Рациональность выбора архитектурного решения.

При разработке алгоритма автоматической сегментации типов местности по ортофотопланам выбор архитектуры нейронной сети определялся комплексом требований к точности распознавания классов и вычислительной эффективности.

Согласно данным, представленным в исследовании Spasev et al. (2024) [1], SegFormer демонстрирует устойчивость к вариациям масштаба объектов и сложным фоновым условиям, характерным для аэрофотоснимков сельскохозяйственных территорий.

Архитектурные особенности SegFormer обеспечивают эффективную обработку изображений высокого разрешения. Архитектура ResNeXt101_32x16d демонстрирует повышенную способность к извлечению мультискалярных признаков при обработке изображений высокого разрешения, характерных для ортофотопланов. Иерархическая структура сети с прогрессивным уменьшением разрешения и увеличением количества каналов позволяет одновременно сохранять пространственную детализацию для распознавания мелких объектов и захватывать контекстные зависимости для корректной классификации крупных сельскохозяйственных структур.

Экспериментальная оценка показала, что архитектура SegFormer превосходит альтернативные решения UnetPlusPlus и DeepLabV3+ по метрике среднего пересечения-объединения (mIoU) и попиксельной точности при сопоставимых вычислительных затратах.

Эксперементальная оценка эффективности рассматриваемых архитектур проводилась на открытых данных датасета DeepGlobe Land Cover Classification Dataset. DeepGlobe Land [2] Cover Classification Dataset – это набор данных, обучающая выборка которого состоит из 702 оптических изображений размером 2448×2448 пикселей, полученных со спутника DigitalGlobe. Валидационная выборка содержит 101 изображение. Эти изображения имеют разрешение пикселя 50 см и охватывают районы в Таиланде, Индонезии и Индии. Такое разрешение пикселей позволяет детализировать объекты на земной поверхности до уровня отдельных строений или деревьев. Данная степень детализации отражена в попиксельной аннотации, которая включает семь классов местности: «город» (бирюзовый), «сельскохозяйственный район» (жёлтый), «пастбища» (фиолетовый), «лес» (зелёный), «водоём» (синий), «бесплодная земля» (белый) и «неизвестно» (чёрный). Пример изображения с аннотацией из набора данных DeepGlobe Land Cover Classification Dataset представлен на рисунках 1 и 2.

