НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-5-1065-1072
УДК 004.032.26

Метод повышения разрешения изображений с использованием референсных изображений на основе трансформерной архитектуры

Александр Викторович Яковенко[1]✉, Алексей Константинович Денисов[2], Дмитрий Сергеевич Ватолин[3], Павел Валерьевич Кустарев[4]

[1] Московский государственный университет им. М.В. Ломоносова, Москва, 119991, Российская Федерация
[2,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[3] Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем МГУ имени М.В. Ломоносова, Москва, 119192, Российская Федерация

[1] alexander.yakovenko@graphics.cs.msu.ru✉, https://orcid.org/0000-0003-3105-512X
[2] denisov@itmo.ru, https://orcid.org/0000-0001-8135-1723
[3] dmitriy@graphics.cs.msu.ru, https://orcid.org/0000-0002-8893-9340
[4] kustarev@itmo.ru, https://orcid.org/0000-0001-9326-0837

Аннотация

Введение. Рассмотрена задача повышения разрешения изображений с использованием референсных изображений, в которой к входному кадру низкого разрешения дополнительно подается референс высокого разрешения со схожим содержанием, что позволяет переносить реальные детали из референса вместо их генерирования. Применяемые в настоящее время методы декомпозируют задачу на отдельные этапы поиска соответствий, выравнивания и переноса деталей, представленные модулями, обучаемыми последовательно, что приводит к накоплению ошибок между этапами и ограничивает возможности модели по выучиванию необходимых преобразований из данных. В работе предлагается единая обучаемая модель, в которой поиск соответствий и перенос текстур выполняются совместно в рамках применения механизма внимания. **Метод.** Предложена единая архитектура, в которой токены входного и референсного изображений конкатенируются и обрабатываются трансформером, неявно реализующим как поиск соответствий между изображениями, так и перенос деталей. Для пространственного кодирования предложено бинокулярное двумерное позиционное кодирование (2D Rotary Positional Embedding) с общей системой координат для входного и референсного изображений, а также схема адаптивного масштабирования логитов внимания, зависящего от длины обрабатываемой последовательности и позволяющего применять модель к разрешениям, превышающим использовавшиеся при обучении. **Основные результаты.** Показано, что на эталонных наборах данных CUFED5 и WR-SR при бикубическом четырехкратном понижении разрешения предложенный метод превосходит лучшие из известных подходов с использованием референсных изображений на 0,33 дБ по метрике Peak Signal-to-Noise Ratio (PSNR) на наборе данных CUFED5 и впервые опережает сильные методы повышения разрешения по одному изображению на WR-SR с увеличением на 0,29 дБ PSNR относительно лучшей из сравниваемых моделей. **Обсуждение.** В отличие от многоэтапных решений, предложенный метод представляет собой единую обучаемую модель, не требующую предобученных модулей. Полученные результаты показывают, что операция внимания над объединенной последовательностью токенов является достаточным механизмом для решения задачи повышения разрешения с использованием референсных изображений. Потенциальной областью применения предложенного подхода является восстановление изображений во всех задачах, где доступен референс со схожим содержанием.

Ключевые слова

повышение разрешения изображений, референсные изображения, трансформерная архитектура, механизм внимания, позиционное кодирование, обработка изображений, глубокое обучение, восстановление изображений, компьютерное зрение, нейронные сети

Благодарности

Работа выполнялась с использованием суперкомпьютера «МГУ-270» МГУ имени М.В. Ломоносова.

© Яковенко А.В., Денисов А.К., Ватолин Д.С., Кустарев П.В., 2026

Ссылка для цитирования: Яковенко А.В., Денисов А.К., Ватолин Д.С., Кустарев П.В. Метод повышения разрешения изображений с использованием референсных изображений на основе трансформерной архитектуры // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1065–1072. doi: 10.17586/2226-1494-2026-26-5-1065-1072

Vision transformer for reference-based super-resolution

Aleksandr V. Yakovenko[1]✉, Aleksei K. Denisov[2], Dmitry S. Vatolin[3], Pavel V. Kustarev[4]

[1] Lomonosov Moscow State University, Moscow, 119991, Russian Federation
[2,4] ITMO University, Saint Petersburg, 197101, Russian Federation
[3] MSU Institute for Artificial Intelligence, Moscow, 119192, Russian Federation

[1] alexander.yakovenko@graphics.cs.msu.ru✉, https://orcid.org/0000-0003-3105-512X
[2] denisov@itmo.ru, https://orcid.org/0000-0001-8135-1723
[3] dmitriy@graphics.cs.msu.ru, https://orcid.org/0000-0002-8893-9340
[4] kustarev@itmo.ru, https://orcid.org/0000-0001-9326-0837

Abstract

Reference-based Super-Resolution (RefSR) aims to restore a low-resolution image by transferring real textures from a high-resolution reference of similar content, rather than hallucinating missing details as in single-image super-resolution. Most existing RefSR methods decompose the problem into separate matching, alignment, and fusion stages, implemented as sequentially trained modules. This decomposition causes errors to accumulate across stages and imposes inductive biases that limit what the model can learn from data. Instead, we propose to handle correspondence and texture transfer jointly through a single attention operation. We propose an end-to-end transformer that operates on the concatenated sequence of input and reference tokens, with self-attention implicitly performing both matching and texture transfer. The architecture introduces binocular 2D Rotary Positional Embedding, a shared coordinate frame for input and reference patches, and a resolution-adaptive attention scaling scheme in which the logits are rescaled by the length of the processed sequence, allowing the model to operate at resolutions beyond those seen at training time. On the standard CUFED5 and WR-SR benchmarks under bicubic ×4 downsampling, our method improves over the best previously published RefSR approach by 0.33 dB Peak Signal-to-Noise Ratio (PSNR) on CUFED5 and, for the first time, surpasses strong single-image super-resolution models on WR-SR by 0.29 dB PSNR. Unlike multi-stage pipelines, the proposed model is end-to-end trainable and requires no pretrained components, showing that a single attention operation over the joint token sequence is sufficient for RefSR. The approach is broadly applicable to image restoration settings where a high-resolution reference of similar content is available.

Keywords

image super-resolution, reference images, transformer architecture, attention mechanism, positional encoding, image processing, deep learning, image restoration, computer vision, neural networks

Acknowledgements

The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

For citation: Yakovenko A.V., Denisov A.K., Vatolin D.S., Kustarev P.V. Vision transformer for reference-based super-resolution. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 1065–1072 (in Russian). doi: 10.17586/2226-1494-2026-26-5-1065-1072

Введение

Качество цифровых изображений играет ключевую роль во многих задачах их последующей обработки и восприятия. При этом в реальных условиях получение изображений достаточного разрешения непосредственно с устройства съемки часто оказывается затруднительным: ограничения оптических систем и невысокая разрешающая способность сенсоров приводят к тому, что доступные изображения имеют разрешение, недостаточное для целевого применения. Это делает актуальной задачу алгоритмического восстановления изображений высокого разрешения по входным данным более низкого разрешения, известную как задача повышения разрешения изображений (Super-Resolution, SR), в рамках которой под разрешением изображения понимается его пространственное разрешение — число пикселов вдоль каждой из осей.

