



**МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ,
КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ/MATHEMATICAL SOFTWARE FOR COMPUTERS,
COMPLEXES AND COMPUTER NETWORKS**

DOI: <https://doi.org/10.60797/IRJ.2026.171.59> EDN: FYPVDZ

**АРХИТЕКТУРНОЕ ОБОСНОВАНИЕ И ПИЛОТНАЯ ОЦЕНКА МОДИФИЦИРОВАННОГО HTR-МОДУЛЯ
ДЛЯ РАСПОЗНАВАНИЯ ПЛОТНЫХ РУКОПИСНЫХ МЕДИЦИНСКИХ ПОЛЕЙ**

Научная статья

Ильичев В.Ю.^{1,*}, Русин Я.Ю.², Федоров В.О.³

¹ORCID : 0000-0003-1017-5544;

^{1, 2, 3}Московский государственный технический университет имени Н.Э. Баумана, Калуга, Российская Федерация

* Корреспондирующий автор (patrol8[at]yandex.ru)

Предложена: 29.07.2026; Принята: 04.09.2026; Опубликовано: 17.09.2026

Аннотация

Актуальность работы обусловлена высокой трудоёмкостью ручного ввода данных из рукописных карт вызова бригад скорой медицинской помощи (СМП): при суточном объёме в несколько тысяч вызовов на обработку одного документа уходит от нескольких до десятков минут. Существующие универсальные решения оптического распознавания символов недостаточно эффективны для медицинских рукописных текстов из-за специфической лексики, аббревиатур, латинских терминов и неоднородного качества изображений.

Целью работы является обоснование и пилотная оценка гибридной нейросетевой архитектуры ResNet18 TE CTC (CNN блок ResNet18, Transformer Encoder, декодер CTC) в качестве HTR модуля системы MedForm HTR. Ключевым этапом выступает аналитическое и экспериментальное подтверждение целесообразности модификации CNN блока на базе ResNet18 под CTC HTR, поскольку именно он во многом определяет точность распознавания.

В ходе исследований подтверждён доменный сдвиг между эталонными данными и картами СМП, показана необходимость доменного дообучения и продемонстрировано преимущество предложенной архитектуры для распознавания плотных гетерогенных рукописных полей. Выявлены факторы, ограничивающие качество распознавания; их устранение предполагает дальнейшие работы по предподготовке документации, обучению HTR модуля и интеграции других компонентов MedForm HTR (в том числе медицинского словаря), реализуя систему как конвейер обработки медицинских документов.

Ключевые слова: HTR, MedForm HTR, ResNet18, Transformer Encoder, CTC, доменная адаптация, карты вызова СМП.

**ARCHITECTURAL BASIS AND PILOT EVALUATION OF A MODIFIED HTR MODULE FOR THE
RECOGNITION OF DENSELY HANDWRITTEN MEDICAL DOCUMENTS**

Research article

Ilichev V.Y.^{1,*}, Rusin Y.Y.², Fedorov V.O.³

¹ORCID : 0000-0003-1017-5544;

^{1, 2, 3}Bauman Moscow State Technical University, Kaluga, Russian Federation

* Corresponding author (patrol8[at]yandex.ru)

Suggested: 29.07.2026; Accepted: 04.09.2026; Published: 17.09.2026

Abstract

The relevance of the work is due to the high labour intensity involved in manually entering data from handwritten emergency medical service (EMS) call records: with a daily volume of several thousand calls, it takes anywhere from a few minutes to tens of minutes to process a single document. Existing universal optical character recognition (OCR) solutions are not sufficiently effective for handwritten medical texts due to their specific vocabulary, abbreviations, Latin terms and inconsistent image quality.

The aim of the work is to substantiate and conduct a pilot evaluation of the ResNet18 TE CTC hybrid neural network architecture (ResNet18 CNN block, Transformer Encoder, CTC decoder) as the HTR module of the MedForm HTR system. A key stage is the analytical and experimental validation of the feasibility of modifying the ResNet18-based CNN block for the CTC HTR, as it is this block that largely determines recognition accuracy.

The research confirmed a domain shift between the reference data and the EMS records, highlighted the need for domain-specific fine-tuning, and demonstrated the advantages of the proposed architecture for recognising dense, heterogeneous handwritten text blocks. Factors limiting recognition quality have been identified; addressing these will require further work on document pre-processing, training the HTR module and integrating other MedForm HTR components (including the medical dictionary), thereby implementing the system as a medical document processing pipeline.

Keywords: HTR, MedForm HTR, ResNet18, Transformer Encoder, CTC, domain adaptation, EMS call records.

Введение

Существующие универсальные решения оптического распознавания символов демонстрируют недостаточную точность при работе с рукописными медицинскими текстами: специфическая профессиональная лексика, латинские термины, аббревиатуры и неоднородное качество исходных изображений существенно снижают качество распознавания. Отсутствие предметно-адаптированных инструментов вынуждает операторов выполнять фактически полный ручной ввод данных, не получая от систем распознавания практической пользы.

Актуальность настоящей работы определяется необходимостью разработки специализированной интеллектуальной системы MedForm-HTR (рисунок 1) в виде конвейера, ориентированного на распознавание рукописных медицинских документов с применением технологий HTR (Handwritten Text Recognition). Приоритетной предметной областью MedForm-HTR являются карты вызовов бригады скорой медицинской помощи (СМП) с последующим применением системы для обработки других медицинских форм. При суточном объёме в 1,5–2,0 тыс. вызовов бригад скорой медицинской помощи для медицинских подразделений среднего города, трудозатраты на обработку карт вызова СМП очень значительные, учитывая, что на ввод одного документа в реестр приходится от нескольких до десятков минут в зависимости от объёма и разборчивости почерка.

