DOI: 10.15514/ISPRAS-2025-37(3)-6

# Нейросетевой метод для стабильного во времени матирования видеопоследовательностей с людьми

[1,2] *И.А. Молодецких, ORCID: 0000-0002-8294-0770 <ivan.molodetskikh@graphics.cs.msu.ru>*
[1] *М.В. Ерофеев, ORCID: 0000-0001-7786-4358 <merofeev@graphics.cs.msu.ru>*
[1] *А.В. Москаленко, ORCID: 0000-0003-4965-0867 <andrey.moskalenko@graphics.cs.msu.ru>*
[1,2] *Д.С. Ватолин, ORCID: 0000-0002-8893-9340 <dmitriy@graphics.cs.msu.ru>*

[1] *Московский государственный университет имени М.В. Ломоносова, Россия, 119991, Москва, Ленинские горы, д. 1.*
[2] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

**Аннотация.** Предлагается новый метод матирования видео с людьми на основе нейросетей, не требующий дополнительных входных данных, таких как тернарные карты. Разработанная нейросетевая архитектура достигает стабильности получаемых карт прозрачности во времени за счёт свёрточных LSTM-модулей в прямых соединениях U-Net в сочетании со сглаживанием карт сегментации на основе поблочной оценки движения. Также предлагается алгоритм генерации искусственного движения, генерирующий обучающие видеоклипы для сети матирования видео, используя фотографии с эталонными картами прозрачности и фоновые видео. К фотографиям и их картам прозрачности применяются случайные карты сдвигов, имитирующие движение в реальных видео. Затем производится композиция результата с фоновыми клипами. Искусственное движение позволяет обучать глубокие нейросети, работающие с видео, в отсутствие большого размеченного набора видеоданных, и предоставляет эталонный оптический поток переднего плана обучающих видео для использования в функциях потерь.

**Ключевые слова:** матирование видео; семантическое матирование людей; сглаживание во времени; глубокое обучение.

**Для цитирования:** Молодецких И.А., Ерофеев М.В., Москаленко А.В., Ватолин Д.С. Нейросетевой метод для стабильного во времени матирования видеопоследовательностей с людьми. Труды ИСП РАН, том 37, вып. 3, 2025 г., стр. 85–106. DOI: 10.15514/ISPRAS–2025–37(3)–6.

**Благодарности:** Исследование поддержано грантом центрам искусственного интеллекта Аналитического центра при Правительстве РФ, соглашение № 000000D730321P5Q0002, и соглашением с ИСП РАН от 02.11.2021 № 70-2021-00142.

# Temporally Coherent Person Matting Trained on Fake-Motion Dataset

[1,2] *I.A. Molodetskikh*, ORCID: 0000-0002-8294-0770 <*ivan.molodetskikh@graphics.cs.msu.ru*>
[1] *M.V. Erofeev*, ORCID: 0000-0001-7786-4358 <*merofeev@graphics.cs.msu.ru*>
[1] *A.V. Moskalenko*, ORCID: 0000-0003-4965-0867 <*andrey.moskalenko@graphics.cs.msu.ru*>
[1,2] *D.S. Vatolin*, ORCID: 0000-0002-8893-9340 <*dmitriy@graphics.cs.msu.ru*>

[1] *Lomonosov Moscow State University,*
*GSP-1, Leninskie Gory, Moscow, 119991, Russia.*
[2] *Ivannikov Institute for System Programming of the Russian Academy of Sciences,*
*25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

**Abstract.** We propose a novel neural-network-based method to perform matting of videos depicting people that does not require additional user input such as trimaps. Our architecture achieves temporal stability of the resulting alpha mattes by using motion-estimation-based smoothing of image-segmentation algorithm outputs, combined with convolutional-LSTM modules on U-Net skip connections. We also propose a fake-motion algorithm that generates training clips for the video-matting network given photos with ground-truth alpha mattes and background videos. We apply random motion to photos and their mattes to simulate movement one would find in real videos and composite the result with the background clips. It lets us train a deep neural network operating on videos in an absence of a large annotated video dataset and provides ground-truth training-clip foreground optical flow for use in loss functions.

**Keywords:** video matting; semantic person matting; temporal smoothing; deep learning.

**For citation:** Molodetskikh I.A., Erofeev M.V., Moskalenko A.V., Vatolin D.S. Temporally coherent person matting trained on fake-motion dataset. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025. pp. 85-106 (in Russian). DOI: 10.15514/ISPRAS-2025-37(3)-6

**Acknowledgements.** This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement [agreement identifier 000000D730321P5Q0002] and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 [agreement number 70-2021-00142].

## *1. Введение*

Матирование, или нахождение карты прозрачности объекта переднего плана, является одной из основных операций обработки изображений и видеопоследовательностей. Задача матирования – определить точные значения прозрачности пикселей объекта на переднем плане, а также цвета переднего и заднего плана. Результатом матирования является «вырезанное» изображение объекта, которое можно вставить на другой, произвольный фон при помощи композиции.

Помимо систем видеоконференций, замена фона, а значит определение прозрачности пикселей переднего плана, находит широкое применение в проведении прямых трансляций, в задачах дополненной реальности и в киноиндустрии в области специальных эффектов. С развитием технологий компьютерного зрения и компьютерной графики значительно возросло число сцен, для съёмки которых быстрее и выгоднее создать искусственный задний план при помощи компьютерного моделирования и прочих подобных техник, чем строить специальные павильоны и декорации. Для таких сцен актёры снимаются отдельно, а затем вставляются на искусственный фон при помощи композиции.

Наиболее надёжным и, до недавних пор, популярным способом выделения объектов переднего плана при съёмке является техника хромакея, или съёмки на заднем плане, окрашенном в определённый ключевой цвет, как правило, зелёный. Далее в отснятом таким образом видео пиксели цвета, близкого к ключевому, меняются на прозрачные с помощью

программного инструмента редактирования видео. Именно хромакей использовался для замены фона в ранних версиях системы Zoom.

Однако техника хромакея накладывает ряд ограничений. Самым существенным из них является необходимость покупки и установки зелёного фонового экрана. Для большинства пользователей систем видеоконференций это не рационально в финансовом плане и представляет значительные неудобства в связи с размерами экрана, который нужно разместить в комнате перед рабочим местом. Также необходимо помнить про ограничения на цвета одежды. Дополнительно усложняет использование хромакея необходимость в однородном освещении экрана.

Матирование помогает решить эти проблемы, поскольку не требует специального заднего плана и какого-либо оборудования со стороны пользователя. Однако, в связи с этим матирование – гораздо более сложная задача, чем выделение объекта при помощи хромакея, потому что в общем случае неизвестно, какие из объектов являются объектами переднего плана и должны быть выделены.

Для указания объектов в задаче матирования часто используются тернарные карты. На них каждый пиксель отмечен одним из трёх цветов: принадлежность к переднему плану, принадлежность к заднему плану или неопределённая принадлежность. Алгоритму матирования в таком случае нужно найти значения прозрачности у неопределённых пикселей. Примеры подобных алгоритмов описаны в работах [1-3]. Однако существуют другие ручные и автоматические способы указания принадлежности объектов переднему или заднему плану, например, мазки или нейросетевые признаки.

Конкретно для систем видеоконференций необходим полностью автоматический поиск объектов переднего плана, например, с использованием нейросетевого алгоритма. В этом случае в качестве первого шага используется нейросеть для сегментации изображений, определяющая пиксели, принадлежащие определённому классу, в данном случае «человек». Так алгоритм матирования будет автоматически выделять людей на видеопотоке без каких-либо дополнительных входных данных.

Дополнительной сложностью в задаче матировании видео является поддержание стабильности полученных карт прозрачности во времени. Даже при наличии дополнительных входных данных, покадровое применение большинства алгоритмов матирования изображений даёт карты прозрачности, которые заметно вибрируют и мерцают при преобразовании их обратно в видеоформат, что является неприятным и отвлекающим артефактом.

Особенно часто мерцание проявляется на мелких деталях, таких как кончики волос людей на переднем плане. Одним из способов решения проблемы стабильности во времени является матирование кадров пониженного разрешения и последующее повышение разрешения полученных карт прозрачности обратно до исходного. Таким образом, алгоритм матирования выполняет основную задачу выделения человека в кадре, а алгоритм повышения разрешения далее восстанавливает мелкие детали, основываясь на исходных кадрах видео, что позволяет уменьшить или избежать мерцания.

В данной работе для решения проблемы мерцания предлагается алгоритм темпорального сглаживания результатов работы метода сегментации, а также свёрточные LSTM-слои [4] в основном методе матирования.

При обучении нейросетевых алгоритмов, одним из основных факторов, влияющих на финальное качество, является количество и качество данных для обучения. На момент написания, публично доступные наборы видео с эталонными картами прозрачности людей сильно ограничены в размерах, и таких наборов немного. При этом, существует несколько больших наборов фотографий людей с эталонными картами прозрачности [5-6]. Использование данных наборов фотографий потенциально может сильно повысить качество и обобщаемость обучаемой нейросети матирования.

Поэтому, в данной работе также предлагается новый алгоритм генерации искусственного движения, с помощью которого из фотографии человека на переднем плане и видео заднего плана можно получить видео обучающий пример, путём плавного искажения переднего плана в течение нескольких кадров. В данной работе показано, что предложенное искусственное движение может быть использовано для обучения нейросети, которая впоследствии корректно работает на настоящих входных видео и выдаёт стабильные во времени карты прозрачности.

