

УДК 004.89

doi: 10.15622/rcai.2025.070

НЕЙРОСЕТЕВАЯ МОДЕЛЬ ТЕМПОРАЛЬНОГО ОБЪЕДИНЕНИЯ КАДРОВ ВИДЕОЛЕКЦИИ ДЛЯ РЕШЕНИЯ ЗАДАЧИ РЕКОНСТРУКЦИИ ИЗОБРАЖЕНИЯ

М.Е. Исмагулов (*m_ismagulov@ugrasu.ru*)^A

А.В. Мельников (*melnikovav@uriit.ru*)^{A,B}

^A Югорский государственный университет, Ханты-Мансийск

^B Югорский научно-исследовательский институт
информационных технологий, Ханты-Мансийск

В связи с ростом популярности формата видеолекций, возникают задачи мультимодальной обработки видеолекции в процессе которой можно получить конспект лекции или краткое содержание видеоматериала. При мультимодальной обработке видеолекции извлечение данных из видеоряда осложняется перекрытием контента лектором. Для решения этой задачи существует метод темпорального объединения кадров, данный метод широко применяется для реконструкции изображений в случаях движущихся объектов в кадре. Цель исследования разработка нейросетевой модели темпорального объединения видеокадров для восстановления областей доски. Отражен процесс разработки и обучения нейросетевой модели, в качестве метода обучения выбрано обучение с учителем. Основной датасет выбраны кадры видеолекции разбитые на сэмплы с эталонными изображениями доски. Модель построена на основе гибридной архитектуры, сочетающей сверточную нейронную сеть и рекуррентный слой с долгой краткосрочной памятью (LSTM). По результатам обучения были получены значения метрик accuracy: 0.7711, loss: 0.0773, метрика PSNR достигла 35 децибел, что является хорошим показателем восстановления изображений.

Ключевые слова: темпоральное объединение видеокадров, сверточные нейронные сети, рекуррентные нейронные сети, обучение нейронной сети, реконструкция изображения, обработка кадров видеолекции.

Введение

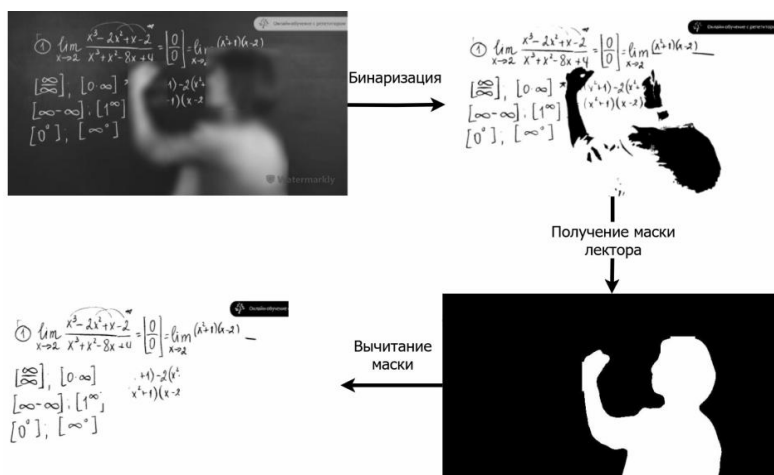
В настоящее время формат видеолекций приобретает большую популярность и активно развивается. Это проявляется в росте числа массовых открытых онлайн-курсов (MOOC), вебинаров, а также в распространении практики видеозаписи очных лекций. Помимо роста популярности, увеличивается и объем такого контента. В связи с этим все острее ощущается потребность в сервисах, автоматически преобразующих видеолекции в текстовый формат для создания конспектов или аннотированных документов. В данной работе, как часть решения этой комплексной задачи, рассматривается алгоритм темпорального объединения кадров, направленный на реконструкцию изображения.

В процессе мультимодальной обработки видеолекции с лектором и доской, необходимо извлекать информацию с каждой из модальностей, как из аудиодорожки, так и из видеоряда, в процессе извлечения информации существует проблема того, что часть контента доски скрывается за лектором, и периодически открывается при перемещении лектора [Wang et al., 2022]. Получить единое представление доски без пробелов в автоматическом режиме становится нетривиальной задачей [Urala Kota et al., 2019]. Единое представление доски необходимо для последующего оптического распознавания символов (далее OCR), и получения текстового содержания видеомодальности. Если в изображениях, подаваемых на модель OCR, будут пробелы или незаполненные участки, то информация будет не полной и иметь ошибки. Таким образом, получить точное содержание видеомодальности становится труднодостижимым.

