iairas.ru/mag/2026/abst2.php#abst11
18. *Ren J. et al.* MINIMA: Modality invariant image matching. arXiv:2412.19412 [cs.CV], 2025. URL: https://www.researchgate.net/publication/387512339_MINIMA_Modality_Invariant_Image_Matching
19. *Oquab M. et al.* DINOv2: Learning robust visual features without supervision // Transactions on Machine Learning Research, 2024. URL: https://www.researchgate.net/publication/370058767_DINOv2_Learning_Robust_Visual_Features_without_Supervision
20. *Xu W. et al.* UAV-VisLoc: A large-scale dataset for UAV visual localization. arXiv:2405.11936 [cs.CV], 2024. URL: https://github.com/IntelliSensing/UAV-VisLoc
21. Esri. World Imagery, 2024. URL: https://www.esri.com/ru-ru/arcgis/products/data-location-services/data/basemaps-imagery

NAUCHNOE PRIBOROSTROENIE, 2026, Vol. 36, No. 4, pp. 105–117 *ISSN 0868-5886*

Московский физико-технический институт (национальный исследовательский университет), Москва

Контакты: *Татаринова Елена Александровна*, tatarinova.ea@mipt.ru

Материал поступил в редакцию 16.08.2026

PRELIMINARY SELECTION OF GEOREFERENCED MAP FRAGMENTS USING GLOBAL DESCRIPTORS IN A UAV VISUAL NAVIGATION SYSTEM

A. S. Fateev, E. L. Kiselev, M. V. Terentyev, E. V. Bokshansky, N. V. Saushkin, E. A. Tatarinova, D. N. Shchelkunov, S. O. Russkin

Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Russia

The paper addresses the problem of preliminary selection of georeferenced map fragments for visual navigation of an unmanned aerial vehicle (UAV) under growing position uncertainty following prolonged absence of global corrections. A two-stage scheme is proposed, in which candidates are first ranked by global descriptors and then verified using a full matching model. The descriptor is formed by aggregating the intermediate feature map of a frozen DINOv2 encoder used within the matching model, thereby eliminating the need for a separate neural network model. For search, we compare exact indexing with approximate IVF and HNSW indices. Experiments on the UAV-VisLoc dataset and the Esri World Imagery map show that approximate search preserves the recall of exhaustive search with negligible query time, while preliminary selection substantially reduces the number of matching model invocations. Additionally, the effects of search area size, map data source, and input image distortions are investigated.

Keywords: unmanned aerial vehicles, visual navigation, satellite image matching, cross-modal matching, approximate nearest neighbor search, global image descriptors, preliminary map fragment selection, georeferenced imagery

REFERENCES

1. Gyagenda N. et al. A review of GNSS-independent UAV navigation techniques. *Robotics and Autonomous Systems*, 2022, vol. 152, id. 104069. DOI: 10.1016/j.robot.2022.104069
2. Choudhary G. et al. Internet of drones (iod): Threats, vulnerability, and security perspectives. *arXiv, preprint arXiv:1808.00203*, 2018, pp. 1–13. URL: https://arxiv.org/abs/1808.00203
3. Murhij Y. et al. Visual localization system for GPS-blind environments in unmanned aerial vehicles. *Proceedings of the 2024 8th International Conference on Information, Control, and Communication Technologies (ICCT)*, 2024, pp. 431–435. DOI: 10.1109/ICCT62929.2024.10875019
4. De Tone D., Malisiewicz T., Rabinovich A. SuperPoint: Self-supervised interest point detection and description. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 224–236. DOI: 10.1109/CVPRW.2018.00060
5. Sarlin P.-E. et al. SuperGlue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947. DOI: 10.1109/CVPR42600.2020.00499
6. Lindenberger P., Sarlin P.-E., Pollefeys M. LightGlue: Local feature matching at light speed. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17627–17638. DOI: 10.1109/ICCV51070.2023.01616
7. Sun J. et al. LoFTR: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8922–8931. DOI: 10.1109/CVPR46437.2021.00881
8. Wang Y. et al. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21666–21675. URL: https://ieeexplore.ieee.org/abstract/document/10655054
9. Edstedt J. et al. RoMa: Robust dense feature matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024,