Основной вклад данной работы состоит в следующем:

- Предложен новый метод для матирования видео с людьми на переднем плане без дополнительных входных данных на основе глубокой нейросети с архитектурой U-Net, блоками LSTM и модулем внимания на прямых соединениях.
- Предложен новый алгоритм искусственного движения для генерации обучающих видеоклипов для нейросети, используя изображения с эталонными картами прозрачности и видео заднего плана.
- Предложен новый метод сглаживания во времени карт вероятности метода сегментации изображений, который значительно увеличивает их временную стабильность.

## *2. Обзор литературы*

В данном разделе приведён обзор существующих подходов к задаче матирования изображений и видео, уделяя основное внимание полу- и полностью автоматическим семантическим методам.

### 2.1 Методы матирования изображений

В настоящее время активно развивается область матирования изображений – подзадача матирования видеопоследовательностей. Это в значительной степени связано со скачком в развитии нейросетей: матирование изображений – одна из задач, в которых нейросетевые подходы показали себя намного лучше классических. К тому же, алгоритмы матирования изображений менее ресурсозатратные по сравнению с алгоритмами матирования видео, что упрощает их интеграцию в системы видеоконференций, где они должны работать на пользовательских устройствах. Следующие работы наиболее релевантны разработанному методу; подробный обзор классических алгоритмов можно найти в работах [7-8].

Первым алгоритмом автоматического матирования фотографий людей без необходимости дополнительных входных данных был алгоритм, предложенный в [6]. Авторы создали набор данных, состоящий из 2000 портретов людей с эталонными картами прозрачности, и обучили нейросеть, предсказывающую тернарную карту по изображению. Затем изображение и тернарная карта подавались в классический алгоритм матирования для нахождения значений прозрачности в неизвестных областях. Также нейросети на вход подавалась «средняя» карта прозрачности по всем фотографиям из набора данных, выровненная относительно входной фотографии, что уменьшило число фоновых пикселей, ошибочно отмечаемых нейросетью как пиксели переднего плана.

Авторы [9] также предлагают решение задачи автоматического матирования портретов людей. Для этого используется архитектура из двух нейросетей: первая сеть предсказывает тернарную карту, а вторая сеть использует её в качестве дополнительных входных данных для предсказания карты прозрачности. Затем карта прозрачности комбинируется с тернарной картой для получения результирующей карты. Нейросети обучаются сначала по отдельности, а затем совместно. Кроме того, большим вкладом данной работы является набор из 34 425 портретов людей с эталонными картами прозрачности, опубликованный осенью 2019 года в открытом доступе [5].

В работе [10] предлагается новый нейросетевой блок для повышения разрешения карт признаков на этапе декодирования. Авторы заметили, что все используемые на текущий момент операции повышения разрешения (билинейная интерполяция, метод ближайшего соседа, транспонированная свёртка, повышение разрешения с использованием индексов с соответствующего слоя с этапа кодирования) можно обобщить на последний случай и предложили использовать обучаемый свёрточный слой для получения карты индексов. Использование подобных блоков для повышения разрешения привело к повышению качества работы ряда нейросетевых методов матирования изображений.

Метод, предложенный в [1] использует две нейросети. Первая, построенная по архитектуре кодировщик-декодировщик, выдаёт изначальное предсказание карты прозрачности, а вторая, состоящая из небольшого числа свёрточных слоёв, улучшает изначальное предсказание. За счёт отсутствия операций понижения и повышения разрешения, вторая сеть лучше обрабатывает мелкие детали изображения и, таким образом, увеличивает детализацию результирующей карты прозрачности.

Нейросетевой подход, предложенный в статье [11], ориентирован на быстрое выполнение на мобильных устройствах. Авторы использовали легковесную архитектуру с поканальными свёртками и квантование нейросети, что позволило им достичь скорости работы алгоритма, равной 30 кадрам в секунду, на современных мобильных устройствах. Данный подход ориентирован на автоматическое матирование портретов, и, соответственно, не принимает на вход тернарные карты. Вместо этого информация о принадлежности пикселей фону или переднему плану определяется самой нейросетью на основании обученной классификации людей в кадре. Основным ограничением данного подхода является низкое качество карт прозрачности, особенно для фотографий высокого разрешения.

Метод, предложенный авторами [12], может автоматически генерировать карты прозрачности для изображений с разными объектами переднего плана. В нём используется нейросеть из трёх частей: модули предсказания вероятности переднего и заднего плана и объединяющая сеть, которая выдаёт карту коэффициентов смешивания для вычисления итоговой карты прозрачности.

Авторы [13] предложили алгоритм, состоящий из нескольких нейронных сетей, что позволило им использовать как грубые, так и детальные эталонные карты прозрачности портретных фотографий. Грубые карты легко размечать, в то время как детальные карты размечать очень затратно по времени, но они сильно помогают в обучении хорошей нейросети для матирования. Сначала, нейросеть предсказания маски выдаёт грубую семантическую маску. Затем, нейросеть унификации качества увеличивает качество маски, чтобы сделать её похожей на детальные эталонные карты. Наконец, нейросеть улучшения матирования использует полученную маску как ориентир для генерации финальной карты прозрачности.

## 2.2 Методы матирования видеопоследовательностей

В отличие от задачи матирования изображений, задача матирования видеопоследовательностей исследуется менее активно. Наибольший интерес здесь представляют методы, принимающие на вход тернарную карту не на каждый кадр видеопоследовательности, а только на отдельные, ключевые кадры. Здесь всё ещё преобладают классические подходы, «протягивающие» карту прозрачности с ключевого кадра на последующие.

Метод [14] использует два ключевых кадра в начале и конце фрагмента видеопоследовательности, и обеспечивает стабильность полученной карты прозрачности во времени между этими двумя кадрами. Авторы предложили метод построения оптического потока между кадрами на основе алгоритма PatchMatch [15] и способ «протягивания» карт прозрачности с использованием построенного оптического потока. Отдельным плюсом

данного метода является возможность использовать произвольный базовый алгоритм матирования изображений для получения карт прозрачности на ключевых кадрах.

Авторы [16] предлагают классический подход, основанный на представлении пикселей в областях переднего и заднего плана произведением матрицы словаря и разреженной кодовой матрицы. Словарь обучается на входной видеопоследовательности с помощью оптимизационного процесса, принимающего во внимание стабильность во времени. К сожалению, в данном подходе требуются дополнительные входные данные в виде тернарных карт на каждый кадр входного видео, что ограничивает его применимость.

В работе [17] предлагается метод матирования портретных видео без использования дополнительных входных данных в виде тернарных карт или мазков. Однако, метод требует снятых вручную изображений-примеров фона видео, которые помогают нейросети обрабатывать сложные фоновые детали.

Для работы метода, предложенного в [18], от пользователя вместо тернарной карты требуется одна фотография заднего плана. Точное выравнивание обрабатываемого видео и фотографии заднего плана не требуется. В этом методе используется нейросеть для сегментации изображений. Одним из ограничений является динамический задний план, например, водопад, или сильное движение камеры при съёмке. В этих случаях алгоритм может ошибочно отметить какие-то элементы фона как передний план.

Авторы [19] предлагают метод сегментации объектов видео, использующий генерацию синтетических обучающих примеров при помощи случайных геометрических и цветовых преобразований как к композиции передних и задних планов, так и к передним и задним планам по отдельности. В следующей статье [20] авторы улучшили свой метод путём генерации коротких видеоклипов по похожему принципу.

## 2.3 Наборы данных

Существует несколько общедоступных наборов фотографий и видео людей с эталонными картами прозрачности.

- Набор данных AISegment.com [5], собранный авторами [9], содержащий 34 425 портретных фотографий людей с эталонными картами прозрачности.
- 2000 портретных фотографий с эталонными картами прозрачности, собранных авторами [6].
- Наборы данных PhotoMatte85 и VideoMatte240K [21], собранные авторами [22], которые содержат 85 портретных фотографий и 484 видеоклипов с людьми с эталонными картами прозрачности.

Эти наборы данных были сняты на зелёном фоне с помощью хромакея, что означает, что у них нет «родного» фона, и их можно использовать только для композиции с другим задним планом.

## 2.4 Выводы

Задача матирования изображений с использованием тернарных карт в последние годы активно исследуется. Это связано с возросшей популярностью нейросетевых алгоритмов для решения задач компьютерного зрения. Все наилучшие алгоритмы матирования в эталонном тесте https://alphamatting.com основаны на нейросетях.

Постепенно начинает исследоваться задача автоматического матирования без необходимости дополнительных входных данных в виде тернарных карт или мазков. В основном методы в данной области направлены на матирование портретов людей, что может быть связано с возрастающей популярностью мобильных приложений для наложения эффектов на фотографии.

Задача матирования видео исследуется менее активно. Это может быть связано с вычислительной сложностью в применении нейросетей к видеопоследовательностям и отсутствию достаточно больших наборов данных для обучения. Однако, в последнее время начинают появляться многообещающие подходы с использованием ограниченных входных данных, например, одного изображения фона, такие как [17-18].

Многие из подходов матирования изображений [23-25, 9-10] можно адаптировать для применения к задаче автоматического матирования видео. Также большую ценность для данной работы составляют наборы портретов людей с эталонными картами прозрачности [6, 9, 21], которые можно использовать для создания видеопоследовательностей для обучения нейросети.

