Правительство Российской Федерации
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Направление 01.03.02 «Прикладная математика и информатика»
Кафедра технологии программирования

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:

ОЦЕНКА ПОЗЫ ЧЕЛОВЕКА НА ИЗОБРАЖЕНИИ С ПОМОЩЬЮ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Научный руководитель — д-р физ.-мат. наук, проф. М.В. Сотникова

Работу выполнил — гр. 20.Б08-пу, И.А. Пивнев

Санкт-Петербург 2022

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время компьютерное зрение (Computer Vision, CV) применяется повсеместно [1–3]. Начиная с помощи аниматору или на промышленном производстве и заканчивая анализом спортивной деятельности, возникает потребность распознать человека на изображении, выделить его позу и определить совершаемое им действие. Такую задачу принято называть Human Pose Estimation [4], или сокращенно HPE.

В силу активного развития машинного обучения (Machine Learning, ML), за последние несколько лет было представлено множество работ, точность которых и скорость вычисления показывают удовлетворительные результаты, но при этом подходы сильно разнятся [5–8].

Целью данной работы является изучение, сравнение и освещение существующих методов выделения человека на изображении. Для достижения этого необходимо:

– проанализировать актуальные подходы решения задачи HPE;

– сравнить их по скорости обработки и требовательности к ресурсам;

– продемонстрировать работу выбранных архитектур.

Стоит отметить, что в последнее время всё чаще рассматривается представление полученных данных в трех измерениях, однако данная работа сконцентрирована на двух в качестве базового случая задачи Human Pose Estimation. Так в дальнейшем можно будет перейти от простого к более сложному, используя уже полученные знания.

1 Постановка задачи

В ходе научно-исследовательской работы был проанализирован ряд статей, посвященных выделению человека на изображении методами ML и, в частности, глубокого обучения (Deep Learning, DL).

1.1 Входные данные и модель тела

Модели машинного обучения чаще всего на вход подается трехцветное красно-зелёно-синее изображение (RGB), обычно размером 256 на 192 пикселя. Иногда вместо RGB кадра используют RGBD (где D обозначает depth, то есть глубину) или инфракрасную фотографию, но такие модели создаются редко в силу малого количества данных для обучения. На выходе ожидается граф, вершины которого определяют суставные точки человека на изображении, а рёбра представляют собой естественные соединения, такие как кисть — локоть или колено — ступня, пример представлен на Рисунке 1.1. Число узлов может варьироваться в зависимости от архитектуры модели, но обычно оно в пределах 13-30 штук [1].

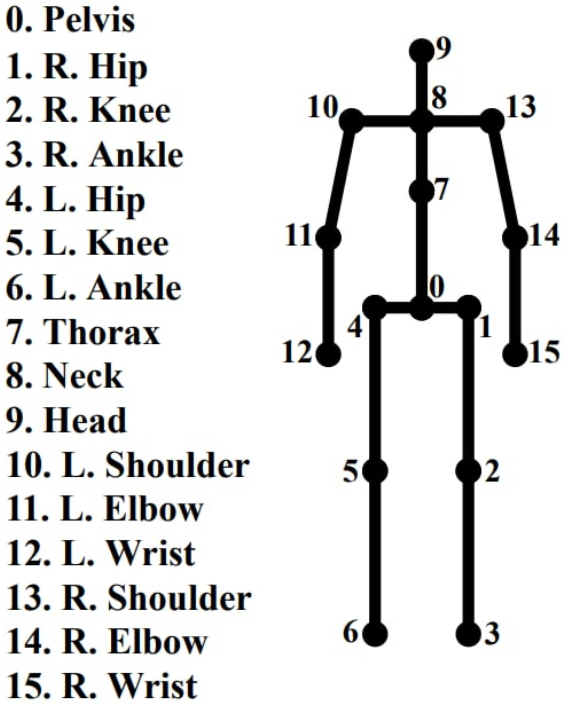
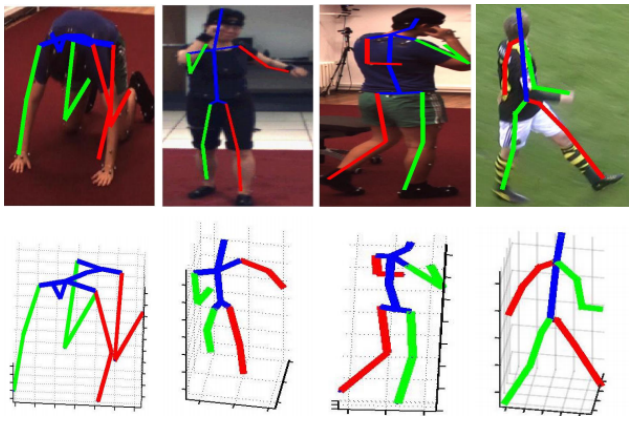


