Министерство образования и науки Российской Федерации
Московский физико-технический институт (национальный исследовательский университет)

Физтех-школа радиотехники и компьютерных технологий
Кафедра интеллектуальных информационных систем и технологий

Выпускная квалификационная работа магистра

Исследование методов доменной адаптации для улучшения распознавания ключевых точек на теле человека

Автор:
Студент М01-205а группы
Токарев Андрей Сергеевич

Научный руководитель:
Доктор технических наук
Назаров Алексей Николаевич
Научный конультант:
Ст. Преподаватель
Воронков Илья Михайлович

Москва 2024

Аннотация

Исследование методов доменной адаптации для улучшения распознавания ключевых точек на теле человека

Токарев Андрей Сергеевич

Краткое описание задачи и основных результатов, мотивирующее прочитать весь текст

Содержание

1 Введение

Современные технологии машинного обучения и компьютерного зрения продолжают активно развиваться, находя применение в самых разнообразных областях. Одно из направлений, активно развивающихся в последние годы, является решение задачи распознавания ключевых точек на теле человека (Keypoint Detection) или оценка позы человека (Human Pose Estimation). В настоящее время методы решения этой задачи могут иметь разнообразные практические применения.

Одной из возможностей использовать распознавание позы человека является виртуальная реальность. Оцифровка позы человека с помощью неросетей позволяет сэкономить на закупке дорогостоящий костюмов. Можно установить несколько камер, которые будут восстанавливать позу человека и переносить ее компьютерное пространство. При добавлении генеративных алгоритмов можно создавать всевозможные аватары и погрузиться в "Оазис"из фильма Стивена Спилберга "Первому игроку приготовиться".

Другим, уже вполне реальным, применением данной технологии является использование её в качестве рефери на спортивных соревнованиях. Уже сейчас система полуавтоматического определения офсайда активно помогает судьям футбольных матчей по всему миру. Эта система функционирует на основе распознавания ключевых точек частей тела футболистов, которыми они могут играть в мяч, что позволяет определить, были ли нарушены правила или гол был забит чисто [1]. Таким образом, технология существенно повышает точность и объективность судейства, уменьшая количество ошибок и спорных моментов в игре. (ССЫЛКА НА ЦИФРЫ ПРИ НАЛИЧИИ)

Продолжая тему спорта, следует отметить, что оценка позы может быть использована для анализа тренировок и создания персональных по-

мощников. Уже существуют несколько решений, направленных на анализ вашей игры в большой теннис, которые способны оценивать текущие результаты, указывать на области, требующие улучшения, и предлагать рекомендации по коррекции техники [2, 3]. Существует также проект MediaPipe от Google, предоставляющий публичные интерфейсы для анализа спортивной активности на основе распознавания ключевых точек [4]. Этот проект не ограничивается только оценкой и классификацией асан йоги. Он также включает функции для подсчёта количества повторяющихся упражнений, таких как отжимания, подтягивания и приседания.

Применений данной технологии можно придумать множество, но для их реализации необходима модель, которая будет работать быстро, поддерживая режим реального времени, а также демонстрировать высокие показатели точности своей работы. Однако обучение модели и разработка алгоритма её работы представляют собой чрезвычайно сложный и трудоемкий процесс. Этот процесс требует значительных ресурсов, как со стороны специалистов, так и в плане вычислительной мощности. Сначала необходимо собрать и подготовить данные, затем обучить модель, настроить её параметры и протестировать на различных наборах данных, чтобы убедиться в её точности и надёжности. Это часто занимает много времени и требует значительных финансовых вложений.

В тоже время, датасеты, на которых обучаются модели часто имеют общий характер и могут вносить сильную погрешность в результаты при изменении общих характеристик входных данных. Например, модель, обученная распознавать объекты на дневных фотографиях, может показывать низкую точность на ночных снимках из-за разницы в освещении и визуальных характеристиках. Поэтому важно проводить дополнительное обучение модели на данных, соответствующих конкретной задаче.

Этот процесс также требует значительных усилий, аналогичных созданию универсального решения. В связи с этим ученые задумались над тем, как можно уменьшить объем дополнительных работ по приспособлению модели к новым проблемам, не теряя в ее производительности. Таким образом появились алгоритмы доменной адаптации (англ. domain adaptation) и переноса обучения (англ. transfer learning).

Суть данных подходов состоит в том, чтобы преодолеть разрыв между исходным и целевым доменами данных. Это достигается путем выравнивания распределений данных, адаптации признаков и применения обученных моделей к новому домену с минимальной дообработкой. Таким образом, модели становятся более гибкими и способны эффективно работать в различных условиях, не требуя значительного объема новых данных или переработки архитектуры. В конечном итоге этоне только ускоряет процесс внедрения, но и снижает затраты на разработку, поскольку уменьшает необходимость в полном цикле обучения новой модели.

Оценив все вышеизложенное, можно сделать вывод, что методы доменной адаптации для моделей распознавания ключевых точек на теле человека имеют значительный потенциал. Различные приложения требуют адаптации к специфическим условиям съемки, будь то освещение, ракурс или качество изображения. А в разнообразных видах спорта, акцент может быть смещен на разные части тела: в футболе важны ноги и корпус, а в баскетболе — руки и верхняя часть туловища.

Стоит отметить, что возможность адаптировать нейронную сеть к целевому набору данных имеет особую важность в задаче распознавания ключевых точе на теле человека. Различные приложения требуют адаптации к специфическим условиям съемки, будь то освещение, ракурс или

качество изображения. А в разнообразных видах спорта, акцент может быть смещен на разные части тела: в футболе важны ноги и корпус, а в баскетболе — руки и верхняя часть туловища. Отсюда требуется возможность быстрого и дешевого улучшения качества работы нейроной сети.

С учетом вышесказанного, напрашивается вывод, что тема данной работы является полезной и важной в нынешних реалиях. В разд. 5 произведен обзор как различных моделей распознавания ключевых точек на теле человека (см в разд. 5.1), так и некоторых методов доменной адаптации (см в разд. 5.2). Также проведен эксперимент по применению *Progressive unsupervised learning (PUL)*, который хорошо себя показал в задачах детекции объектов и повторной реидентификации, к оценке позы. Для него был собран и размечен целевой набор данных, описанный в разд. 6. Результаты эксперимента дают ход дальнейшим исследованиям применения PUL, за счет вариативности способов отбора псевдоразметки.

ПОПРАВИТЬ ОКОНЧАНИЕ. КАК БУДТО НЕКРАСИВОЕ

2 Сверточные нейронные сети

Добавить этот раздел при наличии времени и только при необходимости увеличения объема работы.

3 Распознавание ключевых точек

С развитием технологий человечество начало ставить все более разнообразные задачи, для решения которых применялись сверточные нейронные сети (convolutional neural network, CNN). Одной из таких задач оказалось распознавание ключевых точек (Keypoint Detection). При этом распознавание ключевых точек на теле человека выделилось в отдельный раздел, известный как оценка позы (Pose Estimation). Далее рассмотрим формулировку этих задач.

3.1 Ключевые точки

Задача распознавания ключевых точек заключается в том, чтобы обнаружить и точно локализовать определенные точки или места внутри изображения или кадра видео. Эти точки, называемые ключевыми, могут быть определены для различных объектов, таких как лица, тела человека или других структур. Например, в случае распознавания лица ключевые точки могут включать углы глаз, кончик носа, уголки рта и другие характерные особенности лица. В контексте человеческого тела ключевыми точками могут быть суставы, такие как локти, колени, плечи и так далее. Для других объектов ими могут быть выбраны уникальные элементы, которые помогают идентифицировать и анализировать заданный предмет.

Другими словами, ключевые точки являются структурными, которые используются для определения положения и/или местонахождения объекта в пространстве. Они играют важную роль в задачах компьютерного зрения, таких как отслеживание движений, 3D-моделирование, анимация, медицинская визуализация и другие. Они могут использоваться для создания каркасов объектов, анализа их формы, измерения расстояний между различными частями и выполнения других аналитических

задач.