## *3. Предложенный метод*

Предложенный подход основан на свёрточной нейронной сети для предсказания карты прозрачности на основе карты вероятности, получаемой предобученным методом семантической сегментации изображений. Отсутствие полностью связных слоёв в нейросети даёт возможность подавать на вход кадры произвольного размера. Схема архитектуры нейросетевого алгоритма представлена на рис. 1.

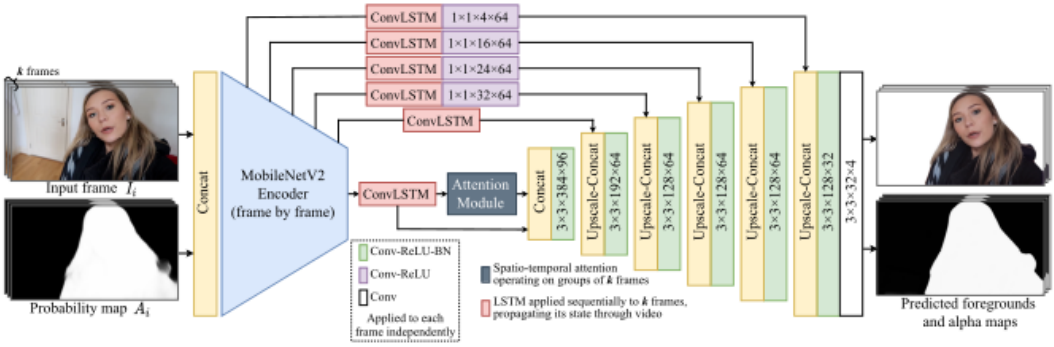


*Рис. 1. Архитектура предложенной нейросети.*
*Fig. 1. Architecture of the proposed neural network.*

Для учёта стабильности во времени в архитектуре используются модули долгосрочной кратковременной памяти (LSTM) [26], а именно их свёрточная адаптация [4]. Модули помещаются на пути прямых соединений между слоями кодировщика и декодировщика. Использование данных модулей позволяет нейросети передавать информацию о предсказываемых картах прозрачности вперёд во времени, что даёт возможность добавить в функцию потерь слагаемое, штрафующее за отсутствие стабильности.

На каждом шаге во времени $s$, нейросеть принимает на вход набор $k$ последовательных кадров видео $\{I_i\}_{i=s}^{s+k}$, соответствующие карты вероятности $\{A_i\}_{i=s}^{s+k}$ и скрытое состояние LSTM с предыдущего кадра $H_{s-1}$. Результатом шага нейросети являются предсказанные карты прозрачности $\{\alpha_i\}_{i=s}^{s+k}$, пиксели переднего плана $\{\mathrm{FG}_i\}_{i=s}^{s+k}$, а также новое скрытое состояние LSTM $H_{s+k}$.

Временна́я информация содержится в скрытом состоянии модулей LSTM. Это состояние инициализируется нулевыми значениями, и затем сохраняется между запусками нейросети на последовательных кадрах видео.

Предложенный подход может быть адаптирован для матирования других классов объектов путём использования соответствующего выходного класса объекта метода семантической сегментации изображений и переобучения предложенной нейросети на наборе данных матирования нужного класса объекта. Однако, фокус данной работы именно на матировании

людей, так как эта задача на текущий момент имеет много практических применений, и соответственно много наборов данных для обучения и экспериментальной оценки.

## 3.1 Архитектура нейросети

Нейросеть построена по модели кодировщик-декодировщик с прямыми соединениями между слоями кодировщика и декодировщика на одинаковой глубине. Подобный способ построения («U-Net») используется для решения большого количества различных задач компьютерного зрения. Прямые соединения помогают нейросети лучше восстанавливать мелкие детали изображений, теряющиеся при сжатии информации в размерность самого глубокого представления.

В качестве основы для кодировщика взята MobileNetV2 [27], так как данная архитектура хорошо себя показывает в задачах матирования видео [2][1]. Число каналов первого слоя сети увеличено с трёх до четырёх для передачи на вход карты вероятности. Для прямых соединений берутся выходы 2-го, 4-го, 7-го, 14-го и 18-го блоков сети в реализации TorchVision [28]. Прямые соединения проходят через свёрточные LSTM-модули, которые дают результирующим картам прозрачности стабильность во времени.

Кроме того, на самом глубоком прямом соединении используется модуль внимания, основанный на предложенном в работе [29]. Сначала, карта признаков размерностью в 320 каналов конкатенируется с тремя дополнительными каналами, содержащими для каждого пикселя его номер кадра, x-координату и y-координату. Эти дополнительные каналы часто называют картой индексов. Затем, два полносвязных слоя уменьшают число каналов «выпрямленной» карты признаков с 323 до 64 и генерируют карты запроса и ключа для вычисления внимания. Для карты значения также используется полносвязный слой, но здесь используется 320-канальная карта признаков сама по себе, без координат пикселей. Далее применяется масштабированная операция внимания с использованием скалярного произведения, и восстанавливается двумерная размерность результата.

В то время как модули LSTM обрабатывают кадры последовательно и накапливают информацию во времени, модуль внимания оперирует на группах по $k$ кадров за раз, распространяя информацию внутри каждой группы.

## 3.2 Определение положения людей

Первым шагом в работе предложенного алгоритма является определение положения людей на входных кадрах видео – получение карты вероятности $A_i$. Для этого используется широко применяемая на практике нейросетевая архитектура для сегментации изображений DeepLabv3+ [30] с базовой архитектурой Xception [31], предобученная на наборе данных PASCAL VOC 2012 [32]. Конкретно, была взята реализация TorchVision [28].

Перед подачей в DeepLabv3+, входное изображение изменяется в размере до 520 × 520 пикселей при помощи билинейной интерполяции. Далее к выходной карте признаков применяется поканальная операция softmax. Результат – карта вероятностей принадлежности каждого пикселя изображения к каждому из 21 класса набора данных PASCAL VOC. Из этой карты выделяется слой, соответствующий классу «человек», и увеличивается или уменьшается до исходного размера входного кадра, используя билинейную интерполяцию. Таким образом, значение каждого пикселя в выходной карте данного шага можно рассматривать как вероятность того, что данный пиксель изображения принадлежит человеческой фигуре.

DeepLabv3+ является составной частью предложенного метода и используется как при обучении, так и при применении предложенной нейросети.

## 3.3 Генерация обучающих видеопоследовательностей

Для обучения глубокой нейронной сети для матирования видео требуется большой набор данных видео с эталонными картами прозрачности. К сожалению, наборы данных в открытом доступе ограничены в размерах. Однако, существуют большие наборы портретных фотографий людей с эталонными картами прозрачности. В этом разделе предлагается алгоритм для генерации обучающих видеопоследовательностей, используя фотографии переднего плана с эталонными картами прозрачности и фоновые видео. На рис. 2a представлен пример сгенерированной видеопоследовательности.

Для генерации обучающей последовательности длиной в $N$ кадров, сначала выбирается случайное фоновое видео длиной $N$ и два случайных портрета для переднего плана. Процедура искусственного движения, описанная ниже, обрабатывает портреты и выдаёт два видео переднего плана длиной $N$ кадров, а также два соответствующих эталонных оптических потока. В половине случаев выходным видео является композиция обоих передних планов и фонового видео; в остальных случаях же на выход выдаётся одно из видео переднего плана как есть, а второе видео и фоновое видео отбрасываются. Наконец, в 5% случаев на выход выдаётся фоновое видео без каких-либо изменений, что сделано для улучшения качества обработки предложенной нейросетью видео без людей. В качестве последнего шага, к кадрам сгенерированного видео применяется сжатие алгоритмом JPEG со случайным параметром качества от 30% до 80%.

Использование двух видео переднего плана вместо одного улучшает качество работы нейросети матирования на видео, где на переднем плане присутствует несколько человек. На рис. 2a показан пример сгенерированного видео с двумя передними планами друг поверх друга. Одновременно с этим, 50%-я вероятность использования одного переднего плана как есть без композиции и JPEG-сжатие на последнем шаге предотвращает переобучение нейросети матирования на кадры, полученные искусственной композицией. Это происходит благодаря тому, что в половине случаев кадры вообще не являются результатом композиции, а в другой половине артефакты JPEG-сжатия не дают нейросети легко «найти» границы композиции.