Рисунок 1 – Оптическое изображение датасета DeepGlobe Land Cover Classification

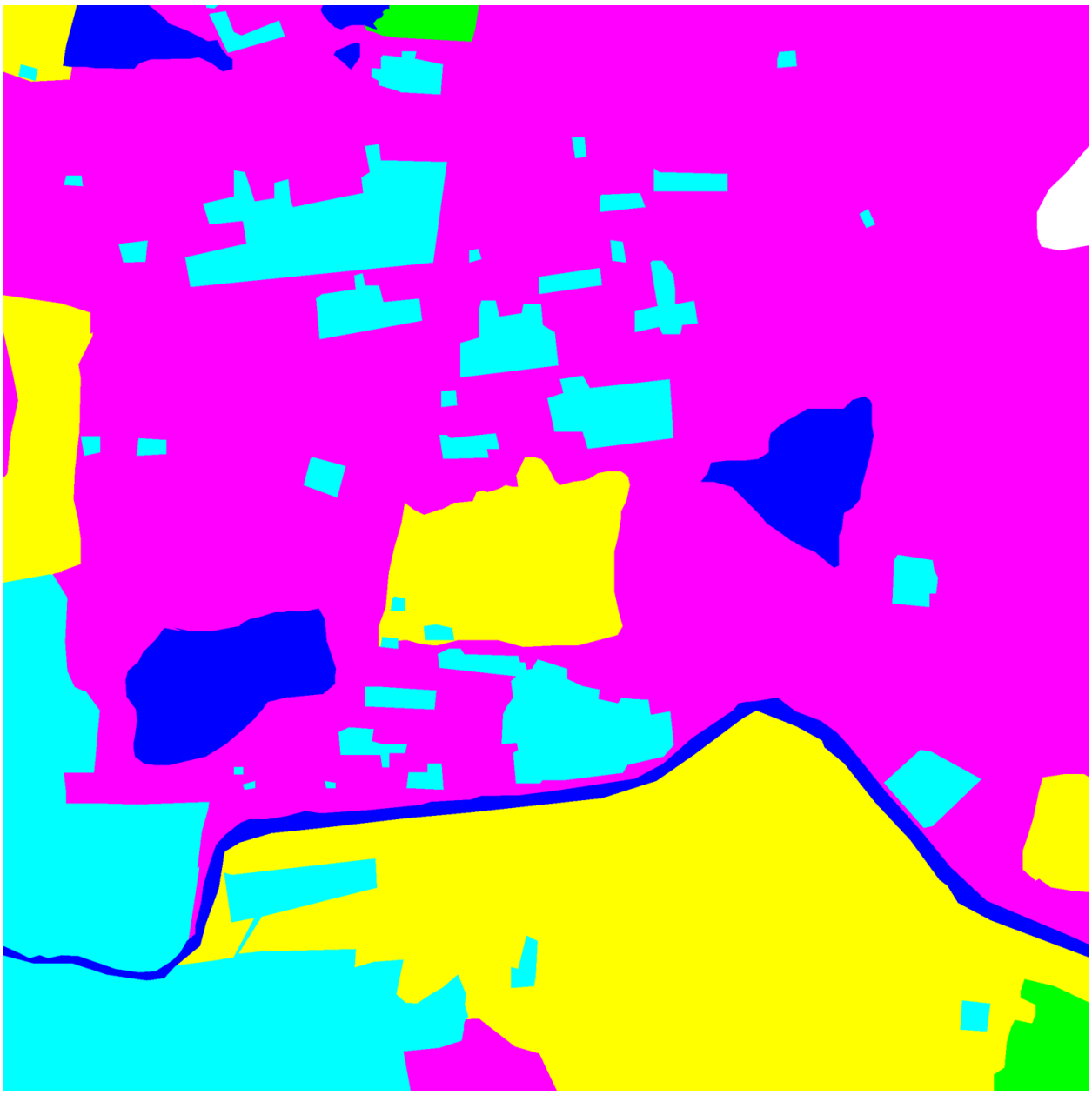

Рисунок 2 – Анотация для изображения датасета DeepGlobe Land Cover Classification

Классы в наборе данных переименовывались или объединялись в рассматриваемый класс по схожим признакам. Например, классы «город» и «деревня» были объединены в общий класс «здания (строения)». Таким образом в течении экспериментального анализа была реализована сегментация следующих классов: отсутствие сцены, здания (строения), сельскохозяйственный район, водоём, природная территория. После разделения исходных изображений датасета на изображения размером 512×512 пикселей, для обучения модели были сформированны 17550 оптических изображений. Характеризация модели осуществлялась на совокупности тестовых изображений размером 512×512 пикселей, включающей в себя 2525 изображений.В таблице 1 представлены показатели точности для каждого класса.

Таблица 1 – Точность сегментации каждого типа местности на совокупности тестовых изображений представленных наборов данных

Наименование метрики	SegFormer	DeepLabv3+	UnetPlusPlus
mPA, %	91.92	90.28	86.81
mIoU, %	78.80	67.18	58.80
IoU для класса «Отсутствие сцены», %	61.89	23.11	0.19
IoU для класса «Здания (строения)», %	75.98	74.85	70.12
IoU для класса «Сельскохозяйственный район», %	89.51	87.37	82.61
IoU для класса «Водоём», %	87.68	73.95	71.96
IoU для класса «Природная территория», %	78.98	76.63	69.15

Описание архитектуры нейронной сети.

Выбранная на основе проведённых экспериментов конфигурация модели Segformer представляет собой гибридную архитектуру, где энкодер на основе ResNeXt-101 обеспечивает эффективное извлечение иерархических признаков (рисунок 3), а декодер, вдохновленный трансформерными механизмами, восстанавливает пространственную детализацию сегментационной карты. Такой подход позволяет достичь оптимального баланса между глобальным контекстным пониманием и точностью локальных границ объектов.