Целью работы является обоснование и пилотная оценка нейросетевой архитектуры CNN — Transformer Encoder — CTC в качестве кандидатной линии модуля MedForm-HTR для распознавания рукописных полей, включая карты вызова СМП. Ключевым этапом в работе выступает аналитическое и экспериментальное подтверждение адаптации CNN-блока на базе ResNet18 под CTC-HTR, так как именно этот блок во многом определяет уровень точности и практическую применимость архитектуры.



Рисунок 1 - Диаграмма контекста разрабатываемого конвейера MedForm-HTR

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.1>

Основная часть

Карта вызовов СМП представляет собой учетную форму №110/у, утвержденную Приказом Министерства здравоохранения и социального развития Российской Федерации от 02.12.2009 №942 (рисунок 2). Документ содержит десятки структурированных полей, в том числе данные о пациенте, время и адрес вызова, диагноз по МКБ-10, применённые лекарственные средства и дозировки, а также отметки о транспортировке и передаче пациента. Карта вызова СМП является типичным представителем широкого класса первичной медицинской документации и характеризует такие ключевые проблемы предметной области, как высокая вариативность почерка, смешанный текст (прописной и печатный), содержание на нескольких языках, а также необходимость обеспечения юридической и клинической значимости данных.

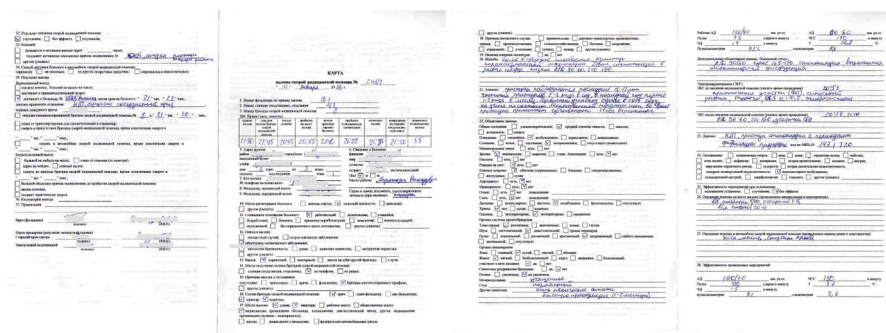


Рисунок 2 - Обезличенная карта вызовов скорой медицинской помощи
DOI: <https://doi.org/10.60797/IRJ.2026.171.59.2>

Общеизвестно, что в практике машинного обучения практически невозможно обучение моделей без качественно подготовленных данных, соответственно допущенные при этом ошибки влияют на эффективность функционирования самих алгоритмов. Необходимо отметить, что при тиражировании карт вызовов СМП используется формат бумаги А5, из-за чего высота поля одной строки составляет всего 3 мм, что обуславливает чрезмерное уплотнение рукописного текста и, соответственно, снижает качество распознавания.

Карта вызовов СМП состоит из 37 полей, около 70% которых заполняются рукописным текстом, остальные поля представлены чекбоксами и стандартными подчеркиваниями (таблица 1). В MedForm-HTR чекбоксы могут обрабатываться отдельным простым классификатором OMR (Optical Mark Recognition), в связи с чем они были исключены из задачи HTR.

Таблица 1 - Характеристика репрезентативных полей карты вызова СМП

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.3>

Номер поля	Название поля по форме № 110/у	Категория данных для HTR	Особенности визуального и текстового распределения
Поле 20	Жалобы	Свободный текст	Рукописный текст (до 6 строк, до 30–40 символов в строке). Высокая вариативность.
Поле 21	Анамнез	Свободный текст	Сверхплотный рукописный текст (до 6, до 30–40 символов в строке), обилие латинских терминов и сокращений (например, «ФП» — фибрилляция предсердий, «Luxatio» — вывих).
Поле 22/28	Объективные данные	Клинические метрики, смешанный тип	Рукописный текст, строгие паттерны: цифры, дроби (130/90), единицы измерения (до 25 строк, до 10–30 символов в строке). Критически важна точность распознавания знаков.
Поле 33/35	Результат выезда / Направление	Смешанный тип	Чекбоксы + короткие рукописные вставки названия больниц, отделений, например, «КОКБ, отд. сложных нарушений ритма сердца и электрокардиостимуляц

Номер поля	Название поля по форме № 110/y	Категория данных для HTR	Особенности визуального и текстового распределения
			ии», до 12 строк, 10–20 символов в строке).

HTR-модель должна быть устойчивой к вариативности исходных изображений: различным почеркам врачей, бледным или размытым сканам, наклону текста, пересечению рукописи с линиями бланка, исправлениям. Следовательно, выбранная архитектура должна допускать применение процедур доменного дообучения, регуляризации и аугментации данных, направленных на повышение устойчивости модели к реальным условиям эксплуатации. Кроме этого, модель должна обеспечивать контролируемый характер распознавания, особенно для значимых полей, где ошибка в одном символе или цифре может приводить к искажению клинического смысла. Поэтому результаты работы HTR-модуля должны оцениваться не только по средним текстовым метрикам, но и по метрикам полного совпадения поля, точности числовых значений и объёму последующей операторской правки.

Эволюция современных методов HTR демонстрирует переход от изолированного посимвольного анализа к сквозным (End-to-End) гибридным глубоким нейросетевым архитектурам [1], [2], [3], [4]. При проектировании системы распознавания для гетерогенных документов, критически важным является выбор оптимального баланса между способностью модели извлекать мелкозернистые визуальные признаки и её вычислительной сложностью при обучении на специфических медицинских датасетах ограниченного объема.

С учетом вышеназванных требований были рассмотрены основные гибридные нейросетевые архитектуры, имеющие потенциальную возможность работать со смешанными гетерогенными текстами (таблица 2).