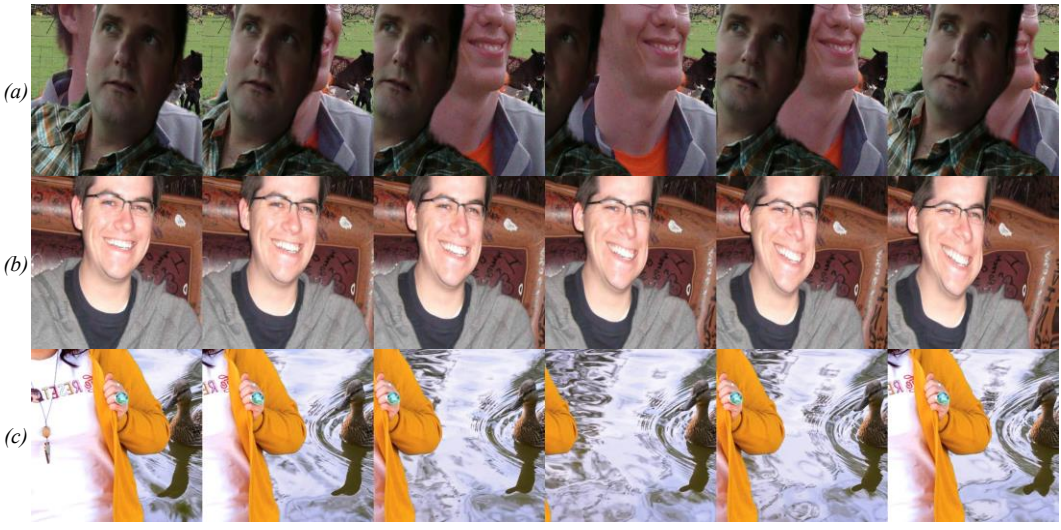


*Рис. 2. Примеры сгенерированных видеопоследовательностей: (a) итоговая обучающая последовательность; (b) последовательность, сгенерированная с помощью только основного компонента искусственного движения; (c) пример работы компонента сдвига за границу кадра.*
*Fig. 2. Examples of generated clips: (a) final training clip; (b) clip from the base fake-motion algorithm; (c) example of the fake-motion shift component's effect.*

## 3.4 Искусственное движение

Процедура искусственного движения генерирует случайные карты оптического потока в трёх масштабах. Сначала, генерируется один вектор оптического потока путём выбора двух значений ($x$ и $y$) из нормального распределения $\mathcal{N}(0,32)$. Этот вектор можно рассматривать как карту оптического потока размером $1 \times 1$. Данная карта представляет собой сильное глобальное движение. Далее, генерируется карта оптического потока размерностью $\lfloor\frac{H}{128}\rfloor \times \lfloor\frac{W}{128}\rfloor$ из распределения $\mathcal{N}(0,16)$. Эта карта представляет собой более детальное движение. Наконец, генерируется карта размерностью $\lfloor\frac{H}{32}\rfloor \times \lfloor\frac{W}{32}\rfloor$ из распределения $\mathcal{N}(0,4)$, представляющая собой наиболее детальную компоненту движения.

Все сгенерированные карты увеличиваются до размера входного изображения ($H \times W$ пикселей) при помощи бикубической интерполяции, суммируются и разделяются $N$, желаемое число кадров генерируемой последовательности, для получения покадрового оптического потока. На рис. 2б показан пример клипа, сгенерированного описанным алгоритмом.

Для повышения устойчивости разрабатываемой нейросети к видео, где человек временно покидает границы кадра, к результирующему оптическому потоку с вероятностью $\frac{1}{3}$ добавляется компонента, сдвигающая человека на половину кадра и обратно на протяжении видео. Выбор направления для сдвига случаен с условием, что стороны, противоположной направлению сдвига, не касаются непрозрачные пиксели человека переднего плана. Это условие позволяет избежать артефактов повторения граничных пикселей. На рис. 2в показано влияние этой компоненты.

Также, так как в процедуре используется два передних плана, к оптическому потоку добавляется компонента, сдвигающая передний план на половину кадра на первом кадре. Этот шаг уменьшает вероятность полного перекрытия одного переднего плана другим при композиции.

Полученные покадровые карты оптического потока используются для смещения входных изображений и карт прозрачности, чтобы получить обучающую видеопоследовательность с необходимым количеством кадров. Оптический поток также используется в дальнейшем при вычислении функции потерь, описанных в следующем разделе.

## 3.5 Функции потерь

Для обучения предложенной нейросети используется четырёхкомпонентная функция потерь:

$$L = L_\alpha + L_G + L_L + \frac{1}{4}L_{\text{FG}}.$$

Первая компонента – попиксельное $L_1$-расстояние между картами прозрачности $\alpha'_i$, предсказанными нейросетью, и эталонными картами прозрачности $\alpha_i$:

$$L_\alpha = \frac{\sum_{i=1}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W}\left|\alpha'_{ijk} - \alpha_{ijk}\right|}{N \cdot H \cdot W},$$

где $N$ – число кадров в обучающей видеопоследовательности, $H$ и $W$ – её высота и ширина в пикселях соответственно.

При обучении нейросетей для матирования применяется $L_1$-расстояние, а не часто используемое в других задачах $L_2$-расстояние, так как обучение с помощью $L_2$-расстояния приводит к нейросети, предсказывающей сильно размытые карты прозрачности.

Вторая и третья компоненты отвечают за стабильность результирующих карт во времени. Они представляют из себя попиксельное $L_1$-расстояние между картами прозрачности различных кадров из обучающей последовательности, сдвинутых в соответствии с исходным оптическим потоком, созданным в процессе генерации искусственного движения. Вторая

компонента «сравнивает» карту прозрачности первого кадра с картой прозрачности каждого последующего кадра, а третья компонента – карты прозрачности между последовательными кадрами:

$$L_G = \frac{\sum_{i=2}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W}\left|\alpha'_{ijk} - \alpha'^{\to i}_{1jk}\right|V^{\to i}_{1jk}}{(N-1)\cdot H\cdot W}, L_L = \frac{\sum_{i=2}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W}\left|\alpha'_{ijk} - \alpha'^{\to i}_{(i-1)jk}\right|V^{\to i}_{(i-1)jk}}{(N-1)\cdot H\cdot W}.$$

Здесь $\alpha'^{\to i}_{l}$ обозначает карту прозрачности, предсказанную для $l$-го кадра, сдвинутую к $i$-му кадру в соответствии с эталонным оптическим потоком, получаемым в ходе процедуры генерации искусственного движения, описанной в разделе 3.4. $V^{\to i}_{l}$ – маска пригодных для сравнения пикселей, равная 1 в точках, где, согласно оптическому потоку, пиксели $i$-го кадра соответствуют пикселям внутри $l$-го кадра и 0 в точках, где пиксели $i$-го кадра соответствуют пикселям за границей $l$-го кадра. Описанные карты проиллюстрированы на рис. 3.

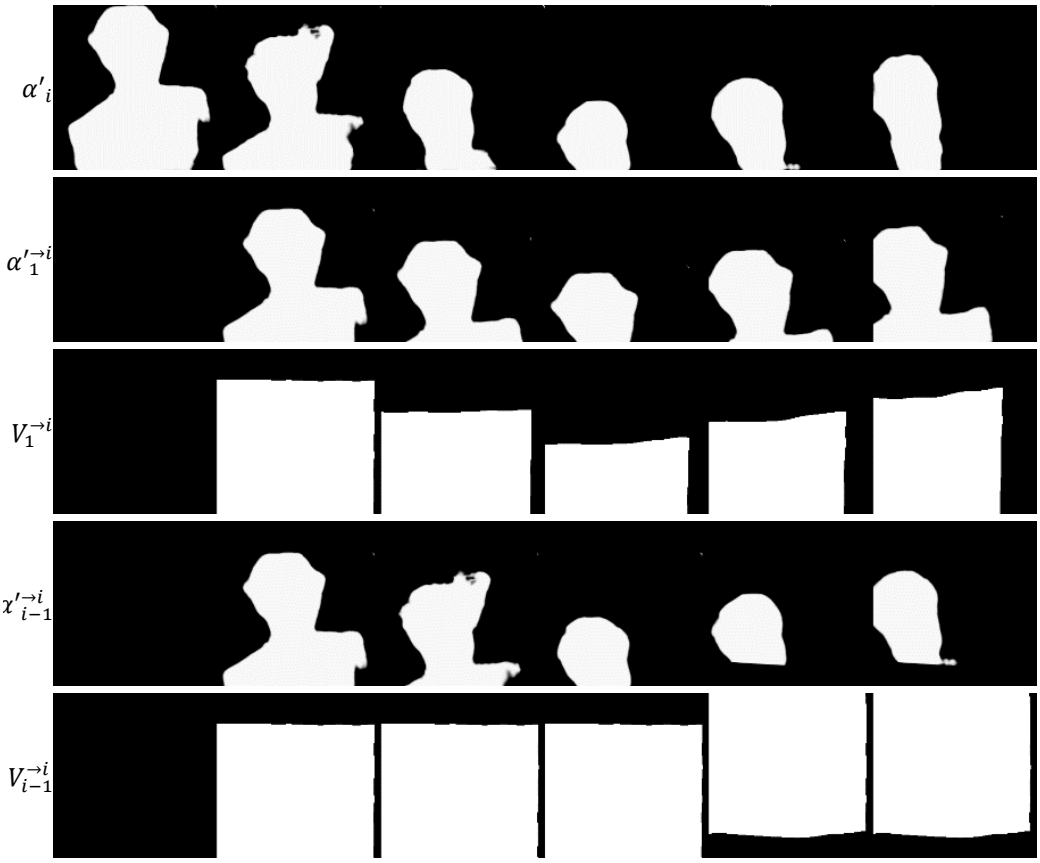


*Рис. 3. Иллюстрация карт, использующихся в функции потерь.*
*Fig. 3. Loss function component map visualization.*

Последний компонент, $L_{\text{FG}}$, – это попиксельное $L_1$ расстояние между предсказанным передним планом $\text{FG}'_i$ в цветовом пространстве RGB и эталонным передним планом $\text{FG}_i$, умноженное на предсказанную карту прозрачности:

$$L_{\text{FG}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W}\alpha'_{ijk}\sum_{c=1}^{3}\left|\text{FG}'_{ijkc} - \text{FG}_{ijkc}\right|}{N\cdot H\cdot W\cdot 3}.$$

## 3.6 Набор данных

Фоновые видео для процедуры генерации искусственных видео были взяты из набора DAVIS [33]. Конкретно, были выбраны 27 видео, не содержащие людей. Их средняя длина составила 80 кадров.