pp. 19790–19800. DOI: 10.1109/CVPR52733.2024.01871
10. Arandjelović R. et al. NetVLAD: CNN architecture for weakly supervised place recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307. DOI: 10.1109/CVPR.2016.572
11. Radenović F., Tolias G., Chum O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, vol. 41, no. 7, pp. 1655–1668. DOI: 10.1109/TPAMI.2018.2846566
12. Berton G., Masone C., Caputo B. Rethinking visual geo-localization for large-scale applications. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4878–4888. DOI: 10.1109/CVPR52688.2022.00483
13. Jégou H., Douze M., Schmid C. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, vol. 33, no. 1, pp. 117–128. DOI: 10.1109/TPAMI.2010.57
14. Muja M., Lowe D.G. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, vol. 36, no. 11, pp. 2227–2240. DOI: 10.1109/TPAMI.2014.2321376
15. Malkov Yu.A., Yashunin D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, vol. 42, no. 4, pp. 824–836. DOI: 10.1109/TPAMI.2018.2889473
16. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2021, vol. 7, no. 3, pp. 535–547. DOI: 10.1109/TBDATA.2019.2921572
17. Fateev A.S. et al. Structure-based preselection of unmanned aerial vehicle video frames for cross-modal matching with satellite imagery. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2026, vol. 36, no. 2, pp. 103–120. (In Russ.). URL: http://iairas.ru/mag/2026/abst2.php#abst11
18. Ren J. et al. MINIMA: Modality invariant image matching. *arXiv:2412.19412 [cs.CV]*, 2025. URL: https://www.researchgate.net/publication/387512339_MINIMA_Modality_Invariant_Image_Matching
19. Oquab M. et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL: https://www.researchgate.net/publication/370058767_DINOv2_Learning_Robust_Visual_Features_without_Supervision
20. Xu W. et al. UAV-VisLoc: A large-scale dataset for UAV visual localization. *arXiv:2405.11936 [cs.CV]*, 2024. URL: https://github.com/IntelliSensing/UAV-VisLoc
21. *Esri. World Imagery*, 2024. URL: https://www.esri.com/ru-ru/arcgis/products/data-location-services/data/basemaps-imagery

Contacts: *Tatarinova Elena Alexandrovna,*
tatarinova.ea@mipt.ru

The article was received by the editorial office on 16.08.2026

INTRODUCTION

In modern unmanned aerial vehicle (UAV) navigation systems, position is estimated using data from multiple, complementary sources. Inertial and other local navigation systems provide high update rates but their errors accumulate over time. To limit this error, periodic global corrections are used, obtained from the global navigation satellite system (GNSS) or by visually matching frames with a georeferenced map [1–3].

When performing visual matching, an onboard camera frame is compared with a satellite or aerial image. Based on the matches found, a geometric transformation is estimated and the position of the observed area is determined. This measurement is directly linked to the map, so its error does not accumulate from frame to frame.

However, before matching, suitable map fragments must be selected. In normal mode, the search area is bounded by the navigation filter prediction and contains a small number of fragments. In the absence of global correction for a long time, the uncertainty of position increases, and the search space can cover thousands of fragments. Since its area increases proportionally to the square of the radius, exhaustive search quickly becomes computationally infeasible. Modern matching models use detectable keypoints [4–6] or construct semi-dense and dense correspondence fields [7–9]. Processing a single pair of images on an onboard computer can take from hundreds of milliseconds to several seconds. Therefore, sequentially checking thousands of fragments does not ensure the timely formation of a coordinate measurement. Therefore, before running the full matching model, the number of candidates must first be reduced.

This paper proposes a two-stage method for selecting georeferenced map fragments. First, candidates are ranked by global descriptors used in terrain recognition problems [10–12], after which the best of them are passed to the full matching model for geometric verification. The descriptor database is searched using approximate nearest neighbor (ANN) search methods [13–16].

To avoid using a separate onboard global description model, the descriptor is formed from the intermediate feature map of the matching model. Thus, the pre-selection and subsequent geometric verification use a single feature encoder.

This paper addresses the following objectives:

1) a two-stage procedure for selecting map fragments was developed, combining spatial restriction of the search area and ranking by global descriptors;

2) the construction of a global descriptor by aggregating the intermediate feature map of the matching model was investigated;

3) the influence of the search area size, the source of cartographic data, and input image distortions on the selection recall was assessed, and the computational costs of the method were measured.