Для решения подобных проблем можно использовать методы темпорального объединения кадров видео [Urala Kota et al., 2018]. Темпоральное объединение кадров — это метод обработки видео или изображений, при котором кадры, близкие во времени (например, в пределах одной сцены), объединяются на основе их временной последовательности, чтобы восстановить, усилить или дополнить информацию, недоступную на отдельных кадрах [Zhang et al., 2024].

Идея заключается в следующем, если исключить (абстрагировать) лектора из кадра путем вычитания маски и оставить только надписи на доске, то получится область, в которой содержимое доски отсутствует, подробнее можно ознакомиться если обратиться к рис. 1. Абстрагирование необходимо для сосредоточения внимания на содержимом видеолекции, в данном случае, на содержимом доски [Kumar et al., 2020], [Исмагулов, 2024]. Получив очищенное, полное представление изображения доски, в дальнейшем можно передать изображение на модель OCR в результате обработки которой, можно получить текст, математические выражения в формате LaTeX или MathML. Объединив текстовое содержание видеомодальности и текстовое содержание аудиомодальности можно получить

документ, отражающий полное содержание видеолекции. Вне зависимости, от содержащихся на доске видов информации, таких как – текст, формулы, графики и т.д., задача модели состоит в корректном объединении кадров на уровне сцен, где каждая сцена отражает содержание доски на определенном этапе видеолекции. Корректность заключается в том, что на полученных изображениях доски отсутствуют артефакты, мешающие применению OCR, а также в заполнении пустых участков изображения.



Источник: видео TutorOnline (<https://youtu.be/19TSR9rtvwxQ>), авторские права принадлежат TutorOnline

Рис. 1. Схема процесса абстрагирования лектора

Как видно из рис. 1, на последнем этапе обработки в области доски образуется пустая область, в процессе воспроизведения ролика лектор перемещается в кадре открывая разные области. За счет открывающихся областей, возможна реконструкция изображения доски в полном объеме. Для объединения кадров с разными открытыми областями доски и применяется метод темпорального объединения видеок кадров.

Построение датасета

В качестве основы датасета были выбраны несколько видеолекций по математике, с помощью библиотеки FFMPEG и языка Python из видеолекций были извлечены кадры, затем к каждому кадру была применена функция бинаризации изображения из библиотеки OpenCV.

Далее кадры были распределены по видеолекциям, для каждой видеолекции вручную производилась сортировка по сценам, где каждая сцена, это изменение положение кадра. Например, зуммирование, изменение ракурса камеры и т.д. Затем сцены были разбиты на сэмплы по 15 кадров,

для каждого сэмпла было подготовлено изображение истинности для реализации метода обучения с учителем [Feichtenhofer et al., 2018], [Wang et al., 2017]. Размер сэмпла в 15 кадров был выбран экспериментально для избежания ошибки out of memory, поскольку входной тензор формируется из кадров 1920 на 1080 пикселей. Подготовка изображения истинности была выполнена вручную, путем наложения бинаризированных кадров друг на друга в редакторе изображений GIMP 2.

Для обучения также были использованы маски лектора, полученные в результате обработки видеолекции моделью YOLO11 Small, для указания областей, которые необходимо восстанавливать. Структура датасета представлена на рис. 2, здесь представлена схема, отражающая иерархию каталогов датасета.