В качестве передних планов были взяты 40 074 портретных фотографий с эталонными картами прозрачности: 34 425 из набора [5], 2000 из набора [6] и 3649 изображений, размеченных в ходе данной работы. Из этих изображений, 4000 были отложены для валидации, а оставшиеся 36 074 были использованы для обучения.

При обучении нейросети использовались следующие аугментации. Для передних планов:

- случайный поворот кадра на ±15%,
- случайная обрезка кадра вплоть до 20% исходного размера,
- случайное отражение относительно вертикальной оси,
- случайное изменение яркости и контраста.

Для фоновых видео:

- случайная обрезка кадра вплоть до 8% исходного размера,
- случайное отражение относительно вертикальной оси,
- случайное изменение яркости и контраста.

Для композиции передних планов и фоновых видео: сжатие JPEG со случайным параметром качества от 30% до 80%. Эти аугментации, вместе со случайными сдвигами, описанными в разделе 3.4, дают процедуру генерации обучающих видеопоследовательностей, схожую с описанной в работах [19-20]. Предлагаемый основной компонент искусственного движения работает совместно с этой процедурой и ещё больше улучшает качество обучающих примеров.

При обучении и валидации, каждые две фотографии переднего плана используются для генерации одной обучающей видеопоследовательности, как описано в разделе 3.3. Следовательно, в одной эпохе обучения содержалось 18 037 обучающих видео и 2000 валидационных видео.

Описанный набор данных использовался только для обучения и валидации нейросети. Итоговая экспериментальная оценка проводилась на отдельном наборе видеопоследовательностей, описанной в разделе 4.

## 3.7 Разметка эталонных карт прозрачности

На сервисе Яндекс.Толока [34] пользователям предлагалось скорректировать карты прозрачности, полученные при помощи автоматического метода, схожего с описанным в разделе 3.2. Карта признаков на выходе нейросети DeepLabv3+ подавалась на вход алгоритму разреза графа GrabCut [35] для получения пикселей, принадлежащих человеку на переднем плане фотографии. Затем, при помощи морфологического расширения и сужения, границы выделенной области помечались, как пиксели с неизвестным значением прозрачности, и полученная тернарная карта подавалась вместе с исходной фотографией на вход алгоритму матирования изображений [1] для создания карты прозрачности. От пользователей требовалось исправить ошибки в карте прозрачности, используя мазки трёх типов – передний план, задний план и полупрозрачная область. Первые два типа мазков передавались в алгоритм GrabCut в качестве вспомогательных данных, а мазок типа полупрозрачная область переносился на тернарную карту. Изображение пользовательского интерфейса можно увидеть на рис. 4.

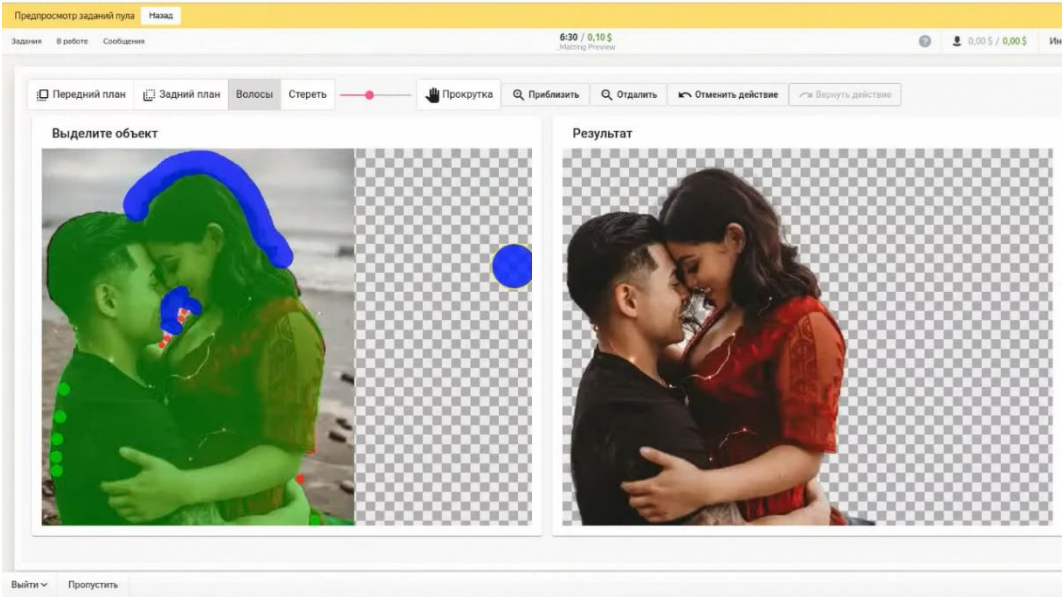


*Рис. 4. Пользовательский интерфейс для разметки фотографий на сервисе Яндекс.Толока.*
*Fig. 4. User interface for photo alpha matte annotation on Yandex.Toloka.*

## 3.8 Обучение нейросети

Нейросетевая модель была обучена на четырёх видеокартах Nvidia Tesla P100. В каждом обучающем наборе было по 8 видеоклипов длиной по шесть кадров разрешения 520 × 520 пикселей каждый. Для градиентного спуска использовался оптимизатор Adam [36] с темпом обучения, равным 0,001. После каждой эпохи темп обучения уменьшался в 0,7 раз. В процессе разработки, тестовые модели как правило сходились через семь эпох, или около 9,5 часов обучения.

## 3.9 Применение нейросети

При применении нейросети используется только выход карты прозрачности, а выход цвета переднего плана игнорируется. Это сделано в связи с тем, что на нём часто содержатся цветовые артефакты. Тем не менее, наличие этого выход и соответствующей компоненты функция потерь улучшило результат обучения.

Применение обученной нейросети напрямую на длинных видеопоследовательностях даёт высокую стабильность во времени, однако ведёт к появлению артефактов: LSTM-слои «запоминают» положение людей на переднем плане и не могут достаточно быстро адаптироваться к движению человека. Для решения этой проблемы используется следующая схема применения нейросети. Обработка каждого кадра входного видео начинается со сброса состояния LSTM-слоёв. Затем нейросети на вход подаются 3 предшествующих кадра видео, после чего нейросеть обрабатывает текущий кадр видео.

Подобная схема применения с одной стороны не ведёт к значительной потере стабильности во времени результирующих карт прозрачности, а с другой стороны позволяет сети очень быстро адаптироваться к резким движениям человека и предотвращает накопление ошибки.

Во время применения нейросети на вход $A_i$ передаются не исходная карта вероятности, полученная из алгоритма определения положения людей, а сглаженная карта вероятности, полученная алгоритмом, описанным в разделе 3.9. Это приводит к ещё большему повышению

стабильности во времени карт прозрачности на выходе нейросети, что подтверждается результатами проведённых объективных и субъективных оценок (раздел 4).

Передача сглаженных карт вероятности на вход нейросети во время обучения не дала видимого улучшения качества работы. Возможно, это связано с малой длиной обучающих примеров – сглаживание «не успевает» значительно изменить карты вероятности.

Скорость работы предложенного метода составляет около 1,7 кадров разрешения 1280 × 720 в секунду на машине с одной видеокартой Nvidia TITAN Xp и процессором Intel Xeon E5-2683 v3 с тактовой частотой 2 ГГц.

## 3.10 Алгоритм сглаживания карт вероятности

Покадровое применение алгоритма определения положения людей на видео даёт в результате сильно «дрожащую» последовательность масок. Для улучшения работы нейросети матирования, предлагается следующий алгоритм сглаживания во времени результатов работы сети определения положения людей при помощи векторов движения.

Пары последовательных кадров входной видеопоследовательности $\{I_i\}_{i=1}^{N}$ подаются в блочный алгоритм поиска векторов движения [37] в прямом и обратном порядке для получения прямых и обратных векторов движения. Векторы движения представляются в виде карты оптического потока: для каждого пикселя второго входного кадра известен вектор смещения в соответствующий ему пиксель из первого входного кадра, выраженный разницей координат по осям $x$ и $y$. Таким образом, получаются две последовательности карт: $\{F_i\}_{i=2}^{N}$ и $\{B_i\}_{i=1}^{N-1}$ прямого и обратного оптического потока, соответственно.

Далее, вычисляется согласованность прямого и обратного потока способом, похожим на предложенный в работе [38].

1. Карты обратного оптического потока сдвигаются в соответствии с картами прямого оптического потока. Обозначим получившиеся карты как $\left\{B_{i-1}^{\rightarrow i}\right\}_{i=2}^{N}$.
2. Карты сдвинутого обратного оптического потока вычитаются попиксельно из карт прямого оптического потока: $D_i = F_i - B_{i-1}^{\rightarrow i}$.
3. В каждом пикселе полученных карт вычисляется $L_2$-норма: $L_{ijk} = \sqrt{D_{ijkx}^2 + D_{ijky}^2}$,

   где $i = \overline{2, \mathrm{N}}$, $j = \overline{1, \mathrm{H}}$, $k = \overline{1, W}$, $H$ – высота кадра в пикселях, $W$ – ширина кадра в пикселях, $x$ и $y$ – компоненты векторов оптического потока.
4. Для получения карт согласованности от значений полученных карт попиксельно вычисляется экспонента: $C_i = e\left(\frac{-L_i}{100}\right)$.

Карта согласованности принимает значения, близкие к единице, при близости прямых и соответствующих им обратных векторов движения (что говорит о правильности вектора движения в данном пикселе) и значения, близкие к нулю, в противоположном случае.