Таблица 2 - Основные гибридные нейросетевые архитектуры, используемые в HTR

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.4>

Архитектура	Извлечение признаков	Блок последовательности/контекста	Механизм преобразования признаков в текст	Уровень распознавания	Применение	Статус для MedForm-HTR
CNN - RNN - CTC	VGG-like CNN, ResNet, другие CNN	LSTM, BiLSTM, GRU, BiGRU	CTC + greedy/beam decoding	слово/строка, поле	исторические строки, анкеты, отдельные поля	базовый уровень
CNN - Transformer Encoder - CTC	CNN, ResNet, ConvNeXt	Transformer Encoder	CTC + greedy/beam decoding	слово/строка, поле	HTR при ограниченной разметке, анкеты, медицинские поля	основной кандидат для исследования
Encoder-decoder/TrOCR	ViT / BEiT-like image encoder	cross-attention между image encoder и text decoder	авторегрессионный Transformer decoder	слово/строка, поле	печатный, рукописный текст	сильный, базовый уровень
VAN/page-level attention models	CNN encoder	vertical attention / implicit line segmentation	attention-based decoder	абзац/страница	рукописные абзацы, исторические документы	перспективная модель для распознавания рукописного текста на уровне абзаца и страницы

На основании проведенного анализа в настоящей работе в качестве базовой архитектурной линии рассматривается гибридная модель CNN-Transformer Encoder-CTC. Семейство CNN-Transformer Encoder-CTC является перспективным для HTR, поскольку объединяет локальную визуальную экстракцию CNN, глобальное контекстное моделирование Transformer Encoder (TE) и CTC-обучение без посимвольного выравнивания [5], [6], [7].

Свёрточная часть архитектуры (CNN-фронтенд) используется для извлечения локальных визуальных признаков рукописного текста: штрихов, петель, соединений букв, наклона и толщины линии. TE (первая ключевая часть

архитектуры трансформера) применяется для контекстного моделирования полученной последовательности признаков вдоль оси письма. В отличие от рекуррентных блоков, TE позволяет учитывать более широкий контекст и выполнять обработку последовательности параллельно [8], [10], [11], [12]. CTC (Connectionist Temporal Classification, метод обучения для работы с последовательными данными), как часть нейросетевой архитектуры, используется для обучения модели по парам «изображение — текстовая транскрипция» без необходимости задавать координаты каждого символа. В качестве CNN-backbone (backbone — «основная сеть» для извлечения признаков) был выбран ResNet-18 (Residual Network), который отличается практической реализуемостью и архитектурной прозрачностью, что позволяет исследовать влияние горизонтального понижения разрешения (downsampling) на длину последовательности, обрабатываемой модулем CTC. Актуальность данного исследования обусловлена тем, что прямое использование классификационных CNN-backbone (включая ResNet18), разработанных для задач классификации изображений, не является оптимальным решением для HTR. Архитектуры такого типа изначально создавались с учётом существенного пространственного сжатия признаковой карты, что может негативно влиять на качество распознавания текста. В задаче CTC-HTR горизонтальная ось выходной карты признаков интерпретируется как временная ось последовательности, поэтому чрезмерное сжатие ширины может приводить к уменьшению числа временных шагов T и ухудшению условий CTC-выравнивания, обуславливая снижение точности распознавания.

Таким образом, выбранная архитектурная линия сочетает три свойства, важные для MedForm-HTR: локальную чувствительность CNN к графемным признакам, контекстное моделирование TE и возможность обучения без посимвольного выравнивания за счёт CTC. В статье исследуется архитектура ResNet18-TE-CTC с модификацией горизонтального downsampling в ResNet18. Такой дизайн позволяет определить влияние исследуемого фактора на CTC-распознавание.

Учитывая вышесказанное, в работе рассматриваются две конфигурации одной архитектурной линии: стандартная ResNet18-TE-CTC, используемая как контрольный вариант, и модифицированная ResNet18-TE-CTC, в которой в третьем сверточном слое (layer3) и в четвертом сверточном слое (layer4) горизонтальный шаг свертки (stride) заменяется с 2 на 1. Сохранение горизонтальной структуры признаковой карты является распространённым принципом проектирования специализированных HTR-архитектур [13], [14], [15], [16]. Однако при применении стандартных классификационных CNN-backbone, таких как ResNet18, данное свойство не обеспечивается автоматически. В связи с этим в настоящей работе исследуется возможность адаптации ResNet18-фронтенда к задаче HTR путём изменения горизонтального stride в поздних residual-блоках (остаточные блоки) в рамках конкретной архитектурной модели с проверкой и оценкой данной модификации на конкретных рукописных кириллических датасетах, включая гетерогенные, сверхплотные медицинские документы. Поэтому данная модификация не рассматривается как изолированный инженерный приём, а проверяется как аналитически обоснованная гипотеза о влиянии сохранения горизонтального разрешения CNN-фронтенда на качество CTC-распознавания.

Сравнение различных CNN-backbone, включая, например, VGG-like, VGG-11, VGG-19, ConvNeXt и TrOCR, а также VAN выходит за рамки данной работы. В настоящей статье рассматривается архитектурная линия ResNet18-TE-CTC с экспериментальной вариацией геометрии горизонтального downsampling.