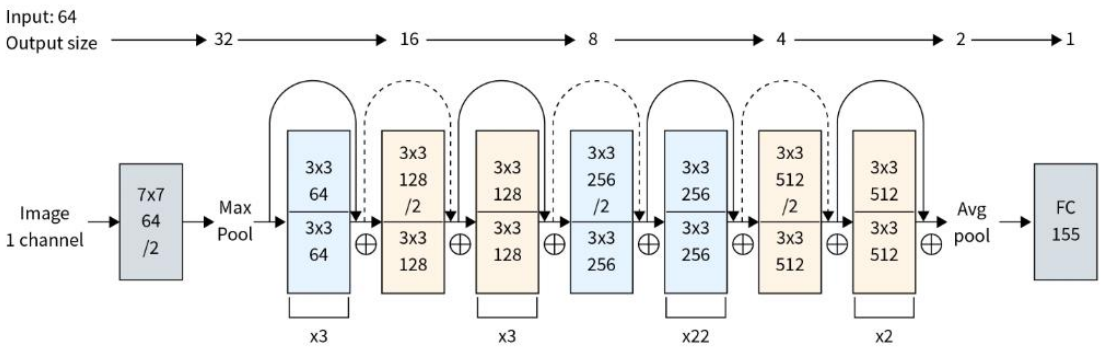

Рисунок 3 – Общая архитектура ResNet-101

Архитектура энкодера resnext101_32x16d» основана на фундаментальной архитектуре ResNet-101, которая повышает эффективность глубокого обучения благодаря введению остаточных соединений. ResNet-101 решает проблему затухающих градиентов в глубоких сетях путем использования skip-соединений, которые позволяют градиентам напрямую проходить через слои, обеспечивая стабильное обучение сетей с сотнями слоев. Структура типичного ResNet блока представлена на Рисунке 4 а, где видно, как входной тензор разделяется на два пути: один проходит через последовательность сверток, а другой напрямую передается через skip-соединение, после чего оба пути суммируются.

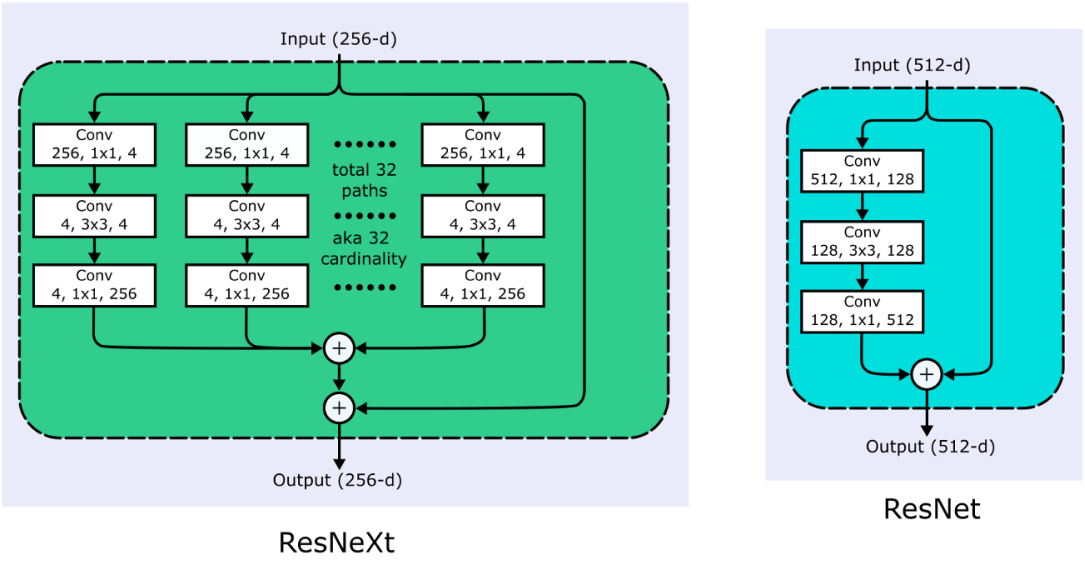

Рисунок 4 – Сравнение архитектурных блоков: (*а*) классический ResNet блок с последовательностью 3×3 сверток, (*б*) ResNeXt блок с групповыми свертками в конфигурации 32x16d

На Рисунке 4 б детально показана трансформация ResNet блока в ResNeXt блок: последовательность двух 3×3 сверток заменяется на 32 параллельных ветви, каждая из которых обрабатывает 16-канальную группу входных признаков. Этот подход обеспечивает

более эффективное использование параметров, поскольку вместо полносвязной обработки всех каналов, модель фокусируется на извлечении специализированных признаков в каждой группе.