Рисунок 1.1 — Ключевые точки, ожидаемые от модели и соединённые в единый граф

Нередко изображение представляет собой трудность не только для модели, но и для человека: части тела перекрыты другими объектами, находятся в необычных положениях или вовсе отсутствуют как в примерах на Рисунке 1.2. При этом принято различать две отдельные подзадачи: выделение только одного (Single Person Pose Estimation, SPPE) или нескольких людей (Multiple Person Pose Estimation, MPPE) на изображении.



а) скрытые части тела

б) близкие узловые точки и MPPE

Рисунок 1.2 — Примеры трудных задач на определение положения человека в кадре

1.2 Метрики

Оценочная метрика часто является одной из нижеперечисленных [2]:

– Percentage of Correct Parts (PCP) — процент корректно обнаруженных соединений; евклидово расстояние между предсказанной и истинной парами ключевых точек должно быть меньше половины длины этой конечности:

$$\frac{\sum_{i=1}^{n} bool(d_i < l_i)}{n} * 100\%, \tag{1.1}$$

где

– $bool$ — функция, равная 1, если условие выполнено, иначе 0,

– d_i — расстояние между предсказанным и истинным частями тел,

– l_i — истинная длина конечности,

– n — число соединённых пар точек на изображении.

– Percentage of Detected Joints (PDJ) — процент обнаруженных соединений: евклидово расстояние между предсказанной и истинной парами ключевых точек должно быть меньше доли диаметра туловища:

$$\frac{\sum_{i=1}^{n} bool(d_i < fraction * diameter)}{n} * 100\%, \tag{1.2}$$

где

– $bool$ — функция, равная 1, если условие выполнено, иначе 0,

– d_i — расстояние между предсказанным и истинным частями тел,

– $fraction$ — пороговая доля, равная числу от 0 до 1,

– $diameter$ — длина туловища, истинное расстояние от верхней точки головы до дальней точки ног,

– n — число соединённых пар точек на изображении.

– Percentage of Correct Key-points (PCK) — процент корректно обнаруженных ключевых точек: евклидово расстояние между предсказанной и истинной парами точек должно находиться в пределах определенного порога, который обычно берут равным 150 мм, доле от длины торса или головы (например, PCKh0.5 обозначает порог, равный половине длины головы); формула имеет схожий вид, что и PDJ, меняя лишь порог в условии функции bool.

Перечисленные выше метрики показывают, насколько хорошо модель находит отдельные ключевые точки или их соединения в парах (таких как предплечье или бедро). Результат тем лучше, чем ближе он будет к 100%.

Существуют и другие метрики, цель которых также показать, насколько две позы совпадают. Например, Object Keypoint Similarity (OKS) — степень схожести (от 0 до 1) двух объектов, которая тем лучше, чем ближе окажутся предсказанные ключевые точки к истинным:

$$\frac{\sum_{i=1}^{n} exp(-\frac{d_i^2}{2s^2k_i^2})}{n}, \tag{1.3}$$

где

– d_i — евклидово расстояние между обнаруженной ключевой точкой и соответствующей ей истиной,

– s — коэффициент масштаба, равный квадратному корню из площади сегмента изображения, на котором находится человек,

– k_i — уникальная константа для каждого сустава,

– n — число ключевых точек на изображении.