Специфика распознавания рукописных документов накладывает жесткие ограничения на геометрию извлекаемых признаков. Это приводит к повышенной плотности информационных элементов на единицу длины горизонтальной оси W. Соответственно в классической архитектуре ResNet18 последовательное применение слоев Layer 3 и Layer 4 с двумерным шагом свертки stride = (2,2) приводит к изотропному сжатию пространственных размерностей входного тензора признаков:

$$X \in R^{C \times H \times W}$$

обуславливая уменьшение его разрешения в k раз по обеим осям с генерацией тензора

$$V_{std} \in R^{C' \times \frac{H}{K} \times \frac{W}{K}}$$

Для плотного и вариативного рукописного текста свертка с параметром stride = 2 по горизонтальной оси может быть критичной, обуславливая избыточную субдискретизацию графем и связей с их необратимым смещением в рамках одного и того же вектора признаков. Учитывая возможность такого отклика в предлагаемой архитектуре MedForm-HTR в модуле ResNet18 Layer 3 и Layer 4 были модифицированы путем перехода к анизотропному шагу свертки: stride = (2,1). Это позволяет сохранить исходную дискретизацию по горизонтальной оси, формируя тензор признаков:

$$V_{mod} \in R^{C' \times \frac{H}{K} \times W}$$

Последующая операция вертикального пулинга трансформирует данный тензор в последовательность токенов для модуля TE, которые, в свою очередь, должны влиять на последовательность временных шагов T, увеличивая пространственную дискретизацию вдоль оси написания текста и, соответственно, влиять на качество последующего

СТС-распознавания. Данное влияние проявляется не напрямую, а через изменение геометрии карты признаков, поступающей из ResNet18 в ТЕ и далее в СТС-блок.

Действительно, пусть

W_{in} — исходная ширина изображения рукописного поля;

r_x — суммарный коэффициент горизонтального сжатия сверточного фронтеда;

T — длина последовательности визуальных признаков, поступающей в ТЕ и СТС-блок.

Тогда приближенно можно записать:

$$T \approx \frac{W_{in}}{r_x}$$

То есть чем больше суммарное горизонтальное сжатие r_x , тем меньше временных шагов T получает СТС-блок. В стандартной конфигурации ResNet18 поздние блоки Layer 3 и Layer 4 уменьшают горизонтальное разрешение карты признаков. При замене горизонтального stride в этих блоках с 2 на 1 суммарное горизонтальное сжатие уменьшается, поэтому длина выходной последовательности признаков возрастает.

Это можно записать так:

$$T_{mod} > T_{std}$$

где:

T_{std} — длина последовательности признаков в стандартной архитектуре ResNet18;

T_{mod} — длина последовательности признаков после модификации stride в Layer 3 и Layer 4.

Для СТС-декодирования увеличение T имеет принципиальное значение. Если целевая транскрипция содержит $|Y|$ символов, то отношение количества символов к количеству временных шагов можно рассматривать как плотность символов на один временной шаг:

$$\rho = \frac{|Y|}{T},$$

где:

ρ — плотность символов на временной шаг;

$|Y|$ — длина правильной текстовой строки в символах;

T — число временных шагов на выходе сверточного фронтеда.

При плотном рукописном заполнении медицинских полей значение ρ в стандартной архитектуре может становиться слишком высоким. Это означает, что на один временной шаг приходится слишком большая часть текстовой информации. В результате возрастает риск слияния соседних графем, пропуска символов и ошибок СТС-выравнивания.

После модификации stride длина последовательности увеличивается:

$$T_{mod} > T_{std}$$

Следовательно, плотность символов уменьшается:

$$\rho_{mod} < \rho_{std}$$

Иными словами, модифицированная архитектура предоставляет СТС-блоку больше временных позиций для размещения символов целевой строки. Это расширяет пространство допустимых СТС-выравниваний и создает более благоприятные условия для корректного сопоставления визуальных признаков с текстовой транскрипцией.

Дополнительно изменение stride влияет на геометрию рецептивного поля сверточного фронтеда. В стандартной конфигурации поздние блоки ResNet18 увеличивают не только размер рецептивного поля, но и эффективный шаг между соседними токенами по горизонтальной оси. В результате один визуальный токен может агрегировать информацию сразу от нескольких соседних рукописных символов. Для плотного медицинского почерка это повышает риск смешения графемных признаков.

В модифицированном варианте горизонтальный эффективный шаг уменьшается, поэтому признаки становятся более локализованными вдоль оси письма. ТЕ получает более детализированную последовательность токенов, а СТС-блок — больше возможностей для построения корректной траектории выравнивания между визуальными признаками и символами.

В обобщенном виде аналитическая цепочка влияния модификации stride может быть представлена следующим образом:

stride_x: 2 → 1 в Layer 3 и Layer 4
 ↓
 уменьшается горизонтальное сжатие g_x
 ↓
 увеличивается длина последовательности признаков T
 ↓
 уменьшается плотность символов ρ
 ↓
 CTC получает больше допустимых вариантов выравнивания
 ↓
 снижается риск слияния и пропуска графем
 ↓
 может снижаться CTC-Loss
 ↓
 может снижаться CER (Character Error Rate, доля ошибок на уровне символов) и WER (Word Error Rate, доля ошибок на уровне слов).

Данная зависимость не является строгой гарантией улучшения качества для любого входного изображения, поскольку итоговые метрики зависят от обученных параметров модели, качества разметки, вариативности почерка и условий обучения. Однако она формирует аналитически обоснованную гипотезу, которая проверена в данной работе.

Кроме изменения шага сверток была проведена также адаптация ResNet18 под задачу извлечения признаков (а не классификации целых изображений): из оригинальной архитектуры backbone были исключены слои глобального усреднения (global average pooling) и полносвязный слой (fully connected), что позволило передавать пространственные карты признаков напрямую в блок TE и декодировать их с помощью CTC-Loss. Модифицированная архитектура ResNet18-TE-CTC представлена на рисунке 3.