Существующие методы SR принято разделять на два класса. Первый — методы, работающие с единственным входным изображением (Single Image Super-Resolution, SISR), которые опираются исключительно на информацию, содержащуюся в самом изображении низкого разрешения. Второй — методы, использующие референсные изображения (Reference-based Super-Resolution, RefSR), в которых наряду с входным изображением низкого разрешения модели подается дополнительное референсное изображение высокого разрешения со схожим содержанием. Фундаментальное ограничение SISR состоит в том, что утраченные при понижении разрешения детали невозможно достоверно восстановить, имея только вход низкого разрешения: модель вынуждена либо сглаживать неоднозначные области, либо генерировать правдоподобные, но не соответствующие действительности детали. Методы RefSR позволяют обойти это ограничение, привлекая в качестве источника фактической информации референсное изображение, из которого реальные текстуры могут быть перенесены на входное изображение вместо их генерирования. Под референсным изображением здесь понимается изображение, содержащее сцену или объект, частично совпадающие с содержанием целевого

изображения, например, снимок того же объекта, полученный из базы изображений, или кадр той же сцены, снятый с большим оптическим увеличением, например, объективом с большим фокусным расстоянием или на телефотокамеру смартфона.

Большинство современных методов RefSR декомпозируют задачу на отдельные этапы: поиск соответствий между входным и референсным изображениями, выравнивание референса относительно входного изображения и перенос текстур, реализованные отдельными модулями, обучаемыми последовательно и нередко опирающимися на предобученные компоненты. Такая организация приводит к накоплению ошибок между этапами и ограничивает возможности модели по выучиванию необходимых преобразований из данных. В данной работе предлагается отказаться от декомпозиции и решать задачу единой обучаемой моделью, в которой поиск соответствий и перенос текстур выполняются совместно в рамках применения механизма внимания. Основные результаты работы состоят в следующем: предложена единая трансформерная архитектура для решения задачи RefSR; введены бинокулярное двумерное позиционное кодирование (2D Rotary Positional Embedding, 2D-RoPE) с общей системой координат для входного и референсного изображений и схема адаптивного масштабирования логитов внимания, зависящего от длины обрабатываемой последовательности; экспериментально показано, что предложенный метод превосходит по качеству известные подходы.

Обзор научных публикаций

Современные методы SISR в основном представлены глубокими нейронными сетями, обучаемыми прямому преобразованию изображения низкого разрешения в изображение высокого разрешения. В работе [1] представлена сеть Super-Resolution Convolutional Neural Network и впервые продемонстрирована применимость сверточных нейронных сетей для решения задачи SR, после чего глубокое обучение стало основным инструментом разработки методов в этой области. Последующие архитектуры, такие как Very Deep Super-Resolution [2] и Enhanced Deep Super-Resolution [3], углубляли сверточные сети и использовали остаточные соединения, что позволило существенно повысить качество восстановления. В последние годы более широкое распространение в задачах SISR получили трансформерные архитектуры и модели на основе пространств состояний (State-Space Models, SSM): методы SwinIR [4] и HAT [5] на основе оконного механизма внимания, а также SSM-модель MambaIRv2 [6] продемонстрировали дальнейшее улучшение качества восстановления по сравнению со сверточными подходами. Тем не менее, методы SISR обладают фундаментальным ограничением: восстановление изображения высокого разрешения по единственному изображению низкого разрешения представляет собой некорректную обратную задачу, что вынуждает модели либо сглаживать неоднозначные области, либо генерировать правдоподобные, но не соответствующие действительности детали.

Методы RefSR направлены на преодоление указанного ограничения за счет привлечения дополнительного источника фактической информации, референсного изображения высокого разрешения со схожим содержанием, из которого реальные текстуры могут быть перенесены на входное изображение. В работе [7] рассмотрен метод Super-Resolution by Neural Texture Transfer и предложен перенос текстур по фрагментам в пространстве признаков предобученной сверточной сети. В методе C²-Matching [8] выделен поиск соответствий в отдельный этап и обучена сеть извлечения признаков входного и референсного изображений с использованием контрастивного обучения, далее получена сеть восстановления изображения на основе найденных соответствий. В методе Reference-based Image Super-Resolution with Deformable Attention Transformer (DATSR) [9] использован механизм деформируемого внимания для динамической агрегации признаков референсного изображения. Метод Task Decoupled Framework for Reference-based Super-Resolution (TDF) [10] разделяет задачу на этап повышения разрешения и этап уточнения деталей, при этом для повышения разрешения используется отдельный предобученный SISR-блок, а перенос деталей из референса выполняется на его выходе. Метод Multi-Reference Super-Resolution (MRefSR) [11] расширяет постановку задачи на случай нескольких референсных изображений и предлагает модуль внимания с пространственно-зависимой фильтрацией для их совместного использования. Несмотря на различия в конкретной реализации, общим для этих подходов является декомпозиция задачи RefSR на отдельные этапы поиска соответствий, выравнивания и переноса деталей, реализованные в виде модулей, обучаемых последовательно и часто опирающихся на предобученные компоненты. Такая организация обработки приводит к накоплению ошибок между этапами и ограничивает класс соответствий между входным и референсным изображениями, который модель в принципе способна выучить из данных.

Параллельно с развитием специализированных методов решения задач восстановления изображений в последние годы трансформерные архитектуры получили широкое распространение в задачах компьютерного зрения. В работе [12] впервые представлена архитектура Vision Transformer (ViT) и предложена идея представления изображения как последовательности блоков (также называемых визуальных токенов) и применения к ним *механизма внимания*. Данная идея получила развитие в моделях, обрабатывающих одновременно несколько изображений путем конкатенации их токенов в единую последовательность и применения механизма внимания к объединенной последовательности, таких как Large View Synthesis Model [13] в задаче синтеза новых ракурсов сцены и Visual Geometry Grounded Transformer [14] в задачах оценки трехмерной геометрии. Эти работы показывают, что механизм внимания над объединенной последовательностью токенов нескольких изображений способен неявно выучивать соответствия между ними, что мотивирует его применение для решения задачи RefSR. Существенным элементом трансформерной архитектуры является способ

кодирования пространственного положения токенов: переход от абсолютного позиционного кодирования, плохо обобщающегося на разрешения, не встречавшиеся при обучении, к позиционному кодированию на основе вращения (RoPE) [15] и его двумерному обобщению 2D-RoPE [16] позволил трансформерам устойчиво работать с входами различных размеров и пространственной структуры.

Описание предлагаемого метода

Предложенный метод представляет собой трансформер, обрабатывающий входное и референсное изображения совместно как единую последовательность токенов. Общая схема архитектуры приведена на рис. 1.

Опишем ключевые элементы разработанного метода.

Пусть $\mathbf{I}_{\mathrm{LR}}$ — входное изображение низкого разрешения; $\mathbf{I}_{\mathrm{Ref}}$ — референсное изображение высокого разрешения. Входное изображение сначала повышается до целевого разрешения с помощью бикубической интерполяции, что дает изображение $\mathbf{I}_{\mathrm{LR}\uparrow}$. Полученное изображение и референс разбиваются на неперекрывающиеся блоки размером 8 × 8 пикселов, которые линейно проецируются в визуальные токены:

$$\mathbf{X}_{\mathrm{LR}} = \mathrm{Patchify}(\mathbf{I}_{\mathrm{LR}\uparrow}), \quad \mathbf{X}_{\mathrm{Ref}} = \mathrm{Patchify}(\mathbf{I}_{\mathrm{Ref}}),$$

где $\mathbf{X}_{\mathrm{LR}} \in \mathbb{R}^{L_{\mathrm{LR}} \times D_{\mathrm{emb}}}$ и $\mathbf{X}_{\mathrm{Ref}} \in \mathbb{R}^{L_{\mathrm{Ref}} \times D_{\mathrm{emb}}}$ — последовательности токенов для входного и референсного изображений; L_{Ref} и L_{LR} — длины последовательностей токенов; D_{emb} — размерность токена. Полученные последовательности конкатенируются:

$$\mathbf{X} = [\mathbf{X}_{\mathrm{LR}}, \mathbf{X}_{\mathrm{Ref}}] \in \mathbb{R}^{(L_{\mathrm{LR}}+L_{\mathrm{Ref}}) \times D_{\mathrm{emb}}}$$

и подаются на вход набору трансформерных блоков. После прохождения всех блоков последовательность разделяется обратно на части, соответствующие входному и референсному изображениям, и токены входного изображения линейно проецируются в блоки выходного изображения высокого разрешения. Такая организация обработки позволяет механизму внимания напрямую сопоставлять токены входного изображения с токенами референса и аккумулировать из них необходимую информацию, объединяя поиск соответствий и перенос текстур в рамках одной операции.

