ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 66 за 2023 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

А.В. Соловьев, А.В. Анциферова, Д.С. Ватолин, В.А. Галактионов

Разработка метода обработки видео, повышающего оценку метрики качества VMAF на основе её дистилляции

Рекомендуемая форма библиографической ссылки: Разработка метода обработки видео, повышающего оценку метрики качества VMAF на основе её дистилляции / А.В. Соловьев [и др.] // Препринты ИПМ им. М.В.Келдыша. 2023. № 66. 11 с. https://doi.org/10.20948/prepr-2023-66
https://library.keldysh.ru/preprint.asp?id=2023-66

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.В.Соловьев, А.В.Анциферова, Д.С.Ватолин, В.А.Галактионов

Разработка метода обработки видео, повышающего оценку метрики качества VMAF на основе ее дистилляции

Москва — 2023

Соловьев А.В., Анциферова А.В., Ватолин Д.С., Галактионов В.А.

Разработка метода обработки видео, повышающего оценку метрики качества VMAF на основе её дистилляции

В данной работе рассматривается задача создания метода предобработки видео, повышающего оценку его качества методом Video Multimethod Assessment Fusion (VMAF). В работе описан нейросетевой метод для автоматической предобработки входного видео, работающий в режиме реального времени. Предобработка осуществляется глубокой нейронной сетью с архитектурой на базе U-Net. В ходе обучения сети используется обучаемая аппроксимация VMAF. В работе описаны способы улучшения качества работы итогового метода, а именно использование SSIM в функции потерь и фильтрация обучающей выборки. Итоговая версия метода повышает VMAF исходного видео в среднем на 18% после предобработки. Разработанный метод позволяет оценить надежность метода оценки качества видео VMAF и демонстрирует уязвимости метода, которые могут использоваться разработчиками алгоритмов обработки видео для повышения рейтингов их методов в ходе автоматического сравнения результатов работы по оценке качества VMAF.

Ключевые слова: оценка качества видео, анализ устойчивости, дистилляция нейронной сети, состязательные атаки

Aleksei Valerievich Solovev, Anastasia Vsevolodovna Antsiferova, Dmitriy Sergeevich Vatolin, Galaktionov Vladimir Aleksandrovich

Development of neural network-based video preprocessing method to increase the VMAF score relative to source video using distillation

In this work, we consider the problem of creating a video preprocessing method that improves video’s quality score measured by the Video Multimethod Assessment Fusion (VMAF) metric. The paper describes a neural network method for automatic preprocessing of input video, operating in real time. Preprocessing is carried out by a deep neural network based on U-Net architecture. In the course of network training, a trained VMAF approximation is used. The paper describes ways of improving the quality of the final method, namely, adding neural network compression, using SSIM in the loss function, and filtering the training set. The final version of the method increases the VMAF score of the original video by an average of 18% after preprocessing. The developed method demonstrates the flaws of the VMAF quality assessment method that can be used by developers of video processing algorithms to improve the ratings of their methods during automatic comparison carried out using VMAF quality assessment method.

Key words: video quality, robustness analysis, neural network distillation, adversarial attacks

Введение

Разработка методов обработки и сжатия видео требует проведения контроля визуального качества. Проведение субъективной оценки качества видео рядом экспертов является долгим и дорогостоящим процессом, в связи с чем получили широкое распространение разнообразные алгоритмы объективной оценки качества видео. Однако современные методы оценки качества видео основаны на машинном обучении и нейронных сетях, что делает их уязвимыми к состязательным атакам путем различной обработки видео. Состязательные атаки, направленные на повышение оценок, выдаваемых объективным методом измерения качества видео, выгодны производителям видеокодеков и других алгоритмов, так как их разработка быстрее и дешевле, чем разработка нового алгоритма сжатия. Атаки помогают получить более высокие позиции в рейтингах открытых сравнений алгоритмов обработки и сжатия видео, что позволяет привлечь большее число инвесторов и потребителей. Одним из наиболее популярных сегодня методов оценки качества видео является Video Multimethod Assessment Fusion (VMAF) [1], разработанный компанией NETFLIX. Для анализа устойчивости VMAF к состязательным атакам в данной работе предлагается новый метод предобработки входного видео, повышающий оценки VMAF. Задача осложняется тем, что оценка VMAF не является дифференцируемой по входным видео.