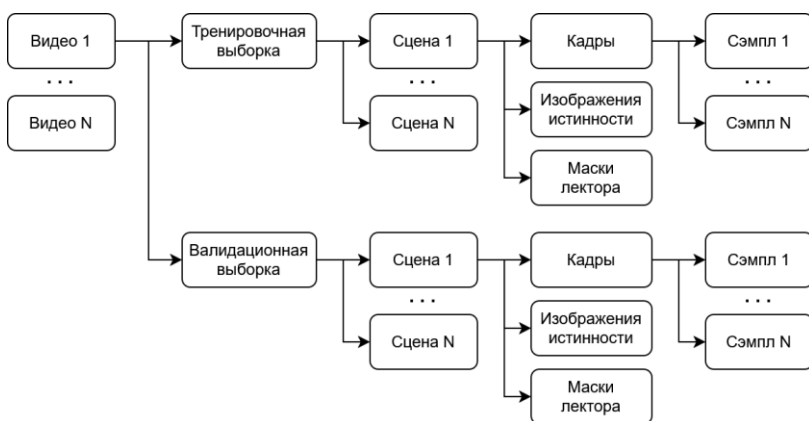


Рис. 2. Схема иерархии каталогов датасета

Помимо структуры датасета необходимо определить количественные и качественные характеристики, датасет состоит из 5 видео в каждом из которых 80 тренировочных сэмплов по 15 кадров и 16 валидационных сэмплов также по 15 кадров, если в сцене не хватает 15 кадров, то недостаточность компенсируются дополняющими изображениями (padding images), пустыми по содержанию. Разрешение изображений в датасете составляет 1920 на 1080 пикселей (исходное разрешение видеолекции). В итоге для каждого видео насчитывается по 1300 кадров.

Разработка архитектуры модели, обучение и инференс модели

Модель написана на языке Python в качестве библиотеки машинного обучения была использована Tensorflow-Keras. Для работы с векторами многомерными матрицами была использована библиотека NumPy. Для визуализации полученных данных использовалась библиотека Matplotlib.

Архитектура модели представляет собой нейронную сеть, сочетающую сверточные и рекуррентные слои с долгой краткосрочной памятью (LSTM). Применение LSTM-сетей обосновано тем, что они хорошо справляются с темпоральной обработкой данных [O'Donncha et al., 2022], [Pham, 2021].

Формально темпоральное объединение кадров можно определить следующим образом, пусть существует:

- $\{F_0, F_1, \dots, F_n\}, F_i \in \{0, 1\}^{H \times W}$ – последовательность бинаризованных кадров от 0 до n , и каждый элемент этой последовательности, кадр F_i это бинарная матрица размером $H \times W$, т.е. кадр из чёрно-белых пикселей.
- $\{M_0, M_1, \dots, M_n\}, M_i \in \{0, 1\}^{H \times W}$ – последовательность бинаризованных масок от 0 до n , и каждый элемент этой последовательности, кадр M_i это бинарная матрица размером $H \times W$, т.е. маска из чёрно-белых пикселей.

Тогда:

$$F_{\text{объединенное}}(x, y) = \max_{t \in \{0, \dots, n\}} \{F_t(x, y) | M_t(x, y) = 0\},$$

где $x \in [0, W]$, $y \in [0, H]$ – координаты пикселя в кадре.

Для обработки пространственных данных активно применяются сверточные нейронные сети, поскольку они хорошо справляются с этим классом задач [Li et al., 2020], [Melnikov et al., 2017]. Алгоритм объединения имеет особенность в виде двухступенчатого способа объединения кадров, то есть, на первом этапе кадры объединяются на уровне сэмплов, а на втором этапе на уровне сцен видеолекции, архитектура модели отражена на рис. 3. Алгоритм двухступенчатого объединения также был выбран для избежания ошибки out of memory.

Основой архитектуры выступает слой ConvLSTM2D с 16 фильтрами и ядром размером (3, 3), данный слой объединяет возможности сверточных нейронных сетей и LSTM-сетей. Для введения нелинейности, применяется функция активации relu.

За этим слоем следует Dropout с коэффициентом 0.3, который, случайным образом, обнуляет часть выходных данных во время обучения, для предотвращения переобучения, и улучшения обобщающей способности модели.

Финальный этап обработки осуществляется с помощью слоя Conv2D, содержащего один фильтр с ядром (3, 3). Этот слой выполняет заключительную свертку для предсказания маски, сохраняя размерность благодаря параметру padding='same'. Активационная функция sigmoid нормализу-

ет выходные значения в диапазон от 0 до 1. Выходной слой использует тип данных float32. Модель компилируется (обучается) с оптимизатором Adam (скорость обучения 0.0001, clipnorm=1.0 для стабилизации градиентов), функцией потерь binary_crossentropy и метриками accuracy.