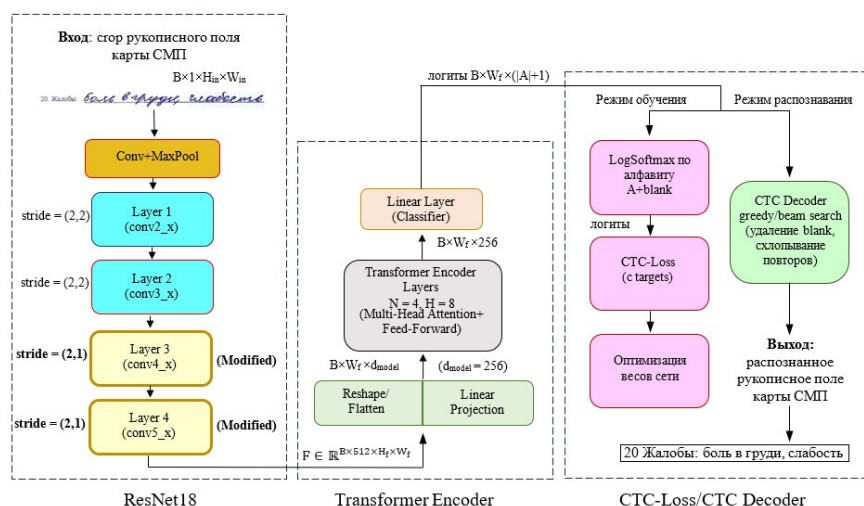


Рисунок 3 - Базовая модифицированная архитектура ResNet18-TE-CTC

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.5>

Экспериментальная часть работы включала два этапа. На первом этапе проводилась проверка стандартной и модифицированной конфигураций HTR-модуля ResNet18-TE-CTC на открытом кириллическом рукописном датасете HKR. Целью данного этапа была оценка влияния сохранения горизонтального разрешения в ResNet18-фронте на качество CTC-распознавания рукописного кириллического текста.

На втором этапе планировалась пилотная проверка точности распознавания лучшей конфигурации модели на обезличенных фрагментах полей карт вызова СМП. Данный этап рассматривался как предварительная оценка переносимости архитектурной модификации на целевой медицинский домен, характеризующийся высокой плотностью рукописного текста, медицинскими сокращениями и числовыми клиническими показателями.

В настоящем разделе основное внимание уделяется первому этапу эксперимента, поскольку именно на нём проверялась базовая архитектурная гипотеза: уменьшение горизонтального сжатия в ResNet18 должно увеличивать длину признаковой последовательности, поступающей в TE и CTC-блок, и тем самым улучшать условия CTC-выравнивания.

2.1. Первый этап исследований. Датасет HKR и подготовка данных

Для первого этапа эксперимента использовался датасет HKR (Handwritten Kazakh and Russian database), содержащий рукописные слова и строки на кириллице. Датасет HKR содержит более 1400 заполненных форм, примерно 63 000 предложений и свыше 715 000 символов, написанных около 200 разными людьми. Слова, формы и

строки датасета представляют смесь русского (около 95%) и казахского (около 5%) языков на кириллице. Поскольку целевой областью исследования является русскоязычная медицинская документация, из набора данных были исключены образцы, содержащие специфические казахские символы. Это позволило сформировать русскоязычный кириллический поднабор, более близкий к будущей задаче распознавания карт вызова СМП (таблица 3). В эксперименте использовались строки, содержащие отдельные словосочетания (word-level) датасета НКР. Такой выбор позволил сосредоточиться именно на задаче НТР-распознавания рукописного фрагмента, не смешивая её с отдельной задачей сегментации строк, полей или страниц. Полные изображения форм НКР на данном этапе не использовались.

Таблица 3 - Количественное распределение датасета НКР по выборкам

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.6>

Режим	Тренировочная выборка	Валидационная выборка	Тестовая выборка
Обучение	50000	6000	6000

Для всех запусков эксперимента применяли единое разбиение датасета на обучающую, валидационную и тестовую выборки. Это было необходимо для корректного сравнения стандартной и модифицированной конфигураций модели. Обучающая выборка применялась для оптимизации параметров модели, валидационная — для контроля сходимости и выбора лучшего сохраненного состояния модели (checkpoint), тестовая — для финальной оценки качества распознавания.

Предварительная обработка изображений включала перевод в полутоновый формат, масштабирование по высоте с сохранением аспектного соотношения, дополнение до унифицированного размера (padding) по ширине и нормализацию значений пикселей. Для совместимости с ImageNet-инициализированным ResNet18 одноканальные изображения при необходимости дублировались в три канала. Аугментации применялись только к обучающей выборке и не использовались на validation/test, чтобы не исказить оценку качества.

В качестве аугментаций использовались слабые преобразования, сохраняющие читаемость текста: аффинные преобразования, изменение яркости и контраста, слабое размытие, гауссовский шум и эластическая деформация. Все аугментации выполнялись в режиме «на лету» (on-the-fly) при формировании обучающих батчей (рисунок 4).

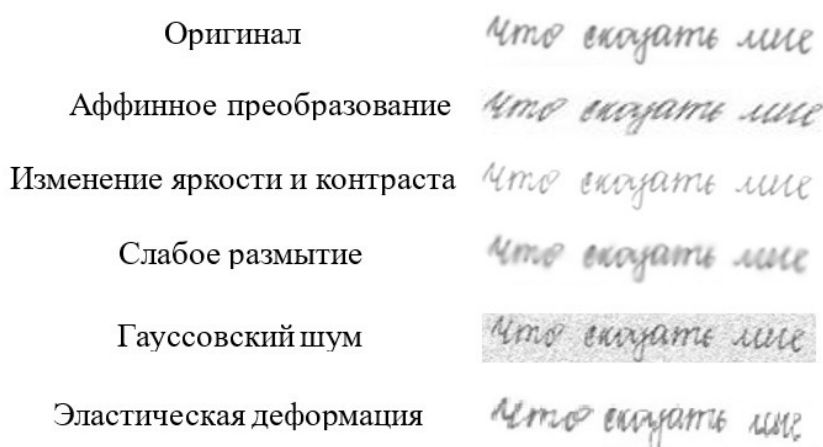


Рисунок 4 - Пример аугментированных изображений НКР

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.7>

2.1.1. Обучение модели. Результаты тестирования

В эксперименте сравнивались две принципиальные конфигурации модели:

- стандартная ResNet18-TE-СТС (далее «стандарт»);
- модифицированная ResNet18-TE-СТС, в которой в Layer 3 и Layer 4 горизонтальный stride был изменён с 2 на 1 (далее «модифицированная»).