Обзор существующих методов

Многие объективные оценки качества являются дифференцируемыми по входным параметрам. Из примеров можно привести LPIPS [2], SSIM [3]. Состязательная атака путем использования градиента оценки качества по исходному изображению широко распространена. Изначально она была предложена в работе [4], и существуют стандартные библиотеки, предлагающие широкое разнообразие таких преобразований (например, [5]).

Однако в случае метода VMAF у нас нет доступа к градиентам рассматриваемой оценки качества по входным параметрам, поэтому применение методов на основе вычисления градиентов рассматриваемой функции по входным данным невозможно. Альтернативным подходом может быть использование аппроксимации градиента. Метод [6] решает задачу оптимизации кодирования видео с учетом оценки VMAF. Авторы одновременно обучали нейросетевое приближение значений оценки VMAF и модель кодировщика, добиваясь кодирования с учетом повышения значений VMAF путем включения приближения VMAF в функцию потерь кодировщика. В данной работе авторы предлагали метод кодирования изображений, а не метод предобработки видео для повышения значений VMAF. Также существуют две известные состязательные атаки на VMAF, основанные на

методе черного ящика. Данные методы не аппроксимируют и не используют градиенты оценки VMAF по исходному видео. Метод [7] использует набор стандартных преобразований для изображений (преобразование повышения резкости, выравнивания гистограмм) для получения преобразований, практически не отражающихся на визуальном качестве и оценке качества SSIM, но повышающих значение оценки VMAF. Заявленное авторами повышение VMAF составляет 5-7% в среднем. Метод [8] развивает идею из предыдущей статьи, применяя генетические алгоритмы для автоматического итеративного подбора параметров стандартных преобразований, комбинация которых должна повысить значение оценки VMAF исходного видео. Авторы заявляют о получении увеличения в 60% в среднем (против 5-7% в прошлой работе), однако оптимизация осуществляется отдельно для каждого видео, и время работы алгоритма не позволяет использовать его для задач увеличения VMAF в ходе кодирования видео в реальном времени.

Предложенный метод

Так как в жизненных сценариях атака должна работать достаточно быстро для возможности покадрового использования в видео, подходы на основе генетических алгоритмов для нее слишком медленные. Предлагаемый метод основан на обучении промежуточной аппроксимации рассматриваемой оценки качества VMAF, при котором атакующее преобразование видео осуществляется другой нейронной сетью в один проход. За основу была взята процедура обучения, предложенная в статье [6]. В исходной статье авторы решают задачу обучения нейросетевого кодировщика для сжатия видео с потерями. Авторы ставят задачу максимизации оценки качества по методу VMAF при сохранении качества кодирования и с помощью обучения нейронной сети осуществляют дистилляцию метода VMAF. Такая нейронная сеть позволяет осуществлять дифференцируемую аппроксимацию оценки качества VMAF и может напрямую быть использована в функции потерь в ходе обучения основной нейронной сети. Архитектура нейронной сети, осуществляющей предобработку, была заменена на архитектуру U-Net [9], не содержащую в себе частей, отвечающих за энтропийное кодирование. Последний слой и его активация были заменены на один сверточный слой с тремя выходными каналами. Итоговая архитектура предобрабатывающей нейронной сети показана на рис. 1. Архитектура нейронной сети, осуществляющей аппроксимацию VMAF, была взята из статьи [6] без изменений.

