Автоматическая Вертикализация Коротких Видео

Егор Домнин, Иви

Меня зовут Егор Домнин

Мои прошлые выступления

Синтетические постеры для кино

Как обрезать логотип телеканала,
хардсабы и чёрные грани

Habr

Команда Иви

Генерация постеров к фильмам
и мультикам в онлайн-кинотеатре

ODS Datafest 2022

Иви сегодня

№1

Среди работодателей

- Тренинги и воркшопы Иви
- Внешние ресурсы обучения
- Статьи

830+

Экспертов и профессионалов

- Собственная технологическая платформа
- Highload
- Десятки релизов ежедневно

15

Лет на рынке

- Лояльная вовлечённая аудитория
- Сильный бренд
- Производство собственного контента

CV в Иви

Компьютерное зрение — 4 CV инженера

Проекты:

- Синтетические постеры
- Синтетические интересные видеофрагменты
- Поиск контекста в видео
- Апскейл видео
- Вертикализация видео

Прошлогодний доклад на VideoTech:

Направить все ресурсы
в бесконечность моментов,
Александр Коншин

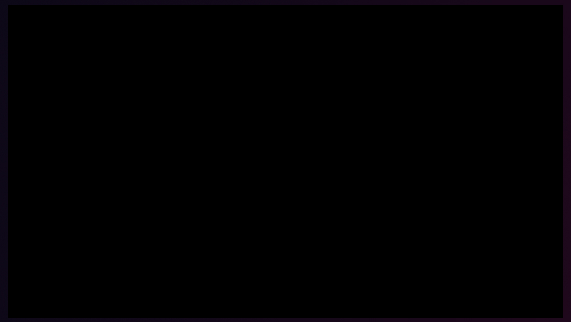
Оглавление

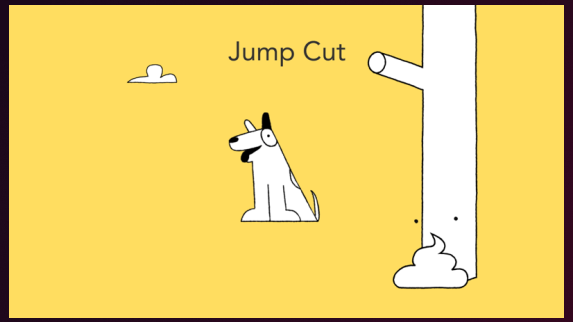
Что такое Шот?

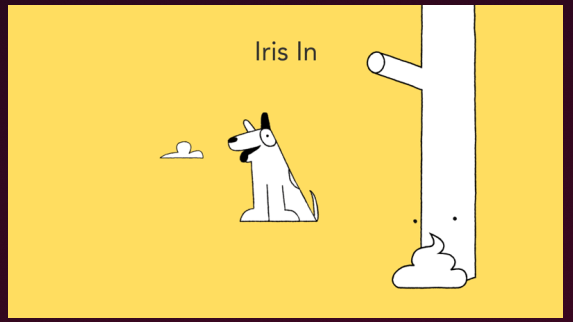
Терминология

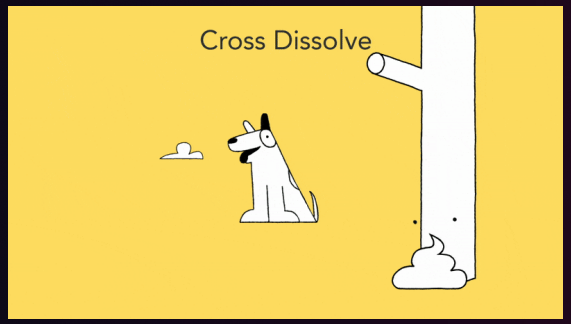
Шот — последовательность кадров от одной монтажной склейки до другой

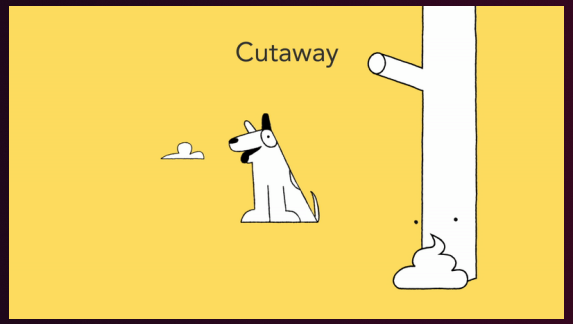
Примеры монтажных склеек











Что такое Момент?

Терминология

Момент — это короткое видео из фильма

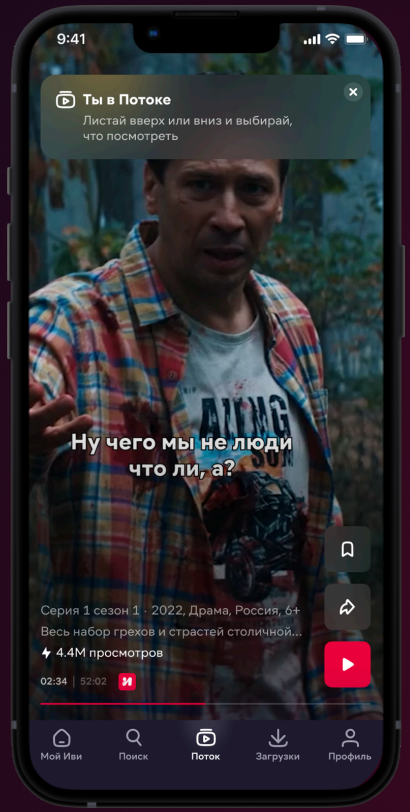
Момент состоит из множества шотов

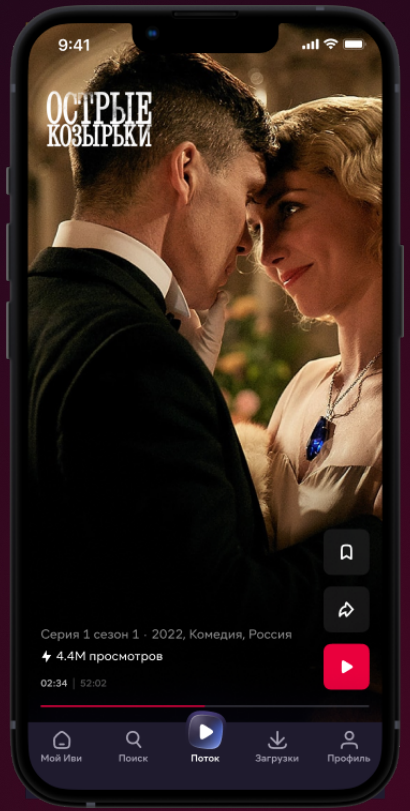
Что такое Поток?

Терминология

Поток — это лента моментов, подобная Youtube Shorts / TikTok / Instagram Reels

Нужна для конвертации в просмотр контента





Что такое Кроп?

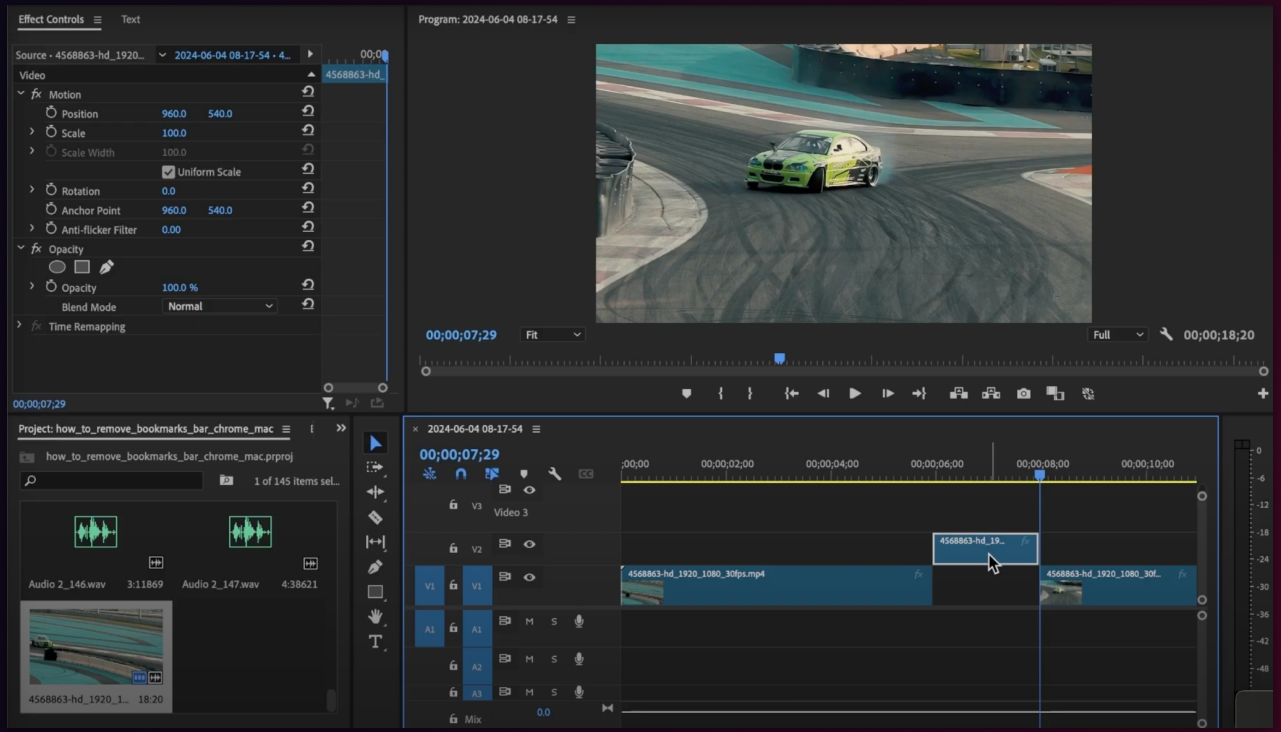
Терминология

Многие люди открывают приложение с телефонов в вертикальной ориентации, поэтому для улучшения их опыта моменты из фильмов нужно вертикализировать