Для того чтобы механизм внимания мог корректно сопоставлять токены входного и референсного изображений, каждому токену необходимо сопоставить информацию о его пространственном положении. Для входного и референсного изображений строятся раздельные двумерные координатные сетки $\mathbf{G}_{\mathrm{LR}} \in \mathbb{R}^{L_{\mathrm{LR}} \times 2}$ и $\mathbf{G}_{\mathrm{Ref}} \in \mathbb{R}^{L_{\mathrm{Ref}} \times 2}$, в которых каждая строка содержит координаты (x, y) соответствующего блока в его изображении. Обе сетки начинают отсчет координат в левом верхнем углу, который имеет координаты (0, 0), благодаря чему блоки в одинаковых пространственных позициях входного и референсного изображений получают одинаковые координаты, а области, не имеющие соответствия, естественным образом занимают непересекающиеся диапазоны. Сетки конкатенируются в общую сетку $\mathbf{G} = [\mathbf{G}_{\mathrm{LR}}, \mathbf{G}_{\mathrm{Ref}}]$, согласованную с объединенной последовательностью токенов $\mathbf{X}$.

В каждом блоке внимания применяется нормализация query и key векторов (QK-Norm), которая стабилизирует процесс обучения [17]:

$$\mathbf{Q} = \mathrm{LN}(\mathbf{X}\mathbf{W}_{\mathrm{Q}}), \quad \mathbf{K} = \mathrm{LN}(\mathbf{X}\mathbf{W}_{\mathrm{K}}),$$

где $\mathbf{W}_{\mathrm{Q}}$ и $\mathbf{W}_{\mathrm{K}}$ — матрицы линейных проекций; LN — операция LayerNorm [18]. Затем к запросам и ключам применяется 2D-RoPE [16] с использованием общей координатной сетки $\mathbf{G}$:

$$\tilde{\mathbf{Q}} = \text{2D-RoPE}(\mathbf{Q}, \mathbf{G}), \quad \tilde{\mathbf{K}} = \text{2D-RoPE}(\mathbf{K}, \mathbf{G}).$$

Такое позиционное кодирование, далее называемое *бинокулярным*, рассматривает входное и референсное изображения как два «вида» одной сцены, заданных в

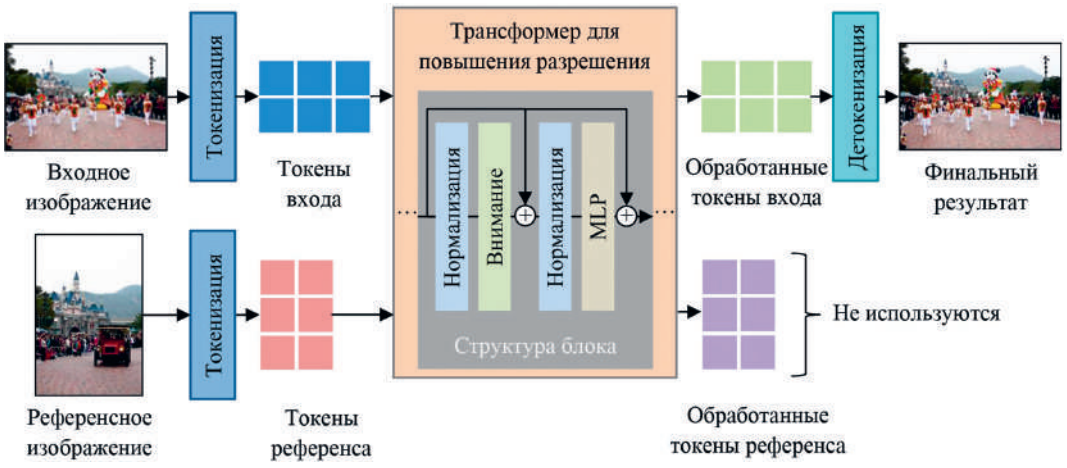

Рис. 1. Архитектура предложенного метода.
MLP — многослойный персептрон

Fig. 1. Architecture of the proposed method.
MLP — multilayer perceptron

общей системе координат, и при этом не накладывает жестких требований на их размеры и соотношения сторон.

При применении модели к изображениям, разрешение которых превышает использовавшиеся при обучении, общее количество токенов в объединенной последовательности увеличивается. Это приводит к смещению распределения логитов внимания: при большем числе токенов нормирующая операция SoftMax вынуждена распределять вероятностную массу между большим числом значений, что эффективно «размывает» внимание и приводит к деградации качества восстановления, проявляющейся в виде блочных артефактов на выходном изображении.

Для устранения этого эффекта без необходимости дообучения модели применяется масштабирование логитов внимания, зависящее от длины обрабатываемой последовательности. Подобный прием ранее предлагался в задачах обработки длинных последовательностей в области обработки естественного языка [19] и в задачах оценки оптического потока [20]; в настоящей работе он впервые применяется к задаче восстановления изображений. Пусть $L = L_{\mathrm{LR}} + L_{\mathrm{Ref}}$ — текущее количество токенов в объединенной последовательности, а L_{train} — максимальное значение этой величины, встречавшееся при обучении. Тогда операция внимания принимает вид:

$$\mathrm{Attention}(\tilde{\mathbf{Q}}, \tilde{\mathbf{K}}, \mathbf{V}) = \mathrm{SoftMax}\left(\frac{\ln L}{\ln L_{\mathrm{train}}} \cdot \frac{\tilde{\mathbf{Q}}\tilde{\mathbf{K}}^{\mathrm{T}}}{\sqrt{D_{\mathrm{K}}}}\right)\mathbf{V},$$

где $\mathbf{V}$ — значения; D_{K} — размерность ключа. При $L = L_{\mathrm{train}}$ масштабирующий множитель равен единице, и формула сводится к стандартной операции внимания; при $L > L_{\mathrm{train}}$ множитель превышает единицу, чем компенсирует увеличение энтропии распределения в механизме внимания при росте длины последовательности, т. е. делает максимумы этого распределения более выраженными, не «размазывая» вероятностную массу по всем токенам последовательности.

Экспериментальная оценка

Предложенный метод реализован в виде трансформера, содержащего 24 блока с размерностью скрытого представления 768 и 16 головами внимания. Обучение проводилось в две стадии на наборе данных Large-Scale Multi-Reference Dataset (LMR) [11]. На первой стадии модель обучалась на парах входных и референсных изображений размером 224 × 224 пикселов в течение 200 000 итераций с использованием оптимизатора AdamW [21] с параметрами $\beta_1 = 0{,}9$ и $\beta_2 = 0{,}95$ и коэффициентом регуляризации весов 0,05. Темп обучения линейно увеличивался до значения $5 \cdot 10^{-4}$ в течение первых 10 000 шагов разогрева, после чего убывал по косинусному закону. Размер мини-батча составлял 384 примера. На второй стадии модель дообучалась на парах изображений размером 480 × 480 пикселов в течение 100 000 итераций с размером мини-батча 96 и максимальной скоростью обучения $5 \cdot 10^{-5}$ с 2000 шагов разогрева; остальные параметры оптимизатора оставались неизменными. В качестве функции потерь на обеих стадиях использовалась L1-норма попиксельной разности между восстановленным и эталонным изображениями. Обучение проводилось на одном вычислительном узле с 8 графическими процессорами с 80 ГБ видеопамяти и занимало около четырех суток.