Более точно эти объекты можно определить следующим образом: *ключевые точки (КТ)* - это специфические, заранее определенные части распознаваемого объекта, которые имеют особое значение для дальнейшего анализа местоположения объекта на изображении. Каждая ключевая точка обычно соответствует определенной анатомической или структурной особенности, которая легко распознается и может служить ориентиром для алгоритмов обработки изображений. Иными словами, КТ необходимо обладать следующими характеристиками, чтобы можно было использовать их в качестве референсных для заданного объекта:

1. *Уникальность*

   Точки должны быть уникальными и отличаться от других точек на изображении

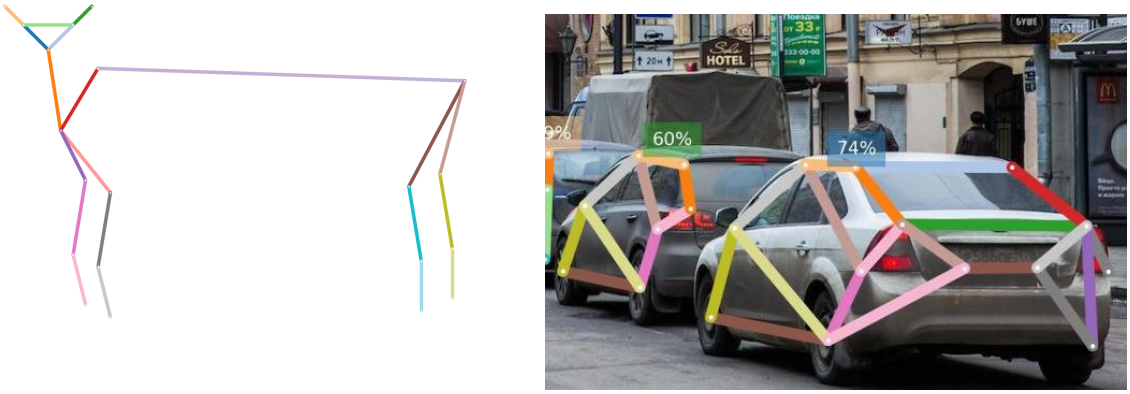
2. *Инвариантность*

   Точки должны сохранять свою идентичность при общих преобразованиях изображения, таких как вращение, масштабирование и изменения условий освещения

3. *Повторяемость*

   Точки должны быть обнаруживаемы в разных экземплярах одного и того же объекта или сцены

Для анализа структуры объекта и взаимосвязи между его различными КТ часто применяется схематичное описание, которое обеспечивает более наглядное визуальное представление. Этот метод помогает лучше понять анатомическую и функциональную структуру распознаваемого предмета. При этом саму схему, представляющую собой своего рода «скелет», принято называть *топологией*. Примеры различных топологий представлены на рис. 1.



(a) Animal Keypoints (b) Car keypoints

Рис. 1: Примеры топологий объектов от OpenPifPaf [5]

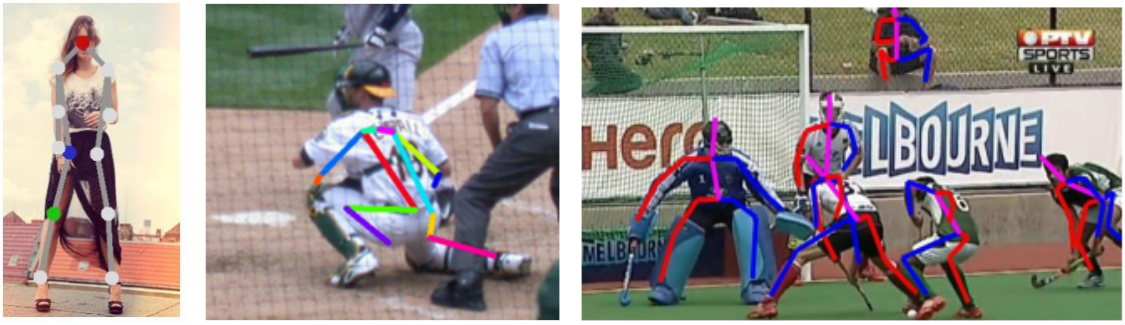
Математическая постановка задачи

СДЕЛАТЬ ПРИ НАЛИЧИИ ВРЕМЕНИ

3.2 Распознавание ключевых точек на теле человека. Скелет человека

На теле человека тоже можно выделить несколько ключевых точек, информация о которых дает возможность цифровизовать позу человека и использовать ее для аналитики с помощью методов машинного обучения. Именно для этого и была разработана задача распознавания ключевых точек на теле человека или, как ее часто называют в англоязычной литературе, задача оценки позы человека (Human Pose Estimation).

Основной вопросом для HPE стал выбор набора ключевых точек и топологии, по которой они будут соединяться. В первых работах были представлены различные примеры топологий, некоторые примеры представлены на рис. 2. Если точки туловища имеют большое количество пересечений, то точки головы сильно отличались: где-то учитывалось только положение головы (то есть добавлена верхняя точка головы), где-то рассматривались некоторые точки на лице, где-то голова вообще не учи-



(a) PDJR [8] (b) DeepPose [9] (c) AlphaPose [10]

Рис. 2: Примеры различных топологий у первых решений задачи распознавания ключевых точек на теле человека

тывалась. Все зависело от задачи и возможностей исследователей. Позде в 2015 году Microsoft выпустило набор данных с детальным описанием 17 точек на теле человека и запустила соревнование по распознаванию этих точек [6, 7]. Исследователей это заинтересовало и они начали адаптировать свои модели под топологию, описанную в датасете COCO [6]. Хотя данные в наборах перестали обновлять после 2017 года, многие новые можели до сих пор оценивают свои можелли по набору данных COCO. Отсюда и получилось, что данная топология стала основной для задачи оценки позы.

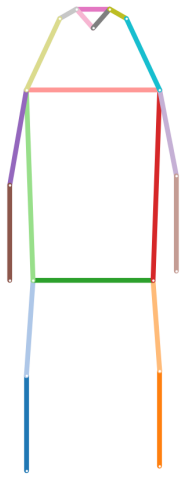


Рис. 3: Топология COCO

Другим вопросом для описываемой задачи стал подход к распознаванию. Делать детекцию человека и уже на кропнутом изображении производить поиск или искать все возможные точки, а потом собирать их в скелет. Эти две идеи и сформировали два направления развития методов HPE. Немного о них:

1. *Подход сверху-вниз (англ, top-down)*

   Для данного этапа вам понадобится дополнительная модель детекции, которая локализует человека на изображении, выделяя его в прямоугольную область. Затем этот прямоугольник передается в модель распознавания ключевых точек. Этот подход обеспечивает высокую точность предсказания КТ, но может быть чувствителен к ошибкам на этапе обнаружения и ориентации человека в кадре, что требует дополнительной предобработки изображений.

2. *Подход внизу-вверх (англ, bottom-up)*

   Для данного подхода вам не нужны помощники в виде детектора, но все равно имеет два этапа в своей работе. Первоначально модель распознает все ключевые точки на полученном на вход изображении, получая таким образом карту распределения КТ по фото. А вторым шагом модель собирает все полученные точки единый скелет. При детекции нескольких людей необходимо верно сопоставить их части тела. Для этого есть несколько способом, один из которых является построение полей сходства частей тела (part affinity fields), используемый в проекте OpenPose [11].

Но в результате обоих подходов к решению задачи на выход модели получаем массив данных [N x K x 3] в двумерном пространстве, где третьей размерностью идет предсказание видимости точки на изображении.

СТОИТ ВНИМАТЕЛЬНО ПЕРЕЧИТАТЬ ПОСЛЕДНИЙ АБЗАЦ. ВОЗМОЖНО ТРЕБУЕТ ПЕРЕДЕЛКИ

4 Domain Adaptation

В современном мире машинного обучения и искусственного интеллекта способность моделей адаптироваться к новым условиям и данным стала одной из ключевых задач. Традиционные методы обучения моделей предполагают, что данные, используемые для обучения и тестирования, имеют схожие характеристики и распределения. Однако в реальных приложениях часто возникает необходимость применять модели на данных, которые существенно отличаются от тех, на которых они были изначально обучены. Это приводит к снижению точности и эффективности моделей, что ставит под угрозу их практическое применение.

Для решения данной проблемы была разработана задача доменной адаптации, целью которой является создание методов для акклиматизации модели к целевым данным, отличающимся от исходных. В рамках этой задачи были разработаны подходы и техники, направленные на уменьшение расхождений между исходным и целевым доменами, что позволяет сохранять точность и производительность моделей в новых условиях. Эти методы способствуют переносу знаний, накопленных в исходном домене, на целевой, что относит данную задачу к области методов transfer learning.

4.1 Перенос знаний

Перенос обучения (англ. transfer learning) позводлил уже существующим решениям выйти за пределы первоначально заданных задач. Этот подход позволил использовать различные архитектуры для решения новых, разнообразных проблем. К примеру, это позволило перенести опыт использования трансформеров из задач обработки естественного языка в задачи компьютерного зрения.