По нашим замерам вертикальные моменты с телефона смотрят на 30% дольше

Откуда берутся моменты? ч.1

Редакторы — полностью ручная нарезка интересных моментов из фильмов



Откуда берутся моменты? ч.2

Хайлайтер — наш сегментатор фильмов.
Потом модерируется редакторами

Кадры

Шоты

Сцены

Фильм

Частично про него
рассказывали тут

Стоимость ручной вертикализации

Польза вертикализации

На 5 октября 2023

4 272

Единичных тайтлов

67 143 серий

В 2060 сериалах

20 минут

На вертикализизацию
и субтитры в ручном режиме

500

Роликов в месяц делает
один монтажер

143 человеко-месяца

Чтобы обработать
71 415 единиц контента

160 руб

Стоимость одного ролика

Что такое хорошая вертикализация?

Хорошая вертикализация держит в фокусе наиболее важную для понимания информацию и плавно передвигает камеру

Один объект интереса — фокус на нём

Что такое хорошая вертикализация? ч.2

Если объектов 2+ не должна теряться суть ролика из-за отрезания части информации / не должно возникать ощущения, что чего-то в кадре не хватает

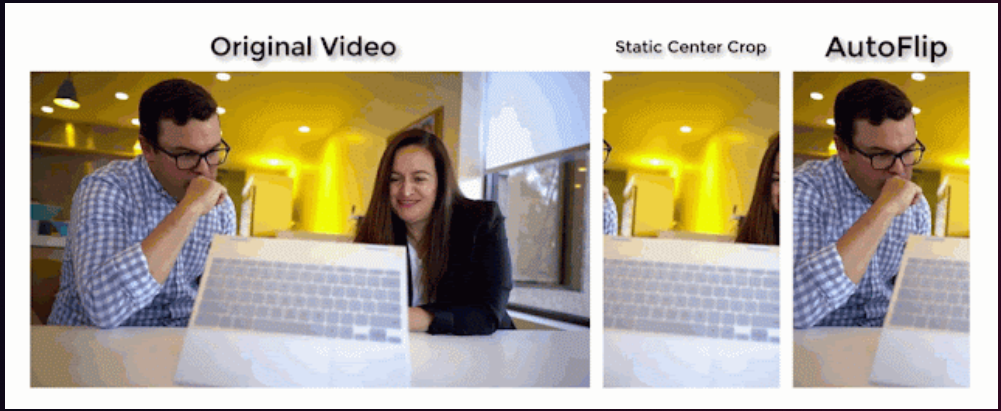
Смена фокуса может происходить двумя способами:

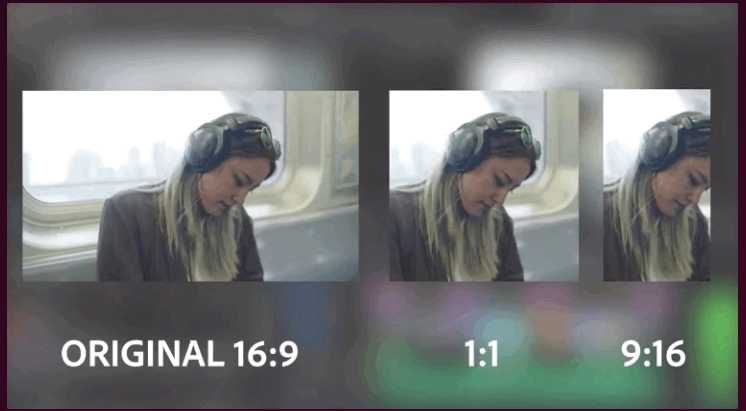
1. Через склейку
2. Через перемещение

Существующие решения

Исследование способов вертикализации

Google AutoFlip и Adobe Premier Auto Reframe не устраивают по качеству





Датасет редакторских видео

Датасет

План действий такой:

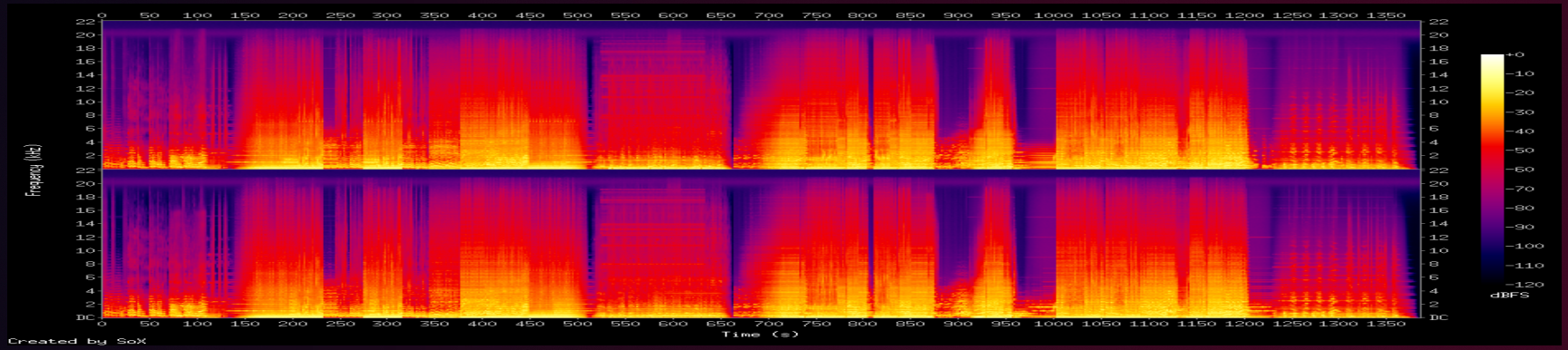
1. Получить список вертикальных моментов из потока,
   ID соответствующих им единиц контента и таймстемпов начала / конца момента
2. Скачать видео вертикальных моментов и горизонтальных оригиналов
3. Восстановить frame-level таймстемпы начала и конца моментов,
   потому что в базе записаны таймстемпы в секундах
4. Восстановить координаты кропа для каждого кадра момента, используя пары
   из горизонтального и вертикального time-aligned видео,
5. Выкинуть некачественный матчинг по таймстемпам, такого было 10% — осталось 5 012 видео

Как матчили?

Датасет

1. Для «Восстановления frame-level таймстемпы начала и конца моментов» использовали:
   a. Примерное время совпадения по звуку
   b. И уточнение по кадрам низкого разрешения
2. Для Восстановление координат вертикального кропа использовали кадры высокого разрешения

Никакой магии, просто scipy.signal.correlate и cv2.matchTemplate



Метрики

Датасет

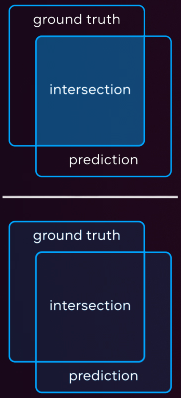
Точность кропа

Mean Average Error (MAE)

$$MAE = \sum_{i=1}^{N} |x\_gt_i - x\_pred_i|$$

Intersection Over Union (IOU)

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} =$$



Стабильность кропа

Jitter Degree Ratio (JDR)

$$JDR(\mathbf{y}) = \sum_{t=1}^{T-1} \left(\frac{\|y_{t+1} - y_t\|_2}{w}\right),$$

Stable Movement (SM)

$$SM = \sum_{i=2}^{N} |x_i - x_{i-1}|$$

Бейзлайн — статичный кроп по центру

Больше лучше:

"iou": 0.428 — область пересечения кропа с ожидаемой позицией
Меньше лучше

"mae": 0.131 — средняя ошибка сдвига кропа относительно ожидаемого

"jdt": 0.0 — степень малого ‘дрожания’ кропа

"sm": 0.0 — оценка резких сдвигов кропа

Выбор моделей для пайплайна вертикализации

Шаг 1. Определяем границы шотов

Далее обрабатываем каждый шот по отдельности

Используем TransNet V2

Выбор моделей для пайплайна вертикализации

Шаг 1. Определяем границы шотов

Шаг 2. Детектируем лица

Мы давно используем для различных проектов YoLo v5s дообученную на мультиках

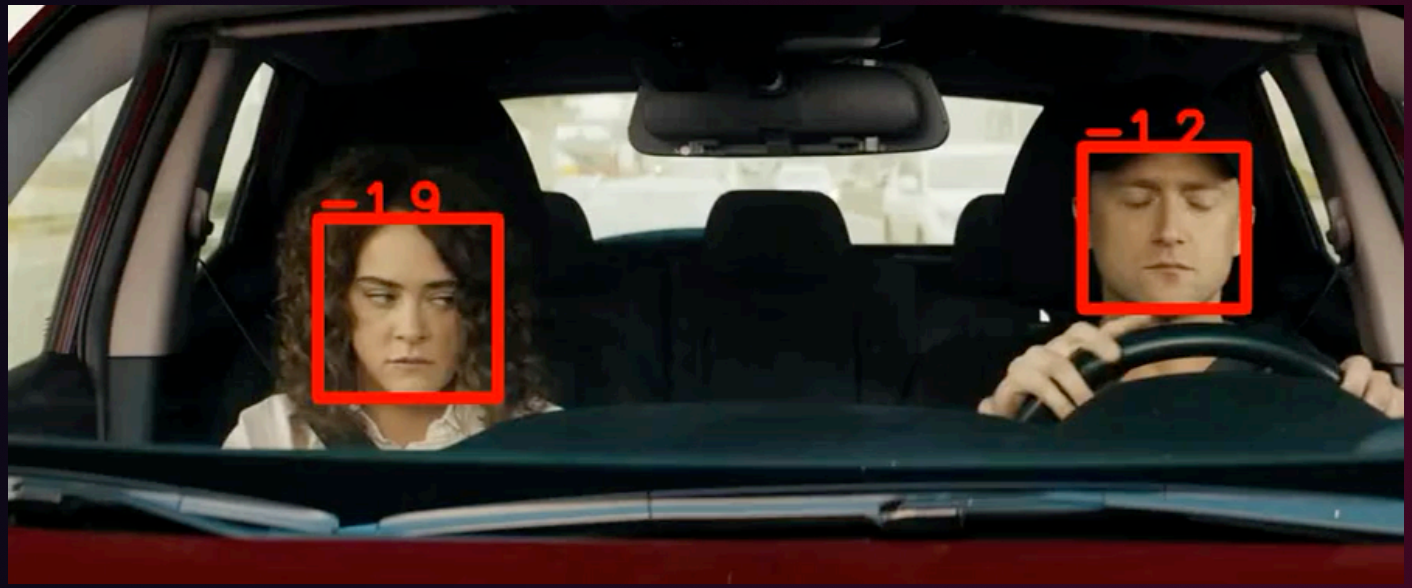
Выбор моделей для пайплайна вертикализации

Шаг 1. Определяем границы шотов

Шаг 2. Детектируем лица

Шаг 3. Используем модель Active Speaker, определяем говорящих

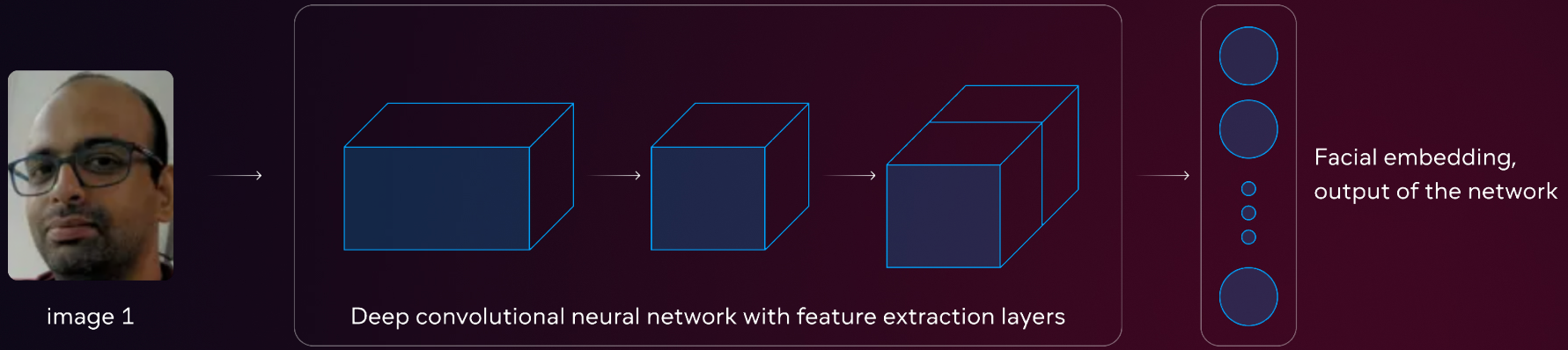
A Light Weight Model for Active Speaker Detection (Light-ASD)



Выбор моделей для пайплайна вертикализации

Шаг 1. Определяем границы шотов
Шаг 2. Детектируем лица
Шаг 3. Используем модель Active Speaker, определяем говорящих
Шаг 4. Извлекаем эмбединги лиц, присваиваем id разным персонажам

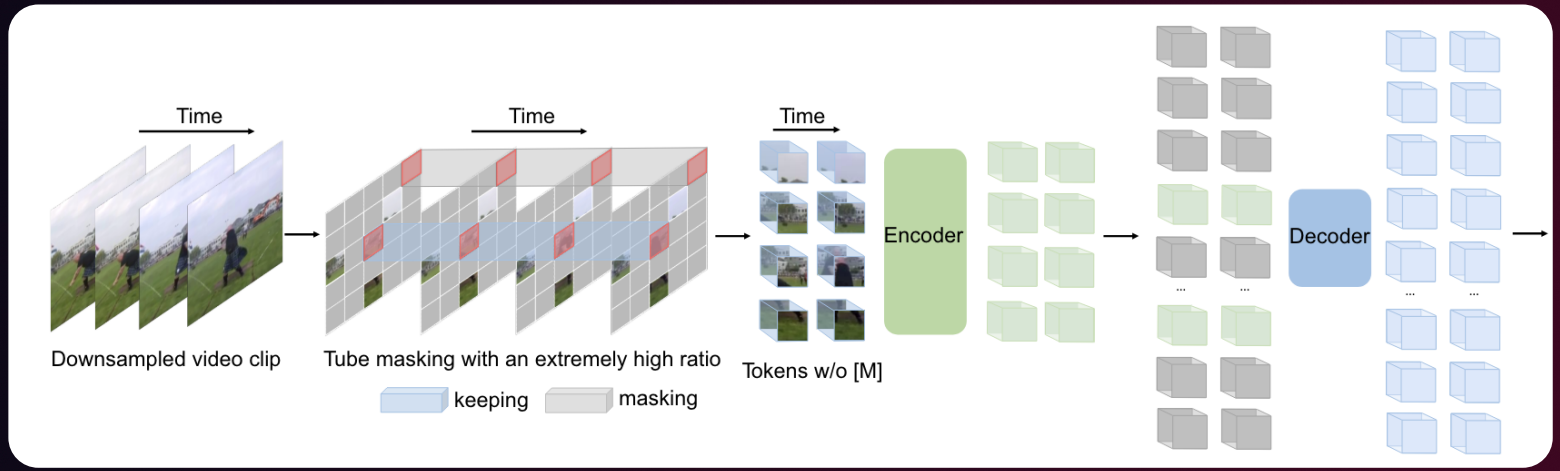
Извлекаем эмбединги с помощью FaceNet (InceptionResnetV1)



Выбор моделей для пайплайна вертикализации

Шаг 1. Определяем границы шотов
Шаг 2. Детектируем лица
Шаг 3. Используем модель Active Speaker, определяем говорящих
Шаг 4. Извлекаем эмбединги лиц, присваиваем id разным персонажам
Шаг 5. Определяем тип шота, движущийся или статичный

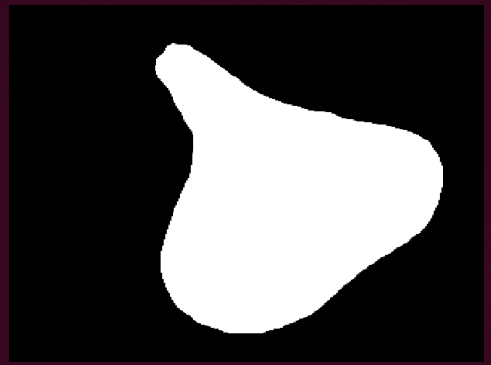
Взяли модель VideoMAE (Video Masked AutoEncoder), обучили её на датасете MovieNet, где имеются данные о типах шотов



Выбор моделей для пайплайна вертикализации

Шаг 1. Определяем границы шотов
Шаг 2. Детектируем лица
Шаг 3. Используем модель Active Speaker, определяем говорящих
Шаг 4. Извлекаем эмбединги лиц, присваиваем id разным персонажам
Шаг 5. Определяем тип шота, движущийся или статичный
Шаг 6. Определяем области фокуса / блюра в кадрах