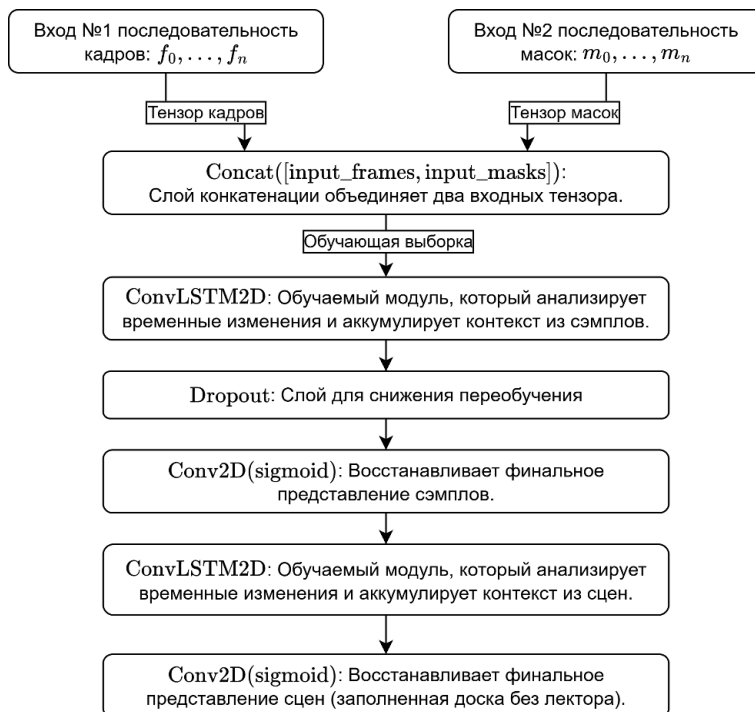


Рис. 3. Архитектура нейронной сети

Для предотвращения переобучения, применяется функция EarlyStopping, которая, отслеживает валидационную потерю, и восстанавливает лучшие веса после 10 эпох без улучшения. Количество эпох обучения 50. Для ускорения инференса модели, использовалась библиотека ONNXRuntime, так как, эта библиотека имеет встроенный инструментарий оптимизации [Someki et al., 2022], [ONNX Runtime, 2025].

В качестве среды разработки была выбрана среда Kaggle Notebook основанная на Jupiter Notebook, модель обучалась на двух графических ускорителях серии NVidia Tesla T4 с 16 гигабайт GDDR6 видеопамати для каждого ускорителя, обучение производилось с использованием стратегии MirroredStrategy, для использования двух графических ускорителей.

По результатам обучения имеем следующие показатели по метрикам loss и ассигасу, лучшее значение loss 0.0773, лучшее значение ассигасу 0.7711 для тестовой выборки, лучшее значение loss 0.0787, лучшее значение ассигасу 0.773 для тестовой выборки. Более подробно с процессом обучения можно ознакомиться на рис. 4 и 5, отражающие графики обучения.