Каждая из метрик имеет свои преимущества и недостатки. Например, PCP нередко штрафует за неправильно расположенные суставы небольшой длины, так как их сложнее точно определить. OKS учитывает масштаб изображения и использует отдельные константы под каждый сустав, что учитывает широкий спектр длин конечностей, потому эту метрику и PCK выбирают чаще всего.

1.3 Данные для обучения

Модели глубокого машинного обучения используются для выделения скрытых признаков и закономерностей, присущих данной задаче, что требует большого количества данных для обучения.

В настоящий момент существует два больших набора изображений (датасета) и целое множество поменьше:

– MPII Human Pose [9] — содержит примерно 25 тыс. фотографий с более чем 40 тыс. отмеченных людей (задачи SPPE и MPPE), которые представляют 410 видов различных активностей;

– COCO (Microsoft Common Objects in Context) [10] — содержит 39 тыс. изображений, на которых отмечено примерно 56 тыс. людей.

Каждый такой датасет представляет из себя набор изображений и аннотаций к ним, содержащих координаты (в пикселях) узловых точек отмеченных людей. И если раньше авторы работ концентрировались на одном из этих двух хранилищ данных для обучения, то более поздние архитектуры используют их объединение, дополненное другими датасетами с так же размеченными изображениями.

Данные для разметки обычно получают: из открытых источников Интернета, с помощью специальных съемок с привлечением профессиональных актеров или применяя другие архитектуры по определению человека на изображении с последующей проверкой и корректировкой разметки.

2 Основные подходы

2.1 Сверточные и рекуррентные сети

Первоначально интерес к задаче Human Pose Estimation появился еще в 2014 году. Тогда одной из первых работ стала глубокая нейронная сеть DeepPose [11], в которой последовательно применялись свёрточные слои с последующими полносвязными для извлечения из изображения ключевой информации. Основной идеей стало использование рекуррентного блока, вывод которого подается на вход вновь для уточнения координат точек.

Передовой идеей стало использование тепловых карт признаков, которые попиксельно показывают вероятность присутствия сустава в этой области изображения. Одной из первых работ, в которой внедрили данный принцип, стала ConvNet [12]. Пример тепловых карт, которые используются в этой нейросети, представлен на Рисунке 2.1.

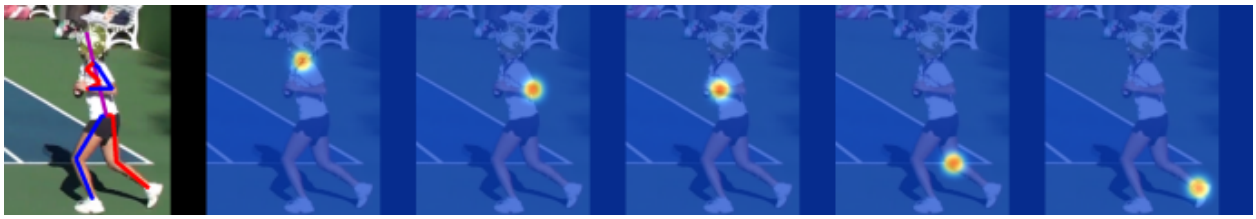


Рисунок 2.1 — Тепловые карты признаков узловых точек

Многие последующие работы [13; 14] всячески развивали эту идею, комбинируя рекуррентные, полносвязные и свёрточные слои для достижения наилучших результатов. Некоторые из них [7] предварительно выделяют положения людей на кадре в рамки, чтобы последующие слои могли сконцентрироваться на конкретных частях изображения (подход сверху вниз), в то время как другие сразу находят отдельные узлы и уже затем пытаются их соединить суставами (снизу вверх).

Принцип песочных часов [8] с последующим его масштабированием [15] полностью полагается на способность свёрточных слоев выделять скрытые признаки, а последующая развертка этой информации раскры-

вает её на всю площадь изображения. Многократное применение данного механизма дает значительный прирост в точности предсказания положения узловых точек, что позволило перейти за порог в 90%-а по метрике PCKh0.5.

Особое место занимает вывод не только соединенных узловых точек суставов, но и целостное закрашивание всего человека, что можно увидеть на Рисунке 2.2 с примером работы нейросети Mask R-CNN [16].

