*Информационные процессы, Том 24, № 2, 2024, стр. 163–172.*


**ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ**

# Практическое приложение методов многокритериальной оптимизации к задаче релокализации камеры

**Б.Р. Габдуллин**[*,**], **Д.С. Сенюшкин**[**], **А.С. Конушин**[**]

[*] *Национальный исследовательский университет «Высшая школа экономики»*
[**] *Институт искусственного интеллекта AIRI, Москва, Российская Федерация*

Поступила в редколлегию 21.05.2024 г. Принята 06.08.2024 г.

**Аннотация**—Релокализация камеры — фундаментальная задача в области компьютерного зрения. Сверточные нейронные сети показали впечатляющие результаты в решении этой задачи. В этой работе мы предлагаем рассмотреть исходную задачу совместной оценки сдвига и поворота камеры с точки зрения поиска Парето оптимального решения, доставляющего лучшее качество каждой из них. Мы рассматриваем практический аспект применения методов многокритериальной градиентной оптимизации для обучения нейросетевой модели, решающей задачу релокализации. Целью работы является апробация существующих методов и поиск наиболее подходящего решения для этого сценария. Мы демонстрируем, что применение подобного рода оптимизационных подходов при обучении нейронной сети повышает ее качество, при этом не требуя дополнительных данных на обучение и не увеличивая количество ее параметров. Исследование проведено в рамках методологии экспериментального машинного обучения. Представленные результаты получены на публично доступном наборе данных Microsoft 7SCENES.

**КЛЮЧЕВЫЕ СЛОВА:** Релокализация камеры, сверточная нейронная сеть, многокритериальная оптимизация.

**DOI:** 10.53921/18195822_2024_24_2_163

## 1. ВВЕДЕНИЕ

Релокализация камеры, или локализация на основе изображений, является фундаментальной проблемой в робототехнике и компьютерном зрении. Это относится к процессу определения положения камеры на основе визуального представления сцены и имеет важное значение для многих приложений, таких как навигация автономных транспортных средств, structure from motion (SfM), дополненная реальность (AR) и одновременная локализация и построение карты (SLAM). В связи с важностью этой задачи было предложено множество методов ее решения.

Существуют методы локализации на основе ключевых точек, которые находят соответствия между локальными объектами, извлеченными из изображения путем применения дескрипторов изображения (SIFT, ORB и т. д. [1, 2, 3]), и трехмерными облаками точек сцены, полученными из SfM. В свою очередь, такой набор 2D-3D соответствий позволяет восстановить полную позу камеры с 6 степенями свободы (местоположение и ориентация). Однако этот низкоуровневый процесс поиска совпадений работает робастно и точно не во всех сценариях, например, в случае сцен без текстур, больших изменений освещения, окклюзий и повторяющихся структур.

В последнее время различные методы машинного обучения [4, 5, 6], в частности, лес регрессии координат сцены (SCoRF [5, 6]), были успешно применены для решения проблемы

локализации камеры. SCoRF использует предсказанное трехмерное местоположение четырех пикселей входного изображения для генерации начального набора гипотез о положении камеры, которые впоследствии уточняются с помощью цикла RANSAC. Однако все эти методы требуют карт глубины, связанных с входными изображениями во время обучения, поэтому применимость таких подходов ограничена.

В связи с успехом в классификации изображений [7, 8], семантической сегментации [9, 10] и поиске изображений [11, 12], сверточные нейронные сети (CNN) также использовались для оценки позы камеры на основе визуальных данных [13, 14]. Они рассматривают перемещение камеры как проблему регрессии, где местоположение камеры напрямую оценивается с помощью CNN, предварительно обученной на данных классификации изображений [15]. Хотя подходы, основанные на обучении, преодолевают многие недостатки методов на основе ключевых точек, они все же имеют определенные ограничения. Непосредственная регрессия абсолютной позы камеры ограничивает возможности обучения и оценки текущих моделей машинного обучения, когда сцены регистрируются в разных системах координат. Причина этого в том, что обученная модель изучает сопоставление изображения (пикселей) с позой, которое зависит от системы координат обучающих данных, принадлежащих конкретной сцене. Это вызывает сложности, особенно, если требуется локализация одновременно в нескольких сценах, а также препятствует передаче полученных знаний о геометрических отношениях между сценами. Вторая проблема заключается в явно ограниченной масштабируемости для больших сред, поскольку конечная нейронная сеть имеет верхнюю границу физической области, которую она может выучить, как указано в [14]. В работе [16] было предложено отделить процесс обучения от системы координат сцены. То есть вместо непосредственной регрессии абсолютной позы, как в [13, 14], предлагается обучение сиамской архитектуры CNN для задачи регрессии относительной позы между парой входных изображений. Однако в [16] обучение нейронной сети рассматривалось как минимизация суммы функций потерь для задачи регрессии сдвига и регрессии поворота, что является частным случаем многокритериальной оптимизации — **uniform scaling**. В данной работе предлагается рассмотреть более продвинутые методы многокритериальной оптимизации для задачи релокализации камеры.