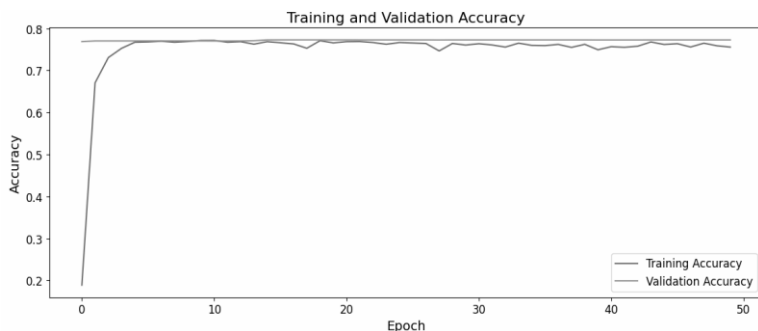


Рис. 4. График Ассигасу обучения модели

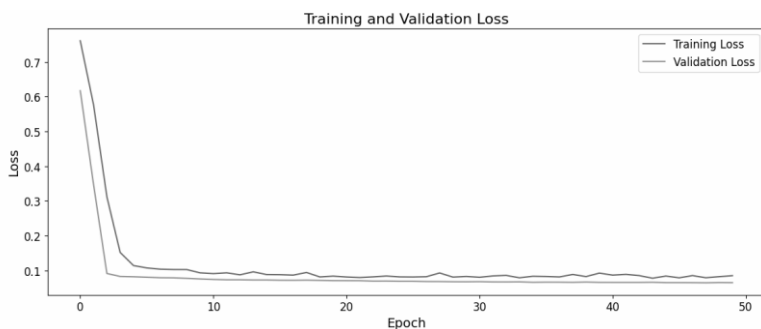


Рис. 5. График Loss обучения модели

Качественная метрика восстановления изображения PSNR

Для оценки качества реконструкции изображения, принято использовать меру пикового отношения сигнала к шуму (англ. peak signal-to-noise ratio или PSNR). Данная мера характеризует соотношение между максимумом возможного значения сигнала и мощностью шума, искажающего значения сигнала [Shen et al., 2024], [Keleş et al., 2021]. Измеряется в децибелах, и характеризует разницу оригинального изображения и изображения полученного в ходе генерации при инференсе модели.

Определяется по следующей формуле:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right),$$

где MAX_I – максимальное значение пикселя (например, 255 для 8-битных изображений);

MSE (*Mean Squared Error*) – среднеквадратичная ошибка;

MSE определяется по следующей формуле:

$$MSE = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n [I(i, j) - K(i, j)]^2,$$

где I – оригинальное изображение;

K – сгенерированное изображение;

m и n – размерность изображения.

Хорошим результатом считается если мера PSNR полученного изображения достигла диапазона от 32 до 40 децибел, если выше, то считается, что изображение не отличимо от оригинала [Huynh-Thu et al., 2008], [Lim et al., 2017]. В ходе эксперимента, мера PSNR достигла 35,011 децибел, что является хорошим показателем восстановления изображений. В контексте бинаризованных изображений, мера PSNR равная 35,011 характеризует практически полное совпадение положения пикселей [Корешев и др., 2020].

В качестве Baseline взята модель из статьи [Park et al., 2020], в ходе исследования, модель BVDNet достигла значения по мере PSNR 34,7055, что также является хорошим результатом. В исследовании стояла задача восстановления цветных кадров видео.

Если сравнивать изображения визуально (рис. 6), то можно сделать следующие выводы. При создании Ground Truth изображений, были пропущены некоторые части математических выражений (выделено сплошной рамкой), но модель восстановила не только пропущенные части, но и то, что изначально не было указано в эталонах. Также, на восстановленных изображениях, присутствуют артефакты оставшиеся от процесса вычитания маски, с той лишь разницей, что эти артефакты на оригинальных изображениях носят четкий характер с резкими границами, а в восстановленном изображении, эти артефакты размытые (выделено пунктирной рамкой). Также, из особенностей реконструкции, можно выделить многократное наложение водяного знака онлайн школы (выделено штриховой рамкой), и утолщение контуров надписей.

Изображение эталон

$$\textcircled{1} \lim_{x \rightarrow 2} \frac{x^3 - 2x^2 + x - 2}{x^3 + x^2 - 8x + 4} = \left[\frac{0}{0} \right] = \lim_{x \rightarrow 2} \frac{(x^3 + 1)(x - 2)}{(x - 2)(x^2 + 2x + 4)} =$$

$$\left[\frac{\infty}{\infty} \right], [0 \cdot \infty] \neq x(x^2 + 1) - 2(x^2 + 1) =$$

$$[\infty - \infty], [1 \cdot \infty] \neq x^3 + x - 2x^2 - 2 =$$

$$[0^0], [\infty^0] \neq \frac{x^3 + x - 2x^2 - 2}{x^2 + 2x + 4} \Big|_{x=2} = \frac{8 + 2 - 8 - 2}{4 + 4 + 4} = \frac{0}{12} = 0$$