Помимо переноса знаний между различными областями нейронных

сетей, transfer learning предоставляет возможность использовать опыт, накопленный в процессе обучения модели, для работы с новыми данными. Эти данные могут представлять собой не только новые классы в задачах классификации или кластеризации, но и иметь значительные структурные различия, такие как язык и жанр для текстовых данных или стиль и качество для изображений. Такие методы, называемые адаптацией к новым доменам данных, позволяют существенно сократить время и ресурсы, направленные на решение новых, узкоспециализированных задач.

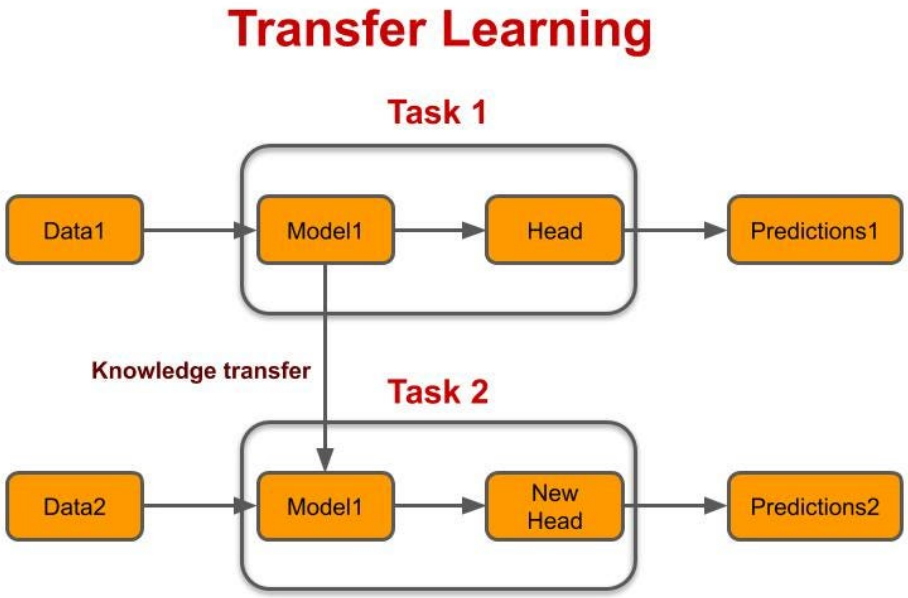


Рис. 4: Основная идея переноса знаний

ИСТОЧНИК

Основная идея состоит в том, чтобы взять модель, которая уже научилась извлекать и интерпретировать общие признаки из большого набора данных, может эффективно адаптироваться к новым данным, требующим более специфических знаний (см рис. 4). Это аналогично тому, как человек, обладая общими знаниями в одной области, может быстрее освоить смежную сферу деятельности. Таким образом получается не

только ускорить процесс обучения, но и повысить точность и эффективность модели при условии ограниченности в ресурсах.

4.2 Основные определения

Для дальнейшего повествования введем формальные обозначения для данной задачи.

Определение домена

Пусть существует пространство признаков χ и в нем задан набор данных X и частотное распределение вероятностей на нем $P(X)$:

$$X = X_1, ..., X_n \in \chi.$$

В таком случае *доменом* D называют совокупность пространства признаков и частотного распределения веротности на заданном наборе данных:

$$D = \{\chi, P(X)\} \tag{1}$$

Определение задачи

Пусть на пространстве признаков χ задан набор данных

$$X = \{X_1, ..., X_n\} \in \chi. \tag{2}$$

Пусть задано пространство меток γ и для заданного набора данных существует соответственный набор меток

$$Y = \{Y_1, ..., Y_n\} \in \gamma \tag{3}$$

Тогда *задача* T определяется как

$$T = \{Y, f(X)\}, \tag{4}$$

где f - прогностическая функция зависимости целевой переменной, которую можно рассматривать как условное вероятностное распределение $P(Y|X)$.

Определение доменной адаптации

В задаче доменной адаптации подразумевается наличие не менее двух наборов данных:

1. *Исходный домен (Source)*

   Представляет собой универсальный набор данных большого объема. Обозначается следующим образом:

   $$D^s = \{\chi^s, P(X^s)\}$$

   На домене определена задача:

   $$T^s = \{Y^s, f^s(X^s)\}$$

2. *Целевой домен (Target)*

   Представляет собой набор данных маленького объема, к которому планируется адаптировать модель. Обозначается следующим образом:

   $$D^t = \{\chi^t, P(X^t)\}$$

   На домене определена задача:

   $$T^t = \{Y^t, f^t(X^t)\}$$

В зависимости от того, как соотносятся между собой домены D^S и D^T и задачи T^s и T^t, ставятся различные задачи:

1. Самым распространенным является случай совпадения задач и доменов, когда $D^s = D^t$ и $T^s = T^t$. В таком случае говорят о постановке задачи классического машинного обучения. Домен D^s называют

обучающей выбркой, которая используется для обучения решения T, а домен D^t - тестовой выборкой, на которой производится оценка правильности полученного решения.

2. При совпадении данных $D^S = D^t$, но различной постановке задачи $T^s \neq T^t$ говорят о мультизадачном обучении. Часто такая постанвка задачи подходит для адаптации модели классификации к новым меткам классо, которые представлены в целевой выборке, но отсутствуют в исходной.

3. При совпадении задач $T^s = T^t$, но различии в данных $D^s \neq D^t$ говорят о методах кросс-доменной адаптации. В таком случае ставится задача улучшения предсказания T^t с использованием информации, полученной в T^s.

4. При полном несовпадении данных $D^S \neq D^t$ и задач $T^s \neq T^t$ подразумается, что пары (D, T) решают разные проблемы и адаптация происходит в индивидуальном порядке. Например, использование архитектуры трансформер, зарекомендовавшей себя в задачах обработки естественного языка, для задач компьютерного зрения.

В текущей работе рассмотривается третий вариант: $D^s \neq D^t$, но $T^s = T^t$. Схематически постановка задачи описана на рис. 5, а формулировка задачи математическим языком выглядит следующим образом:

$$D^s, T^s \rightarrow D^t, T^t$$

4.3 Классификация методов доменной адаптации

Ранее методы, которые решают задачу *доменной адаптации (англ. domain adaptation)* были определены как методы переноса знаний, направленные на на приспособление моделей $T^s = T^t = T$, обученных на данных из

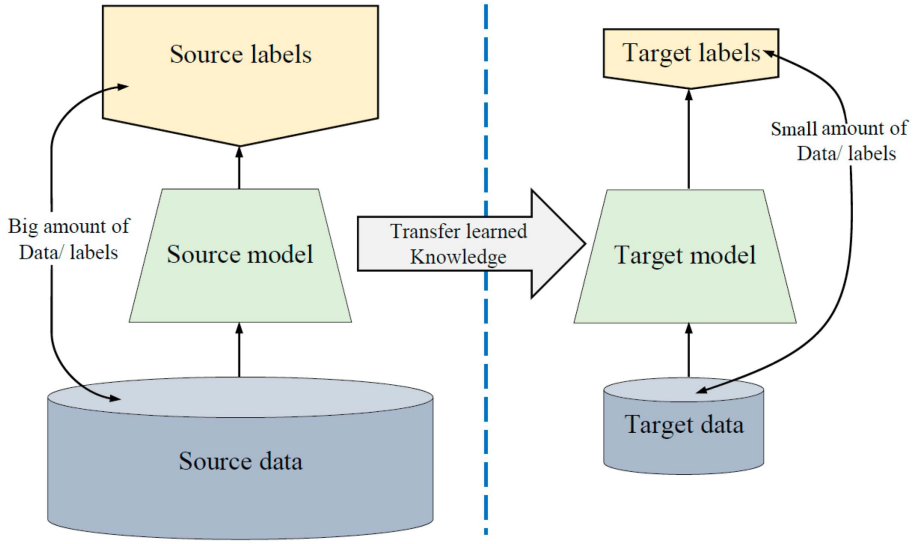


Рис. 5: Схематическое представление работы алгоритмов доменной адаптации

домена D^s, к данным из домена D^t, причем $D^s \neq D^t$. Данные могут различаться по различным характеристикам и в зависимости от различают несколько типов методов, которые схематически представлены на рис. 6.

В зависимости от того, насколько сопоставляются распределения наборов данных $P(X)$ различают гомогенные и гетерогенные методы.

1. *Гомогенные методы (Homogeneous DA)*

   Данный тип методов применяется, когда исходный и целевой домены имеют одинаковое пространство признаков $\chi^s = \chi^t = \chi$, но их распределения отличаются $P(X^s) \neq P(X^t)$. Методы гомогенной адаптации фокусируются на выравнивании распределений признаков между доменами, чтобы модель, обученная на исходных данных D^s , могла эффективно работать на целевых данных D^t.