При многокритериальной оптимизации (Multi Task Learning — MTL) [17, 18] несколько задач оптимизируются одновременно с использованием единой модели и за счет использования общей информации между задачами для улучшения обобщения и повышения производительности для всех задач. MTL является ключевым компонентом реальных приложений, таких как обработка естественного языка [19], компьютерное зрение [20, 21] и обучение с подкреплением [22, 23]. Однако эти системы сложно обучать, поскольку различные задачи необходимо правильно сбалансировать, что бывает затруднительным, когда в градиенте многозадачности доминирует градиент одной из задач.

Новизна представленной работы заключается в следующем:

- Задача релокализации камеры была сформулирована как задача многокритериальной оптимизации.
- Были применены методы многокритериальной оптимизации для задачи релокализации камеры и показаны их преимущества по сравнению с использованием классического алгоритма однородного взвешивания функций потерь.
- Произведен сравнительный анализ работы методов MTL в рамках выбранной задачи, а также произведена оценка их асимптотических сложностей.

## 2. ОБЗОР ЛИТЕРАТУРЫ

### *2.1. Релокализация камеры*

Существуют различные подходы к решению задачи релокализации камеры, в частности, визуальное распознавание места, локализация на основе структуры, а также подходы, основанные на методах машинного обучения.

Визуальное распознавание места рассматривает локализацию на основе изображений как задачу поиска изображений, используя такие методы, как дескрипторы изображений (SIFT, ORB, SURF [1, 2, 3]), быстрое пространственное сопоставление [24], набор визуальных слов [25,6] для поиска представления неизвестной сцены (запрос-изображение) в базе данных изображений с геотегами. Затем геотег наиболее релевантного извлеченного изображения базы данных рассматривается как приближение местоположения запроса. Основным ограничением методов визуального распознавания является то, что изображения в базе данных часто разрежены, поэтому в ситуациях, когда запрос находится далеко от изображений базы данных, оценка будет неточной.

Подходы локализации на основе структуры используют представление трехмерной сцены из SfM, находят соответствия между трехмерными точками и локальными объектами из изображения запроса для установления совпадений 2D-3D. В таких подходах зачастую используется RANSAC в сочетании с алгоритмом Perspective-n-Point [26] для определения позы камеры. Однако сопоставление дескрипторов — дорогостоящая и трудоемкая процедура, что усложняет задачу релокализации камеры для крупномасштабных сцен, таких как городская среда. Чтобы ускорить этот процесс, некоторые методы [27, 28] исключают поиск соответствий, как только найдено достаточное количество совпадений, или [29,30] предлагают сопоставление с трехмерными точками изображений, извлекаемых чаще всего. А также в работе [31] было показано, что сочетание визуального распознавания места с локальным SfM улучшает эффективность локализации.

Подходы, основанные на методах машинного обучения, такие как регрессионные леса [5,6] и CNN [13, 14], предлагают основу для эффективных решений проблемы оценки позы. Кроме того, сети LSTM [32, 33] были применены для определения места, из которого была сделана фотография. Подходы на основе CNN, в том числе PoseNet [14] и HourglassPose [34], показали наилучшие результаты по сравнению с двумя рассмотренными выше подходами при решении задачи локализации камеры на основе изображений. В частности, в работе [14] была показана эффективность использования совместной оптимизации предсказания сдвига и поворота. В текущей работе используется подход, предложенный в [16] — использовать сиамскую сверточную нейронную сеть для оценки относительного изменения позы между двумя кадрами.