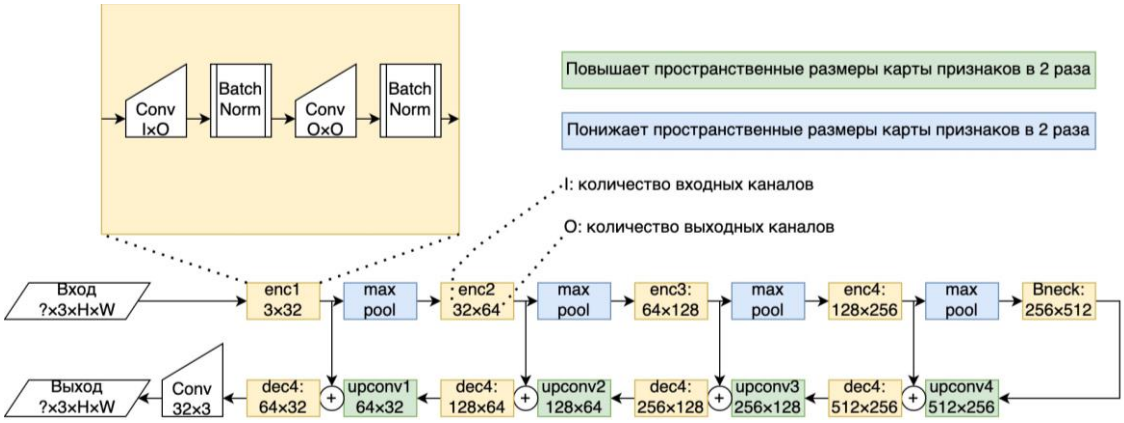

Рис. 1. Использованная архитектура на базе U-Net

В ходе обучения оптимизировалась функция потерь, состоящая из трех компонент: компонента для вспомогательной нейронной сети (1), компонента для основной нейронной сети (2) и стабилизирующая компонента (3).

$$L_{proxy} = mean\left(\left(P(B_{out}, B\ \ ) - VMAF(B_{out}, B\ \ )\right)^2\right), B_{out} = G_{frozen}(B\ \ ). \quad (1)$$

$$L_{vmaf} = M - P_{frozen}(G(B\ \ ), B\ \ ) \quad (2)$$

$$L_{stable} = \|G(B\ \ ) - B\ \ \|_2 \quad (3)$$

Итоговая функция потерь представляет собой взвешенную сумму этих компонент. В ходе обучения нейронной сети эти коэффициенты могут меняться; так, на ранних этапах увеличение стабилизационного коэффициента может увеличить скорость обучения, а на поздних этапах его уменьшение может увеличить итоговое качество работы обученной нейронной сети. В отличие от статьи [6], в функции потерь нет части, описывающей качество сжатия основной нейронной сетью, поскольку основная нейронная сеть не решает задачу кодирования.

Модификация обучающей выборки

В ходе анализа результатов работы предложенного метода на обучающей выборке было обнаружено, что выигрыш заметно отличается от изображения к изображению. Некоторый процент примеров оказался сложным для модели, и предобработка не увеличивает, а уменьшает значение оценки качества VMAF, как можно видеть на рис. 2. Фильтрация некоторой части наиболее и наименее удачных примеров в ходе обучения повышает качество работы нейронной сети на валидационных наборах.

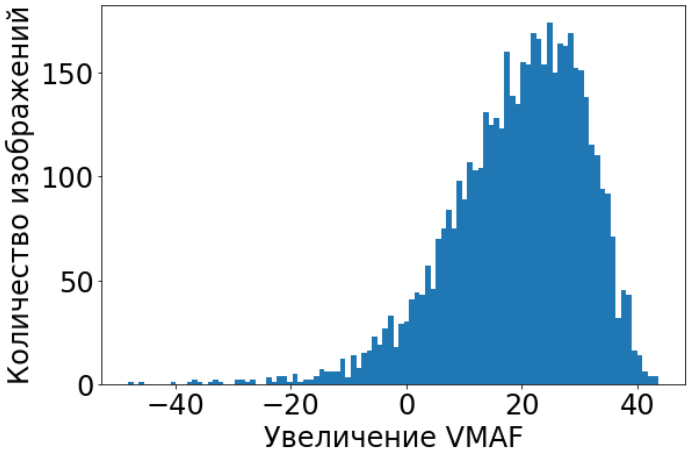

Рис. 2. Повышение значений VMAF на тестовой выборке после применения предложенного метода

Экспериментальная оценка

Полученные в ходе проведения экспериментов версии метода были оценены несколькими способами. Субъективная оценка включала в себя поиск и описание вносимых методами искажений на нескольких наборах данных разной природы; влияние наиболее заметных из них на объективное качество изображения или видео было продемонстрировано с помощью других, наиболее известных оценок качества. Объективная оценка включала в себя численную оценку работы метода в сценарии с последующим сжатием предобработанного видео и без сжатия, усредненную по большому набору данных.

Наборы данных

В данной работе рассмотрена процедура обучения на двух различных наборах данных. Первый из них создан случайным сэмплированием и масштабированием изображений из обширного разнообразного набора данных для классификации. В качестве такого набора был выбран Pascal VOC [10], размер получившегося набора данных составил 5000 изображений в обучающей и 1000 изображений в валидационной части. Исходные изображения были взяты из обучающей и валидационной частей набора данных Pascal VOC; соответственно, эти множества изображений не пересекаются, и не происходит утечки информации.