Сглаживание происходит по следующей процедуре. На входе – карты вероятности $\{p_i\}_{i=1}^{N}$, полученные из DeepLabv3+, как описано в разделе 3.2. Для карт вероятности $p_i$ попиксельно вычисляются значения уверенности: $s_i = 4\left(p_i - \frac{1}{2}\right)$.

Финальные значения сглаженных карт вероятности $\{A_i\}_{i=1}^{N}$ вычисляются следующим образом:

$$A_1 = p_1,$$
$$A_i = s_i \cdot p_i + (1 - s_i)\left(C_i \cdot A_{i-1}^{\rightarrow i}\right),$$

где $i = \overline{2, \mathrm{N}}$, а $A_{i-1}^{\rightarrow i}$ – карты вероятности $A_{i-1}$, сдвинутые в соответствии с картами оптического потока $F_i$.

Интуитивно, в пикселях, где вероятность присутствия человека близка к нулю или единице (нейросеть «уверена» в своём ответе), в качестве значения прозрачности берётся эта вероятность, а в пикселях, где вероятность близка к $\frac{1}{2}$ (нейросеть «не уверена»), в качестве значения прозрачности берётся скомпенсированное по векторам движения значение прозрачности предыдущего кадра, отмасштабированное в соответствии с согласованностью векторов.

## *4. Экспериментальная оценка*

В процессе разработки различные модификации алгоритма оценивались как субъективно на выборке тестовых видео, так и с помощью объективных метрик. Сравнение проходило между предложенным нейросетевым методом и следующими методами.

- Deep Image Matting [1]: своя реализация, обучена на наборе данных Deep Image Matting.
- FBA Matting [2]: официальная реализация и модель, обученная на наборе данных Deep Image Matting.
- COSNet [39]: официальная реализация и модель, обученная на наборе данных DAVIS16.
- MMNet [11]: официальная реализация, обученная на наборе данных AISegment [5].
- Semantic Human Matting [9]: неофициальная реализация и модель, обученная на закрытом наборе данных разработчика.
- Late Fusion Matting [12]: официальная реализация и модель, обученная на наборе данных Deep Image Matting и закрытом наборе данных разработчика.
- CRGNN [40]: официальная реализация и модель, обученная на наборе данных Deep Image Matting и наборе данных разработчика.

Deep Image Matting и FBA Matting – это методы матирования изображений, занимающие первые места на эталонном сравнении https://videomatting.com [41], COSNet – это метод сегментации видео без учителя, занимающий высокое место на эталонном сравнении DAVIS16 [33], CRGNN – метод матирования видео, и MMNet, Semantic Human Matting и Late Fusion Matting – это методы семантического матирования фотографий с людьми. В сравнении также участвовали карты вероятности $A_i$, сглаженные предложенным в разделе 3.9 алгоритмом, для чего они интерпретировались как итоговые карты прозрачности.

Тернарные карты, необходимые для работы алгоритмов [1-2, 40], генерировались при помощи процедуры, аналогичной описанной в разделе 3.2: к местоположению людей в кадре, полученному при помощи нейросети DeepLabv3+, применялось морфологическое расширение и сужение для пометки граничных областей как пикселей с неизвестным значением прозрачности.

Пример результатов предложенных и участвовавших в сравнении методов приведён на рис. 5.

## 4.1 Субъективное сравнение

Для субъективной оценки алгоритма был использован сервис проведения субъективных исследований https://subjectify.us. Пользователям показываются пары видеопоследовательностей, полученных путём композиции тестовых видео на статический фон, используя карты прозрачности, полученные сравниваемыми алгоритмами. Задача пользователей формулировалась в терминах сравнения алгоритмов замены фона на видео, и от них требовалось выбрать наилучший, по их мнению, вариант из двух. Интерфейс субъективного исследования со стороны пользователя приведён на рис. 6.

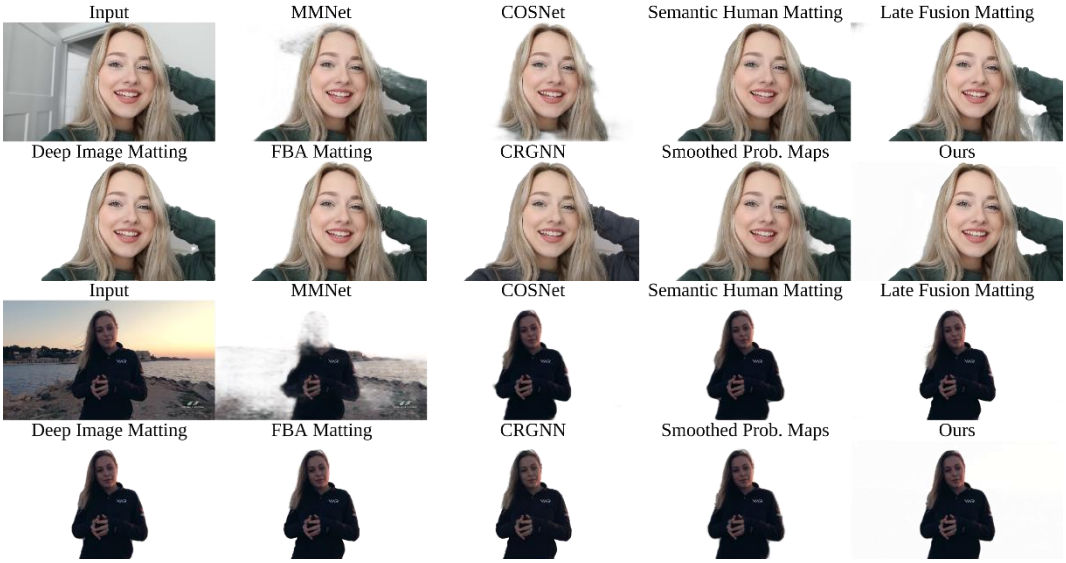


*Рис. 5. Результаты участвовавших в сравнении методов на двух тестовых видео.*
*Fig. 5. Results of the evaluated methods on two of the test clips.*

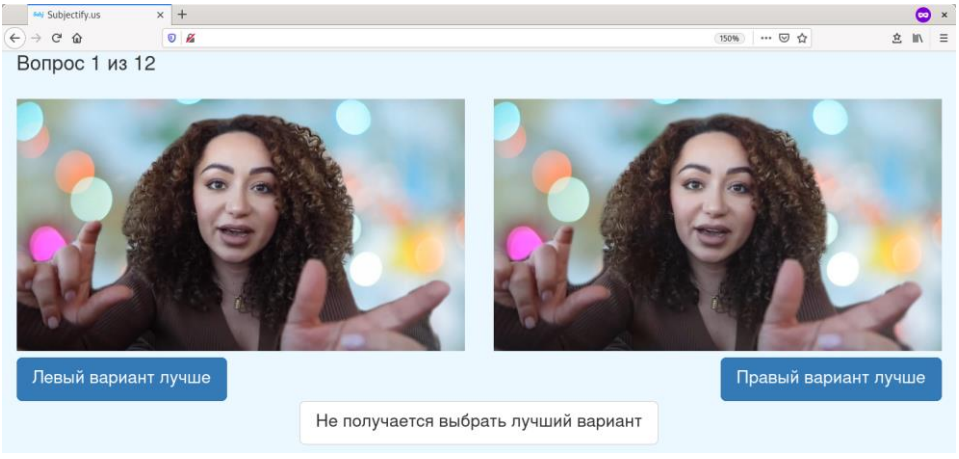


*Рис. 6. Результаты субъективного исследования по всем тестовым видеопоследовательностям. Чёрные метки погрешности показывают 95%-е доверительные интервалы.*
*Fig. 6. Subjective evaluation results on all test sequences. Error bars show 95% confidence intervals.*

Для исследования была составлена выборка из 2 видеопоследовательностей (*city* и *snow*) с эталонного теста https://videomatting.com, попадающих в целевые входные данные для разрабатываемого алгоритма (видео с людьми, но без крупных полупрозрачных областей), 2 видео, снятые в процессе работы, и 55 видео с людьми средней длиной в 371 кадр [42], вырезанных из различных видео любительской съёмки, найденных на видеохостинге YouTube.

В процессе исследования каждому пользователю необходимо было оценить 12 пар видеопоследовательностей. Среди них присутствовали две контрольных видеопоследовательности: результат покадрового применения алгоритма матирования изображения [1] и последовательность, полученная при помощи композиции с эталонной картой прозрачности. Если пользователь неправильно оценивал какую-либо из контрольных

пар, его ответы не учитывались в результатах. Для контрольных пар были использованы видеопоследовательности *city* и *snow*.

Всего таким образом было собрано 32 272 оценок пар видеопоследовательностей от пользователей, правильно оценивших контрольные пары. Попарные оценки, данные пользователями, использовались для настройки параметров модели Брэдли-Терри [43]. Эта модель максимизирует субъективные оценки при условии данных попарных оценок. Чем выше субъективная оценка, данная алгоритму настроенной моделью, тем выше вероятность предпочтения результата этого алгоритма в попарном сравнении.

Общие результаты по всем видео приведены на рис. 7. Методы матирования изображений [9, 2, 1, 12] показали неплохой результат, хотя нестабильность во времени в результирующих картах прозрачности уменьшает предпочтение зрителей этим методам. Результаты COSNet [39] достаточно сильно мерцали, несмотря на то что это алгоритм матирования видео. MMNet [11] показал худший результат, скорее всего, потому что он не предназначен для больших изображений.