### *2.2. Многокритериальная оптимизация*

Формулировка многокритериальной оптимизации хорошо мотивирована в различных областях [35,21,36,37,20,38]. Существуют методы, использующие ручную настройку весов функции потерь для конкретной задачи, в частности, [14]. Однако ручная настройка несет за собой увеличение необходимых вычислений. Чтобы решить эту проблему, [20] предлагают использовать гомоскедастическую неопределенность для расчета взвешенных потерь при многозадачности, а [39] предлагают эвристику для настройки весов потерь на основе величин градиента задачи. Конфликтующие градиенты в многозадачном обучении (MTL) смягчаются с помощью явных методов модуляции градиента, таких как «градиентная хирургия» [40], которая также противопоставляет конфликтующим градиентам модифицированный, неконфликтный градиент. Многообъектная оптимизация (MOO) направлена на одновременную оптимизацию контрастирующих целей [41, 42, 43, 44, 45]. В работе [46] расширяют классический метод MGDA для

решения многомерных задач, но он может сходиться к неоптимальному решению для невыпуклых функций.

Для решения проблемы сходимости к неоптимальному Парето-стационарному решению в случае невыпуклости рассматриваемой функции, был предложен метод Independent Component Alignment [38]. В представленной работе проведено сравнение качества и сложности основных из вышеперечисленных методов применительно к задаче релокализации камеры.

## 3. ПОСТАНОВКА ЗАДАЧИ И ОБОЗНАЧЕНИЯ

Имея на вход два RGB кадра размера $H \times W$: $X_i, X_j \in \mathbb{R}^{3\times H\times W}$, необходимо оценить относительное изменение положения камеры между этими двумя кадрами, которое задается с помощью вектора сдвига $\vec{t}_{ij} \in \mathbb{R}^3$ и вектора поворота $\vec{q}_{ij} \in \mathbb{R}^4$, представленного в виде кватернионов. Решение задачи ищется с помощью сверточной нейронной сети:

$$\phi(\theta, \theta_t, \theta_R, X_i, X_j) : \mathbb{R}^{3\times H\times W} \times \mathbb{R}^{3\times H\times W} \to \mathbb{R}^3 \times \mathbb{R}^4, \tag{1}$$

где $\theta_t$ — параметры, необходимые только для оценки сдвига, $\theta_R$ — параметры, необходимые только для оценки поворота, $\theta$ — параметры, используемые для обеих задач.

Для обучения модели запишем задачу минимизации в терминах многокритериальной оптимизации:

$$\min_{\theta,\theta_t,\theta_R} \sum_{\substack{i,j\in\{1,\dots,N\}\\ i\neq j}} \Big( w_t \cdot \mathcal{L}_t[\phi(\theta, \theta_t, X_i, X_j), \vec{t}_{ij}] + w_R \cdot \mathcal{L}_R[\phi(\theta, \theta_R, X_i, X_j), \vec{R}_{ij}] \Big), \tag{2}$$

где $\mathcal{L}_t$ — функция потерь для задачи оценки сдвига, $\mathcal{L}_R$ — функция потерь для задачи оценки поворота, $w_t, w_R$ — соответствующие задачам весовые коэффициенты, $N$ — количество изображений из набора данных, используемых при обучении. Поиск минимума функции потерь осуществляется с помощью градиентных методов оптимизации.

## 4. ЭКСПЕРИМЕНТЫ

### *4.1. Архитектура нейронной сети*

Архитектура нейросетевой модели содержит в себе два дескриптора, необходимых для выделения признаков из пары изображений, и голову, оценивающую относительные изменения сдвига и поворота на основе сконкатенированных признаков с двух изображений. В качестве дескриптора, подобно [32, 16], используется сверточная нейронная сеть ResNet-34 [7], предобученная на наборе данных ImageNet [8], без последнего полносвязного классификационного слоя. В качестве модели оценки относительных изменений сдвига и поворота используются две полносвязные нейронные сети, принимающие на вход сконкатенированные признаки, полученные из изображений. Подробное описание архитектуры можно найти в таблице 1 и на рисунке 1. Стоит заметить, что целью представленной работы является улучшение качества оценки нейросетевой модели без добавления новых обучаемых параметров.