В качестве второго набора данных был рассмотрен набор Vimeo 90K [11], а именно, поднабор vimeo_septuplets, содержащий 91701 видеопоследовательностей длиной 7 кадров. Из него случайным образом был выбран поднабор, состоящий из 5000 видео в обучающей и 1000 видео в валидационной выборке.

Субъективная оценка

В ходе работы субъективный анализ будет представлен для следующих моделей:

1. Нейронная сеть, обученная на наборе данных на основе Pascal VOC.
2. Нейронная сеть, обученная на наборе данных на основе Pascal VOC без SSIM-регуляризации, дообученная с использованием SSIM-регуляризации.
3. Нейронная сеть, обученная на наборе данных на основе Pascal VOC с использованием SSIM-регуляризации.
4. Нейронная сеть, обученная на наборе данных на основе Vimeo 90K.

Модели, обученные на наборе данных Pascal VOC, оставляли на изображениях или видео искажения нескольких типов. Первым и наиболее заметным из них являлись яркие цветные пятна в темных областях изображений (рис. 3). Этот эффект связан с тем, что оригинальный метод VMAF использует для оценки исключительно Y-компоненту входного видео, отбрасывая цветовые компоненты целиком.

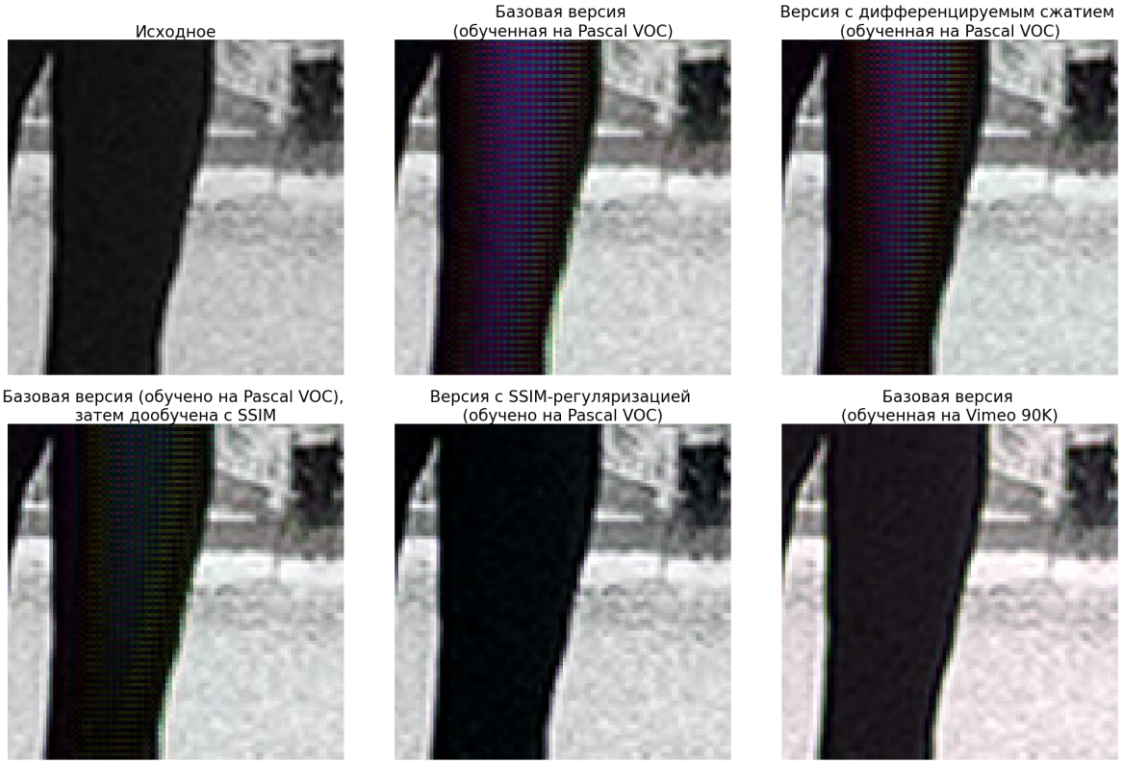

Рис. 3. Визуальные артефакты: фиолетовые области

Другими типами вносимых искажений являлись цветные артефакты, возникающие на некоторых границах изображений. Это ложные цветовые границы, иногда создаваемые методом внутри цветового градиента, а также цветовая коррекция, не приводящая к появлению ложных цветовых границ на изображениях. Ее можно заметить у сети, обученной на Vimeo 90K, на рис. 3.