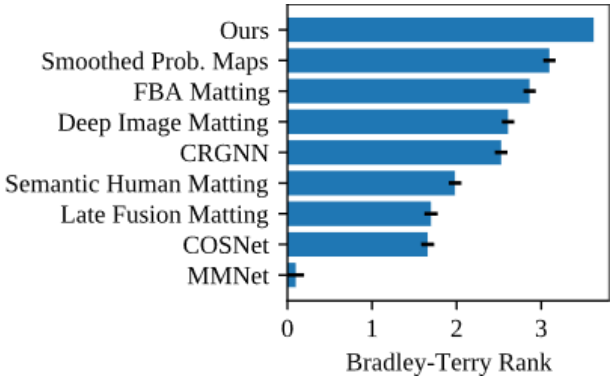


*Рис. 6. Результаты субъективного исследования по всем тестовым видеопоследовательностям. Чёрные метки погрешности показывают 95%-е доверительные интервалы.*
*Fig. 7. Subjective evaluation results on all test sequences. Error bars show 95% confidence intervals.*

Сглаженные карты вероятности, интерпретированные как карты прозрачности, показали неожиданно хороший результат. Основной недостаток данного подхода – плохая обработка границ и «след», остающийся за движущимися объектами.

Предложенный в данной работе нейросетевой подход показал наилучший результат, и в большинстве сравнений был предпочтён другим алгоритмам.

## 4.2 Объективная оценка

Для объективной оценки алгоритма использовались метрики эталонного теста алгоритмов матирования видеопоследовательностей https://videomatting.com [41]. Они определяются следующим образом:

$$\text{SSDA} = \frac{\sum_{i=1}^{N}\sqrt{\sum_{p}^{P}\left(\alpha'_{i,p} - \alpha_{i,p}\right)}}{N \cdot P},$$

$$\text{dtSSD} = \frac{\sum_{i=2}^{N}\sqrt{\sum_{p}^{P}\left(\left(\alpha'_{i,p} - \alpha'_{i-1,p}\right) - \left(\alpha_{i,p} - \alpha_{i-1,p}\right)\right)^2}}{(N-1) \cdot P},$$

$$\text{MESSDdt} = \frac{\sum_{i=2}^{N}\sum_{p}^{P}\left|\left(\alpha'_{i-1,p} - \alpha_{i-1,p}\right)^2 - \left(\alpha'^{\rightarrow i-1}_{i,p} - \alpha^{\rightarrow i-1}_{i,p}\right)^2\right|}{(N-1) \cdot P},$$

где:

- $N$ – число кадров в видео,
- $P = W \times H$ – число пикселей в одном кадре видео,
- $\alpha_i$ – эталонная карта прозрачности $i$-го кадра,
- $\alpha'_i$ – карта прозрачности $i$-го кадра на выходе оцениваемого алгоритма матирования,
- $\alpha_i^{\to i-1}$ – карта прозрачности $i$-го кадра, сдвинутая к $i-1$-му кадру в соответствии с оптическим потоком.

Все три метрики принимают значения, большие либо равные нулю, и равны нулю в случае полного совпадения предсказанной и эталонной карты прозрачности. Чем ниже значение каждой метрики, тем ближе предсказанная карта прозрачности к эталонной. Выбор данных метрик для оценки мотивирован их использованием в эталонном сравнении и нескольких предыдущих работах, предлагающих методы матирования видео [40, 14].

Оценка проводилась на тех же двух видеопоследовательностях из эталонного теста https://videomatting.com, что использовались в субъективном сравнении: *city* и *snow*. Результаты представлены в табл. 1. Предложенный нейросетевой метод даёт либо наилучший результат, либо входит в четвёрку лучших методов вместе с [12, 2, 40]. Различие с субъективной оценкой скорее всего объясняется недостаточной чувствительностью существующих объективных методов оценки к нестабильности результатов во времени.

Объективная оценка также была проведена на первых 50 кадрах видео тестовой части набора VideoMatte240K [22] (всего пять видео). Так как видео из этого набора не содержат «родного» фона, для оценки они были наложены на пять случайных фоновых видео без людей. Результаты представлены в табл. 2. Предложенный нейросетевой метод показывает второй или третий лучший результат после [12] и [2].

*Табл. 1. Результаты объективной оценки на двух видеопоследовательностях https://videomatting.com. Лучший результат выделен жирным, второй подчёркнут, третий показан курсивом.*
*Table 1. Objective evaluation results on http://videomatting.com clips.*
*The best result is shown in bold, the second-best is underlined and the third-best is shown in italics.*

| Метод | SSDA | dtSSD | MESSDdt | SSDA | dtSSD | MESSDdt |
|---|---|---|---|---|---|---|
| | *city* | | | *snow* | | |
| Предложенный нейросетевой метод | *69,651* | **15,314** | **0,695** | 56,338 | 30,240 | *0,662* |
| Сглаженные карты вероятности | 91,577 | 17,609 | 1,291 | 65,772 | 35,133 | 1,264 |
| FBA Matting | 57,700 | *30,825* | *1,613* | 27,113 | 20,881 | 0,423 |
| Deep Image Matting | 97,506 | 47,107 | 3,258 | 59,648 | 41,463 | 2,128 |
| CRGNN | 76,456 | 35,591 | 2,354 | *34,735* | *27,183* | 1,244 |
| Semantic Human Matting | 108,393 | 53,086 | 5,696 | 71,844 | 43,689 | 2,595 |
| Late Fusion Matting | **44,766** | 31,621 | 4,152 | **24,602** | **19,484** | **0,341** |
| COSNet | 271,878 | 62,798 | 22,387 | 156,617 | 58,536 | 9,424 |
| MMNet | 154,656 | 62,439 | 13,580 | 347,065 | 143,696 | 58,429 |

## *5. Заключение*

В данной работе предложен нейросетевой метод семантического матирования видео с людьми на переднем плане, не требующий дополнительных входных данных. Для генерации обучающих видеопоследовательностей для нейросети, используя портретные фотографии людей с эталонными картами прозрачности, предложена процедура создания искусственного движения. Также предложен алгоритм сглаживания карт вероятности, получаемых на выходе из алгоритмов сегментации изображений.

*Табл. 2. Результаты объективной оценки на пяти тестовых видеопоследовательностях VideoMatte240K. Лучший результат выделен жирным, второй подчёркнут, третий показан курсивом.*
*Table 2. Objective evaluation results on five VideoMatte240K test clips.*
*The best result is shown in bold, the second-best is underlined and the third-best is shown in italics.*

| Метод | SSDA | dtSSD | MESSDdt |
|---|---|---|---|
| Предложенный нейросетевой метод | *84,710* | *46,792* | 2,080 |
| Сглаженные карты вероятности | 117,253 | 53,836 | 4,452 |
| FBA Matting | 62,679 | 40,996 | *2,901* |
| Deep Image Matting | 165,094 | 100,595 | 15,039 |
| CRGNN | 140,095 | 84,434 | 13,900 |
| Semantic Human Matting | 166,325 | 113,648 | 22,300 |
| Late Fusion Matting | **29,146** | **25,459** | **0,845** |
| COSNet | 226,256 | 74,765 | 17,506 |
| MMNet | 445,550 | 156,778 | 75,171 |

Для оценки качества разработанного метода проведено субъективное сравнение с 7 существующими методами матирования и сегментации на 59 видеопоследовательностях, в рамках которого собрано 32 272 попарных оценок. Субъективное сравнение показало превосходство разработанного метода над аналогами. Дополнительно, было проведено объективное сравнение, в котором предложенный подход также оказался в числе лучших.

Также было проведено дополнительное субъективное сравнение 9 вариантов предложенного метода для анализа его избыточности, в рамках которого собрано 22 360 попарных оценок. Это сравнение показало положительный вклад компонентов предложенного метода, в том числе сильное положительное влияние на результат предложенных процедуры генерации искусственного движения и алгоритма сглаживания карт вероятности.