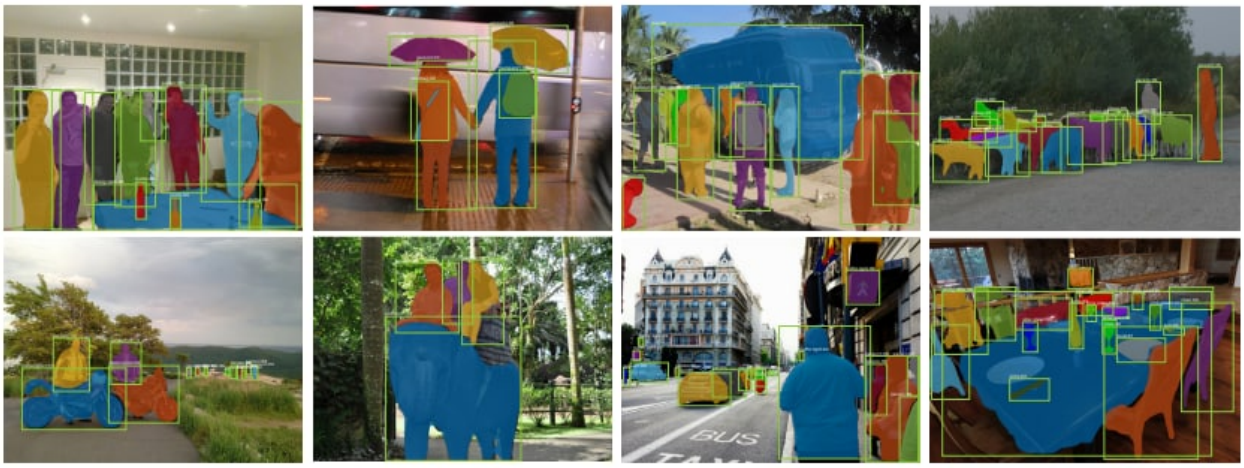


Рисунок 2.2 — Вывод нейросети Mask R-CNN с масками, наложенными поверх найденных на кадре людей

Исследования возможных способов решения данной задачи на этом не закончились, однако новые методы начали приходить из других областей. Так, например, подход под названием Long-Short Term Memory (LSTM) лёг в основу нейросети UniPose [6]. Его идея первоначально использовалась в области обработки естественного языка (Natural Language Processing, NLP), но рекуррентный подход с запоминанием лишь небольшой информации за несколько последних итераций хорошо себя показал и в задачах Computer Vision.

2.2 Трансформеры

Концепция attention [17], то есть внимания, представленная инженерами компании Google в 2017 году, изначально была направлена на задачи

NLP, однако со временем её стали применять и для целей компьютерного зрения. Эволюционной идея стала благодаря тому, что в её основе лежит желание найти скрытую взаимосвязь между всеми частями объекта, что и является аналогом привычного нам внимания.

Самой передовой архитектурой на данный момент является ViTPose [5], которая достигла 94%-го значения по метрике PCKh0.5. Примечательно, что авторы не стали использовать свёрточные сети для извлечения карт признаков изображения, остановившись лишь на многоразовом использовании Multi-Head Self-Attention (MSHA), то есть параллельного применения принципа attention. Схему архитектуры данной работы можно увидеть на Рисунке 2.3.