Было проведено более подробное исследование для модели, обученной на Pascal VOC без применения SSIM-регуляризации в функции потерь. В ходе субъективной оценки порядка 150 изображений были вручную размечены на наиболее заметный тип искажения, после чего визуализировано соотношение между наиболее заметным типом искажения, оценкой VMAF и другими объективными оценками качества. В качестве них использовались PSNR и SSIM [3]. Результаты представлены на графике на рис. 4. На нем видно, что изображения с внесенными искажениями типа "яркие цветовые пятна" выделяются (по SSIM) в худшую сторону. Так как данное искажение является наиболее заметным и метод SSIM является дифференцируемым по входным изображениям, было предложено использование SSIM в качестве стабилизирующей компоненты в функции потерь в ходе обучения модели. Обученная с нуля модель с использованием SSIM в качестве регуляризации не производит цветовых пятен и ложных цветовых границ, однако проводит заметную глазу цветовую коррекцию и немного повышает резкость видео.

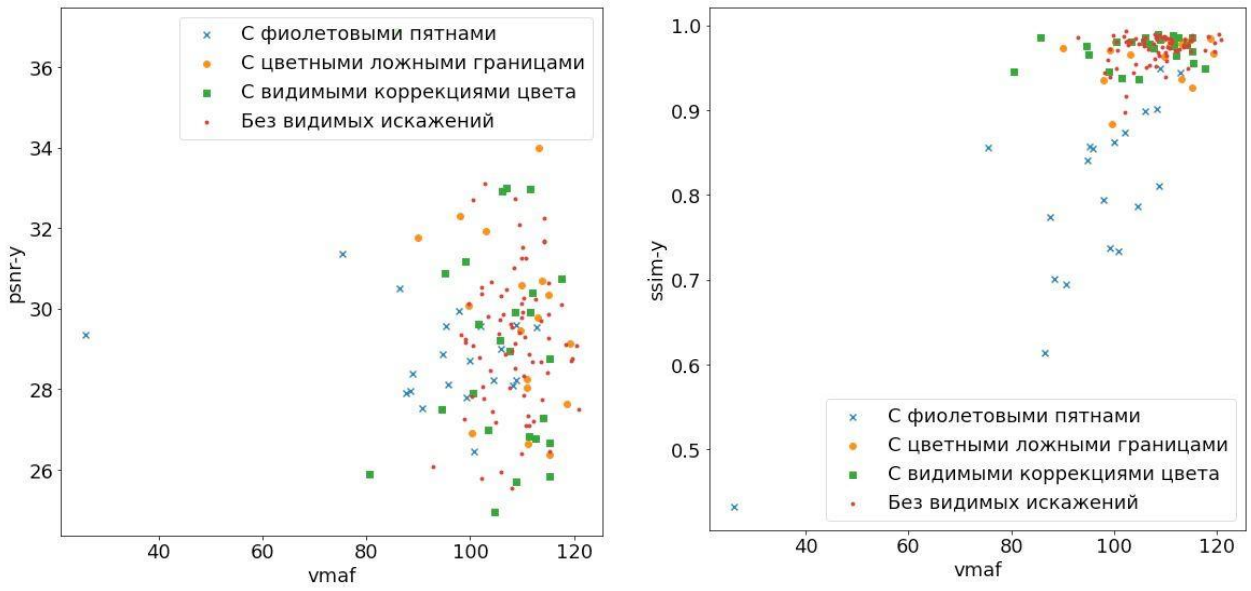

Рис. 4. Влияние рассматриваемых типов визуальных артефактов на PSNR и SSIM

Объективная оценка

Анализ эффективности предложенных вариаций метода был проведен на наборах данных Pascal VOC, Vimeo 90K и xiph [12]. Для первых двух наборов данных измеряемой величиной является среднее увеличение оценки VMAF, достигаемое после предобработки всего рассматриваемого набора данных конкретной моделью. Для набора данных xiph, который содержит видео длиной по крайней мере 60 кадров, оценивание качества работы было проведено с использованием BSQ-Rate [13]. Набор целевых битрейтов, с которыми производилось кодирование каждого видео, был зафиксирован: 60 кбит/с, 250 кбит/с, 1000 кбит/с, 2000 кбит/с, 4000 кбит/с, 8000 кбит/с. В качестве кодека взят x264 со стандартным набором параметров “medium”. Скорость работы метода составила 120 кадров в секунду при использовании графического ускорителя Nvidia Titan RTX (и центрального процессора Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz).