2. *Гетерогенные методы (Heterogeneous DA)*

   В отличие от предыдущих способов, в текущем случае отличаются и признаковые пространства можду доменами $\chi^s \neq \chi^t$. Это пред-

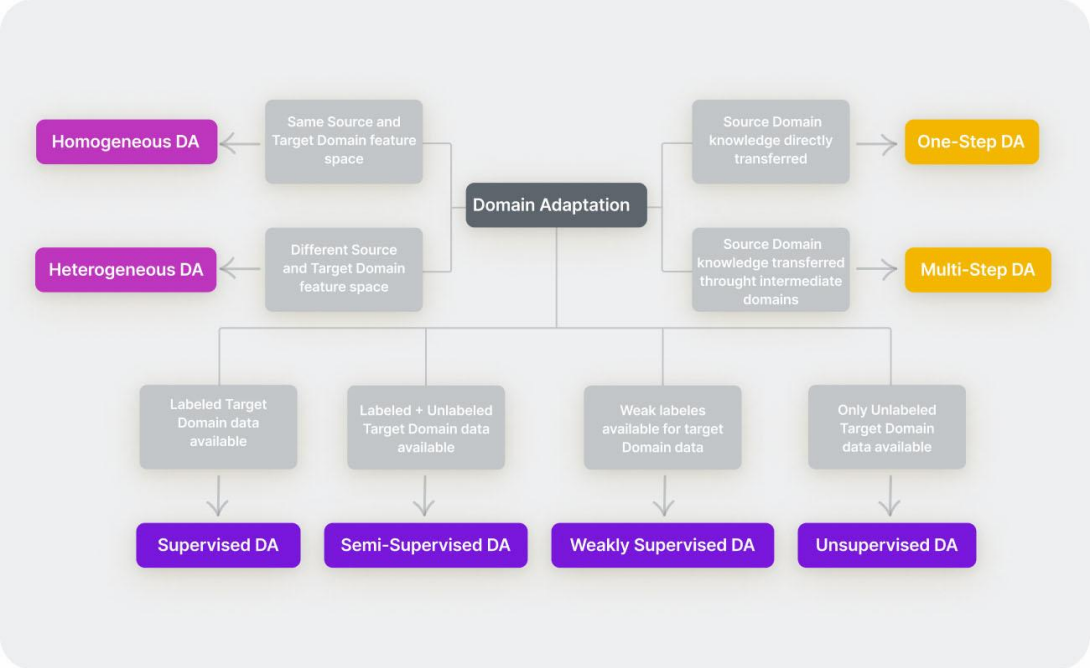


Рис. 6: Схематическое представление работы алгоритмов доменной адаптации

ставляет собой более сложный сценарий, так как необходимо либо преобразовывать целевые данные в такое представление, которое будет сопоставимо с исходными, либо выуживать обобщенные признаки из обоих доменов, на которых модель будет считать данные похожими.

Также стоит обратить внимание насколько сильно отличаются распределения данных, так как от этого зависит сколько шагов необходимо будет производить между доменами:

1. *Одношаговая доменная адаптация (One-step DA)*

   При небольших различиях между распределениями можно использовать прямой перенос знаний из D^s в D^t. В таком случае опыт, накопленный моделью в исходном домене, напрямую используется при адаптации модели T на целевом домене. Преимущества этого

метода заключаются в его простоте и быстроте внедрения, так как он не требует промежуточных шагов.

2. *Многошаговая доменная адаптация (Multi-step DA)*

   Однако эффективность One-Step DA может снижаться, когда различия между исходным и целевым доменами слишком велики для прямого переноса знаний. Для этих случаев и используются методы многошаговой адаптации. Они подразумевают поэтапное выполнение: знания из исходного домена сначала переносятся в один или несколько промежуточных доменов, прежде чем достигнуть целевого. Промежуточные шаги помогают постепенному выравниванию распределения данных, что улучшает адаптацию и повышает точность модели в целевом домене.

Последним маркером в классификации методов доменной адаптации является доступность размеченных данных для D^t. По данному признаку методы разделяются сразу на 4 группы:

1. *Supervised domain adaptation*

   В этом случае обучение модели происходит с использованием как исходных, так и целевых данных, имеющих метки. Поэтому данные методы обычно достигают высокой точности, поскольку модель может явно учиться на целевых данных с метками.

2. *Semi-supervised domain adaptation*

   К данным методам прибегают, когда в целевом домене присутствуют как размеченные данные, которые используются для начального обучения, так и неразмеченные данные, на которых есть возможность улучшать работу модели.

3. *Weakly supervised domain adaptation*

   Когда данные целевого домена слабо размечены, то есть использовалась автоматическая система разметки, которая допускает ошибки, используются методы weakly supervised DA. Они направлены либо на улучшение точности слабых меток, либо используют «мягкие» техники регуляризации, что позволяет учитывать возможность ошибки в данных.

4. *Unsupervised domain adaptation*

   Самый сложный случай, когда нет возможности аннотировать домен и надо на сырых данных улучшить качество модели. Для методов этого типа часто используется кластеризация данных и итеративные подходы, которые помогают модели T улучшать собственные предсказания.

5 Обзор существующих решений

5.1 Обзор моделей для распознавания ключевых точек

В данном разделе будут рассмотрены несколько моделей распознавания ключевых точек на теле человека. Некоторые из них довольно старые, но показывают неплохие результаты. Большинство же представлены не более 5 лет назад и являются лидерами направления на сегодняшний день.

DeepPose

Данный представитель является самым старым решением из данной выборки и, одновременно, один из самых первых в целом. Статья "DeepPose: Human Pose Estimation via Deep Neural Networks"[9] была представлена исследователями из GOOGLE на конференции CVPR в 2014 году.

Исследователи разработали модель, представляющую сосбой каскад из DNN-регрессоров для локализации ключевых точек. Так как на тот момент не было выпущено общепринятых топологий, то в роли ключевых точек выступали суставы тела, а поза кодировалась их координатами, нормализованными на размер изображения.

Рис. 7: Архитектура сети DeepPose. [9]

Работа модели делилась на два этапа, которые схематически показаны на рис. 7. На первом производилась локализация точки. Далее ре-

зультаты переходили на второй этап, где пропускались через каскад из DNN, который производил уточнение предсказания. В результате получался отосительно точный результат, который можно было использовать для дальнейших исследований.

OpenPose

OpenPose - это проект от лаборатории перцептивных вычислений университета Карнеги-Меллона в США. Проект включает в себя несколько моделей распознавания ключевых точек: на лице, руках, теле и различные их комбинации. Количество распознаваемых моделями точек доходит до 135 на одном человеке, а модель может распознавать сразу несколько человек на одном кадре. Скорость работы модели позволет использовать ее для распознавания видео в реальном времени через веб-камеру. К сожалению, основной репозиторий проекта не поддерживается с ноября 2020 года.

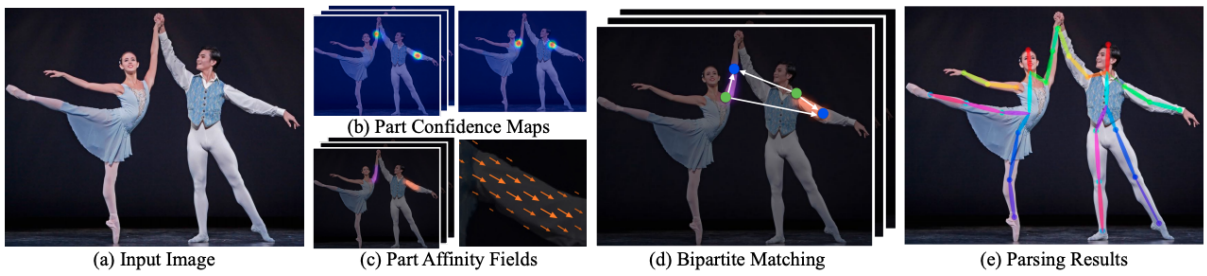


Рис. 8: Последовательность распознавания ключевых точек моделью OpenPose. [11]

На данный момент обратимся к модели, резльтаты которой описываются тополгией из 25 точек. Данная модель одна из немногих в представленном обзоре, которая использует подход снизу вверх. Это достигается за счет использования двухступенчатой архитектуры нейросети (см рис. 8).

В рамках первого этапа (вторая колонка на рис. 8) нейросетью пред-

сказываются такие сущности, как карта достоверности обнаружения точки и карта двумарных векторных полей ориентации конечностей (англ. Part Affinity Fields или PAF). Изначально предсказание сущностей выполнялось паралельно, но взаимный анализ результатов предсказаний позволил увидеть возможность интуитивно предсказывать карты достоверности на основе PAFs. Это позволило значительно ускорить работу алгоритма.