Стандартная модель рассматривалась как контрольная конфигурация, отражающая прямое применение ResNet18-фронтенда к задаче НТР. Модифицированная модель рассматривалась как проверка аналитической гипотезы о необходимости сохранения горизонтального разрешения для СТС-распознавания.

Из классификационной версии ResNet18 были удалены global average pooling и fully connected слой, поскольку в задаче НТР требуется не классификация изображения целиком, а получение пространственной карты признаков. Далее карта признаков преобразовывалась в последовательность токенов вдоль горизонтальной оси изображения и передавалась в Transformer Encoder.

2.1.2. Ключевые метрики измерений

Качество распознавания оценивалось по метрикам CER, WER и Word Exact Match (WEM, доля полностью совпадающих слов). CER использовался как основная метрика, поскольку он отражает долю ошибок на уровне символов и особенно важен для кириллического рукописного текста, где ошибка в одной букве или цифре может существенно изменить значение результата. WER применялась как дополнительная метрика, оценивающая ошибки на уровне целого слова. Для word-level датасета HKR эта метрика позволяет оценить, насколько часто модель ошибается при распознавании всего рукописного слова. Word Exact Match рассчитывался как доля слов, распознанных полностью правильно. Эта метрика является наиболее строгой, поскольку даже одна ошибка в символе делает всё слово неверно распознанным.

2.1.3. Процедура обучения

Обучение моделей проводилось в среде PyTorch. На этапе обучения модель переводилась в режим `model.train()`, а на этапах валидации и тестирования — в режим `model.eval()`. При валидации и тестировании вычисление градиентов отключалось, а аугментации не применялись. Для контроля процесса обучения после каждой эпохи фиксировались значения `train CTC loss`, `validation CTC loss`, `validation CER` и `validation WER`. Лучший checkpoint выбирался по минимальному значению `validation CER`. Такой критерий был выбран потому, что CER является основной метрикой качества для HTR и более устойчиво отражает посимвольное качество распознавания, чем WER на малых выборках.

Для предотвращения переобучения применялся контроль качества на `validation`-наборе. Если `train loss` продолжал снижаться, а `validation loss` или `validation CER` переставали улучшаться, это рассматривалось как признак переобучения. Финальная оценка выполнялась на `test`-наборе только после выбора лучшего checkpoint по `validation CER`.

Кроме основной модификации, для стандартной модели были выполнены несколько контрольных запусков с изменением параметров обучения и регуляризации. Эти запуски были необходимы для проверки возможности улучшения метрик распознавания в стандартной архитектуре за счёт настройки гиперпараметров без изменения геометрии ResNet18-фронтенда. Изменяемые настройки обучения и диапазон их варьирования показан в таблице 4.

Таблица 4 - Основные гиперпараметры обучения исследуемых конфигураций модели

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.8>

Конфигурация модели	Dropout (метод регуляризации)	Learning Rate (скорость обучения)	Warmup epochs (число эпох прогрева)	Weight decay (квадратичная регуляризация весов)	Early Stoppin (ранняя остановка)
Стандартная модель*	0,1	10^{-4} общая для всей модели	отсутствует	10^{-4}	не реализована
Оптимизированная модель	0,3	$5 \cdot 10^{-5}$ для backbone; $2 \cdot 10^{-4}$ для «голов Transformer	3 эпохи	10^{-4}	7 эпох

Примечание: * – в отличие от оптимизированной модели в стандартной модели не применялась аугментация

Таким образом, значительное улучшение результатов обучения оптимизированной стандартной модели обусловлено комплексным подходом к настройке гиперпараметров: дифференцированным learning rate для backbone и head, внедрением warmup-периода, усиленной регуляризацией (dropout, weight decay, early stopping) и использованием аугментации данных. Эти изменения позволили достичь более стабильной сходимости и лучшего качества распознавания.

2.1.4. Результаты первого этапа исследований

В ходе настройки стандартной конфигурации ResNet18-TE-CTC качество распознавания на HKR улучшилось более чем на 40%, тем не менее изменение параметров обучения и регуляризации, хотя заметно и повлияло на динамику `train loss`, `validation loss`, CER и WER, однако не привело к качественному улучшению итоговой посимвольной точности. Лучшее значение CER для оптимизированной стандартной конфигурации составило 0.18. Это подтвердило предположение о том, что основное ограничение стандартной модели связано не только с настройками оптимизации, но и с геометрией признакововой карты, формируемой ResNet18. При стандартном горизонтальном сжатии выходная последовательность признаков оказывается слишком короткой для CTC-выравнивания плотных рукописных фрагментов.

Далее для ResNet18 были использованы настройки ее оптимизированной стандартной модели (базовая модель) и проведена модификация, заключающаяся в замене горизонтального stride в Layer 3 и Layer 4 с 2 на 1, в связи с чем качество распознавания существенно улучшилось (таблица 5). Таким образом, модифицированная модель достигла CER = 0.053 и по сравнению с базовой моделью относительное снижение CER составило около 70% (рисунок 5, таблица 5).