Реконструированное изображение

$$\textcircled{1} \lim_{x \rightarrow 2} \frac{x^3 - 2x^2 + x - 2}{x^3 + x^2 - 8x + 4} = \left[\frac{0}{0} \right] = \lim_{x \rightarrow 2} \frac{(x^3 + 1)(x - 2)}{(x - 2)(x^2 + 2x + 4)} =$$

$$\left[\frac{\infty}{\infty} \right], [0 \cdot \infty] \neq x(x^2 + 1) - 2(x^2 + 1) =$$

$$[\infty - \infty], [1 \cdot \infty] \neq x^3 + x - 2x^2 - 2 =$$

$$[0^0], [\infty^0] \neq \frac{x^3 + x - 2x^2 - 2}{x^2 + 2x + 4} \Big|_{x=2} = \frac{8 + 2 - 8 - 2}{4 + 4 + 4} = \frac{0}{12} = 0$$



Источник: видео TutorOnline (<https://youtu.be/19TSR9mwxQ>), авторские права принадлежат TutorOnline

Рис. 6. Результаты реконструкции изображения

Заключение

В данном исследовании, была разработана, и реализована модель темпорального объединения кадров видеолекций с целью реконструкции содержимого доски, скрываемого лектором. Модель является частью пайплайна агента, который восстанавливает полный образ доски в видеолекции, с последующей обработкой агентом оптического распознавания символов. Метрики обучения показали хорошую сходимость. Лучшее значение loss – 0.0773 (тренировочная выборка), 0.0787 (тестовая выборка). Лучшее значение ассигасу – 0.7711 (тренировочная выборка), 0.773 (тестовая выборка). Качество восстановления (PSNR = 35.011 дБ) ненамного превысило baseline-модель BVDNet (34.7055 дБ), что подтверждает эффективность предложенного подхода. Визуальный анализ показал, что модель не только восстанавливает пропущенные в Ground Truth элементы, но и корректно заполняет области, изначально скрытые лектором, несмотря на наличие незначительных артефактов. Для дальнейшего повышения метрик ассигасу, loss, планируется увеличение обучающей выборки, и применение аугментации данных. Для повышения меры PSNR, планируется улучшение алгоритма бинаризации и других алгоритмов предобработки данных.

Список литературы

- [Исмагулов, 2024] Исмагулов М.Е. Абстрагирование данных в контексте машинного обучения: методы, применения и перспективы // Информационные технологии и математическое моделирование: Труды XXIV Международной конференции ITMM-2024 (Томск, 2024). – Томск: Изд-во Томского гос. ун-та, 2024. – С. 605-607. – URL: https://www.researchgate.net/publication/391833448_1_Conference_proceedings_with_your_article_Ismagulov_M_E_Methods_and_Algorithms_for_Multimodal_Conversion_of_Video_Lectures (дата обращения: 17.05.2025).
- [Корешев и др., 2020] Корешев С.Н., Старовойтов С.О., Смородинов Д.С., Фролова М.А. Методы оценки качества изображений бинарных объектов, восстановленных с помощью синтезированных голограмм-проекторов // Науч.-техн. вестн. инф.-технол., мех. и оптики. – 2020. – Т. 20, № 3. – С. 327-334.
- [Feichtenhofer et al., 2018] Feichtenhofer C., Fan H., Malik J., He K. SlowFast Networks for Video Recognition [Электронный ресурс] // arXiv. – 2018. – arXiv:1812.03982. – URL: <https://arxiv.org/abs/1812.03982> (дата обращения: 18.05.2025). – DOI: 10.48550/arXiv.1812.03982.
- [Huynh-Thu Q., et al., 2008] Huynh-Thu Q., Ghanbari M. Scope of validity of PSNR in image/video quality assessment // Electronics Letters. – 2008. – Vol. 44, No. 13. – P. 800-801.
- [Keleş et al., 2021] Keleş O., Yılmaz M.A., Tekalp A.M., Korkmaz C., Dogan Z. On the Computation of PSNR for a Set of Images or Video [Электронный ресурс] // arXiv. – 2021. – arXiv:2104.14868. – URL: <https://arxiv.org/abs/2104.14868> (дата обращения: 19.05.2025). – DOI: 10.48550/arXiv.2104.14868.
- [Kumar, P., et al., 2020] Kumar, P., Ambati, R., & Raj, L. An Efficient Scene Content-Based Indexing and Retrieval on Video Lectures. // Advances in Intelligent Systems and Computing, vol 1171. – 2020. – P. 521-534.
- [Li, Z., et al., 2020] Li Z., Liu F., Yang W., Peng S., & Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects // IEEE Transactions on Neural Networks and Learning Systems. – 2020. – 33. – P. 6999-7019.
- [Lim B., et al., 2017] Lim B., Son S., Kim H., Nah S., Mu Lee K. Enhanced Deep Residual Networks for Single Image Super-Resolution // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2017. – P. 136-144.
- [Melnikov et al., 2017] Melnikov A., Sochenkova A., Sochenkov I., Makovetskii A., Vokhmintsev A. Convolutional neural networks and face recognition task [Электронный ресурс] // Proc. SPIE Applications of Digital Image Processing XL. – 2017. – Vol. 10396. – P. 103962L. – URL: <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10396/103962L/Convolutional-neural-networks-and-face-recognition-task/10.1117/12.2273624.short> (дата обращения: 18.05.2025).
- [ONNX Runtime, 2025] ONNX Runtime Documentation [Электронный ресурс]. – URL: <https://onnxruntime.ai/docs/> (дата обращения: 19.05.2025).
- [O'Donncha et al., 2022] O'Donncha F., Hu Y., Palmes P., Burke M., Filgueira R., Grant J. A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales // Ecological Informatics. – 2022. – Vol. 69. – P. 101687. – ISSN 1574-9541. – URL: <https://www.sciencedirect.com/science/article/pii/S1574954122001376> (дата обращения: 19.05.2025). – DOI: 10.1016/j.ecoinf.2022.101687.