Для оценки качества предложенного метода использовались эталонные наборы данных CUFED5 [7] и WR-SR [8]. Набор CUFED5 содержит 126 тестовых пар, каждая из которых состоит из входного изображения и 5 референсных изображений различной степени схожести с входным; в соответствии с принятым в научных публикациях протоколом использовался наиболее похожий референс. Набор WR-SR содержит 80 пар входных и референсных изображений. Изображения низкого разрешения формировались посредством бикубического понижения разрешения в четыре раза. В качестве метрик качества применено пиковое отношение сигнала к шуму (Peak Signal-to-Noise Ratio, PSNR) и индекс структурного сходства (Structural Similarity Index Measure, SSIM) [22], вычисленные по каналу яркости в цветовом пространстве YCbCr, а также перцептивная метрика Learned Perceptual Image Patch Similarity (LPIPS) с предобученной сетью AlexNet [23], вычисленная по всем трем цветовым каналам в пространстве RGB.

Предложенный метод сравнивался с современными методами SISR: HAT [5] и MambaIRv2-L [6], а также с методами RefSR: C²-Matching [8], DATSR [9], TDF [10] и MRefSR [11]. Для каждого из сравниваемых методов использовались опубликованные авторами веса, обученные на соответствующем наборе данных: методы SISR обучены на наборе DF2K, методы C²-Matching, DATSR и TDF — на CUFED5, метод MRefSR и предложенный метод — на LMR. Результаты количественного сравнения представлены в таблице.

На наборе данных CUFED5 предложенный метод превосходит лучший из подобных подходов MRefSR на 0,33 дБ по метрике PSNR и достигает наилучших значений по SSIM и LPIPS. На наборе данных WR-SR предложенный метод также превосходит подход MRefSR на 0,33 дБ по метрике PSNR. В ранее опубликованных работах сравнение методов RefSR с методами SISR проводилось, как правило, с переобучением последних на наборах данных задач RefSR, что приводило к ухудшению качества; в настоящей же работе предложенный метод сравнивается с сильными моделями SISR, обученными на крупномасштабном наборе DF2K, существенно превышающем по объему используемые в RefSR наборы, и впервые превосходит их с приростом 0,29 дБ PSNR относительно MambaIRv2-L и наилучшими значениями по SSIM и LPIPS. Полученные результаты подтверждают, что использование механизма внимания над объединенной последовательностью токенов входного и референсного изображений позволяет эффективно использовать дополнительную информацию из референса и компенсировать разрыв в объеме обучающих данных между RefSR и SISR.

Качественное сравнение предложенного метода с подобными подходами представлено на рис. 2 на примерах из набора данных WR-SR. На приведенных фраг-

Таблица. Количественное сравнение предложенного метода с аналогами из научных работ
Table. Quantitative comparison of the proposed approach to prior art

Название метода	Тестовый набор данных	
	CUFED5 PSNR↑ / SSIM↑ / LPIPS↓	WR-SR PSNR↑ / SSIM↑ / LPIPS↓
HAT	27,16 / 0,804 / 0,233	28,52 / 0,806 / 0,291
MambaIRv2-L	27,23 / 0,805 / 0,238	28,56 / *0,807* / 0,289
C^2-Matching	28,24 / 0,841 / 0,170	28,32 / 0,801 / 0,292
DATSR	28,50 / 0,850 / *0,153*	28,34 / 0,805 / *0,279*
MRefSR	*28,66* / *0,851* / 0,165	28,52 / 0,806 / 0,285
Предложенный метод	**28,99 / 0,861 / 0,152**	**28,85 / 0,816 / 0,266**

Примечание: жирным шрифтом отмечены лучшие результаты, курсивом — вторые лучшие.

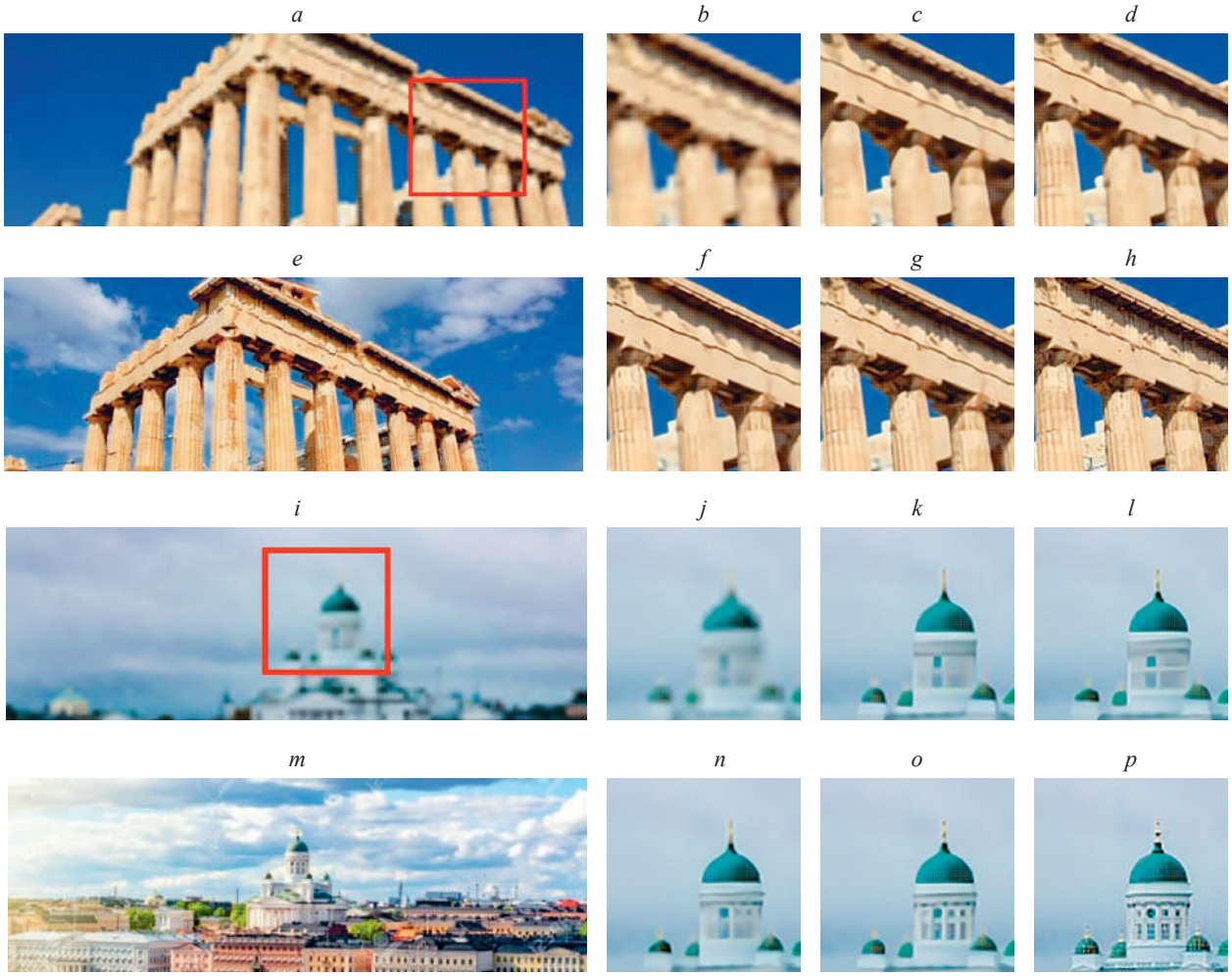

Рис. 2. Качественное сравнение с аналогами на примерах из набора данных WR-SR: входные изображения (*a*, *i*); фрагменты входа (*b*, *j*); методы MambaIRv2 (*c*, *k*) и DATSR (*d*, *l*); референсные изображения (*e*, *m*); методы MRefSR (*f*, *n*) и предложенный (*g*, *o*); эталонное изображение (*h*, *p*).