The proposed method complements a previously developed algorithm for filtering a video stream based on frame information content [17]. Filtering reduces the number of frames processed, and the method under consideration reduces the number of map fragments checked.

PROBLEM STATEMENT

Let the onboard database contain N map fragments. For each fragment, the coordinates of the center c_i and a pre-computed descriptor are specified. For the incoming frame I_q, it is necessary to select fragments with which a full comparison will be performed.

The search area is limited by the navigation system data. Based on the position estimate $\hat{p}_t$ and the uncertainty radius R_t, a set of acceptable candidates is formed:

$$\mathrm{C}_t = \left\{ i : \mathrm{dist}\left(c_i, \hat{p}_t\right) \le R_t \right\}, \qquad (1)$$

where $\mathrm{dist}(\cdot,\cdot)$ denotes the distance on the ground. In the absence of a reliable a priori estimate, the set C_t coincides with the entire image database.

A geometric interpretation of the spatial constraint is shown in Fig. 1. The central circle shows the region of possible UAV position with a radius R_t = 8 km and the bounding square with a side of $2R_t$ = 16 km, used for the initial spatial query. The 2-km grid is included for scale purposes only. With an actual fragment size of 256 × 256 m and a map division step of 128 m, each depicted cell corresponds to approximately 250 fragments.

Fig. 1. An example of a search area (limited by a circle) after the accumulation of coordinate uncertainty

As the coordinate error accumulates, the radius R_t increases, so more and more map fragments are included in the search region. For R_t = 8 km, the area of this region is approximately 200 km^2, corresponding to approximately 12×10^3 fragments.

When $|\mathrm{C}_t|$ is small, a full matching can be performed for all valid fragments. If there are many candidates, a smaller subset K_t is selected from C_t. The ratio shows how much it reduces the number of matching model runs.

$$\rho = \frac{|C_t|}{|K_t|}, \quad (2)$$

here, the numerator and denominator denote the cardinalities of the sets.

The decision to enable preliminary selection is based on the current size of the uncertainty region, not on individual sensor readings. While the global reference remains accurate, the system iterates all fragments from C_t**.** As soon as a correction is delayed and $|C_t|$ increases, a descriptor search is activated.

The pre-selection is required to ensure that a geographically correct fragment is included in K_t with a strict limitation on its size and low computational costs.

METHOD

Characteristics of the method

The proposed method is based on a two-stage selection of georeferenced map fragments. The first stage quickly ranks candidates based on the similarity of their global descriptors, and the second stage performs a full comparison of the onboard frame with only a limited number of fragments that received the highest similarity scores. This separation allows for the elimination of obviously unsuitable map areas without running a computationally complex matching model for each candidate.

Computations are distributed between ground preparation and in-flight operations. Before takeoff, a database of cartographic fragments is compiled, with a global descriptor and geographic coordinates calculated in advance for each. In flight, the same descriptor is constructed for the current frame of the onboard camera, after which the closest elements are searched for in the database. The selected fragments are fed to the main matching model to establish local correspondences and subsequent geometric verification. The general scheme of the method is shown in Fig. 2.

Fig. 2. Accelerated selection scheme for map fragments

Database preparation

The original map image is divided into square fragments of equal size with a specified overlap. Using overlap reduces the dependence of the result on the position of the observed region relative to the grid boundaries and increases the probability that it will fall entirely or predominantly within at least one fragment.

Each fragment T_i passes through the feature encoder of the matching model, yielding a feature map $F(T_i)$. The spatial dimension is reduced by generalized averaging [11]

$$\mathrm{Pool}(F)_c = \left(\frac{1}{|\Omega|} \sum_{u \in \Omega} F_c(u)^p \right)^{1/p}, \quad (3)$$

where c is the channel number, Ω is the set of spatial positions, and p is the generalized averaging parameter. The resulting vector is normalized according to the Euclidean norm:

$$v_i = \frac{\mathrm{Pool}(F(T_i))}{\left\| \mathrm{Pool}(F(T_i)) \right\|_2}. \quad (4)$$

After L2 normalization, the scalar product of two descriptors is equal to the cosine of the angle between them. Therefore, the proximity of a fragment to the query is conveniently measured by $q^{\mathrm{T}} v_i$ and the fragments are ranked in descending order.