Конфигурация 32x16d указывает на использование 32 групп сверток с 16 каналами в каждой группе. Общее количество параметров в такой конфигурации остается сопоставимым с оригинальным ResNet блоком, однако выразительная способность модели значительно возрастает благодаря параллельному извлечению разнородных признаков.

Декодерная часть архитектуры Segformer характеризуется двумя ключевыми параметрами: decoder_segmentation_channels=128 и decoder_attn_channels=128. Первый параметр определяет ширину промежуточных слоев в полносвязном (MLP)-декодере, отвечающем за интеграцию мультимасштабных признаков, полученных от энкодера. Значение 128 каналов представляет собой компромисс между вычислительной сложностью и выразительной способностью, позволяя эффективно объединять признаки различных разрешений. На Рисунке 5 детально представлена архитектура декодера, демонстрирующая процесс объединения признаков с разных уровней энкодера. Второй параметр, decoder_attn_channels=128, определяет емкость механизма внимания в декодере, который играет критическую роль в точном восстановлении границ объектов. Механизм внимания позволяет декодеру динамически взвешивать важность признаков с разных уровней энкодера, уделяя особое внимание тем участкам, которые содержат информацию о границах сегментов.

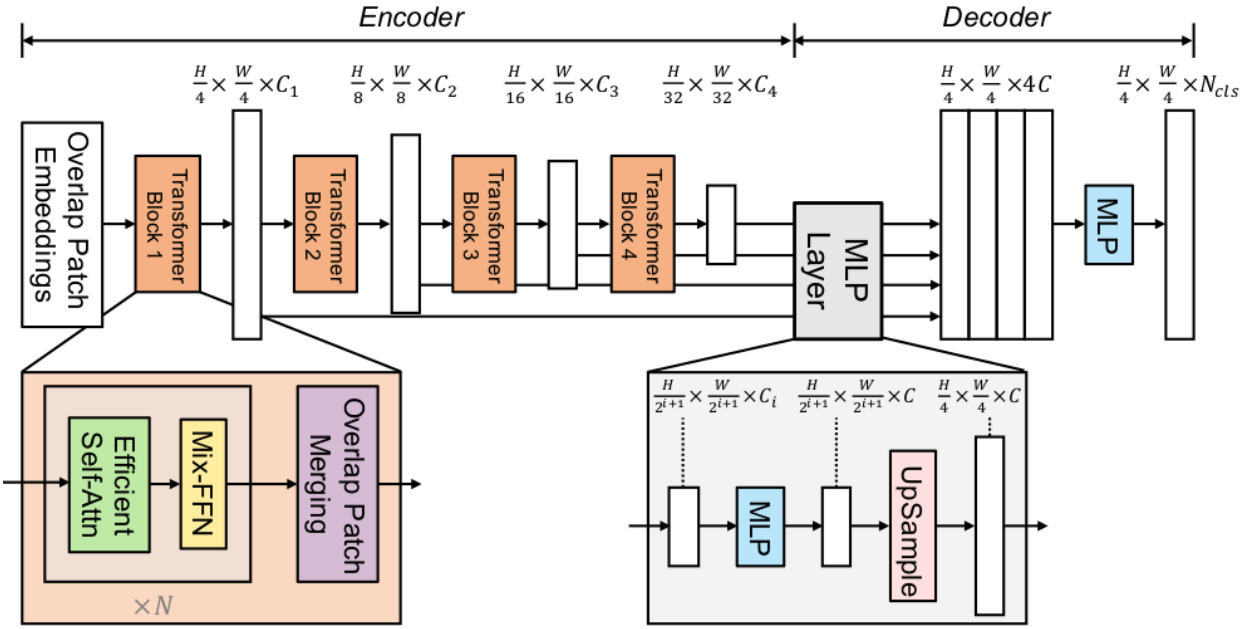

Рисунок 5 – Архитектура энкодера и декодера Segformer

Гибридная природа рассматриваемой архитектуры заключается в комбинации сильных сторон сверточных нейронных сетей и трансформерных механизмов. ResNeXt-101 энкодер