- [Park et al., 2020] Recurrent Temporal Aggregation Framework for Deep Video Inpainting // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2020. – Vol. 42, No. 10. – P. 2547-2561. – URL: https://joonyoung-cv.github.io/assets/paper/20_tpami_recurrent_temporal.pdf (дата обращения: 18.05.2025).
- [Pham, 2021] Pham T.D. Time–frequency time–space LSTM for robust classification of physiological signals // Scientific Reports. – 2021. – Vol. 11. – P. 6936. – URL: <https://www.nature.com/articles/s41598-021-86432-7> (дата обращения: 19.05.2025). – DOI: 10.1038/s41598-021-86432-7.
- [Shen, W., et al., 2024] Shen W., Tian X., Zeng D., & Zhang Y. Multi-scale image compression and reconstruction algorithm for structural health monitoring system // Engineering Structures. – 2024.
- [Someki et al., 2022] Someki M., Higuchi Y., Hayashi T., & Watanabe S. ESPnet-ONNX: Bridging a Gap Between Research and Production // Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). – 2022. – P. 420-427.
- [Urala Kota et al., 2018] Urala Kota B., Davila K., Stone A., Setlur S., Govindaraju V. Automated Detection of Handwritten Whiteboard Content in Lecture Videos for Summarization // Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). – 2018. – P. 19-24. – DOI: 10.1109/ICFHR-2018.2018.00013.
- [Urala Kota et al., 2019] Urala Kota B., Davila K., Stone A., Setlur S., Govindaraju V. Generalized framework for summarization of fixed-camera lecture videos by detecting and binarizing handwritten content // International Journal on Document Analysis and Recognition (IJ DAR). – 2019. – Vol. 22. – P. 221-233. – DOI: 10.1007/s10032-019-00327-y.
- [Wang et al., 2017] Wang X., Girshick R., Gupta A., He K. Non-local Neural Networks [Электронный ресурс] // arXiv. – 2017. – arXiv:1711.07971. – URL: <https://arxiv.org/abs/1711.07971> (дата обращения: 18.05.2025). – DOI: 10.48550/arXiv.1711.07971.
- [Wang et al., 2022] Wang T., He M., Shen K., Liu W., Tian C. Learned regularization for image reconstruction in sparse-view photoacoustic tomography // Biomedical Optics Express. – 2022. – Vol. 13, No. 11. – P. 5721-5737. – DOI: 10.1364/BOE.469460.
- [Zhang et al., 2024] Zhang, T., He, X., Teng, Q., Cheng, J., & Ren, C. Spatio-Temporal Adaptive Weighted Fusion Network for Compressed Video Quality Enhancement. // IEEE Transactions on Circuits and Systems II: Express Briefs. – 2024. – 71. – P. 5064-5068.