### *4.2. Набор данных*

Был выбран набор данных Microsoft 7SCENES [47], который используется для сравнения методов регрессии положения камеры в помещении. Набор данных включает в себя последовательности изображений RGB-D семи различных сцен в помещении, снятых с портативной

| Encoder | | | | | Decoder |
|---|---|---|---|---|---|
| Conv, $7 \times 7$, 64, s=2<br>MaxPool, $3 \times 3$, s=2 | $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$ | $\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$ | $\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$ | $\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$ | FC, 3 ($t$)<br>FC, 4 ($R$) |

**Таблица 1.** Описание архитектуры дескриптора.

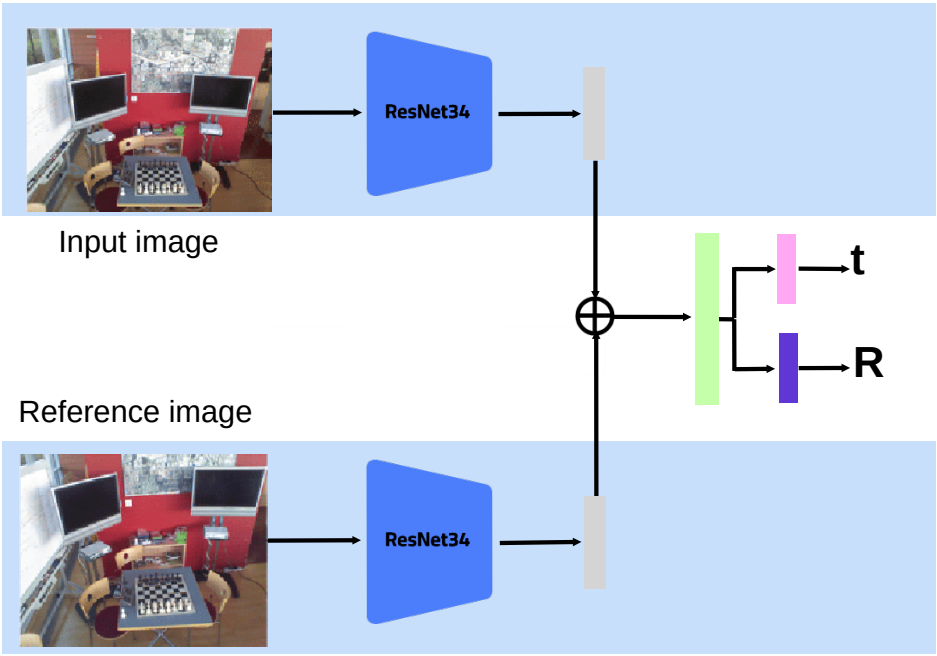


**Рис. 1.** Архитектура нейронной сети. Два изображения подаются в модели-дескрипторы, затем их признаки конкатенируются. На основе сконкатенированных признаков с помощью двух полносвязных нейронных сетей предсказываются относительные сдвиг и поворот.

камеры Kinect. Каждая сцена представляет собой обучающую и тестовую последовательность изображений, состоящую из 1000–7000 кадров с разрешением 640×480. Более подробное описание сцен можно увидеть в таблице 2.

| **Название** | **Размер сцены** | **# кадров** | |
|---|---|---|---|
| | | Train | Test |
| CHESS | $3m^3$ | 4k | 2k |
| FIRE | $4m^3$ | 2k | 2k |
| HEADS | $2m^3$ | 1k | 1k |
| OFFICE | $5.5m^3$ | 6k | 4k |
| PUMPKIN | $6m^3$ | 4k | 2k |
| KITCHEN | $6m^3$ | 7k | 5k |
| STAIRS | $5m^3$ | 2k | 1k |

**Таблица 2.** Описание набора данных 7SCENES. В таблице указаны физические размеры, а также количество изображений в тренировочной и тестовой выборках каждой из 7 сцен.

### *4.3. Методы MTL*

В представленной работе рассматриваются некоторые методы многокритериальной оптимизации.

**Uniform scaling**. Оптимизация равномерно взвешенной суммы функций потерь для каждой задачи: $w_t = w_R = \frac{1}{2}$. Это простейший метод многокритериальной оптимизации, при котором функции потерь усредняются. Этот метод будет рассматриваться в качестве базового, так как целью более продвинутых методов MTL является улучшение относительно суммирования (усреднения) функций потерь разных задач.

**Uncertainty**. При использовании метода [20] поиск оптимальных параметров $w_t, w_R$ осуществляется с помощью вероятностного моделирования.