Вторым этапом (треья колонка на рис. 8) происходит сопоставления точек и конечностей отдельным людям. Для увеличения точности и уменьшения времени построения скелетов используются графы соответсвия, которые позволяют создать целостные и непротиворечивые представления поз людей на изображении.

HRNet

Решение на основе архитектуры HRNet, опубликованной в 2019 году, является одним из первых, представленных в проекте MMPose. Оно использует подход сверху-вниз, в котором для детекции используется модель из проекта MMDetection.

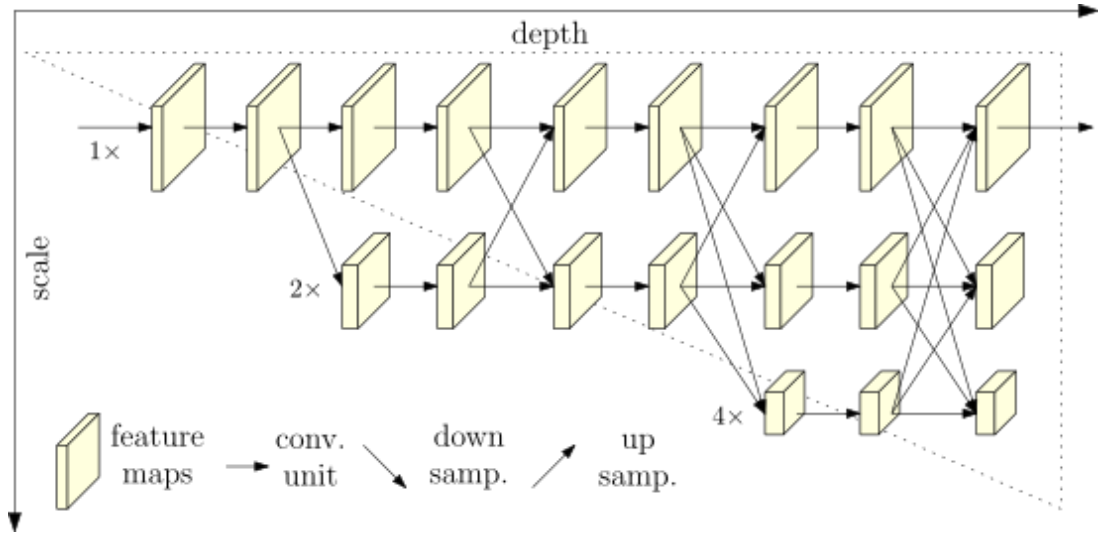


Рис. 9: Архитектура модели HrNet. [12]

Особенность архитекртуры HRNet является возможность поддер-

живать высокое разрешение изобажений на всех этапах обработки, что позволяет достичь высокой точности и детальности предсказаний. Достигается это за счет параллельного использования нескольких ветвей с разным разрешением и постоянного обмена информацией между ними (см. рис. 9).

Все начинается с уровня слоев высокого разрешения, к которому паралельно добавляются слои с пониженным разрешением. Периодически между ветвями происходит обмен информацией с помощью fusion module. Этот процесс обеспечивает сбалансированное и детализированное представление на разных уровнях разрешения, что приводит к улучшению точности получаемого результата.

BlazePose

Архитектура нейросети BlazePose разработана в 2020 году исследователями от GoogleResearch и известна своим использованием в работах проекта MediaPipe [13, 4]. Она предназначена для бытрого и точного распознавания ключевых точек тела в реальном времени, даже на мобильных устройствах и в условиях ограниченных вычислительных ресурсов. Расширенная до 33 точек топология модели помогает исполь-

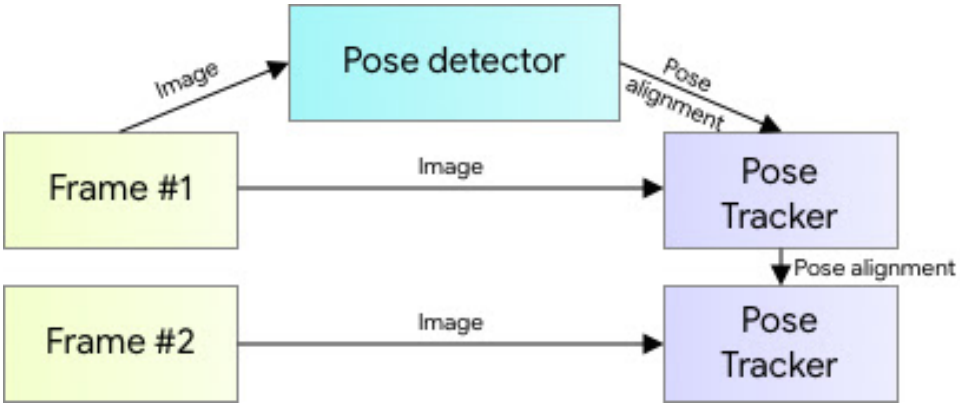


Рис. 10: Структура модели BlazePose для работы в реальном времени. [13]

зовать нейросеть в разнообразных сопртивных приложениях, таких как фитнес-трекеры и анализаторы асан йоги.

BlazePose использует top-down подход оценки позы, который оптимизирован для работы с видеопотоками. Схематично, структура нейросети представлена на рис. 10. На первом этапе необходимо найти человека на входном изображении, чем и занимается PoseDetector. Но он вызывается только для первого кадра и возвращает не только координаты облати с человеком, а информацию об интересующей нас области (region of interest или ROI). Именно ROI дает возможность не звызывать каждый раз детектор, так как она изменяется на втором этапе работы сети и передается сразу для использования в Pose Tracker следующего кадра.

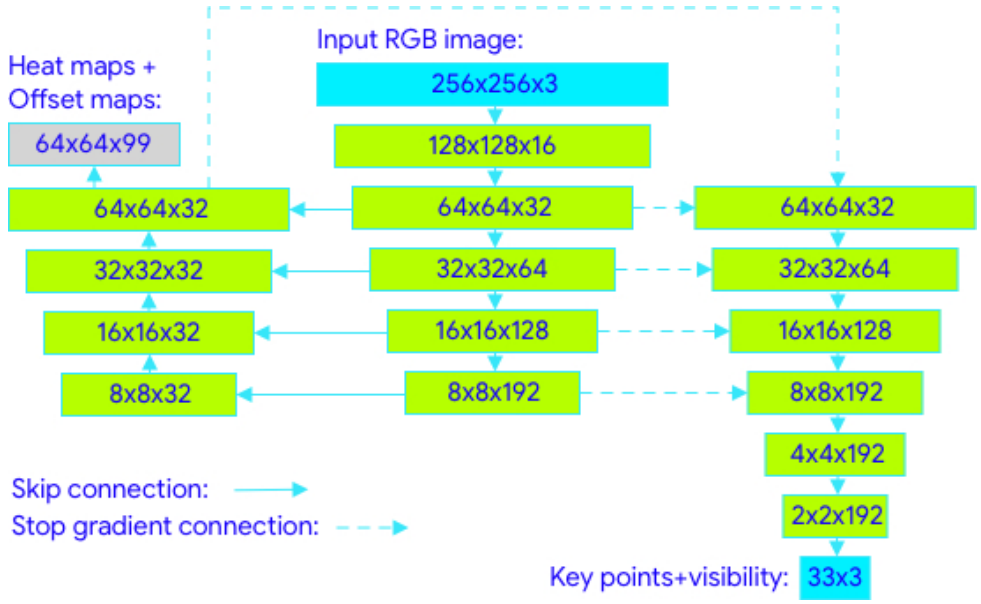


Рис. 11: Архитектура Pose Tracker. [13]

Следующим шагом является распознавание КТ в заданной области интереса. Путем использования нескольких пирамидальных архитектур производится комбинированны анализ тепловой карты и данных о смещении (см. рис. 11).

ViTPose

ViTPose - это модель для распознавания ключевых точек, которая была представлена в 2022 году командой исследователей из университета Сиднея [14]. В ее основе используется архитектура трансформера, которые изначально были разработаны для обработки последовательностей в задачах обработки естественного языка, адаптированного для задач компьютерного зрения (англ. Vision Transformer или ViT) [15]. Данный подход продемонстрировал высокую эффективность и конкурентоспособность по сравнению с традиционными свёрточными нейронными сетями (CNN), что позволило область решений задачи оценки позы.