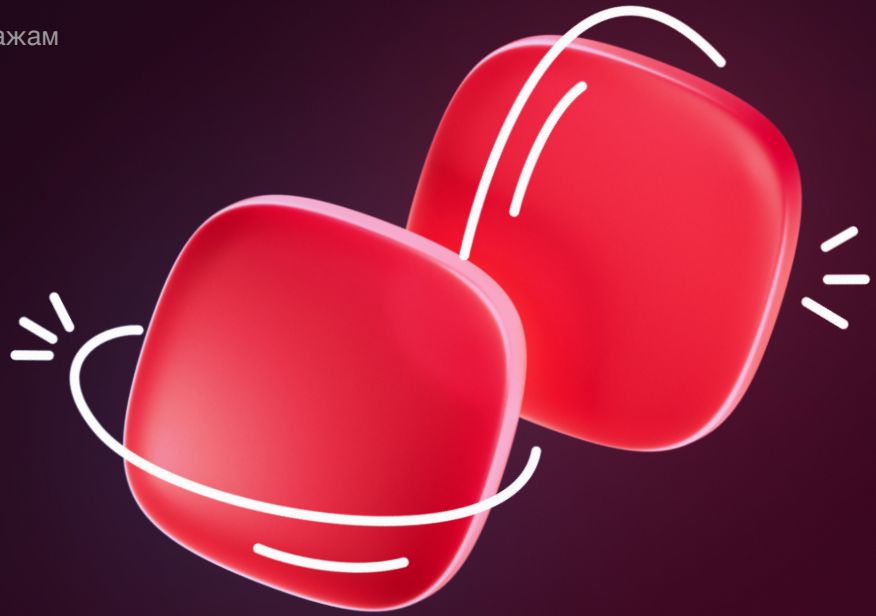
D-DFFNet (Depth and DOF Cues Make A Better Defocus Blur Detector)



Модель сегментации

Пайплайн. Модели

Шаг 1. Определяем границы шотов
Шаг 2. Детектируем лица
Шаг 3. Используем модель Active Speaker, определяем говорящих
Шаг 4. Извлекаем эмбединги лиц, присваиваем id разным персонажам
Шаг 5. Определяем тип шота, движущийся или статичный
Шаг 6. Определяем области фокуса / блюра в кадрах
Шаг 7. Сегментируем изображение на объекты



Нужные категории сегментов:

- Люди
- Машины, мотоциклы, самолёты, автобусы, поезда, траки, лодки
- Остальное

Сегментируем изображение на объекты. OneFormer

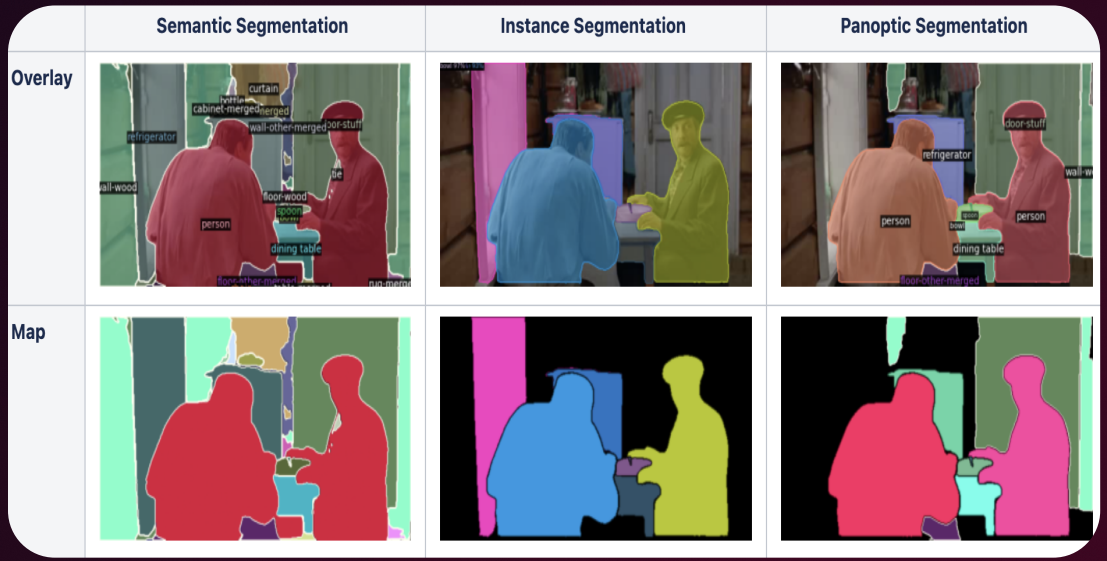
OneFormer — Instance Segmentation

Достоинства:

- Модель делает 3 вида сегментации
- Можно использовать модели, обученные на разных датасетах, под конкретные задачи
- Нетребовательна к видеопамяти

Недостатки:

- Смазывает маски при указании высокого порога
- Не умеет делать фильтрацию по классам и скорам каждого класса
- Медленно работает при создании большого количества масок



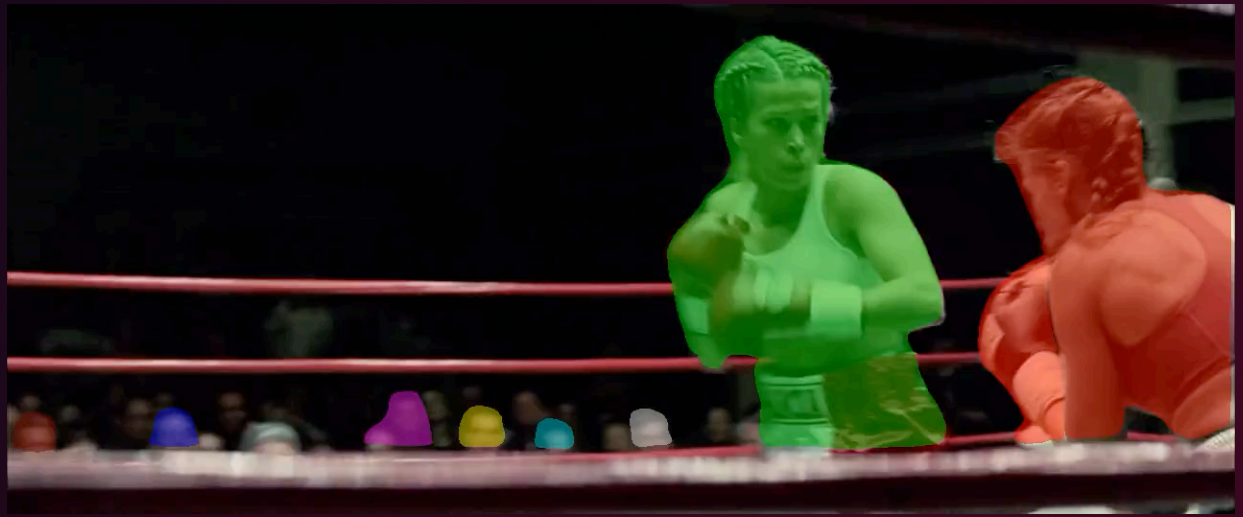
Сегментируем изображение на объекты. Cutie

Достоинства:

- Модель отлично отслеживает объекты, находящиеся в статике или небольшом движении

Недостатки:

- Модель не умеет сегментировать объекты, маски которых не передавались
- Плохо отслеживает быстро движущиеся объекты
- Требовательна к видеопамяти



Модели есть, что дальше?

Пайплайн. Модели

Шаг 1

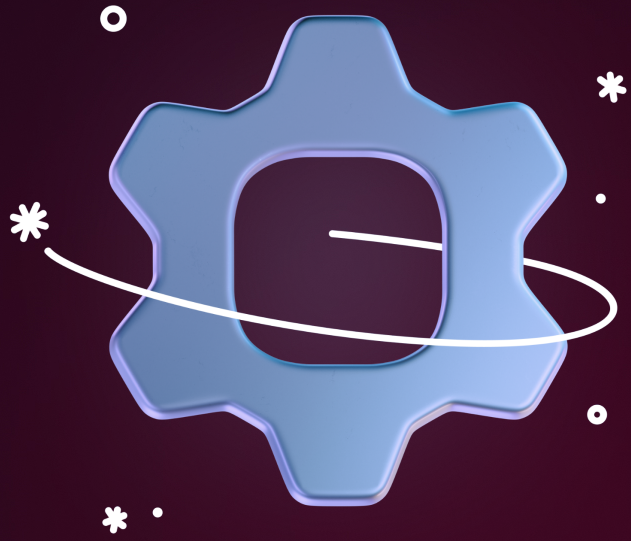
Накручиваем эвристики

Шаг 2

Сравниваем статистику с редакторскими кропами

Шаг 3

Используем сглаживание, чтобы избежать дрожания фокуса от кадра к кадру



Накручиваем эвристики — наивная реализация

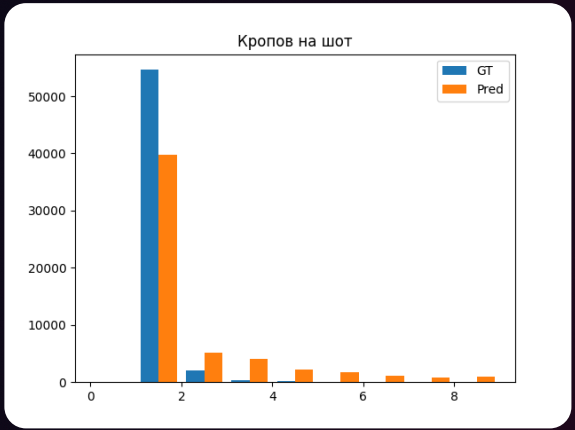
Больше лучше:
"iou": 0.428 -> 0.723

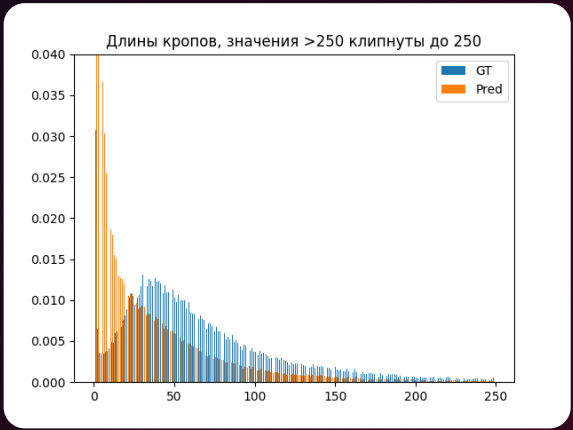
Меньше лучше:
"mae": 0.131 -> 0.040
"jdt": 0.0 -> 0.0023
"sm": 0.0 -> 0.070