Pairs $\{(v_i, c_i)\}_{i=1}^{N}$ form an onboard descriptor database. The search for fragments closest to the query is performed using an approximate nearest neighbor search index [14, 15]. Without an a priori position estimate, the search must be performed across the entire database, and here the index is particularly valuable. It quickly identifies a small set of likely candidates without comparing the query with each descriptor. An approximate search can bias the selection of top candidates compared to an exact iteration, so its impact on the recall of the selection is assessed separately. The correspondence of a fragment to the current frame is confirmed only after a full comparison and geometric verification.

In-flight search

For frame I_q, the query descriptor is calculated using the same encoder and aggregation method.

$$q = \frac{\mathrm{Pool}(F(I_q))}{\left\| \mathrm{Pool}(I_q) \right\|_2}. \quad (5)$$

From the set that passed spatial constraint (1), k vectors with the largest scalar product value are selected:

$$K_t = \text{Top-}K_{i \in C_t} \left(q^{\mathrm{T}}, v_i \right). \quad (6)$$

For each selected fragment, the matching model constructs correspondences and evaluates the transformation between the frame and the map. Inconsistent matches are rejected during geometric verification. If no candidate satisfies the specified criteria,

a coordinate measurement for the current frame is not generated.

Method implementation

The experiments used the RoMa model [9] in its modal-invariant modification MINIMA [18]. To construct the descriptor, only the feature encoder was implemented, without a matching decoder. DINOv2 ViT-L/14 patch tokens [19] with a spatial step of 14 pixels and 1024 channels were used as $F(\ldots)$.

The DINOv2 encoder in RoMa is frozen and does not change during matcher retraining. Therefore, the resulting features match those of the original DINOv2. The advantage of reusing them is not an improvement in the quality of the descriptor, but rather the elimination of the need to employ a separate neural network model. This is also confirmed by the experiments presented below.

The input images were scaled to 518 × 518 pixels and normalized using ImageNet statistics. The parameter p in (3) was not trained and was set to 3.0. The descriptor dimension was d = 1024, and the size of one vector in float32 format was 4 KB. The FAISS library [16] was used for searching. An exact iteration of Index Flat IP served as a benchmark. Index IVFFlat with parameters $n_{\text{list}} = \left\lfloor \sqrt{N} \right\rfloor$ and $n_{\text{probe}} = 16,$, as well as Index HNSW Flat with parameters M = 32, efConstruction = 200, and efSearch = 64, were considered as approximate indices. Since all descriptors are normalized using the L2 norm, their similarity was calculated using the dot product.

When there is no a priori position estimate, the C_t set covers the entire database, and the search is performed directly by index. This mode — with a rough or lost initial approximation — is the primary scenario for the method. If the position estimate is known, the C_t set is narrowed in advance by a spatial query (1). The index is built over the entire database, so ranking within such a geographic subset is simpler directly — by calculating $|C_t|$ scalar products. For several thousand candidates, this takes a fraction of a millisecond and does not require referring to the index.

The gain of approximate search becomes noticeable as the database grows. For 3685 descriptors, exact search takes 0.650 ms — negligible compared to the time $t_{\text{кодир}} \approx 92$ ms required to construct the query descriptor. The exact search time grows linearly and becomes comparable to the encoding time only at $N \approx 5 \cdot 10^5$. With d = 1024 and float32 format, a database of this size occupies approximately 2 GB. Such sizes become noticeable in transition from a single-district map to a regional one — that's where the approximate index truly pays off.

EXPERIMENTAL SECTION

Data and metrics

For the main series of experiments, the UAV-VisLoc dataset [20], region 03 (Taizhou-1), was used. The sample contained all 768 airborne images of the region, acquired from an altitude of approximately 466 m, with dimensions of 3976 × 2652 pixels. A 35092 × 24308-pixel image with a spatial resolution of approximately 0.3 m/pixel was used as the cartographic base. The map was divided into square fragments with a side length of 256 m. The grid step was 128 m, corresponding to an overlap of 50%. A total of 3685 fragments were obtained, each measuring 512 × 512 pixels with a spatial resolution of 0.5 m/pixel. The fragment size was chosen to approximately correspond to the area of the terrain observed by the onboard camera.