Подобласть, на которой приведены результаты работы подходов выделена красной рамкой

Fig. 2. Qualitative comparison with analogues, using examples from the WR-SR dataset: input images (*a*, *i*); input fragments (*b*, *j*); MambaIRv2 (*c*, *k*) and DATSR (*d*, *l*) methods; reference images (*e*, *m*); MRefSR (*f*, *n*) and the proposed method (*g*, *o*); reference image (*h*, *p*).

The subregion where the results of the approaches are presented is highlighted with a red border

ментах видно, что предложенный метод более полно переносит текстуры из референсного изображения и формирует более четкие границы по сравнению с методами RefSR (MRefSR, DATSR) и дает восстановление, сопоставимое по детализации с эталонным изображением.

Заключение

Рассмотрена задача повышения разрешения изображений с использованием референсных изображений. Применяемые в большинстве современных подходов методы декомпозируют задачу на отдельные этапы поиска соответствий, выравнивания и переноса текстур, реализованные в виде последовательно обучаемых модулей и нередко опирающиеся на предобученные компоненты. В качестве альтернативы предложен метод, представляющий собой единую обучаемую трансформерную модель, в которой поиск соответствий и перенос текстур выполняются совместно в рамках применения механизма внимания к объединенной последовательности токенов входного и референсного изображений. Введены бинокулярное двумерное позиционное кодирование 2D Rotary Positional Embedding с общей системой координат для обоих изображений и схема адаптивного масштабирования логитов внимания, зависящего от длины обрабатываемой последовательности. На эталонных наборах данных CUFED5 и WR-SR при бикубическом четырехкратном понижении разрешения предложенный метод превосходит лучшие из известных подходов Reference-based Super-Resolution на 0,33 дБ по метрике Peak Signal-to-Noise Ratio (PSNR) на обоих наборах данных и впервые опережает методы Single Image Super-Resolution, обученные на крупномасштабном наборе DF2K, с приростом 0,29 дБ PSNR на наборе WR-SR. Потенциальной областью применения предложенного подхода является восстановление изображений в задачах, где доступен дополнительный референс со схожим содержанием. Направлениями дальнейших исследований являются адаптация метода к более широкому спектру типов деградаций, включая реальные сценарии съемки, а также построение крупномасштабных обучающих наборов данных Reference-based Super-Resolution, сопоставимых по объему с наборами данных, используемыми в задачах Single Image Super-Resolution, что должно способствовать дальнейшему сокращению разрыва в качестве восстановления между этими двумя классами методов.

Литература

1. Dong C., Loy C.C., He K., Tang X. Learning a deep convolutional network for image super-resolution // Lecture Notes in Computer Science. 2014. Lecture Notes in Computer Science. V. 8692. P. 184–199. doi: 10.1007/978-3-319-10593-2_13
2. Kim J., Lee J.K., Lee K.M. Accurate image super-resolution using very deep convolutional networks // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 1646–1654. doi: 10.1109/CVPR.2016.182
3. Lim B., Son S., Kim H., Nah S., Lee K.M. Enhanced deep residual networks for single image super-resolution // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017. P. 1132–1140. doi: 10.1109/CVPRW.2017.151
4. Liang J., Cao J., Sun G., Zhang K., Van Gool L., Timofte R. SwinIR: Image restoration using Swin Transformer // Proc. of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021. P. 1833–1844. doi: 10.1109/ICCVW54120.2021.00210
5. Chen X., Wang X., Zhou J., Qiao Y., Dong C. Activating more pixels in image super-resolution transformer // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 22367–22377. doi: 10.1109/CVPR52729.2023.02142
6. Guo H., Guo Y., Zha Y., Zhang Y., Li W., Dai T., et al. MambaIRv2: Attentive state space restoration // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025. P. 28124–28133. doi: 10.1109/CVPR52734.2025.02619
7. Zhang Z., Wang Z., Lin Z., Qi H. Image super-resolution by neural texture transfer // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 7974–7983. doi: 10.1109/CVPR.2019.00817
8. Jiang Y., Chan K.C.K., Wang X., Loy C.C., Liu Z. Robust reference-based super-resolution via C^2-Matching // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 2103–2112. doi: 10.1109/CVPR46437.2021.00214
9. Cao J., Liang J., Zhang K., Li Y., Zhang Y., Wang W., et al. Reference-based image super-resolution with deformable attention transformer // Lecture Notes in Computer Science. 2022. V. 13678. P. 325–342. doi: 10.1007/978-3-031-19797-0_19
10. Huang Y., Zhang X., Fu Y., Chen S., Zhang Y., Wang Y., et al. Task decoupled framework for reference-based super-resolution // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 5921–5930. doi: 10.1109/CVPR52688.2022.00584

References

1. Dong C., Loy C.C., He K., Tang X. Learning a deep convolutional network for image super-resolution. *Lecture Notes in Computer Science. 2014. Lecture Notes in Computer Science*, vol. 8692, pp. 184–199. doi: 10.1007/978-3-319-10593-2_13
2. Kim J., Lee J.K., Lee K.M. Accurate image super-resolution using very deep convolutional networks. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654. doi: 10.1109/CVPR.2016.182
3. Lim B., Son S., Kim H., Nah S., Lee K.M. Enhanced deep residual networks for single image super-resolution. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140. doi: 10.1109/CVPRW.2017.151
4. Liang J., Cao J., Sun G., Zhang K., Van Gool L., Timofte R. SwinIR: Image restoration using Swin Transformer. *Proc. of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1833–1844. doi: 10.1109/ICCVW54120.2021.00210
5. Chen X., Wang X., Zhou J., Qiao Y., Dong C. Activating more pixels in image super-resolution transformer. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22367–22377. doi: 10.1109/CVPR52729.2023.02142
6. Guo H., Guo Y., Zha Y., Zhang Y., Li W., Dai T., et al. MambaIRv2: Attentive state space restoration. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 28124–28133. doi: 10.1109/CVPR52734.2025.02619
7. Zhang Z., Wang Z., Lin Z., Qi H. Image super-resolution by neural texture transfer. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7974–7983. doi: 10.1109/CVPR.2019.00817
8. Jiang Y., Chan K.C.K., Wang X., Loy C.C., Liu Z. Robust reference-based super-resolution via C^2-Matching. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2103–2112. doi: 10.1109/CVPR46437.2021.00214
9. Cao J., Liang J., Zhang K., Li Y., Zhang Y., Wang W., et al. Reference-based image super-resolution with deformable attention transformer. *Lecture Notes in Computer Science*, 2022, vol. 13678, pp. 325–342. doi: 10.1007/978-3-031-19797-0_19
10. Huang Y., Zhang X., Fu Y., Chen S., Zhang Y., Wang Y., et al. Task decoupled framework for reference-based super-resolution. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5921–5930. doi: 10.1109/CVPR52688.2022.00584