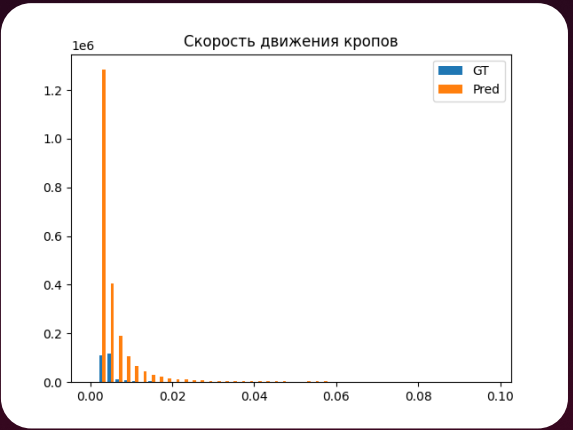
Итог:
Стало точнее, но появилась «дрожь» и скачки фокуса

Анализ редакторских кропов

Пайплайн. Статистический анализ







Сравниваем работу наивной модели с работой редакторов по статистикам кропа

1. Длина кропа
2. Количество кропов на шот
3. Скорость движения координаты кропа от кадра к кадру

Выводы:

1. Надо реже использовать склейки при перемещении внутри шота, редко когда нужно больше 2х кропов
2. Нашли границу допустимой скорости передвижения

1 шот — 1 кроп

Пайплайн. Эвристики

Больше лучше:

"iou": 0.723 -> 0.744

Меньше лучше:

"mae": 0.040 -> 0.0388

"jdt": 0.0023 -> 0.0005

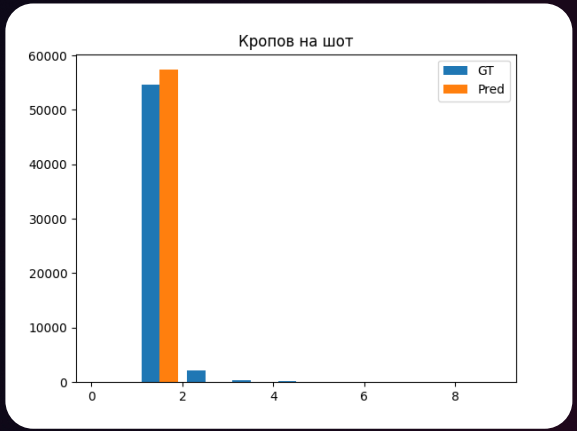
"sm": 0.070 -> 0.034

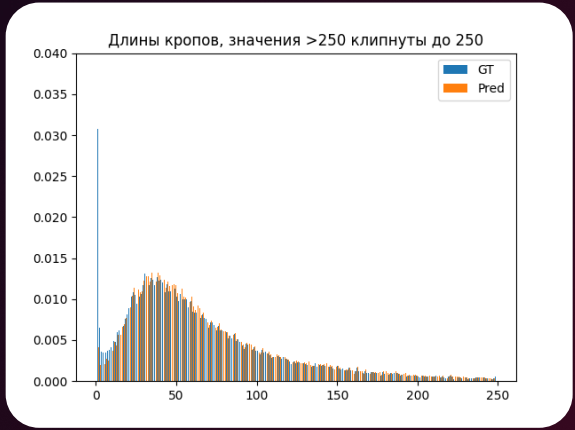
Итог:

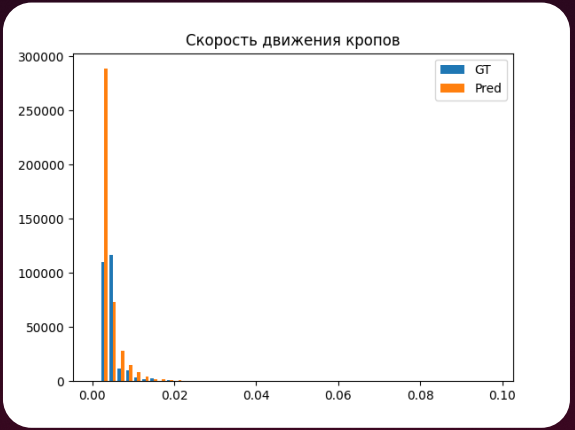
Стало сильно лучше

Анализ редакторских кропов 1 шот — 1 трек

Пайплайн. Статистический анализ





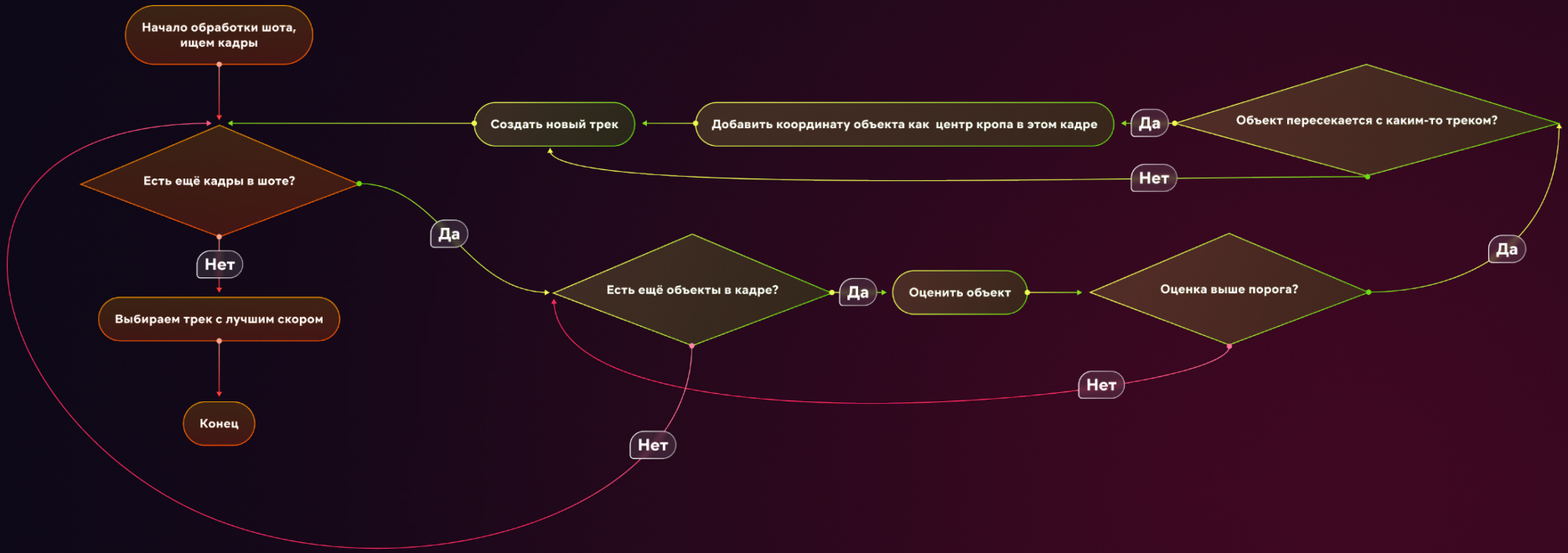


Вывод:

Переход к 1 кропу на шот выровнял и длины, и скорость движения

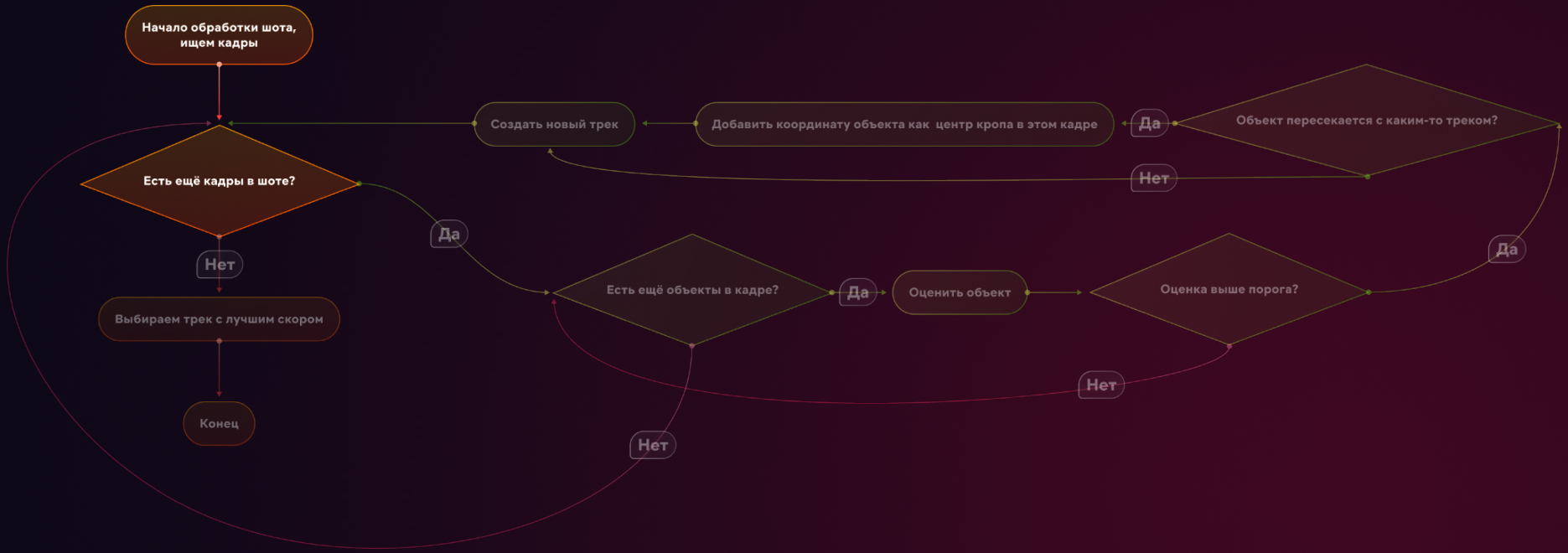
1Шот — 1Кроп ч.2

Пайплайн. Эвристики



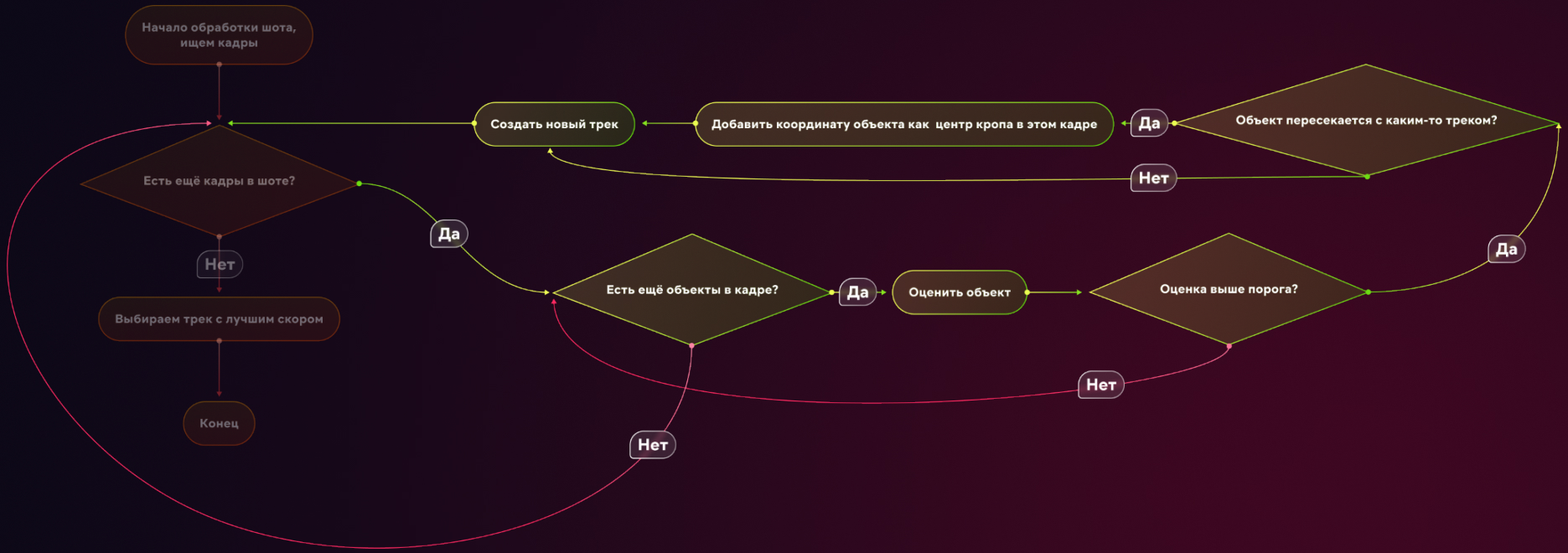
1 шот — 1 кроп ч.2

Пайплайн. Эвристики



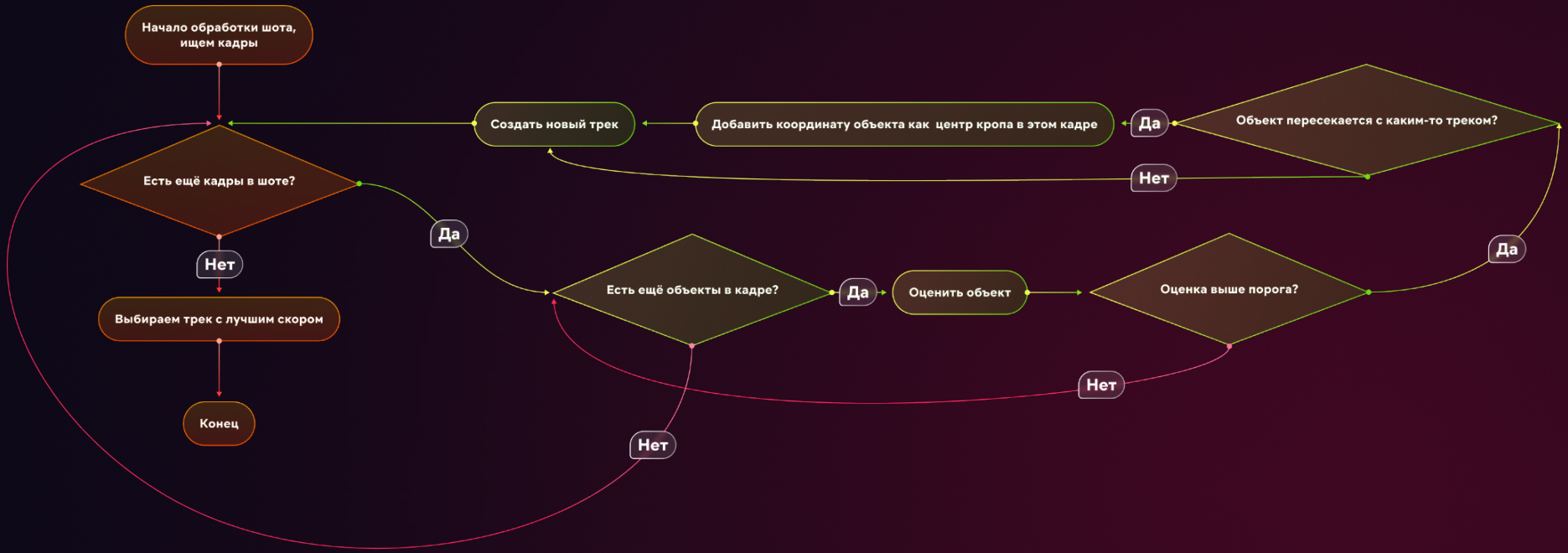
1 шот — 1 кроп ч.2

Пайплайн. Эвристики



1 шот — 1 кроп ч.2

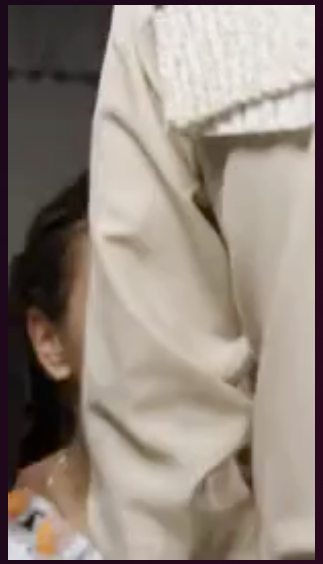
Пайплайн. Эвристики

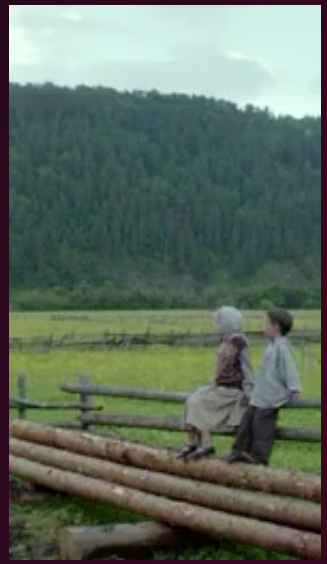


«Дрожание» видео

Пайплайн. Эвристики

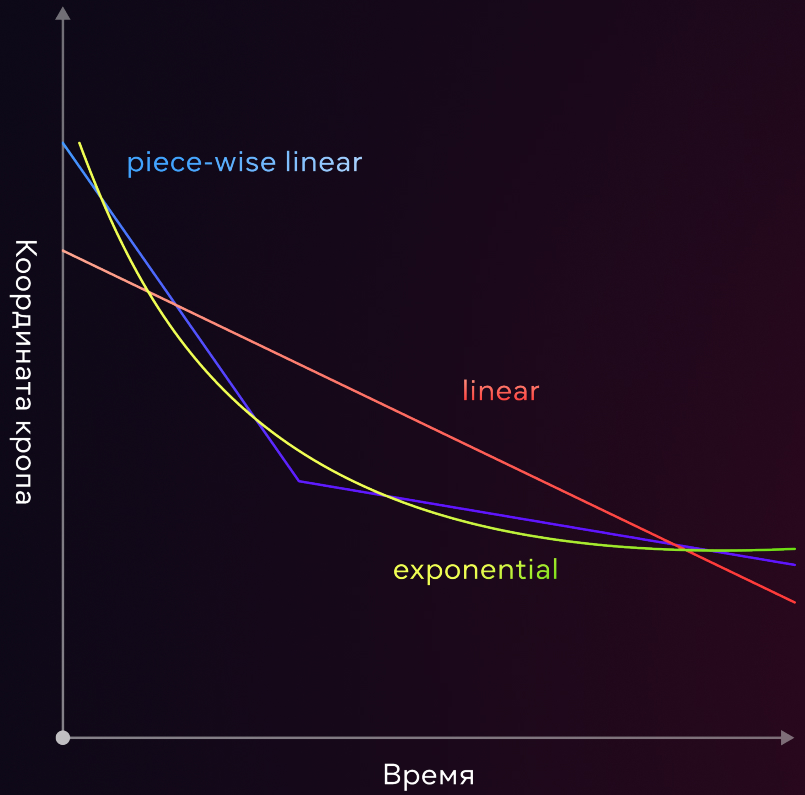
При слишком низкой скорости движения кропа появляется ощущение дрожи картинки,
такой кроп надо делать статичным. Особенно это заметно при обработке в низком разрешении





«Дрожание» видео ч.2

Пайплайн. Эвристики



Фикс

Замена линейной аппроксимации внутри шота на кусочно-линейную и превращение слабо двигающихся сегментов в статичные