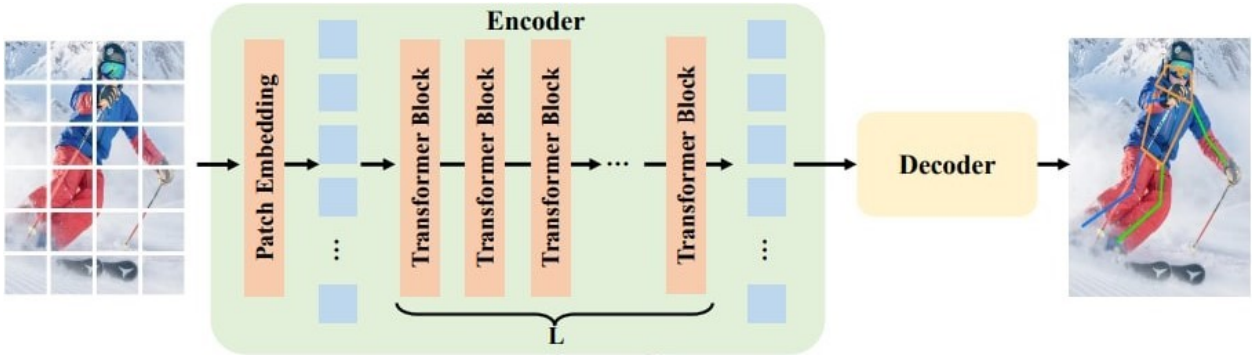


Рисунок 2.3 — Структура нейросети-трансформера ViTPose

Идея attention легла в основу целого ряда статей и архитектур, которые принято называть трансформерами.

Например, другой выдающейся работой стала TokenPose [18]. Её особенностью является первоначальное извлечение карт признаков изображения (матриц информации, используемых в свёрточных слоях) с последующим привлечением идеи attention для связи этих признаков воедино. Такой подход позволяет модели найти скрытую взаимосвязь разных ключевых точек (например, что некоторые из них соединены, даже если конечность скрыта из поля зрения). Интересно, что авторы также используют ряд так называемых токенов — векторов для каждой из узловых точек, что

необходимо для концентрации нейросети на особенностях строения тела человека. Архитектура этой работы представлена на Рисунке 2.4.

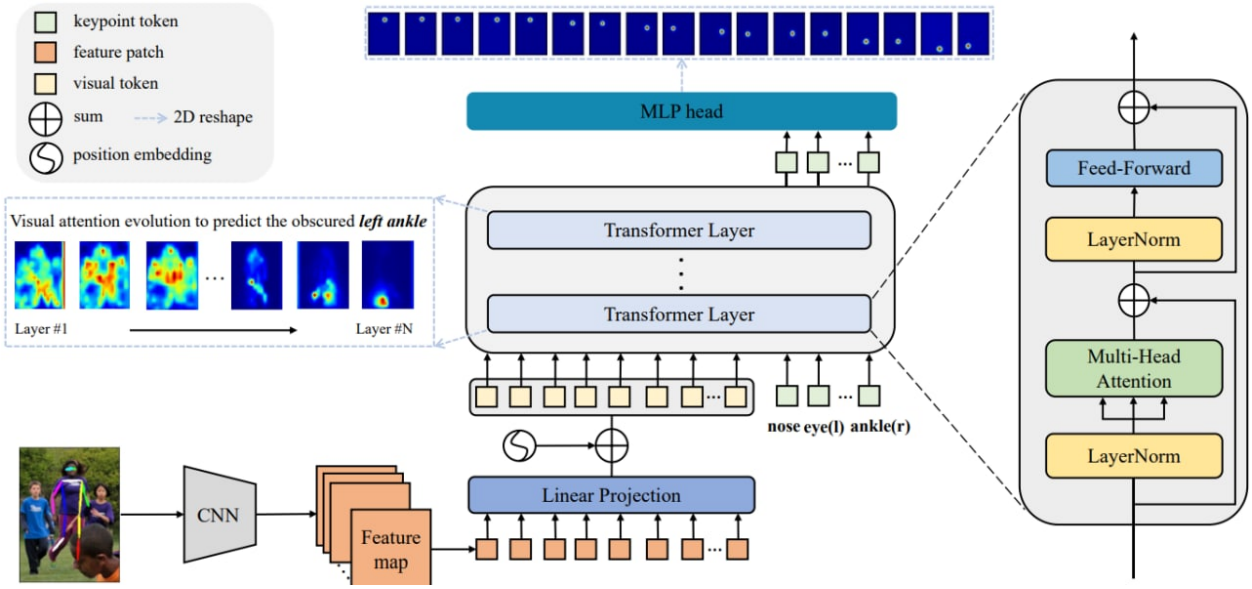


Рисунок 2.4 — Схема строения нейросети-трансформера TokenPose

ЗАКЛЮЧЕНИЕ

В ходе проделанной работы был исследован ряд статей о задаче выделения человека на изображении для двумерного случая, названы отличия между принципиально разными идеями и приведены в пример конкретные архитектуры, показывающие наилучшие результаты в настоящий момент.