## Список литературы / References


[1]. Xu N., Price B., Cohen S., Huang T. Deep Image Matting // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
[2]. Forte M., Pitié F. *F*, *B*, Alpha Matting // arXiv preprint arXiv:2003.07711. 2020.
[3]. Goel A., Kumar M., Sudheendra P., Team V. I. IamAlpha: Instant and Adaptive Mobile Network for Alpha Matting. 2021.
[4]. Shi X., Chen Z., Wang H., Yeung D.-Y., Wong W.-k., Woo W.-c. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting // Advances in Neural Information Processing Systems 28. Curran Associates, Inc., 2015. сс. 802–810.
[5]. AISegment dataset [электронный ресурс]. 2019. URL: https://github.com/aisegmentcn/matting_human_datasets/tree/1829b5f722024d29b780993f06b45ea3f47ba777 (дата обращения: 12.04.2020).
[6]. Shen X., Tao X., Gao H., Zhou C., Jia J. Deep automatic portrait matting // European Conference on Computer Vision. 2016. сс. 92–107.
[7]. Wang J., Cohen M. F. Image and video matting: a survey. Now Publishers Inc, 2008.
[8]. Li X., Li J., Lu H. A survey on natural image matting with closed-form solutions // IEEE Access. IEEE, 2019. т. 7. сс. 136658–136675.
[9]. Chen Q., Ge T., Xu Y., Zhang Z., Yang X., Gai K. Semantic human matting // Proceedings of the 26th ACM international conference on Multimedia. 2018. сс. 618–626.
[10]. Lu H., Dai Y., Shen C., Xu S. Indices Matter: Learning to Index for Deep Image Matting // The IEEE International Conference on Computer Vision (ICCV). 2019.
[11]. Seo S., Choi S., Kersner M., Shin B., Yoon H., Byun H., Ha S. Towards Real-Time Automatic Portrait Matting on Mobile Devices // arXiv preprint arXiv:1904.03816. 2019.
[12]. Zhang Y., Gong L., Fan L., Ren P., Huang Q., Bao H., Xu W. A Late Fusion CNN for Digital Matting // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[13]. Liu J., Yao Y., Hou W., Cui M., Xie X., Zhang C., Hua X.-S. Boosting Semantic Human Matting With Coarse Annotations // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
[14]. Backes M. H., Oliveira M. M. A PatchMatch-based Approach for Matte Propagation in Videos // Computer Graphics Forum. 2019. т. 38, № 7. сс. 651–662.
[15]. Barnes C., Shechtman E., Goldman D. B., Finkelstein A. The Generalized PatchMatch Correspondence Algorithm // Computer Vision – ECCV 2010. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. сс. 29–43.
[16]. Zou D., Chen X., Cao G., Wang X. Unsupervised video matting via sparse and low-rank representation // IEEE transactions on pattern analysis and machine intelligence. IEEE, 2019.
[17]. Shen X., Wang R., Zhao H., Jia J. Automatic real-time background cut for portrait videos // arXiv preprint arXiv:1704.08812. 2017.
[18]. Sengupta S., Jayaram V., Curless B., Seitz S. M., Kemelmacher-Shlizerman I. Background Matting: The World Is Your Green Screen // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
[19]. Oh S. W., Lee J.-Y., Sunkavalli K., Kim S. J. Fast Video Object Segmentation by Reference-Guided Mask Propagation // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
[20]. Oh S. W., Lee J.-Y., Xu N., Kim S. J. Video Object Segmentation Using Space-Time Memory Networks // 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019. т. 0, № . сс. 9225–9234. DOI: 10.1109/ICCV.2019.00932.
[21]. VideoMatte240K and PhotoMatte85 Datasets [электронный ресурс]. 2022. URL: https://grail.cs.washington.edu/projects/background-matting-v2/#/datasets.
[22]. Lin S., Ryabtsev A., Sengupta S., Curless B. L., Seitz S. M., Kemelmacher-Shlizerman I. Real-Time High-Resolution Background Matting // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. сс. 8762–8771.
[23]. Cai S., Zhang X., Fan H., Huang H., Liu J., Liu J., Liu J., Wang J., Sun J. Disentangled Image Matting // The IEEE International Conference on Computer Vision (ICCV). 2019.
[24]. Lutz S., Amplianitis K., Smolic A. Alphagan: Generative adversarial networks for natural image matting // arXiv preprint arXiv:1807.10088. 2018.
[25]. Tang J., Aksoy Y., Oztireli C., Gross M., Aydin T. O. Learning-based sampling for natural image matting // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. сс. 3055– 3063.
[26]. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. 1997. т. 9, № 8. сс. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
[27]. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
[28]. TorchVision [электронный ресурс]. 2019. URL: https://github.com/pytorch/vision/tree/a263704079d9d35db8b0966a65ec628d28998bce/torchvision.
[29]. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in neural information processing systems. 2017. сс. 5998–6008.
[30]. Chen L.-C., Zhu Y., Papandreou G., Schroff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation // Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[31]. Chollet F. Xception: Deep Learning With Depthwise Separable Convolutions // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
[32]. Everingham M., Eslami S. A., Van Gool L., Williams C. K., Winn J., Zisserman A. The pascal visual object classes challenge: A retrospective // International journal of computer vision. Springer, 2015. т. 111, № 1. сс. 98–136.
[33]. Perazzi F., Pont-Tuset J., McWilliams B., Van Gool L., Gross M., Sorkine-Hornung A. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation // Computer Vision and Pattern Recognition. 2016.
[34]. Сервис Яндекс.Толока [электронный ресурс]. 2020. URL: https://toloka.yandex.ru/.
[35]. Rother C., Kolmogorov V., Blake A. "GrabCut" interactive foreground extraction using iterated graph cuts // ACM transactions on graphics (TOG). ACM New York, NY, USA, 2004. т. 23, № 3. сс. 309-314.
[36]. Kingma D. P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. 2014.

[37]. Simonyan K., Grishin S., Vatolin D., Popov D. Video super-resolution using motion compensation and classification-aided fusion // Proceedings of the 24th Spring Conference on Computer Graphics. 2008. сс. 143–148. DOI: 10.1145/1921264.1921294.
[38]. Hur J., Roth S. MirrorFlow: Exploiting Symmetries in Joint Optical Flow and Occlusion Estimation // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017.
[39]. Lu X., Wang W., Ma C., Shen J., Shao L., Porikli F. See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
[40]. Wang T., Liu S., Tian Y., Li K., Yang M.-H. Video Matting via Consistency-Regularized Graph Neural Networks // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. сс. 4902–4911.
[41]. Erofeev M., Gitman Y., Vatolin D., Fedorov A., Wang J. Perceptually Motivated Benchmark for Video Matting // Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, 2015. с. 99. DOI: 10.5244/C.29.99.
[42]. Videos Used for Subjective Evaluation [электронный ресурс]. 2022. URL: https://drive.google.com/drive/folders/1RL9zbn7MPhPyFT4ahW00fjmKlNKdnyy7?usp=sharing.
[43]. Bradley R. A., Terry M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons // Biometrika. JSTOR, 1952. т. 39, № 3/4. сс. 324–345.

## *Информация об авторах / Information about authors*

Иван Андреевич МОЛОДЕЦКИХ получил степень магистра по прикладной математике и информатике в Московском государственном университете имени М. В. Ломоносова в 2020 году. В настоящее время является аспирантом в лаборатории компьютерной графики и мультимедиа. В область его научных интересов входят суперразрешение, семантическое матирование видео и машинное обучение. Иван курировал разработку бенчмарка методов суперразрешения для улучшения качества видео и был одним из организаторов соревнования по оценке качества суперразрешения видео на ECCV-AIM 2024.

Ivan Andreevich MOLODETSKIKH received his master degree in computer science from the Moscow State University in 2020. He is currently a postgraduate student at the MSU Graphics & Media Lab. His research interests include super-resolution, semantic video matting and machine learning. Ivan supervised the development of MSU video upscalers and SR+codec benchmarks and was one of the organizers of the ECCV-AIM 2024 video super-resolution quality assessment challenge.

Михаил Викторович ЕРОФЕЕВ получил степень специалиста по прикладной математике и информатике в Московском государственном университете имени М. В. Ломоносова в 2013 году. В 2017 защитил кандидатскую диссертацию по теме «Преобразование видеопоследовательностей, содержащих объекты с полупрозрачными границами, в стереоскопический формат». В настоящее время является научным сотрудником лаборатории компьютерной графики и мультимедиа факультета ВМК МГУ. В область его научных интересов входят матирование изображений и видео и машинное обучение. Михаил является активным участником проекта по объективному сравнению алгоритмов.

Mikhail Viktorovich EROFEEV – Cand. Sci. (Phys.-Math.). He graduated from the Computer Science department of Lomonosov Moscow State University, and currently he is a researcher at the MSU Graphics & Media Lab. His research interests include video and image matting, and machine learning. Mikhail is one of the major contributors to the video matting methods benchmark.

Андрей Викторович МОСКАЛЕНКО получил степень магистра по прикладной математике и информатике в Московском государственном университете имени М. В. Ломоносова в 2023 году. В настоящее время является студентом аспирантуры в лаборатории компьютерной графики и мультимедиа. В область его научных интересов входят моделирование визуального внимания, интерактивная сегментация, заполнение областей изображений и

оценка качества видео. Андрей является одним из ключевых разработчиков платформы анализа субъективного качества медиаконтента.

Andrey Viktorovich MOSKALENKO received his master degree in computer science from the Moscow State University in 2023. He is currently a postgraduate student at the MSU Graphics & Media Lab. His research interests involve visual attention modeling, interactive segmentation, image inpainting and video quality assessment. Andrey is one of the core contributors to the subjective media quality assessment platform.

Дмитрий Сергеевич ВАТОЛИН закончил ВМК МГУ в 1996 году, кандидат физико-математических наук, заведующий лабораторией компьютерной графики ВМК МГУ, исследователь Центра доверенного искусственного интеллекта ИСП РАН. Читает курсы по компьютерной графике и методам сжатия и обработки видео с 1997 года. Создатель популярных сайтов, посвященных алгоритмам обработки и сжатия видео. Специализируется на исследованиях в области алгоритмов сжатия видео, современных методах измерения качества и обработке цифрового видео. Руководил проектами с компаниями Intel, Cisco, Real Networks, Samsung, Huawei, Broadcom и некоторыми другими. С 2008 года занимается измерением и исправлением артефактов стерео, в том числе организовал проект измерения качества стереофильмов.

Dmitriy Sergeevich VATOLIN – Cand. Sci. (Phys.-Math.) from Moscow State University, head of the MSU Graphics & Media Lab and MSU AI Institute Video Analysis Lab, and a researcher at the Research Centre for Trusted AI of ISP RAS. His research interests include compression methods, video processing, 3D video techniques (depth from motion, focus and other cues, video matting, background restoration, high-quality stereo generation), as well as video quality assessment and robustness of modern video quality metrics. He is a key cofounder of the 3D video quality measurement project, his most known project is the annual MSU Video Codecs Comparison that includes up to 25 modern codecs compared subjectively and objectively in several nominations with detailed 20000+ charts.