**MGDA**. Multiple Gradient Descent Algorithm (MGDA) [43] использует оптимизацию на основе градиента и доказуемо сходится к точке множества Парето. MGDA хорошо подходит для многозадачного обучения глубоких нейросетей.

**MGDA-UB**. Алгоритм Multiple Gradient Descent Algorithm – Upper Bound (MGDA-UB) [46] является аппроксимацией метода MGDA с целью уменьшения вычислительной сложности алгоритма.

**GradNorm**. В методе GradNorm [39] используется нормализация градиентов для балансировки обучения нескольких задач. Параметры $w_t, w_R$ обновляются в процессе обучения в зависимости от градиента.

**PCGrad**. В методе PCGrad [40] градиент, соответствующий определенной задаче, проецируется на нормаль к градиенту по задаче, с которым он (частично) противонаправлен, во избежание негативного взаимодействия между ними.

**Independent Component Alignment**. В работе [38] представлены методы стабилизации процедуры обучения посредством выравнивания главных компонент матрицы градиентов: $\theta$-aligned и $Z$-aligned.

### *4.4. Сравнительный анализ*

**Анализ вычислительной сложности.** Основные вычислительные затраты всех алгоритмов зависят от количества обратных проходов по общим параметрам для каждой задачи и для каждой итерации градиентного спуска. Мы обозначаем количество задач как T, а количество итераций в решателе Франка-Вульфа [48] как $K$. Мы рассматриваем временную сложность разложения SVD константной. При таком определении $T$ и $K$ можно оценить сложность каждого из методов на основе используемых ими алгоритмов. Оценка представлена в таблице 3.

Для каждого метода мы считаем медианные ошибки сдвига $\Delta t$ и поворота $\Delta R$. Также используется показатель роста $\mathcal{R}$ как мера улучшения качества решения задачи по сравнению с методом **uniform scaling**. Этот показатель определяется следующим образом: $\mathcal{R} = 1 - \frac{P_i}{P}$, где $P_i$ — средняя ошибка сдвига (поворота) при использовании метода MTL, $P$ — средняя ошибка сдвига (поворота) при использовании метода **uniform scaling**.

## 5. РЕЗУЛЬТАТЫ

Результаты экспериментов представлены в таблице 4. Для состоятельности оценки было проведено 10 экспериментов для каждого метода, в таблице представлены средние значения ошибок сдвига и поворота. Можно заметить, что методы, не являющиеся Парето-стационарными (GradNorm, PCGrad, Uncertainty, ICA $\theta$-aligned), улучшают качество оценки сдвига, но вместе с этим ухудшают качество оценки поворота. В частности, алгоритм PCGrad, показатель

| Метод | Сложность |
|---|---|
| MGDA | $\mathcal{O}(T + K)$ |
| MGDA-UB | $\mathcal{O}(K)$ |
| PCGGrad | $\mathcal{O}(T)$ |
| GradNorm | $\mathcal{O}(T)$ |
| Uncertainty | $\mathcal{O}(1)$ |
| ICA $\theta$-aligned | $\mathcal{O}(T)$ |
| ICA $Z$-aligned | $\mathcal{O}(1)$ |

**Таблица 3.** Сравнение вычислительной сложности методов многокритериальной оптимизации за шаг градиентного спуска для $T$ задач и $K$ шагов.

улучшения которого наивысший для задачи оценки $t$, имеет значительный отрицательный показатель улучшения по целевой переменной $R$. В то же время алгоритмы MGDA, MGDA-UB улучшают качество оценки сдвига и почти не ухудшают качество оценки поворота. Таким образом, компромиссным решением можно назвать метод ICA $Z$-aligned, который работает заметно лучше остальных Парето-оптимальных алгоритмов в задаче оценки сдвига, при этом не ухудшая качество оценки поворота. Особенный интерес этот метод представляет из-за наилучшей оценки асимптотической сложности — $\mathcal{O}(1)$.