To assess the influence of the cartographic data source, a similar database was constructed using Esri World Imagery [21]. Fragment size, grid step, and spatial resolution (0.5 m/pixel, zoom 18) were maintained constant, allowing for comparison of sources at the same scale of cartographic images.

Images with a resolution of 518 × 518 pixels were fed to the feature encoder. Therefore, the cartographic fragments remained virtually unchanged in size, while the onboard frames were significantly reduced in size. The impact of resolution changes on selection quality was separately studied. The reference set G_q was formed based on the UAV's true position at the time of acquisition. It included fragments with centers no longer than 128 meters from the true position. This tolerance took into account the position of the observed area relative to the grid boundaries and guaranteed the presence of at least one corresponding fragment.

The growth of coordinate uncertainty was modeled by varying the number of candidates $|C|$. Set C included the fragments closest to the true position; it corresponded to a consistent expansion of the search region. To characterize the size of this region, we used an equivalent area equal to the product of the number of candidates and the area of one grid cell:

$$S_c = |C|(0.128\,\text{км})^2 \approx |C| \cdot 0.0164\,\text{км}^2. \qquad (7)$$

Thus, $|C| = 50$ corresponds to an area of approximately 0.8 km^2, and $|C| = 200$ to 32.8 km^2.

The quality of the selection was measured by the proportion of queries for which the correct fragment was in the first k positions:

$$\text{recall}@k = \frac{1}{Q}\sum_{q=1}^{Q} 1\left[\exists\, j \le k:\ i_j^{(q)} \in G_q\right], \qquad (8)$$

where Q is the number of queries, $i_j^{(q)}$ is a candidate with rank j for query q, and $1[\cdot]$ is the condition indicator. At this stage, only the recall of the preliminary selection was assessed. The accuracy of the coordinates after geometric verification was not included in this experiment.

$\text{recall}@k$ is an intermediate metric, which only sets an upper bound on the proportion of successful repeat localizations. The inclusion of a geographically correct fragment in K_t is necessary for the coordinate measurement but does not guarantee it—the matching model may not accumulate a sufficient number of consistent matches.

Expanding the search area

To test the robustness of the selection to increasing coordinate uncertainty, the number of acceptable candidates $|C|$ was increased from 50 to 2000, as a result of which the equivalent search area expanded from 0.8 to 32.8 km². The larger the number $|C|$, the more visually similar fragments were included in the sample, and the more difficult it was for the correct site to remain at the top of the ranked list. The results are presented in Table 1 and Figure 3.

Tab. 1. Recall of selection on UAV-VisLoc when using a map from the dataset

Fig. 3. Dependence of selection recall on the size of the search area

As the number of candidates increased from 50 to 2000, the recall@10 value decreased from 0.97 to 0.71.

Increasing the number of fragments selected partially compensates for this reduction: if $|C| \geq 200$ moving from k = 10 to k = 30 increases recall by 0.10–0.12. Moreover, for k = 10, the number of runs of the full matching model is reduced by a factor of 5 for and by a factor of 200 for $|C| = 2000$.

Figure 4 shows typical ranking examples. For each onboard frame, the top three candidates and the distances from their centers to the UAV's true position are shown.

Fig. 4. Map fragment selection. A frame is on the left, three candidates are on the right: those marked with a number are correct, the rest are incorrect

In the presence of distinct linear features, such as roads and canals, the correct fragment usually occupied the first position. Ranking errors were more common in areas of uniform residential development, where adjacent blocks have similar layouts. There, the correct fragment could end up in the second or third position — this justifies transferring the matching model to not just one, but several top candidates.

Map source impact

To assess the impact of the map source, the database was rebuilt using Esri World Imagery images of the same area, taken on different dates than the airborne survey. The results are presented in Table 2.

Tab. 2. Selection recall on UAV-VisLoc, Esri World Imagery map

On the Esri World Imagery map, recall was lower across the entire studied range: the difference in recall@10 was 0.10 for $|C| = 50$, 0.12 for $|C| = 400$, and 0.18 for $|C| = 2000$. This decrease is understandable: the images were taken in a different season, so the vegetation and fields look different, buildings have changed in places, and the color rendering is different.