Результаты

Результаты объективной оценки отражены в табл. 1. Фильтрация обучающей выборки оказала наибольшее влияние на эффективность предложенного метода. Добавление SSIM-регуляризации позволило избавиться от цветовых пятен, сохранив качество работы нейронной сети. Наилучшая из обученных модель получена с использованием SSIM-регуляризации и фильтрацией 50% наиболее трудной части тренировочной выборки, она показывает BSQ-Rate, равный 0.541, и повышает VMAF в сценарии без дальнейшего сжатия на 19.07 и 17.03 единицы в среднем на двух наборах данных. Разработанный нейросетевой метод повышения оценки VMAF показал лучшую по сравнению с рассмотренными методами повышения оценки VMAF скорость работы. Она является достаточной для работы в режиме реального времени.

Таблица 1

Сравнение предложенных методов

Обучающая выборка	Фильтр обучения	Наличие SSIM	BSQ-rate	Оценка на валидации Pascal VOC*	Оценка на валидации Vimeo 90K*
Pascal VOC	100%	-/-	0.803	10.34	8.95
Vimeo 90K	100%	-/-	0.705	4.42	7.64
Vimeo 90K	0..70%	-/-	1.109	1.73	10.50
Pascal VOC	0..70%	-/-	0.832	15.69	15.31

Обучающая выборка	Фильтр обучения	Наличие SSIM	BSQ-rate	Оценка на валидации Pascal VOC*	Оценка на валидации Vimeo 90K*
Pascal VOC	0..70%	дообу-чение	0.698	12.00	12.00
Pascal VOC	0..40%	да	0.540	18.28	17.06
Pascal VOC	0..50%	да	0.541	19.07	17.13
Pascal VOC	0..60%	да	0.548	17.95	15.35
Pascal VOC	0..70%	да	0.565	20.88	19.95

Список литературы

[1] Li Z. et al. Toward a practical perceptual video quality metric // The Netflix Tech Blog. – 2016. – Т. 6. – №. 2. – С. 2.

[2] Zhang R. et al. The unreasonable effectiveness of deep features as a perceptual metric // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 586-595.

[3] Wang Z. et al. Image quality assessment: from error visibility to structural similarity // IEEE transactions on image processing. – 2004. – Т. 13. – №. 4. – С. 600-612.

[4] Wang Z., Simoncelli E. P. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities // Journal of Vision. – 2008. – Т. 8. – №. 12. – С. 8-8.

[5] Rauber J. et al. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax // Journal of Open Source Software. – 2020. – Т. 5. – №. 53. – С. 2607.

[6] Chen L. H. et al. ProxIQA: A proxy approach to perceptual optimization of learned image compression // IEEE Transactions on Image Processing. – 2020. – Т. 30. – С. 360-373.

[7] Zvezdakova A. et al. Hacking vmaf with video color and contrast distortion // arXiv preprint arXiv:1907.04807. – 2019.

[8] Siniukov M. et al. Hacking VMAF and VMAF NEG: vulnerability to different preprocessing methods // Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference. – 2021. – С. 89-96.

[9] Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation // Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. – Springer International Publishing, 2015. – С. 234-241.

[10] Everingham M., Winn J. The PASCAL visual object classes challenge 2012 (VOC2012) development kit // Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep. – 2012. – Т. 2007. – №. 1-45. – С. 5.

[11] Xue T. et al. Video enhancement with task-oriented flow // International Journal of Computer Vision. – 2019. – Т. 127. – С. 1106-1125.

[12] Montgomery C., Lars H. Xiph. org video test media (derf's collection) // Online, https://media. xiph. org/video/derf. – 1994. – Т. 6.

[13] Zvezdakova A. V. et al. BSQ-rate: a new approach for video-codec performance comparison and drawbacks of current solutions // Programming and computer software. – 2020. – Т. 46. – С. 183-194.

Оглавление