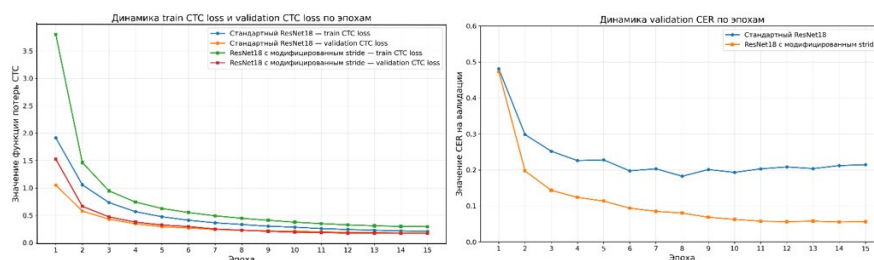


Рисунок 5 - Динамика контролируемых параметров в процессе обучения базовой и модифицированной моделей
DOI: <https://doi.org/10.60797/IRJ.2026.171.59.9>

Примечание: а) функция потерь CTC; б) валидационные значения CER

Таблица 5 - Результаты первого этапа испытаний архитектурного модуля MedForm-HTR

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.10>

Конфигурация	W_{in}	r_x	T	CER	WER	WEM
Базовая ResNet18	1024	32	32	0,18	0,52	0,38
ResNet18 с модифицированным горизонтальным шагом	1024	8	128	0,053	0,22	0,69

Анализ распознавания моделями тестовой выборки показал сильные стороны и зоны уязвимости модифицированной модели: модель хорошо справляется со сложными и длинными топонимами (написанными капсом или с дефисами), а также с целыми фразами (рисунок 6). Ошибки, которые модель все же допускает, носят локальный характер и связаны с оптическим сходством начертания букв (визуальная двусмысленность) или перекрытием буквами соседних зон других букв (лигатура).

Датасет HKR	Транскрипция	Оптимизированная модель (CER=0,18)	Модифицированная модель (CER=0,053)
Западно-Казахстанская	ЗАПАДНО-КАЗАХСТАНСКАЯ	ЗЛАТНО-КАЗАНСКАЯ	ЗАПАДНО-КАЗАХСТАНСКАЯ
то сердечная тоска	то сердечная тоска	т седняя осак	то сердечная тоска
еще не навившиеся	Еще не появившиеся	Еще павишися	Еще не павившиеся
Что добродетель	Что добродетель	Чтородеи	Что добродетель
Слепым волчатам	Слепым волчатам	Слепым валчатам	Слепым волчатам

Рисунок 6 - Примеры распознавания из тестовой выборки датасета HKR
DOI: <https://doi.org/10.60797/IRJ.2026.171.59.11>

Таким образом, результаты эксперимента подтвердили сформулированное предположение о том, что чрезмерное горизонтальное сжатие признаков карты в стандартном ResNet18 ухудшает условия CTC-выравнивания, тогда как сохранение горизонтального разрешения в поздних блоках backbone способствует снижению ошибки распознавания. Модифицированная архитектура формирует более длинную последовательность визуальных признаков вдоль оси письма, что снижает плотность символов на один временной шаг и создаёт более благоприятные условия для CTC-выравнивания. Первый этап исследований подтвердил, что для CTC-HTR важна не только мощность визуального backbone, но и его геометрическая совместимость с задачей последовательного распознавания. Полученные результаты дали все основания продолжить пилотные испытания модифицированной конфигурации ResNet18-TE-CTC на сегментах полей карт СМП.

2.2. Второй этап исследований. Пилотный датасет карт вызовов СМП

Второй этап эксперимента был направлен на оценку эффективности модифицированной архитектуры ResNet18-TE-СТС к доменной адаптации медицинской документации (карты вызовов СМП) и в точности ее распознавания.

Для пилотного датасета было отобрано 630 карт вызовов СМП, которые заполнялись 12 различными бригадами скорой медицинской помощи в среднем в течение 5 рабочих смен каждой из них. Для формирования датасета из карт вызовов СМП было выбрано 4 различных по содержанию поля (таблица 1), включающие как сверхплотный рукописный текст (поле 20. Жалобы) с элементами сокращений и аббревиатур (поле 21. Анамнез), так и медицинскую метрику (поле 22), а также поле, включающее специфические медицинские мероприятия (поле 26. Оказание помощи на месте вызова). Для поля Жалобы и Анамнез, учитывая специфику поля в качестве единиц обучения выбирались отдельные строки, содержащие от 15 до 30 символов, для более коротких полей (поля 22 и 26) бралось поле целиком. Единицей распознавания в пилотном датасете, таким образом, в целом, являлся не весь документ и не всё многострочное поле целиком, а вырезанные фрагменты строк или короткие числовые поля. Такой подход согласует целевой медицинский датасет с последовательной природой СТС-модели и снижает риск чрезмерного увеличения длины входной последовательности.

В результате было подготовлено 2520 строк, включающих по одной строке из каждого поля карты, при этом в валидационную и тестовую выборки не входили поля из карт, подготовленных для тренировочной выборки (таблица 6).

Таблица 6 - Количественное распределение пилотного датасета

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.12>

Режим	Тренировочная выборка (строки)	Валидационная выборка (строки)	Тестовая выборка (строки)
Обучение	2000	260	260

Основные результаты

3.1. Обучение модели. Результаты тестирования

Кроме ранее использованных метрик, применяемых на первом этапе эксперимента для оценки полей, содержащих числовые характеристики, использовалась метрика Digit CER (Character Error Rate, коэффициент ошибок для цифр) и Digit EM (Exact Match, доля полностью совпадающих цифр). Объем и режим аугментации при обучении на тренировочной выборке для пилотного датасета и HKR были идентичными. Кроме основной модифицированной модели ResNet18-TE-СТС, обученной на HKR, в эксперименте участвовала также стандартная модель ResNet18-TE-СТС с оптимизированной настройкой (таблица 4), также обученная на HKR. На втором этапе эксперимента использовались гиперпараметры обучения, соответствующие оптимизированной конфигурации, представленной в таблице 4.

Второй этап эксперимента включал обучение на пилотном датасете предобученных на HKR стандартной и модифицированной моделей. Также была проведена оценка распознавания модифицированной моделью тестовой выборки пилотного датасета без ее предобучения на данном датасете. Результаты данной серии экспериментов показаны в таблице 7.