В дальнейшем следует задаться вопросом обработки видеоряда и получения с него положения человека в реальном времени, что также входит в задачу HPE. Будет необходима оптимизация вычислений, так как многие архитектуры являются требовательными к ресурсам используемых для их запуска устройств. Возможно расширение рассматриваемой задачи до получения данных в трехмерном виде, что может быть реализовано не только в форме графа, но и в виде облака точек (их геометрического множества в пространстве) или mesh-объекта (полигональной сетки трехмерной компьютерной графики).

Наконец, необходимо будет самостоятельно запустить несколько передовых архитектур на выбор и продемонстрировать результат их работы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Human Pose Estimation: Simplified. — Режим доступа: https://towardsdatascience.com/human-pose-estimation-simplified-6cfd88542ab3 (дата обращения: 29.03.2019).

2. A 2019 guide to Human Pose Estimation with Deep Learning. — Режим доступа: https://nanonets.com/blog/human-pose-estimation-2d-guide/.

3. A 2019 Guide to Human Pose Estimation. — Режим доступа: https://heartbeat.comet.ml/a-2019-guide-to-human-pose-estimation-c10b79b64b73 (дата обращения: 05.08.2019).

4. Pose Estimation. — Режим доступа: https://paperswithcode.com/task/pose-estimation.

5. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation / Xu Yufei, Zhang Jing, Zhang Qiming, and Tao Dacheng // arXiv preprint arXiv:2204.12484. — 2022.

6. Artacho Bruno, Savakis Andreas. Unipose: Unified human pose estimation in single images and videos // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — P. 7035–7044.

7. Xiao Bin, Wu Haiping, Wei Yichen. Simple baselines for human pose estimation and tracking // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 466–481.

8. Newell Alejandro, Yang Kaiyu, Deng Jia. Stacked hourglass networks for human pose estimation // European conference on computer

vision / Springer. — 2016. — P. 483–499.

9. 2d human pose estimation: New benchmark and state of the art analysis / Andriluka Mykhaylo, Pishchulin Leonid, Gehler Peter, and Schiele Bernt // Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. — 2014. — P. 3686–3693.

10. Microsoft coco: Common objects in context / Lin Tsung-Yi, Maire Michael, Belongie Serge, Hays James, Perona Pietro, Ramanan Deva, Dollár Piotr, and Zitnick C Lawrence // European conference on computer vision / Springer. — 2014. — P. 740–755.

11. Toshev Alexander, Szegedy Christian. Deeppose: Human pose estimation via deep neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2014. — P. 1653–1660.

12. Efficient object localization using convolutional networks / Tompson Jonathan, Goroshin Ross, Jain Arjun, LeCun Yann, and Bregler Christoph // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — P. 648–656.

13. Human pose estimation with iterative error feedback / Carreira Joao, Agrawal Pulkit, Fragkiadaki Katerina, and Malik Jitendra // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — P. 4733–4742.

14. Convolutional pose machines / Wei Shih-En, Ramakrishna Varun, Kanade Takeo, and Sheikh Yaser // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. — 2016. — P. 4724–4732.

15. Cascade feature aggregation for human pose estimation / Su Zhihui, Ye Ming, Zhang Guohui, Dai Lei, and Sheng Jianda // arXiv preprint arXiv:1902.07837. — 2019.

16. Mask r-cnn / He Kaiming, Gkioxari Georgia, Dollár Piotr, and Girshick Ross // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 2961–2969.

17. Attention is all you need / Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia // Advances in neural information processing systems. — 2017. — Vol. 30.

18. Tokenpose: Learning keypoint tokens for human pose estimation / Li Yanjie, Zhang Shoukui, Wang Zhicheng, Yang Sen, Yang Wankou, Xia Shu-Tao, and Zhou Erjin // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2021. — P. 11313–11322.