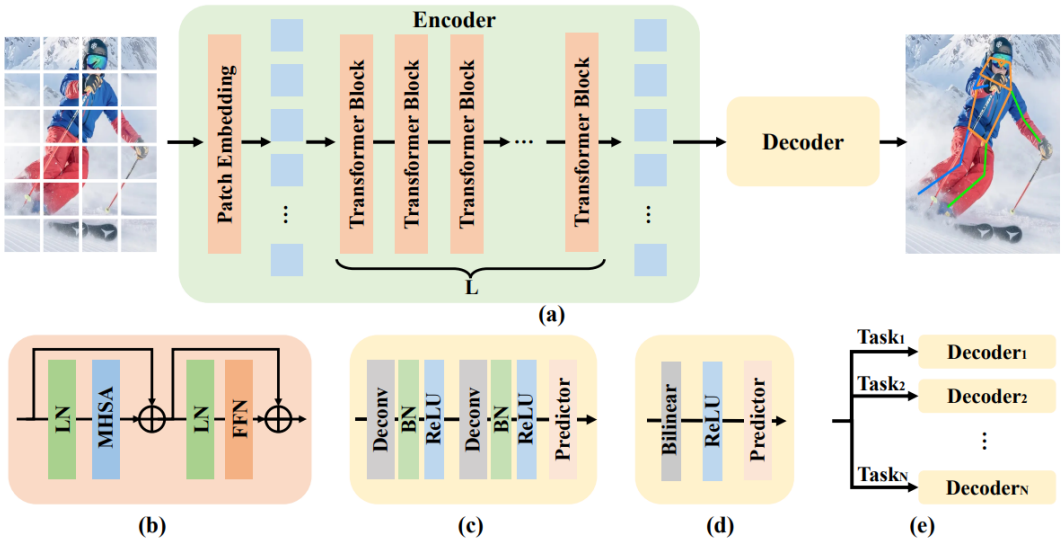


Рис. 12: Структура VitPose (a). Блоск трансфрмера (b). Классический декодер (c). Простой декодер (d). Декодеры для нескольких наборов данных (e). [14]

Для создания анализируемой последовательности изображений разбивается на патчи рамером $16x16$. Они прогоняются через линейные слои нейросети для получения эмбеддинга, к которому дополнительно добавляется позиционный эмбеддинг, чтобы модель могла учитывать пространственное расположение блоков. И потом уже полученная цепочка

векторов передается на вход магистральной части нейронной сети. Используя механизмы самовнимания (англ. self-attention), которые помогают использовать различные взаимосвязи между частями изображения, backbone возвращает признаковое описание изображения.

Для извлечения признаков и локализации ключевых точек из результатов магистральной части испоьзуется два типа декодеров. Классический декодер использует несколько блоков для повышения дискретизации тепловой карты ключевых точек, а для локализации координат используется сверточный слой в качестве выходного слоя. Другой декодер, называемый в работе «простым» [14], использует один слой билинейной интерполяции для повышения детализации результатов, а для получения тепловых карт используется функция активацции RELU. Предсказателем также выступает сверточный слой.

Simcc

Локализация ключевых точек на основе анализа тепловых карт достоверности является весь распространенным подходом в задаче оценки позы. Но даже этот подход имеет некоторые недостатки: плохие результаты на изображениях низкого разрешения, вычислительно тяжелые слои повышения дискретизации тепловых карт и дополнительная постобработка для уменьшения ошибок квантования. Чтобы избежать их исследователями был разработан алгоритм Simple Coordinate Classification или SimCC [16], который предлагает новый подход к оценке поз человека с помощью классификации отдельно горизонтальных и вертикальных координат. Схематическое описание алгоритма представлено на рис. 13.

Для извлечения признаков в данном алгоритме используется магистральная нейронная сеть, которой могут быть как сверточные нейронные сети, так и трансформеры. Она возвращает признаковое описание необходимого количества ключевых точек.

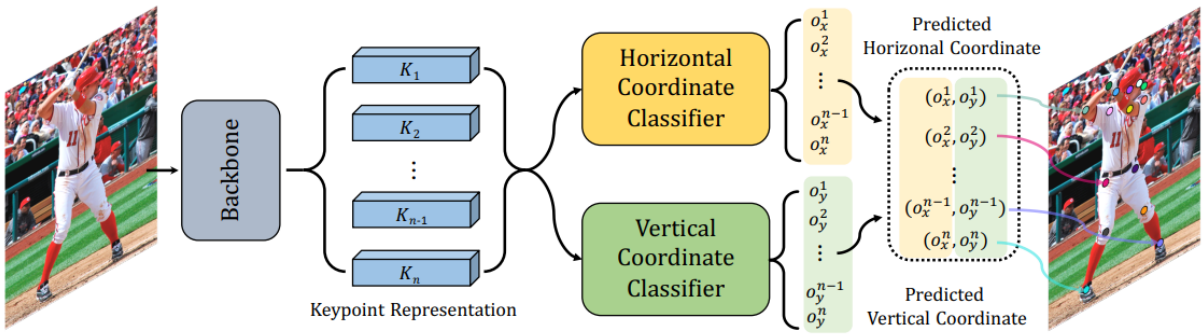


Рис. 13: Структура SimCC [16]

Для дальнейшего шага все координаты дискретезируются на более мелкие бины, чтобы избежать ошибок квантования. Далее классификатор выставляет метки всем бинам, причем делает это независимо для вертикального и горизонтального направления. А для обучения модели используется функция потерь на основе дивергенции Кульбака-Лейблера (англ. Kullback-Leibler divergence).

В итоге алгоритм показывает немного более лучшие результаты, в сравнение с решениями на основе анализа тепловых карт, но сильно уменьшает количество проводимых операций, чем значительно улучшает скорость работы модели.

YoloPose

Yolo-Pose - это семество решений на основе архитектуры Yolo, которое представляет собой новаторский подход для одновременного обнаружения нескольких человек на фото и распознавания их скелета. Алгоритм был предъявлен публике в 2022 году исследователями из Техаса [26]. В своей статье они использовали архитектуру Yolo5, которая показывала хорошие результаты распознавания ключевых точек на датасете COCO. На сегодняшний день к семейству Yolo-Pose подтянули и других представителей популярной архитектуры, таких как Yolo8, YoloX и Yolo-NAS [28, 27].

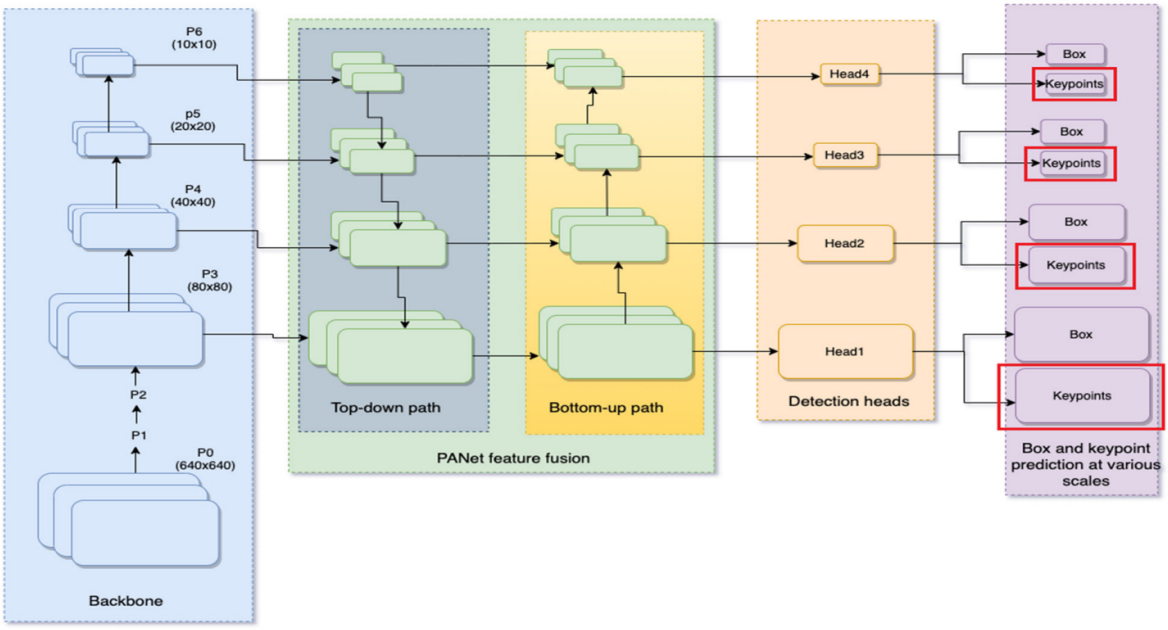


Рис. 14: Архитектура Yolo-Pose на основе модели Yolo5. [26]

Авторы решили пересмотреть использование различных подходов к распознаванию ключевых точек и совместили лучшее из обоих вариантов. Большинство подходов сверху-вниз используют тепловые карты для предсказания координат ключевых точек. Как уже было сказано в разд. 5.1, детализация тепловых карт занимает множество вычислительных ресурсов. А ещё производительность этих решений напрямую зависит от количества человек на изображении, так как все они распознаются по отдельности. В свою очередь подходы снизу-вверх имеют сложность при сопоставлении ключевых точек отдельным людям. В итоге получилось построить решение, которое производит поиск человека на фото, присваивает ему некоторую якорную точку (англ. anchor point), с которой в последвтии ассоциируются все ключевые точки и ограничивающая рамка. А разделение на две сущности: bbox и keypoints происходит на этапе обработки полученных признаков (head) и предсказания результатов.