With the largest search space, the correct fragment made it into the top ten candidates for only 53% of queries.

Choosing an aggregation method and comparison with the original model

To assess the impact of the feature map aggregation method on the quality of preselection, several global descriptor construction options were compared. All experiments maintained the same input image resolution of 518 × 518 pixels, descriptor dimension d = 1024, L2 normalization, and evaluation protocol.

Generalized averaging (3) was considered for p = 1, 2, 3, and 4, as well as aggregation by the maximum value in each feature map component. For p = 1, generalized averaging corresponds to standard averaging over spatial positions.

Additionally, descriptors constructed using features of the DINOv2 ViT-L/14 encoder as part of the matcher and features of the original DINOv2 model were compared. Since the encoder in the matcher is frozen, this comparison allows us to check whether its output features change when used in the main matching pipeline.

For a single image, all aggregation variants were calculated using the same feature map, so rerunning the encoder for each p value was not required. The results are shown in Table 3.

Tab. 3. Recall@10 values for different methods of feature map aggregation

Generalized averaging with $p = 3$ and $p = 4$ provided higher selection recall than the standard averaging and maximum. The best results were obtained with $p = 4$, however, its advantage over $p = 3$ did not exceed 1.7 percentage points. Descriptors constructed using the features of the encoder in the matcher and the original DINOv2 model yielded virtually identical results. Therefore, using a frozen encoder does not improve the quality of selection, but it eliminates the need for a separate model for constructing global descriptors.

Descriptor robustness to distortions

The robustness of the descriptor was assessed using synthetic distortions of cartographic fragments. The original images were subjected to resolution reduction, conversion to grayscale, rotation, gamma correction, blurring, JPEG compression, and cropping, followed by restoration to their original size. The experiment was conducted on a map of the carrier helicopter flight test area. The database included 1172 fragments, 83 of them were related to the test area, and the remaining 1089 were used as distractor candidates. A total of 150 queries were generated. The original fragment and its neighbors within an 80-meter radius were considered correct. The results are presented in Table 4.

Tab. 4. The influence of synthetic distortions on the selection recall

This experiment allows us to compare the impact of individual distortions but does not reproduce the conditions for comparing airborne and cartographic images. The query and database were obtained from the same source, so differences related to season, shooting conditions, and sensor characteristics are absent. Blurring and gamma correction did not affect the results. The recall@10 value remained at least 0.987 after rotation, JPEG compression, conversion to grayscale, and a fourfold reduction in resolution. The most noticeable reduction was observed after a significant change in resolution. After reducing the image by a factor of eight, the recall@1 and recall@10 values were 0.693 and 0.893, respectively. This demonstrates that the spatial resolution of airborne images must be taken into account when preparing a cartographic database.

Computational costs

The average time to construct a descriptor for a single image was approximately 92 ms. Measurements were conducted using an NVIDIA RTX 3090 graphics card for 518 x 518 pixel images with batch processing of 8 images. Time refers only to the DINOv2 ViT-L/14 feature encoder without a matching decoder.

Preparing a database of 3685 mapping fragments took approximately six minutes. This operation is performed once before the take-off. One descriptor (d = 1024, float32 format) is 4 KB, so the entire database of 3685 vectors fits in approximately 14.4 MB, excluding index overhead.

Search time and the impact of approximate indexing on recall were measured using the full database of N = 3685 descriptors — this is the worst case, when there is no a priori position estimate and the C_t set covers the entire database. The search was performed using FAISS in a single thread, and the time was averaged over 768 queries. The results are presented in Table 5.

Tab. 5. Recall and search time for a database of 3685 descriptors

With the chosen parameters, both approximate indices retained their recall: the recall@3, recall@5, and recall@10 values for IVF and HNSW matched those of exact iterating. In terms of query time, IVF was 4.3 times faster, while HNSW was 9.3 times faster. However, with a database of this size, the search itself is already inexpensive compared to descriptor construction: 0.650 ms for exact iterating versus $t_{\text{кодир}} \approx 92$ ms for the encoder. Therefore, the choice of index has almost no effect on the overall query processing time; the gain from approximate search will be noticeable only with a significantly larger database, and this mode was not experimentally studied in this paper.