обеспечивает эффективное извлечение пространственных иерархических признаков благодаря своей глубокой иерархической структуре, в то время как декодер Segformer использует механизмы внимания для точного восстановления пространственных деталей. Такая комбинация позволяет модели преодолевать традиционные ограничения сверточных архитектур, такие как ограниченное receptive field, и одновременно избегать вычислительной неэффективности полных трансформерных моделей.

Обучение нейронной сети.

Таким образом, в представленной реализации архитектура SegFormer использует предобученную backbone-архитектуру ResNeXt101_32x16d, которая была оптимизирована для изображений размером 512×512 пикселей. По этой причине при обучении и применении модели использовались изображения данного размера. В ходе исследования был реализован конвейер обработки геопространственных данных для задачи семантической сегментации. Исходные данные представляли собой снимки в формате GeoTIFF, содержащие как пиксельную информацию, так и геометаданные. На рисунке 6 представлен исходный файл в формате .tif с сохранением всех геопространственных характеристик.

Рисунок 6 – Исходное спутниковое изображение в формате GeoTIFF с сохранением геометаданных (EPSG:4326)

Процесс подготовки данных включал разметку объектов с использованием векторных аннотаций в формате GeoJSON. На основе этих аннотаций было сформировано размеченное изображение, где каждый пиксель классифицирован в соответствии с принадлежностью к определенному классу объектов (рисунок 7).