11. Zhang L., Li X., He D., Li F., Ding E., Zhang Z. LMR: A large-scale multi-reference dataset for reference-based super-resolution // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 13072–13081. doi: 10.1109/ICCV51070.2023.01206
12. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., et al. An image is worth 16×16 words: Transformers for image recognition at scale // Proc. of the International Conference on Learning Representations (ICLR). 2021.
13. Jin H., Jiang H., Tan H., Zhang K., Bi S., Zhang T., et al. LVSM: A large view synthesis model with minimal 3D inductive bias // Proc. of the International Conference on Learning Representations (ICLR). 2025.
14. Wang J., Chen M., Karaev N., Vedaldi A., Rupprecht C., Novotny D. VGGT: Visual Geometry Grounded Transformer // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025. P. 5294–5306. doi: 10.1109/CVPR52734.2025.00499
15. Su J., Ahmed M., Lu Y., Pan S., Bo W., Liu Y. RoFormer: Enhanced transformer with rotary position embedding // Neurocomputing. 2024. V. 568. P. 127063. doi: 10.1016/j.neucom.2023.127063
16. Heo B., Park S., Han D., Yun S. Rotary position embedding for vision transformer // Lecture Notes in Computer Science. 2025. V. 15068. P. 289–305. doi: 10.1007/978-3-031-72684-2_17
17. Dehghani M., Djolonga J., Mustafa B., Padlewski P., Heek J., Gilmer J., et al. Scaling vision transformers to 22 billion parameters // Proc. of the 40th International Conference on Machine Learning. 2023. V. 202. P. 7480–7512.
18. Ba J.L., Kiros J.R., Hinton G.E. Layer normalization // arXiv. 2016. arXiv:1607.06450. doi: 10.48550/arXiv.1607.06450
19. Chiang D., Cholak P. Overcoming a theoretical limitation of self-attention // Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. 2022. P. 7654–7664. doi: 10.18653/v1/2022.acl-long.527
20. Dong Q., Fu Y. MemFlow: Optical flow estimation and prediction with memory // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 19068–19078. doi: 10.1109/CVPR52733.2024.01804
21. Loshchilov I., Hutter F. Decoupled weight decay regularization // Proc. of the International Conference on Learning Representations (ICLR). 2019.
22. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. V. 13. N 4. P. 600–612. doi: 10.1109/TIP.2003.819861
23. Zhang R., Isola P., Efros A.A., Shechtman E., Wang O. The unreasonable effectiveness of deep features as a perceptual metric // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 586–595. doi: 10.1109/CVPR.2018.00068

11. Zhang L., Li X., He D., Li F., Ding E., Zhang Z. LMR: A large-scale multi-reference dataset for reference-based super-resolution. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13072–13081. doi: 10.1109/ICCV51070.2023.01206
12. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., et al. An image is worth 16×16 words: Transformers for image recognition at scale. *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.
13. Jin H., Jiang H., Tan H., Zhang K., Bi S., Zhang T., et al. LVSM: A large view synthesis model with minimal 3D inductive bias. *Proc. of the International Conference on Learning Representations (ICLR)*, 2025.
14. Wang J., Chen M., Karaev N., Vedaldi A., Rupprecht C., Novotny D. VGGT: Visual Geometry Grounded Transformer. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 5294–5306. doi: 10.1109/CVPR52734.2025.00499
15. Su J., Ahmed M., Lu Y., Pan S., Bo W., Liu Y. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024, vol. 568, pp. 127063. doi: 10.1016/j.neucom.2023.127063
16. Heo B., Park S., Han D., Yun S. Rotary position embedding for vision transformer. *Lecture Notes in Computer Science*, 2025, vol. 15068, pp. 289–305. doi: 10.1007/978-3-031-72684-2_17
17. Dehghani M., Djolonga J., Mustafa B., Padlewski P., Heek J., Gilmer J., et al. Scaling vision transformers to 22 billion parameters. *Proc. of the 40th International Conference on Machine Learning*, 2023, vol. 202, pp. 7480–7512.
18. Ba J.L., Kiros J.R., Hinton G.E. Layer normalization. *arXiv*, 2016. arXiv:1607.06450. doi: 10.48550/arXiv.1607.06450
19. Chiang D., Cholak P. Overcoming a theoretical limitation of self-attention. *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 7654–7664. doi: 10.18653/v1/2022.acl-long.527
20. Dong Q., Fu Y. MemFlow: Optical flow estimation and prediction with memory. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19068–19078. doi: 10.1109/CVPR52733.2024.01804
21. Loshchilov I., Hutter F. Decoupled weight decay regularization. *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.
22. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004, vol. 13, no. 4, pp. 600–612. doi: 10.1109/TIP.2003.819861
23. Zhang R., Isola P., Efros A.A., Shechtman E., Wang O. The unreasonable effectiveness of deep features as a perceptual metric. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595. doi: 10.1109/CVPR.2018.00068

Авторы

Яковенко Александр Викторович — аспирант, Московский государственный университет им М.В. Ломоносова, Москва, 119991, Российская Федерация, sc 57226339367, https://orcid.org/0000-0003-3105-512X, alexander.yakovenko@graphics.cs.msu.ru

Денисов Алексей Константинович — кандидат технических наук, ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57210698353, https://orcid.org/0000-0001-8135-1723, denisov@itmo.ru

Ватолин Дмитрий Сергеевич — кандидат физико-математических наук, глава лаборатории интеллектуального анализа видео, Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем МГУ имени М.В. Ломоносова, Москва, 119192, Российская Федерация, sc 6506528788, https://orcid.org/0000-0002-8893-9340, dmitriy@graphics.cs.msu.ru

Кустарев Павел Валерьевич — кандидат технических наук, декан, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 35317916600, https://orcid.org/0000-0001-9326-0837, kustarev@itmo.ru

Authors

Alexander V. Yakovenko — PhD Student, Lomonosov Moscow State University, Moscow, 119991, Russian Federation, sc 57226339367, https://orcid.org/0000-0003-3105-512X, alexander.yakovenko@graphics.cs.msu.ru

Aleksei K. Denisov — PhD, Assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57210698353, https://orcid.org/0000-0001-8135-1723, denisov@itmo.ru

Dmitriy S. Vatolin — PhD (Physics & Mathematics), Head of Intelligent Video Analysis Lab, MSU Institute for Artificial Intelligence, Moscow, 119192, Russian Federation, sc 6506528788, https://orcid.org/0000-0002-8893-9340, dmitriy@graphics.cs.msu.ru

Pavel V. Kustarev — PhD, Dean, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 35317916600, https://orcid.org/0000-0001-9326-0837, kustarev@itmo.ru

Статья поступила в редакцию 18.06.2026
Одобрена после рецензирования 24.08.2026
Принята к печати 24.09.2026

Received 18.06.2026
Approved after reviewing 24.08.2026
Accepted 24.09.2026

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2026-26-5-1073-1076
УДК 004.934

Метод одноканального разделения речевых источников для задач распознавания речи и верификации диктора

Анастасия Александровна Зорькина[1]✉, Никита Анатольевич Хмелев[2]

[1,2] ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация
[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] zorkina@speechpro.com✉, https://orcid.org/0000-0002-9434-8897
[2] khmelev@speechpro.com, https://orcid.org/0009-0006-8662-6750

Аннотация
Предложен метод одноканального разделения двух речевых источников на основе условного генеративного моделирования. Рассматриваемая задача характеризуется неопределенностью порядка двух восстановленных речевых сигналов: одна и та же пара разделенных источников может быть представлена в двух перестановках. Для устранения этой неоднозначности используется замороженный кодировщик диктора: он задает порядок источников при обучении, выбирает лучший вариант среди *N* стохастических генераций и согласует источники при обработке записей произвольной длины по фрагментам. На наборе данных Libri2Mix предложенный вариант на основе архитектуры Transformer U-Net конкурентен по фундаментальным метрикам оценки качества и разборчивости обработанной речи и дает наилучшие результаты при использовании в задачах распознавания речи и верификации диктора.