| | | Сцены | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CHESS | FIRE | HEADS | OFFICE | PUMPKIN | KITCHEN | STAIRS | Mean | $\mathcal{R}$ |
| Uniform | $\Delta t$, m | 0.48 | 1.78 | 0.46 | 0.70 | 0.72 | 0.90 | 0.47 | 0.79 | - |
| | $\Delta R$, ° | 5.83 | 11.57 | 13.04 | **8.43** | 6.79 | 8.88 | **11.22** | 9.39 | - |
| MGDA | $\Delta t$, m | 0.29 | 0.73 | 0.81 | 0.33 | 0.47 | 0.33 | 0.46 | 0.49 | +37.98 |
| | $\Delta R$, ° | 5.90 | 11.84 | 12.59 | 8.67 | **6.37** | 9.08 | 11.46 | 9.42 | -0.32 |
| MGDA-UB | $\Delta t$, m | 0.30 | 0.73 | 0.66 | 0.59 | 0.46 | 0.57 | 0.55 | 0.55 | +30.38 |
| | $\Delta R$, ° | 5.92 | 11.74 | 12.42 | 9.09 | 6.30 | **8.69** | 11.76 | 9.42 | -0.32 |
| GradNorm | $\Delta t$, m | 0.17 | 0.33 | 0.25 | 0.24 | 0.25 | 0.26 | 0.37 | 0.27 | +65.82 |
| | $\Delta R$, ° | 7.08 | 13.54 | 14.70 | 9.65 | 8.35 | 9.08 | 14.07 | 10.92 | -16.29 |
| PCGrad | $\Delta t$, m | 0.17 | 0.30 | 0.23 | 0.23 | 0.26 | 0.25 | 0.37 | 0.26 | +67.09 |
| | $\Delta R$, ° | 9.10 | 13.17 | 16.19 | 10.57 | 9.00 | 10.35 | 13.71 | 11.73 | -24.92 |
| Uncertainty | $\Delta t$, m | 0.19 | 0.33 | 0.30 | 0.25 | 0.25 | 0.28 | 0.48 | 0.30 | +62.03 |
| | $\Delta R$, ° | 9.32 | 16.15 | 17.86 | 10.83 | 11.22 | 10.20 | 14.60 | 12.88 | -37.17 |
| ICA $\theta$-aligned | $\Delta t$, m | 0.18 | 0.46 | 0.69 | 0.27 | 0.34 | 0.29 | 0.61 | 0.41 | +48.10 |
| | $\Delta R$, ° | 5.82 | 12.07 | 18.15 | 9.38 | 8.99 | 10.27 | 12.84 | 11.07 | -17.89 |
| ICA $Z$-aligned | $\Delta t$, m | 0.18 | 0.48 | 0.43 | 0.28 | 0.27 | 0.32 | 0.42 | 0.34 | +56.96 |
| | $\Delta R$, ° | **5.40** | **11.18** | **12.14** | 9.57 | **6.37** | 8.91 | 11.54 | **9.30** | +0.96 |

**Таблица 4.** Сравнение качества методов многокритериальной оптимизации для задачи релокализации камеры на наборе данных 7SCENES.

## 6. ЗАКЛЮЧЕНИЕ

В представленной работе было рассмотрено практическое применение методов многокритериальной оптимизации для решения задачи релокализации камеры на основе изображений. В качестве модели оценки целевых переменных была выбрана нейросетевая модель вида энкодер-декодер со сверточной нейронной сетью ResNet-34 в роли энкодера. Было проведено сравнение методов многокритериальной оптимизации по их асимптотической сложности, а также по относительному улучшению на наборе данных Microsoft 7SCENES.

## СПИСОК ЛИТЕРАТУРЫ

1. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speededup robust features (SURF). Comput. Vis. Image Underst., 110(3):346–359, 2008.
2. D. G. Lowe. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2):91–110, Nov. 2004.
3. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In Proc. ICCV, 2011.
4. E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC - differentiable RANSAC for camera localization. CoRR, abs/1611.05705, 2016.
5. J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In CVPR, 2013.
6. M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In ECCV, 2002.
7. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in NIPS, 2012.
9. S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In NIPS, 2015.
10. H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In ICCV, 2015.
11. A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In Computer Vision - ECCV, 2014.
12. A. Gordo, J. Almazan, J. Revaud, and D. Larlus. Deep image ´ retrieval: Learning global representations for image search. CoRR, abs/1604.01325, 2016.
13. A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. CoRR, abs/1704.00390, 2017.
14. A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In ICCV, 2015.
15. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.
16. Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 920–929, 2017.
17. Richard Caruana. Multitask learning: A knowledge-based source of inductive bias. In Proceedings of the Tenth International Conference on Machine Learning (ICML), pp. 41–48. Morgan Kaufmann, 1993.
18. Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2051–2060, 2017.

19. Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, volume 57. Association for Computational Linguistics, 2019c.

20. Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7482–74911, 2018.

21. Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6129–6138, 2017.

22. Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. In International Conference on Learning Representations (ICLR), 2016.

23. Yee Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In Advances in Neural Information Processing Systems (NIPS), volume 30, pp. 4499–4509. Curran Associates, Inc., 2017.

24. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In Proc.CVPR, 2007.

25. O. G. Cula and K. J. Dana. Compact representation of bidirectional texture functions. In CVPR, 2001.

26. M. Bujnak, Z. Kukelova, and T. Pajdla. New efficient solution to the absolute pose problem for camera with unknown focal length and radial distortion. In ACCV, 2011.

27. Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In Proc. ECCV, 2010.

28. T. Sattler, B. Leibe, and L. Kobbelt. Efficient effective prioritized matching for large-scale image-based localization. IEEE TPAMI, 2016.

29. T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys. Hyperpoints and fine vocabularies for largescale location recognition. In ICCV, 2015. 2

30. T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In BMVC, 2012. 2

31. T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In CVPR, 2017.

32. Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using LSTMs for structured feature correlation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 627–637, 2017

33. R. Clark, S. Wang, N. T. Andrew Markham, and H. Wen. VidLoc: A deep spatio-temporal model for 6-DoF video-clip relocalization. In CVPR, 2017.

34. Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 870–877, 2017.

35. Hakan Bilen and Andrea Vedaldi. Integrated perception with recurrent multi-task neural networks. In Advances in Neural Information Processing Systems (NIPS), volume 29, pp. 235–243. Curran Associates, Inc., 2016.

36. Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3712–3722, 2018.

37. Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shen, and Ian Reid. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In International Conference on Robotics and Automation (ICRA), pp. 7101–7107. IEEE, 2019.

38. Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, Anton Konushin: Independent Component Alignment for Multi-Task Learning In Computer Vision and Pattern Recognition, pp. 20083-20093, 2023.

39. Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Proceedings of the 35th International Conference on Machine Learning (ICML), volume 80 of Proceedings of Machine Learning Research, pp. 794–803. PMLR, 2018

40. Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pp. 5824–5836. Curran Associates, Inc., 2020.

41. Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. Mathematical Methods of Operations Research, 51:479–494, 2000.

42. Stefan Schäffler, Richard R. Schultz, and Konstanze Weinzierl. Stochastic Method for the Solution of Unconstrained Vector Optimization Problems. Journal of Optimization Theory and Applications, 114:209–222, 2002.

43. Jean-Antoine Désidéri. Multiple-gradient descent algorithm for multiobjective optimization. In European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), 2012.

44. Sebastian Peitz and Michael Dellnitz. Gradient-Based Multiobjective Optimization with Uncertainties, pp. 159–182. Springer International Publishing, 2018.

45. Fabrice Poirion, Quentin Mercier, and Jean-Antoine Désidéri. Descent algorithm for nonsmooth stochastic multiobjective optimization. Computational Optimization and Applications, (2):317–331, 2017.

46. Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In Advances in Neural Information Processing Systems (NeurIPS), volume 31, pp. 527–538. Curran Associates, Inc., 2018.

47. Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2930–2937, 2013.

48. Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Proceedings of the 30th International Conference on Machine Learning (ICML), volume 28 of Proceedings of Machine Learning Research, pp. 427–435. PMLR, 2013. 7

# Practical application of Multi Task Learning methods to the camera relocalization problem

**B.R. Gabdullin, D.S. Senushkin, A.S. Konushin**

Camera relocalization is a fundamental problem of computer vision. Convolutional neural networks have shown impressive results in solving this problem. In this work, we propose to consider the original problem of joint estimation of camera translation and rotation from the point of view of searching for a Pareto optimal solution that delivers the best quality for each of them. We consider the practical aspect of using multi task learning methods for training a neural network model that solves the relocalization problem. The goal of the work is to test existing methods and find the most suitable solution for this scenario. We demonstrate that the usage of this kind of optimization approaches during training of neural network improves its quality, without requiring additional training data and without increasing the number of its parameters. The study was conducted within the framework of experimental machine learning methodology. The results presented are based on the publicly available Microsoft 7SCENES dataset.

**KEYWORDS:** camera relocalization, convolutional neural network, multi task learning.