Полученный алгоритм сквозного обучения не мог работать на существующих L1 метриках, поэтому исследователи оптимизировали метрику

OKS для обучения моделей (о метрике рассказано в разд. 6). Это дало возможность учитывать весовые коэфициенты точек и масштаб объекта при обучении.

Но у моделей этого семейства есть и минус. Это требования большого количества вычислительных ресурсов для обучения моделей. Что делает практически невозможным улучшение модели простым обывателем.

RTMPose

Из текущей подборки данная архитектура является самой молодой. Алгоритм был представлен в 2023 году и показывает весьма зорошие результаты. Посмотрим же что они дают)

2023 год. На момент исллледования наиболее актуальная модель от проекта MMPose.

Работает неплохо. Обучается тоже неплохо. Попробуем заюзать в экспериментах.

5.2 Обзор методов доменной адаптации на неразмеченных данных

Progressive Unsupervised Learning

Встречаясь с чем-то новым, люди могут несколько раз рассматривать, пробовать и анализировать новый предмет, прежде чем получат необходимые знания о нем. Похожую схему использует алгоритм прогрессвного обучения без учителя (англ. progressive unsupervised learning или PUL). Его концепция позволяет избавиться от необходимости аннотировать данные для дообучения нейронных сетей на них, тем самым позволяя моделям адаптироваться к новому домену. Первоначально модель

была применена в задаче отслеживания объектов [17], а позже использована в задаче повторной идентификации человка [18].

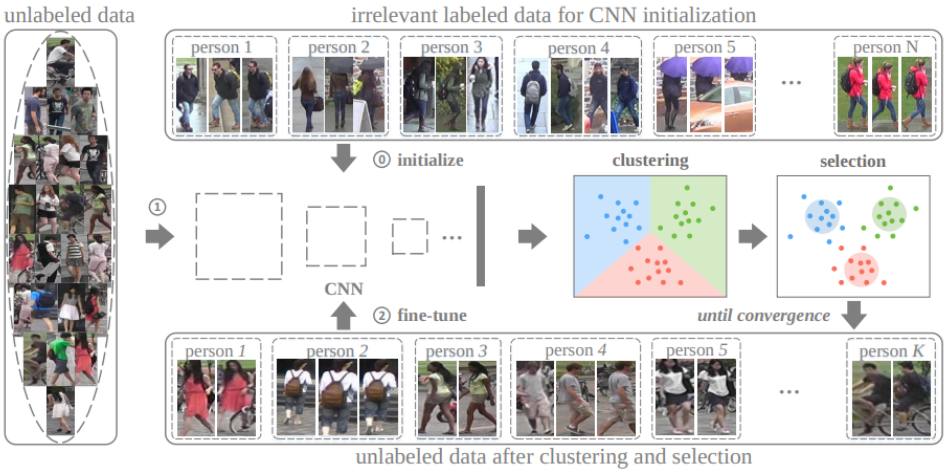


Рис. 15: PUL [18]

Основная концепция алгоритма, как метода unsupervidsed domain adaptation, состоит в отсутствии необходимости размечать данные, а адаптироваться к ним посредством итеративного обучения на псевдоразметке. Опишем действия, которые будут выполняться повторно.

1. Входные данные

   Для начала работы алгоритма необходимо иметь некоторое превариательно обученное состояние модели.

2. Псевдоразметка данных

   Неразмеченные данные прогоняются через модель для получения некоторых результатов. Из-за того, что модель имеет некоторую ошибку, текущие результаты называются псевдо-разметкой.

3. Фильтрация невалидных результатов

   Модель может значительно ошибаться в предсказании псевдо-разметки, поэтому необходимо обозначить некую функцию фильтрации шумных значений. Например, при применении PUL к задаче повторной

идентификации [18], была построена интересная функция фильтрации на основе кластеризации вектора признаков изображения и отбрасывания всех значений, значительно далеких от центра кластера.

4. Обучения модели

   Производится дообучение модели на полученных данных. Состояние модели после дообучения передается на вход следующей итерации алгоритма.

Маркером оценки итеративного процесса следует рассматривать требуемую точность на целевом домене данных. Если требуемая точность не достигается, то следуют остановиться при незначительных улучшениях значений метрик работы модели.

СТОИТ ПЕРЕСМОТРЕТЬ ПОСЛЕДНИЙ АБЗАЦ

RegDA

Интересный алгоритм. На нем базируются все остальные

НАДО БУДЕТ ОСТАВИТЬ ТОЛЬКО ОДИН ИЗ СЛЕДУЮЩИХ МЕТОДОВ.

UDA PoseEstimation

Адаптация от синтетических данных к реальным. Должно бтыь интересно описать. Не использовали, так как до сих пор мы не перешли в 3х мерное пространство.

POST

Похожее на предыдущее

SFDA

Похожее на предыдущее

6 Эксперимент

6.1 Описание эксперимента

Постановка эксперимента. Что планируется сделать и какие результаты хочется получить. Какие метрики будем использовать и по каким метрикам будем сравнивать.

Выбор модели для эксперимента

В рамках поставленного эксперимента поставлена задачи исследования работы алгоритма PUL для задачи распознавания ключевых точек для нескольких моделей:

1. HRNet
2. ViTPose
3. RTMPose
4. SimCC + ResNet ???

Предложенные модели будут обучены на исходном датасете в течение 20 эпох. Таким образом будут получены базовые наборы весов, от которых и будет дпроводиться дальнейшее исследование.

Описание метода доменной адаптации

Отбор точек будет произвед путем сравнения ЗНАЧЕНИЯ (ЗАМЕНИТЬ) с заранее заданным пороговым значением. ЗНАЧЕНИЕ будет отбираться двумя способами:

1. Средняя уверенность в предсказанных значениях

   Для каждго предсказанного результата модель возвращает значение

уверенности в своем предсказании. Усредняя это значении по всем ключевым точкам получаем среднюю уверенность для фотографии

2. Средняя уверенность для всех видимых точек

   Так как увереннсть на точках, которые не видно на фотографии может сильно занижать среднее ЗНАЧЕНИЕ для всего результата, то принято решение отбрасывать эти значения при высчитываании ЗНАЧЕНИЯ

По порогу уверенности будет отбираться набор значений с псевдоразметкой, на которой модель будет дообучаться. Таким образом будет проведено N (УКАЗАТЬ ТОЧНОЕ ЗНАЧЕНИЕ) итераций адаптации и сравнено значение результатов модели при различных способах отбора псевдоразметки. Также в рамках эксперимента произведено полное дообучение модели на размеченном целевом домене и результаты буду предоставлен для сравнения с адаптированными.

Метрики оценки качества распознавания

ТАКЖЕ НЕОБХОДИМО СКАЗАТЬ ОБ МЕТРИКАХ ОЦЕНКИ РАСПОЗНАВАНИЯ

Для проведения количественного анализа результатов эксперимента, необходимо выделить несколько метрик, с помощью которых можно будет отобрать наилучшие результаты.

PCK

ФОРМУЛА И ОБЪЯСНЕНИЕ

OKS

ФОРМУЛА И ОБЪЯСНЕНИЕ

AP/mAP

ФОРМУЛА И ОБЪЯСНЕНИЕ

Используемые ресурсы

ТАКЖЕ НЕОБХОДИМО СКАЗАТЬ ОБ ИСПОЛЬЗУЕМЫХ РЕСУРСАХ

6.2 Данные

По условиям задачи доменной адаптации необходимо найти два набора данных для эксперимента. Далее приведем информацию о выбранных доменах и их характеристиках.

Исходные домен

В качестве исходного домена выбран набор данных Common Objects in Context [6]. COCO — это крупный датасет, широко используемый в области компьютерного зрения для задач распознавания объектов, сегментации, и создания описательных подписей к изображениям. Он был создан Microsoft и с тех пор стал стандартом для обучения и оценки алгоритмов компьютерного зрения.