Processing time estimate

Let $t_{\text{М}}$ denote the time to match an onboard frame with one map fragment. Then, the exhaustive iterating time is estimated as

$$T_{\text{полн}} = |\text{C}| t_{\text{М}}. \quad (9)$$

For the proposed two-stage scheme, the processing time is

$$T_{\text{двух}} = t_{\text{кодир}} + t_{\text{поиск}} + k \cdot t_{\text{м}}, \quad (10)$$

where $t_{\text{кодир}}$ is the query descriptor construction time,

$t_{\text{поиск}}$ is the database search time, and k is the number of fragments passed to the matching model.

The calculation was performed for an NVIDIA RTX 3090 graphics card. The matching time for one pair of 512 × 512 pixel images was assumed to be $t_{\text{м}}$ = 95 ms. Based on the results of previous measurements, $t_{\text{кодир}} = 92$ ms and $t_{\text{поиск}} \le 0.1$ ms. For each $|\text{C}|$ value, the minimum of the studied k values was selected, for which $\text{recall}@k \ge 80\%$. With $|\text{C}| = 2000$, k = 30 and recall@30 = 0.82 correspond this condition (Tab. 1). In this case, exhaustive iterating takes $2000 \cdot 0.095 = 190$ s, and the proposed two-stage scheme is $0.092 + 0.0001 + 30 \cdot 0.095 \approx 2.9$ s.

The estimated speedup is approximately 65 times. The dependence of processing time on the number of candidates is shown in Fig. 5.

Fig. 5. Estimated time for a full iteration and two-stage selection of map fragments

The ratio of the full search time to the two-stage processing time is

$$\frac{T_{\text{полн}}}{T_{\text{двух}}} = \frac{|\text{C}| t_{\text{M}}}{t_{\text{кодир}} + t_{\text{поиск}} + k \cdot t_{\text{M}}} = \frac{|\text{C}| t_{\text{M}}}{t_{\text{кодир}} / t_{\text{M}} + t_{\text{поиск}} / t_{\text{M}} + k}. \quad (11)$$

This estimate applies to the map from UAV-VisLoc. For Esri World Imagery, with the same k = 30, the estimated time remains 2.9 s, but recall decreases to recall@30 = 0.69 (Tab. 2). Therefore, the k value should be selected based on the cartographic source and the required recall.

CONCLUSION

A method for pre-selecting georeferenced map fragments for UAV visual navigation under significant position uncertainty is proposed. Candidates are first ranked by global descriptors, after which a limited number of the best fragments are passed to the full matching model for geometric validation.

The global descriptor is constructed using intermediate features of the frozen DINOv2 ViT-L/14 encoder, which is part of the MINIMA model. Therefore, pre-selection does not require the use of a separate neural network model. Comparison with the original DINOv2 showed virtually identical selection recall. The best results among the studied aggregation methods were obtained using a generalized average with p equal to 3 to 4.

For the UAV-VisLoc dataset, with the equivalent search area increased to 32.8 km^2, the recall@10 value was 0.71, while recall@30 was 0.82. With $|\text{C}| = 2000$ and k = 30, the estimated processing time decreased from 190 to 2.9 s, corresponding to a speedup of approximately 65 times compared to exhaustive search. With the full database of 3685 descriptors, both approximate indices maintained the recall of exact search, reducing the query time from 0.650 ms to 0.150 ms (IVF) and 0.070 ms (HNSW). On such a large map, searching is already much cheaper than constructing a descriptor, so a truly approximate index will only pay off on a significantly larger map.

Switching to the Esri World Imagery map reduced recall by 10 to 18% relative to the map in the dataset: with $|\text{C}| = 2000$ and k = 30, the recall@30 value dropped to 0.69. This means that the number of candidates selected must be tailored to the specific map source and the desired recall level.

In a diagnostic experiment with synthetic distortions, the quality dropped most significantly with a noticeable decrease in frame resolution. Selection errors were also more frequent in areas of uniform development, where adjacent fragments are similar in spatial structure.

Overall, the method provides high recall while significantly reducing computational costs, which is important for UAV visual navigation systems operating with periodic position corrections.

The method will be further tested on a regional scale, an end-to-end experiment will be conducted with geometric verification and coordinate error assessment, and the processing time will be measured on the onboard computer.