Скорость работы

Пайплайн

Железо на проде:

- GPU — Nvidia A100 80gb
- RAM — 1tb
- CPU — 2 x Intel(R) Xeon(R) Gold 5318Y (2.10GHz, 24 cores, 48 threads)

Скорость работы:

- 216 секунд работы на 60 секундное видео в среднем
- Минимальные требования:
- 11gb vram
- 32gb ram
- Время

Ускорение экспериментов

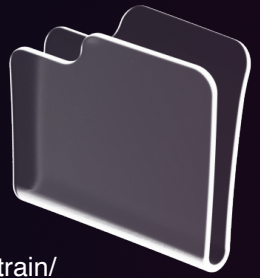
Пайплайн

Модели обсчитываем и кэшируем в DVC

15 минут на проверку эвристик на полном датасете

DVC control

DVC remote



train/

Git control

GitHub

code.py

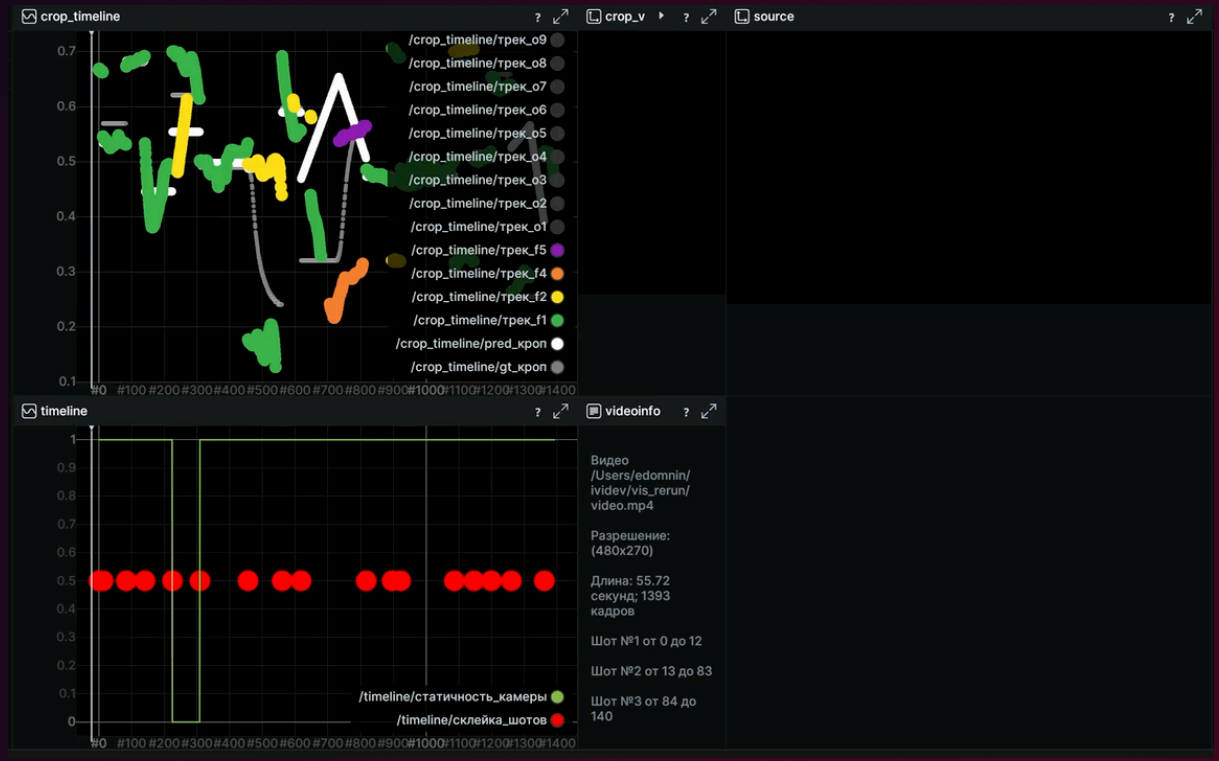
train.dvc

.gitignore

Визуализация ReRun

Пайплайн

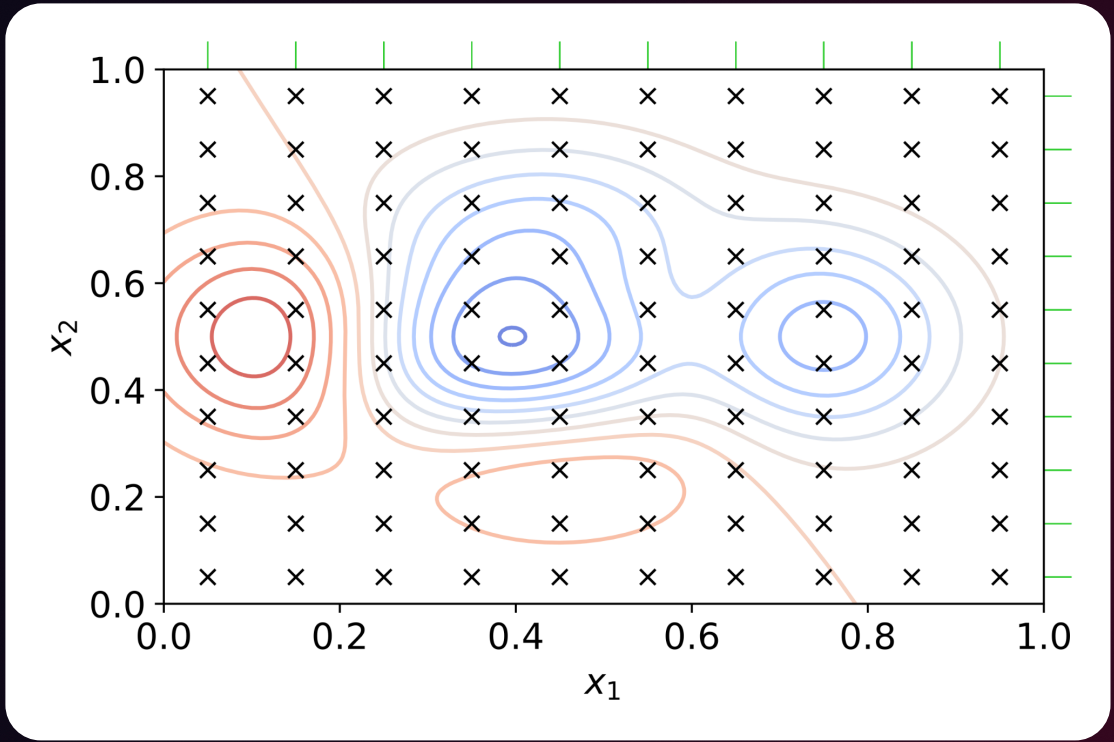
Визуализация очень упрощает отладку



Тюнинг гиперпараметров

Пайплайн

А что если мы руками неправильно подобрали веса в эвристиках?