Рис. 16: Примеры изображений из набора данных COCO. [7]

Учитывая, что в рамках соревнований COCO была и задача детек-

ции ключевых точек (Keypoint detection) [7], то часть этого набора данных была размечена под нее. Если быть точным, то датасет включает более 250 тысяч аннотаций людей на различных изображениях. Формат аннотаций включает в себя:

1. *id* - уникальный номер аннотации;
2. *image_id* - уникальный номер изображения, которому принадлежит данная аннотация;
3. *category_id* - уникальный номер категории, к которой относится данная аннотация. Для задачи оценки позы везде выставляется категория person;
4. *keypoints* - массив из 17 ключевых точек, для каждой их которых указаны координа (x, y) на изобраении, а также информация о видимости. Точки, которые не представлены на изображении заполняются нулями;
5. *num_keypoints* - здесь содержится информация о количестве размеченных точек для данной аннотации;
6. *bbox* - информация об обграничивающем человека прямоугольнике. Значения внутри лежат в следующем формате: $[x, y, width, height]$;
7. *area* площадь сегментированного человека. Значение необходимо при высчитывании метрики OKS;
8. *iscrowd* - информация о том, одиночный человек представлен на изображении или толпа людей.

Также в рамках задачи Keypoint Detection была введена метрика OKS и метрика mAP, о которых было рассказано ранее. Они представ-

ляют собой единые критерии для оценки моделей, что облегчает сравнение и улучшение результатов различных алгоритмов, поэтому регулярно используются для оценки новых методов и технологий.

В рамках задачи были выбраны 8000 аннотаций, которые содержат все 17 ключевых точек то топологии COCO. На них и было произведено обучение моделей для получения бейзлайнов эксперимента.

Целевой домен

В качестве целевого набора данных был собран отдельный набор данных боксеров. В наборе данных представлены 2 человека, снятые с 3 ракурсов: профиль, анфас и 3/4. Датасет состоит из 10 видеозаписей, которые содержат порядка 10 тысяч кадров. Для проведения эксперимента выбрано 2,6 тысячи изображений, которые были впоследствие размечены. Из них для тестовой выборки отобрано около 420 изображений, а оставшиеся 2200 составили обучающую выборку, на которой и проводилась адаптация.

ПРИМЕРЫ ИЗОБРАЖЕНИЙ

В рамках задачи по аннотированию собранных данных была разработана система полуавтоматической разметки изображений pose-markup [19]. Она представляет собой предобученную модель распознавания ключевых точек и инструмент для визуальной корректировки данных экспертом.

Для автоматической части использовалась модель BlazePose от проекта MediaPipe [4]. Выбор сделан благодаря высоким характеристикам скорости инференса результатов и их точности у данного решения. А также для того, чтобы избежать корреляции размеченных данных с предсказаниями, которые будут оцениваться в рамках эксперимента. Результат, возвращаемый моделью был преобразован к формату аннотаций COCO, который был описан выше и сохранен в формате JSON.

Как можно видеть на ИЗОБРАЖЕНИИ, модель имеет неточности, которые необходимо было исправить эксперту. Для этой цели использовалась вторая часть программы - инструмент для визуальной корректировки данных экспертом. В ТАБЛИЦЕ предоставлена сводная информация о количестве изменений, внесенных экспертом, а на ИЗОБРАЖЕНИИ представлена тепловая карта ключевых точек топологии, которые нуждались в корректировке.

6.3 Результаты эксперимента

Предоставить относительно сухо результаты эксперимента. Можно дать базовый анализ ситуации и того, что мы видим.

Необходимо предоставить результаты по времени обучения нейросетей, времени дообучения нейросетей.

Собрать данные по количеству ошибок до-после обучения. Собрать данные по количеству ошибок при изменении домена.

Данные по ресурсам, на которых обучались нейросетки.

7 Заключение

Необходимо рассказать о результатах исследования и о том, что мы получили. Привести сравнения, если таковые будут иметь место. Закинуть удочку для будущих исследований.

Список литературы

[1] *FIFA*. Semi-automated offside technology to be used at FIFA World Cup 2022. — https://www.fifa.com/fifaplus/en/articles/semi-automated-offside-technology-to-be-used-at-fifa-world-cup-202 — 2022.

[2] Player pose analysis in tennis video based on pose estimation / Ryunosuke Kurose, Masaki Hayashi, Takeo Ishii, Yoshimitsu Aoki // 2018 International Workshop on Advanced Image Technology (IWAIT). — 2018. — Pp. 1–4.

[3] *Thorpe, James*. Pose estimation: utilising AI to improve tennis technique. — https://sportretina.com/blog/pose-estimation-utilising-ai-to-improve-tennis-technique/. — 2023.

[4] *google.github.io*. MediaPipe.Home. — https://google.github.io/mediapipe/.

[5] *Kreiss, Sven*. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association / Sven Kreiss, Lorenzo Bertoni, Alexandre Alahi // *IEEE Transactions on Intelligent Transportation Systems*. — 2022. — Vol. 23, no. 8. — Pp. 13498–13511.

[6] *Lin, Tsung-Yi*. Microsoft COCO: Common Objects in Context. — 2014. https://arxiv.org/abs/1405.0312.

[7] *Tsung-Yi Lin Matteo Ruggero Ronchi, Alexander Kirillov*. COCO 2020 Keypoint Detection Task. — https://cocodataset.org/#keypoints-2020.

[8] Human Pose Estimation Using Body Parts Dependent Joint Regressors / Matthias Dantone, Juergen Gall, Christian Leistner, Luc Van Gool // 2013 IEEE Conference on Computer Vision and Pattern Recognition. — 2013. — Pp. 3041–3048.

[9] *Toshev, Alexander*. DeepPose: Human Pose Estimation via Deep Neural Networks / Alexander Toshev, Christian Szegedy // 2014 IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — Pp. 1653–1660.

[10] RMPE: Regional Multi-person Pose Estimation / Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu // ICCV. — 2017.

[11] OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / Z. Cao, G. Hidalgo Martinez, T. Simon et al. // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. — 2019.

[12] Deep High-Resolution Representation Learning for Human Pose Estimation / Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2019.

[13] *Bazarevsky, Valentin*. BlazePose: On-device Real-time Body Pose tracking. — 2020. `https://arxiv.org/abs/2006.10204`.

[14] ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation / Yufei Xu, Jing Zhang, Qiming Zhang, Dacheng Tao // Advances in Neural Information Processing Systems. — 2022.

[15] *Dosovitskiy, Alexey*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. — 2021.

[16] *Li, Yanjie*. SimCC: a Simple Coordinate Classification Perspective for Human Pose Estimation. — 2021.

[17] *Wu, Qiangqiang*. Progressive Unsupervised Learning for Visual Object Tracking / Qiangqiang Wu, Jia Wan, Antoni B. Chan // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2021. — Pp. 2992–3001.

[18] *Fan, Hehe*. Unsupervised Person Re-identification: Clustering and Fine-tuning. — 2017.

[19] *Tokarew, Andrey*. Полуавтоматическая система разметки ключевых точек. — `https://github.com/andrwtokar/pose-markup`.

[20] *Contributors, MMPose*. OpenMMLab Pose Estimation Toolbox and Benchmark. — `https://github.com/open-mmlab/mmpose`. — 2020.

[21] 2D Human Pose Estimation: New Benchmark and State of the Art Analysis / Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2014. — June.

[22] *Johnson, Sam*. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation / Sam Johnson, Mark Everingham // Proceedings of the British Machine Vision Conference. — 2010. — doi:10.5244/C.24.12.

[23] Convolutional pose machines / Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh // CVPR. — 2016.

[24] Deep Learning Models for Classification of Red Blood Cells in Microscopy Images to Aid in Sickle Cell Anemia Diagnosis / Laith Alzubaidi, Mohammed A. Fadhel, Omran Al-Shamma et al. //

Electronics. — 2020. — Vol. 9, no. 3. https://www.mdpi.com/2079-9292/9/3/427.

[25] Domain Adaptation: Challenges, Methods, Datasets, and Applications / Peeyush Singhal, Rahee Walambe, Sheela Ramanna, Ketan Kotecha // *IEEE Access*. — 2023. — Vol. 11. — Pp. 6973–7020.

[26] Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss / Debapriya Maji, Soyeb Nagori, Manu Mathew, Deepak Poddar // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — Pp. 2637–2646.

[27] Yolox: Exceeding yolo series in 2021 / Zheng Ge, Songtao Liu, Feng Wang et al. // *arXiv preprint arXiv:2107.08430*. — 2021.

[28] *Reis, Dillon*. Real-Time Flying Object Detection with YOLOv8. — 2024.