Ключевые слова
разделение речевых источников, генеративное моделирование, сопоставление потоков, распознавание речи, верификация диктора

Ссылка для цитирования: Зорькина А.А., Хмелев Н.А. Метод одноканального разделения речевых источников для задач распознавания речи и верификации диктора // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1073–1076. doi: 10.17586/2226-1494-2026-26-5-1073-1076

Single-channel speech source separation for speech recognition and speaker verification

Anastasia A. Zorkina[1]✉, Nikita A. Khmelev[2]

[1,2] STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation
[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] zorkina@speechpro.com✉, https://orcid.org/0000-0002-9434-8897
[2] khmelev@speechpro.com, https://orcid.org/0009-0006-8662-6750

Abstract
A method for single-channel separation of two speech sources based on conditional generative modeling is proposed. The task is characterized by uncertainty in the order of the two reconstructed speech signals: the same pair of separated sources can be represented in two permutations. To resolve this ambiguity, a frozen speaker encoder is used: it defines the source order during training, selects the best variant among *N* stochastic generations, and aligns the sources during chunk-wise processing of recordings of arbitrary length. On the Libri2Mix dataset, the proposed Transformer U-Net-based variant is competitive in terms of fundamental metrics for evaluating the quality and intelligibility of processed speech and achieves the best results when used in speech recognition and speaker verification tasks.

Keywords
speech source separation, generative modeling, flow matching, speech recognition, speaker verification

© Зорькина А.А., Хмелев Н.А., 2026

For citation: Zorkina A.A., Khmelev N.A. Single-channel speech source separation for speech recognition and speaker verification. *Scientific and Technical Journal of Information Technologies, Mechanics and Optic*s, 2026, vol. 26, no. 5, pp. 1073–1076. doi: 10.17586/2226-1494-2026-26-5-1073-1076

Одноканальное разделение речевых источников является важной задачей для систем автоматического распознавания речи, верификации диктора и речевой аналитики. На практике входной речевой сигнал часто содержит наложение голосов нескольких людей (источников), а ошибки их разделения приводят к уменьшению точности как распознавания речи, так и верификации диктора, поскольку в обработанном речевом сигнале сохраняется остаточная речь собеседника. Исходя из этого, качество разделения речевых источников следует оценивать не только по акустическим метрикам, но и по тому, насколько хорошо после него работают распознавание речи и верификация диктора [1]. Современные нейросетевые разделители часто обучаются с учетом возможных перестановок источников [2]. Однако такой подход устраняет неоднозначность порядка только при вычислении функции потерь и не гарантирует согласованность разделенных сигналов при пофрагментной обработке: один и тот же диктор может оказаться в разных выходных дорожках на соседних фрагментах.

В настоящей работе предложен новый метод разделения наложенной речи двух дикторов, в котором смесь их голосов рассматривается как условие для генерации двух восстановленных раздельных речевых сигналов (рисунок). Задача формулируется в рамках условного моделирования потоков [3]: модель получает спектральные признаки одноканальной записи m, содержащей наложение голосов двух дикторов, и обучается описывать преобразование случайно сгенерированного шума x_0 в целевой разделенный сигнал. Целевой сигнал задается как пара чистых речевых сигналов двух дикторов s_1 и s_2, записанных в одном общем представлении $s_A\|r\|s_B$, где s_A и s_B — речевые сигналы первого и второго дикторов после упорядочивания, а r — фиксированный неречевой разделитель. При этом порядок следования источников определяется с помощью заранее обученного дикторского кодировщика — нейросетевой модели, преобразующей речевой фрагмент в векторное представление диктора. Для каждой смеси векторные представления e_{s_1} и e_{s_2} двух чистых источников сравниваются с заранее заданным фиксированным опорным представлением e_{ref} по косинусной мере близости: более близкий источник записывается первым. Тем самым модель обучается не только удалять взаимную речевую помеху, но и сохранять фиксированную последовательность выходных источников.

На этапе обучения из целевого сигнала формируются его спектральные признаки x_1, а спектральные признаки дублированной смеси $\overline{m} = m\|r\|m$ используются как условие c. Для выбранного момента времени t модель получает промежуточное состояние x_t и условие c и обучается аппроксимировать векторное поле, задающее направление изменения текущего состояния к целевым спектральным признакам разделенных речевых сигналов. Обучение выполняется путем минимизации функции потерь условного моделирования потоков (Conditional Flow Matching, CFM) $\mathcal{L}_{\text{CFM}}$ [3], измеряю-

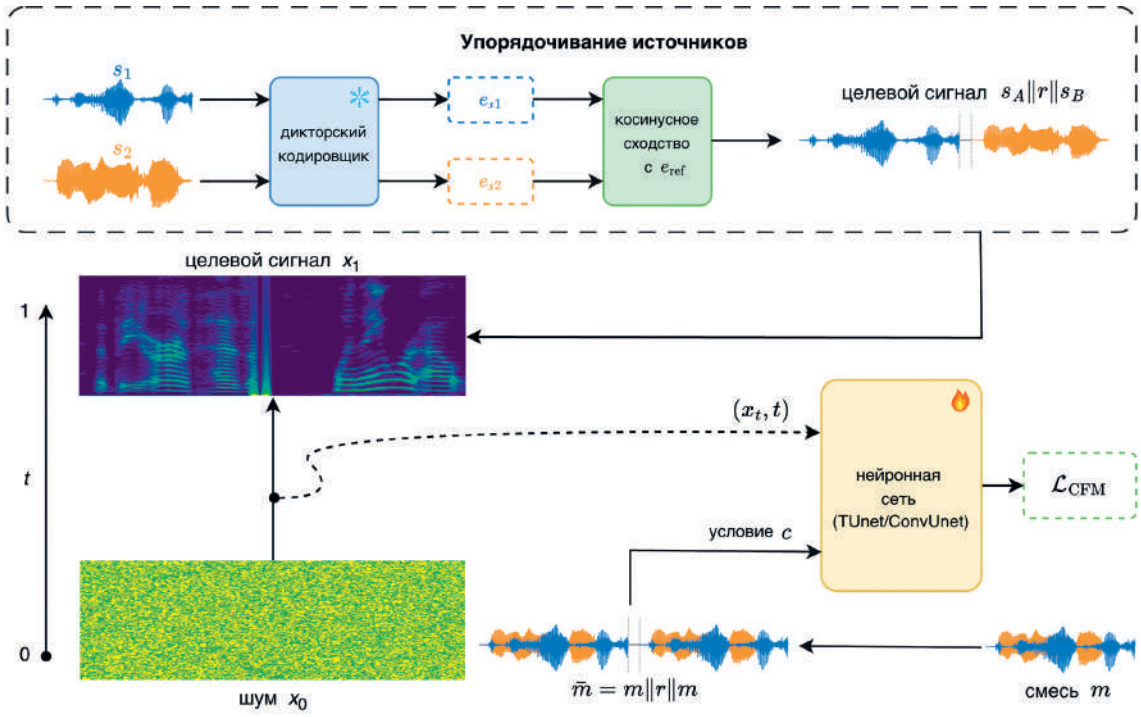

Рисунок. Схема обучения предложенного метода
Figure. Training scheme of the proposed method

щей отклонение предсказанного векторного поля от целевого. В качестве архитектур рассматривались две предобученные модели, ранее применявшиеся в генеративной обработке речи [4]: сверточная сеть NCSN++ (ConvUnet) и Spectrogram Transformer U-Net (TUnet).