Больше лучше:
"iou": 0.744 -> 0.756

Меньше лучше:
"mae": 0.038 -> 0.035
"jdt": 0.0005 -> 0.0005
"sm": 0.034 -> 0.034

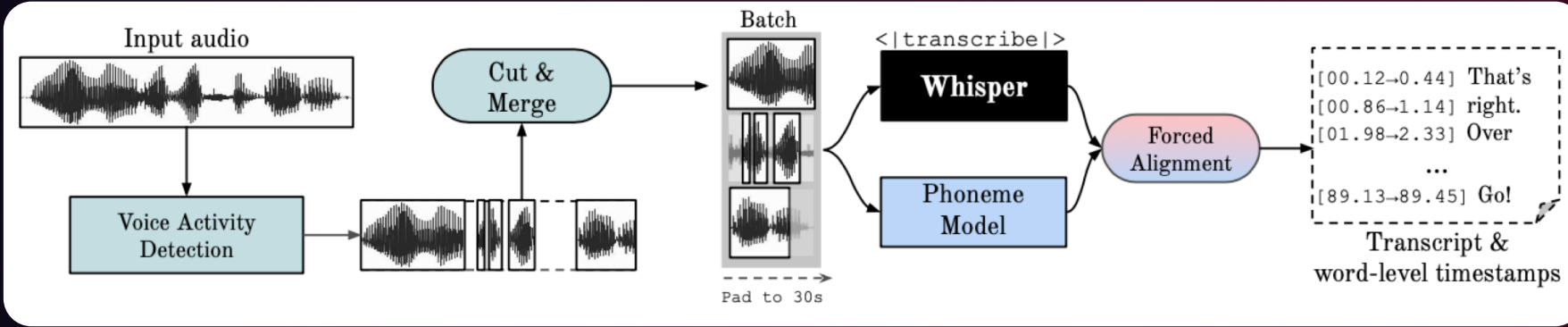
Стало чуууть лучше

Субтитрирование

Пайплайн

Работа редактора состоит из вертикализации и субтитрирования.
Субтитрирование мы тоже частично автоматизировали:

1. Взяли за основу WhisperX
2. Прикрутили ансамбль Voice Activity Detection (Pyannote + Speechbrain + Silero)
3. Подтюнили группировку фрагментов при распознавании
4. Подобрали гиперпараметры



Эксперимент с модерацией

Модерация вертикального контента

100	50	32	64%
Нарезали	Выбрали	Прошли ревью	Успеха

Частые проблемы замеченные при просмотре всех 100 видео:

Не выбран действующий/движущийся персонаж/объект — 10%

Объект интереса в кадре, но кроп не оптимальный — 10%

Экшн сцены это боль — 9%

Отсутствие переключения между активными спикерами в одном шоте — 8%

Выбор не говорящего персонажа в шоте где есть говорящий — 8%

Частые проблемы

1. Не выбран действующий/движущийся персонаж/объект — 10%

Частые проблемы

Модерация вертикального контента

1. Не выбран действующий/движущийся персонаж/объект — 10%
2. Объект интереса в кадре, но кроп не оптимальный — 10%

Частые проблемы

Модерация вертикального контента

1. Не выбран действующий/движущийся персонаж/объект — 10%
2. Объект интереса в кадре, но кроп не оптимальный — 10%
3. Экшн сцены это боль — 9%

Частые проблемы

Модерация вертикального контента

1. Не выбран действующий/движущийся персонаж/объект — 10%
2. Объект интереса в кадре, но кроп не оптимальный — 10%
3. Экшн сцены это боль — 9%
4. Отсутствие переключения между активными спикерами в одном шоте — 8%

Частые проблемы

Модерация вертикального контента

1. Не выбран действующий/движущийся персонаж/объект — 10%
2. Объект интереса в кадре, но кроп не оптимальный — 10%
3. Экшн сцены это боль — 9%
4. Отсутствие переключения между активными спикерами в одном шоте — 8%
5. Выбор не говорящего персонажа в шоте где есть говорящий — 8%

Премодерация на Я.Заданиях — инструкция

Критерии корректного вертикального видео:

- Нет чёрных горизонтальных полос снизу и сверху кадра
- Не должно быть субтитров.Не должно быть рывков и тряски, если их не было в исходном горизонтальном видео
- Должен присутствовать тот объект или персонаж, который является главным по смыслу
- Если главных объектов и персонажей несколько, по возможности они должны помещаться в границы вертикального кадра целиком
- Объект интереса или главный персонаж по возможности должны быть в центре

Подробности по QR

Премодерация на Я.Заданиях — пример задания

Модерация верт. контента

4 946	50%	30 766 ₽	6.2 ₽
Видео	Прошли модерацию	Затрат	Затрат на ролик

Ускорение монтажа. ДО

Модерация вертикального контента

20 минут

Монтаж ролика
без автоматизации

7.8 минут

Монтаж ролика с автоматической
генерацией субтитров

61 %

Ускорение

160 руб

Монтаж ролика
без автоматизации

80 руб

Монтаж ролика с автоматической
генерацией субтитров

50 %

Удешевление

Ускорение монтажа. ПОСЛЕ

Модерация вертикального контента

20 минут

Монтаж ролика
без автоматизации

4.2 минут

Монтаж ролика с автоматической
генерацией субтитров

79 %

Ускорение

160 руб

Монтаж ролика
без автоматизации

40 руб

Монтаж ролика с автоматической
генерацией субтитров

75 %

Удешевление

Итоги

Успех

50%

Прохождение премодерации

58%

Прохождение модерации

29%

Итоговый процент хорошей вертикализации

74 руб

Итоговая средняя стоимость ролика

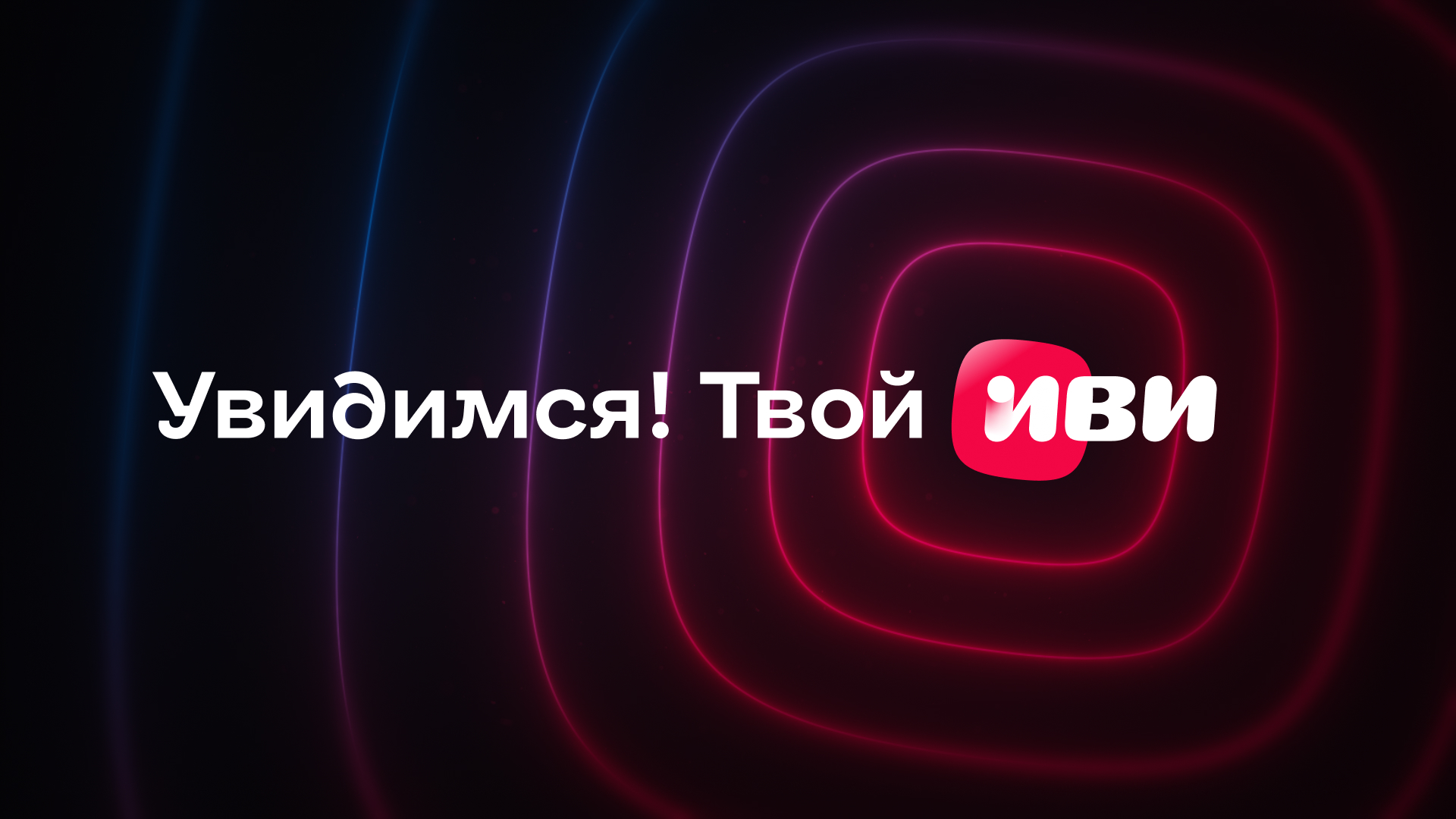
54%

Удешевление

66%

Ускорение

Спасибо за внимание!
Вопросы?


Увидимся! Твой ИВИ