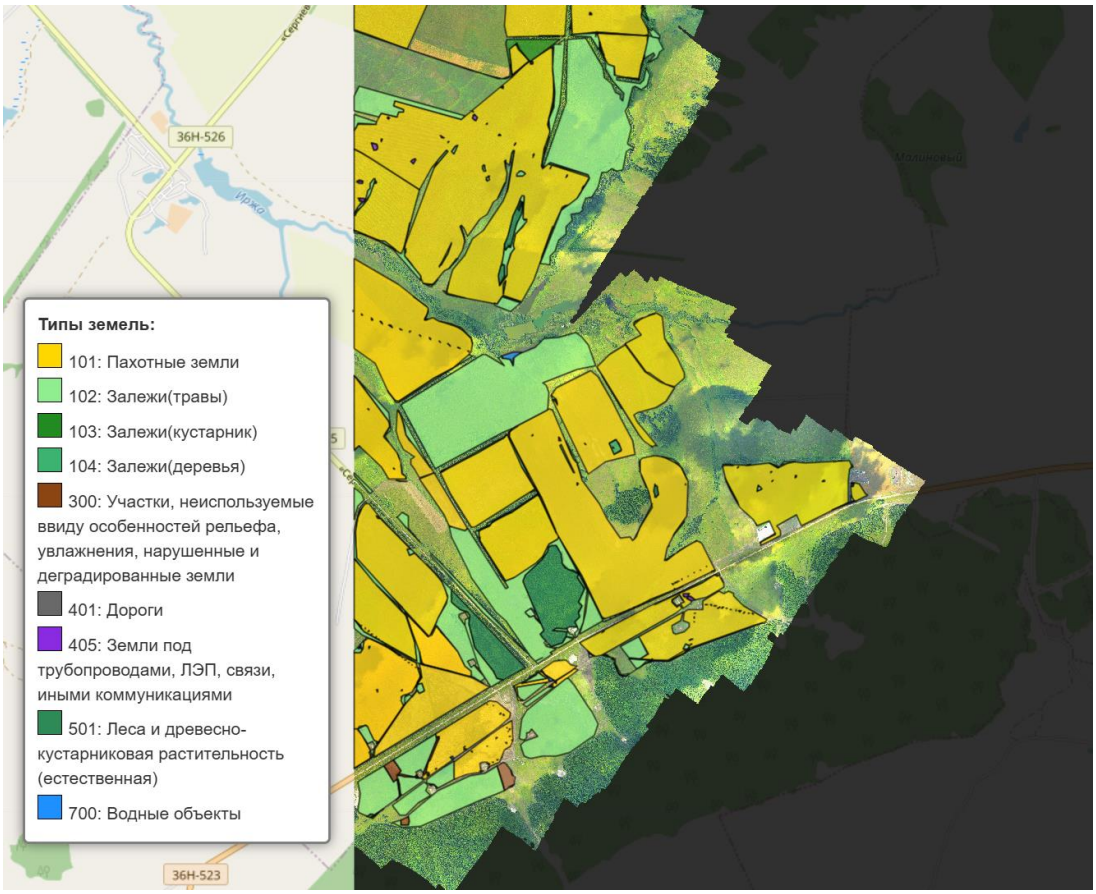

Рисунок 7 – Размеченное изображение после применения аннотаций из файла формата .geojson

Для обучения нейронной сети изображение было разделено на блоки фиксированного размера 512×512 пикселей. Особое внимание уделялось корректному разделению данных на обучающую и тестовую выборки с сохранением пространственной репрезентативности. На рисунке 8 продемонстрирован результат разделения исходного изображения, где блоки, отобранные для тестовой выборки, выделены красным контуром для визуализации процесса отбора данных.

Рисунок 8 – Визуализация разделения изображения на блоки 512×512 пикселей; тестовая выборка выделена красным контуром

Такая стратегия разделения данных обеспечила сбалансированное распределение объектов различных классов между выборками и позволила избежать пространственного дублирования информации. Общий объем обучающей выборки составил 5200 изображений, тестовой — 43 изображения, что соответствует оптимальному соотношению для обучения сверточных нейронных сетей в задачах сегментации изображений.

В процессе обучения Модели для оптимизации параметров модели использовался алгоритм RAdam с начальной скоростью обучения равной 10-4 и коэффициентом регуляризации равным 10-3. Оптимизация осуществлялась с целью минимизации суммы значений таких функций потерь, как: попиксельная кросс-энтропия, DICE Loss и FocalLoss. Обучение проводилось в течение 450 эпох, при этом эпохой считается один полный проход по всему обучающему набору данных. Для управления скоростью обучения был применен

циклический планировщик (CyclicLR), который на каждом шаге обучения обновляет текущую скорость обучения для оптимизатора RAdam согласно заданной циклической функции. Таким образом скорость обучения линейно возрастает от 10-10 до 10-4 на протяжении десяти эпох, а затем линейно уменьшается обратно до 10-10 в течении 215 эпох. После 225 эпохи этот процесс повторяется только максимальная скорость обучения составляет 10-5.

В результате тестирования обученной модели на тестовой выборке были рассчитаны показатели средней попиксельной точности равной 87.25% и mIoU равным 31.9%. В Таблице 2 представлены показатели IoU, рассчитанные для каждого класса в отдельности.

Таблица 2 – Показатели IoU каждого класса местности на тестовых изображениях

Класс	**IoU, %**
Фон	71.81
Водные объекты	0.00
Дороги	7.45
Залежи(деревья)	0.00
Залежи(кустарник)	48.95
Залежи(травы)	45.08
Земли под трубопроводами, ЛЭП, связи, иными коммуникациями	11.96
Леса и древесно-кустарниковая растительность (естественная)	0.00
Пахотные земли	93.37
Участки, неиспользуемые ввиду особенностей рельефа, увлажнения, нарушенные и деградированные земли	40.35

Список источников.

1. Spasev V. et al. Semantic Segmentation of Unmanned Aerial Vehicle Remote Sensing Images Using SegFormer //International Conference on Intelligent Systems and Pattern Recognition. – Cham : Springer Nature Switzerland, 2024. – С. 108-122.

2. Demir I. et al. Deepglobe 2018: A challenge to parse the earth through satellite images //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2018. – С. 172-181.