На этапе применения запись обрабатывается фрагментами длительностью 1 с с шагом 0,5 с. Для каждого фрагмента генерируется N независимых вариантов разделения, поскольку генеративная модель может выдавать несколько правдоподобных решений для одной и той же смеси. Среди полученных вариантов выбирается тот, в котором косинусное сходство дикторских представлений двух выходных каналов минимально. Такой критерий позволяет выбрать пару сигналов, наиболее вероятно соответствующую двум разным голосам, и не требует доступа к эталонным чистым источникам. После выбора локального результата остается согласовать порядок каналов между соседними фрагментами. В результате для каждой из двух восстановленных аудиодорожек накапливаются дикторские представления уже обработанных фрагментов и вычисляется центроид, описывающий голос соответствующего диктора. Для нового фрагмента оцениваются две возможные перестановки выходных каналов: исходная и обратная. Выбирается та перестановка, при которой дикторские представления каналов имеют наибольшую суммарную косинусную близость к центроидам соответствующих дорожек. После согласования фрагменты возвращаются на исходные временные позиции в соответствующих выходных аудиодорожках, а перекрывающиеся участки усредняются. В итоге формируются две непрерывные аудиодорожки, соответствующие двум разделенным речевым источникам.

Экспериментальная апробация подхода выполнена на наборе данных Libri2Mix [5] для условий с чистыми и зашумленными смесями двух голосов. Сравнение проводилось с моделью SepReformer [6] и методом извлечения целевого диктора MeanFlow-TSE [7]. Качество разделения оценивалось по метрикам Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Perceptual Evaluation of Speech Quality (PESQ) и Extended Short-Time Objective Intelligibility (ESTOI) [1]; влияние на прикладные речевые системы — по метрикам concatenated minimum-permutation Word Error Rate (cpWER) для модели Whisper V3 [8] распознавания речи и Equal Error Rate (EER) для модели dWhisper-ASV верификации диктора [9]. Вариант на основе TUnet при $N = 4$ показал конкурентное качество разделения: на чистых смесях — 17,30 дБ для метрики SI-SDR; 3,11 — PESQ и 0,93 — ESTOI. Влияние разделения дикторов на последующие речевые задачи приведено в таблице. Предложенная схема с TUnet дает наименьшие ошибки среди всех сравниваемых систем: 3,84 % cpWER для распознавания речи и 1,05 % EER для верификации диктора.

Таблица. Влияние разделения речи на ее распознавание и верификацию диктора на чистом наборе данных Libri2Mix

Table. Downstream ASR and speaker verification performance on clean Libri2Mix

Модель	Метрика	
	cpWER, %	EER, %
ConvUnet	4,99	1,28
TUnet	3,84	1,05
MeanFlow-TSE	9,05	3,25
SepReformer	5,02	1,29

Таким образом, при сопоставимом качестве разделения речи предложенный подход оказывается более эффективным в сценариях распознавания речи и верификации дикторов по смеси двух речевых источников (голосов дикторов).

Литература

1. Li K., Chen G., Sang W., Luo Y., Chen Z., Wang S., et al. Advances in speech separation: Techniques, challenges, and future trends // arXiv. 2025. arXiv:2508.10830. doi: 10.48550/arXiv.2508.10830
2. Yu D., Kolbaek M., Tan Z.-H., Jensen J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. P. 241–245. doi: 10.1109/icassp.2017.7952154
3. Lipman Y., Chen R.T., Ben-Hamu H., Nickel M., Le M. Flow matching for generative modeling // Proc. of the 11th International Conference on Learning Representations (ICLR). 2023. P. 1–28.
4. Ku P.-J., Liu A.H., Korostik R., Huang S.-F., Fu S.-W., Jukić A. Generative speech foundation model pretraining for high-quality speech extraction and restoration // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025. P. 1–5. doi: 10.1109/icassp49660.2025.10888830
5. Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. LibriMix: An open-source dataset for generalizable speech separation // arXiv. 2020. arXiv:2005.11262. doi: 10.48550/arXiv.2005.11262
6. Shin U.-H., Lee S., Kim T., Park H.-M. Separate and reconstruct: Asymmetric encoder-decoder for speech separation // Proc. of the 38th International Conference on Neural Information Processing Systems. 2024. P. 52215–52240. doi: 10.52202/079017-1655

References

1. Li K., Chen G., Sang W., Luo Y., Chen Z., Wang S., et al. Advances in speech separation: Techniques, challenges, and future trends. *arXiv*, 2025. arXiv:2508.10830. doi: 10.48550/arXiv.2508.10830
2. Yu D., Kolbaek M., Tan Z.-H., Jensen J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. doi: 10.1109/icassp.2017.7952154
3. Lipman Y., Chen R.T., Ben-Hamu H., Nickel M., Le M. Flow matching for generative modeling. *Proc. of the 11th International Conference on Learning Representations (ICLR)*, 2023, pp. 1–28.
4. Ku P.-J., Liu A.H., Korostik R., Huang S.-F., Fu S.-W., Jukić A. Generative speech foundation model pretraining for high-quality speech extraction and restoration. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5. doi: 10.1109/icassp49660.2025.10888830
5. Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. LibriMix: An open-source dataset for generalizable speech separation. *arXiv*, 2020. arXiv:2005.11262. doi: 10.48550/arXiv.2005.11262
6. Shin U.-H., Lee S., Kim T., Park H.-M. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. *Proc. of the 38th International Conference on Neural Information Processing Systems*, 2024, pp. 52215–52240. doi: 10.52202/079017-1655

7. Shimizu R., Jiang X., Mesgarani N. MeanFlow-TSE: One-step generative target speaker extraction with mean flow // arXiv. 2025. arXiv:2512.18572. doi: 10.48550/arXiv.2512.18572
8. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // Proc. of the 40th International Conference on Machine Learning. 2023. P. 28492–28518.
9. Khmelev N., Anikin A., Korenevskaya A., Zorkina A., Novoselov S., Volokhov V., et al. dWhisper-ASV: Distilling knowledge into Whisper encoder for fast and domain-robust speaker recognition // Proc. of the 28th International Conference on Digital Signal Processing and its Applications (DSPA). 2026. P. 1–6. doi: 10.1109/dspa69176.2026.11476749

7. Shimizu R., Jiang X., Mesgarani N. MeanFlow-TSE: One-step generative target speaker extraction with mean flow. *arXiv*, 2025. arXiv:2512.18572. doi: 10.48550/arXiv.2512.18572
8. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. *Proc. of the 40th International Conference on Machine Learning*, 2023, pp. 28492–28518.
9. Khmelev N., Anikin A., Korenevskaya A., Zorkina A., Novoselov S., Volokhov V., et al. dWhisper-ASV: Distilling knowledge into Whisper encoder for fast and domain-robust speaker recognition. *Proc. of the 28th International Conference on Digital Signal Processing and its Applications (DSPA)*, 2026, pp. 1–6. doi: 10.1109/dspa69176.2026.11476749

Авторы

Зорькина Анастасия Александровна — младший научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57807877500, https://orcid.org/0000-0002-9434-8897, zorkina@speechpro.com

Хмелев Никита Анатольевич — научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 58601210700, https://orcid.org/0009-0006-8662-6750, khmelev@speechpro.com

Authors

Anastasia A. Zorkina — Junior Researcher, STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57807877500, https://orcid.org/0000-0002-9434-8897, zorkina@speechpro.com

Nikita A. Khmelev — Researcher, STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 58601210700, https://orcid.org/0009-0006-8662-6750, khmelev@speechpro.com

Статья поступила в редакцию 04.07.2026
Одобрена после рецензирования 28.07.2026
Принята к печати 19.09.2026

Received 04.07.2026
Approved after reviewing 28.07.2026
Accepted 19.09.2026

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»