Таблица 7 - Результаты пилотных испытаний на картах вызовов СМП

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.13>

Модель	Поле	CER	WER	Digit CER	Digit EM
Стандартная, обученная на HKR + СМП	01 — жалобы	0,5052	0,9275	-	-
	02 — цифровые показатели	0,9665	1,0002	1,0000	0,0000
	03 — анамнез	0,5698	0,9898	-	-
	04 — мероприятия	0,6574	0,9987	1,0000	0,0000
Модифицированная, обученная на HKR + СМП	01 — жалобы	0,0924	0,3511	-	-
	02 — цифровые показатели	0,0566	0,2359	0,0776	0,8443
	03 — анамнез	0,0939	0,3568	-	-
	04 — мероприятия	0,1021	0,3879	0,1150	0,7500
Модифицированная, обученная на HKR без	01 — жалобы	0,3895	0,8756	-	-
	02 — цифровые показатели	0,9139	1,0000	1,0000	0,0000

Модель	Поле	CER	WER	Digit CER	Digit EM
предобучения на картах СМП	03 — анамнез	0,4032	0,9695	-	-
	04 — мероприятия	0,5101	1,0000	1,0000	0,0000

Обобщённые показатели качества распознавания на пилотном датасете для модифицированной, обученной на НКР+СМП модели составили CER = 0,0922 и WER = 0,3475, что отражает среднюю точность модели по четырём рукописным полям. Агрегация выполнялась путём суммирования ошибок и эталонных символов/слов по всем полям датасета.

Пилотная проверка на пилотном датасете показала, что модифицированная архитектура ResNet18-TE-CTC значительно превосходит стандартную конфигурацию и модель, обученную только на НКР. Наилучшие результаты были получены после дополнительного обучения на выделенных фрагментах полей карт СМП. Особенно важным является результат по цифровым значениям, где Digit Exact Match достиг 0,8443.

Кроме данных пилотных испытаний была проведена оценка влияния разлиновки карт СМП на точность распознавания. Сравнивались обобщенные показатели CER и WER для модифицированной, обученной на НКР+СМП модели до удаления разлиновки и после удаления. Горизонтальные линии бланка выделялись на основании их цветовых и геометрических характеристик, при этом сформированная маска разлиновки использовалась для удаления соответствующих пикселей с последующим локальным восстановлением фоновой области методом inpainting. Полученные результаты показали, что устранения разлиновки практически не повлияло на точность распознавания: общий CER после разлиновки составил 0,0918, а WER — 0,3367. Сравнение исходных и обработанных изображений также не выявило существенного улучшения показателей распознавания, поэтому процедура удаления разлиновки не будет включена в обязательный вариант конвейера MedForm-HTR.

Анализ допускаемых модифицированной моделью ошибок показал, что большая их часть обусловлена спецификой почерка и его плотностью, так как высота поля строки составляет всего 3 мм, а наличие соседних строк предполагает еще меньшую рабочую высоту строки, иначе соседние строки будут перекрываться, что сделает невозможным прочтение текста даже человеком (рисунок 7).

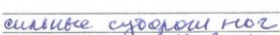
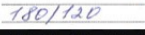
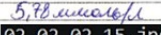
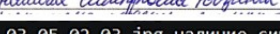
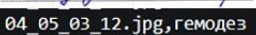
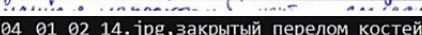
1. 
01_04_01_23.jpg, сильные судороги ног, сильные судороги ног,
2. 
01_04_02_09.jpg, "перекос лица, судороги", "перекос[] лица, судорог[]",
3. 
02_02_02_13.jpg, 180/120, 180/120,
4. 
02_02_02_15.jpg, "5,78 ммоль/л", "5,[]8 ммоль/л", 0
5. 
03_04_01_04.jpg, персистирующая форма ОП, персистирующая форма ОП,
6. 
03_05_02_03.jpg, наличие симптома Ровзинга, наличие сим[]птома []овзинга,
7. 
04_05_03_12.jpg, гемодез 500 мл в/в, гемодез[] 500 мл в/в,
8. 
04_01_02_14.jpg, закрытый перелом костей таза, закрытый перелом[] костей таза,

Рисунок 7 - Типичные ошибки в распознавании строк полей карт вызовов СМП

DOI: <https://doi.org/10.60797/IRJ.2026.171.59.14>

Примечание: строки 1, 2 — поле жалобы; строки 3, 4 — поле цифровые показатели; строки 5, 6 — поле анамнез и диагностика; строки 7, 8 — поле оказание помощи на месте вызова

Одним из выявленных типов ошибок являются замены визуально близких рукописных символов, например «о» и «а», «с» и «е», «и» и «м» и др. (рисунок 8, строки 2, 5, 6, 8). Такие ошибки связаны не только с ограничениями модели, но и с объективной неоднозначностью рукописного текста, наличием лигатур, в связи с чем в ряде случаев отдельный символ не может быть надёжно интерпретирован без учёта контекста слова или поля документа. Вторым значимым фактором, оказывающим влияние на точность распознавания являются ошибки, связанные с наложением отдельных фрагментов символов нижних и верхних строк на целевую строку, в связи с чем модель может их принять за дополнительные символы (строки 2, 5, 6, 8).

Таким образом, пилотные исследования показали, что данные датасета НКР существенно отличаются от реальных карт вызова СМП (так называемый доменный сдвиг). Это подтверждает необходимость дообучения модели на целевом

типе документов и демонстрирует преимущество модифицированной архитектуры ResNet18-TE-CTC для распознавания выделенных фрагментов медицинских полей в картах вызова СМП.

3.2. Результаты второго этапа исследований

Проведённые пилотные исследования подтвердили принципиальную применимость модифицированной архитектуры ResNet18-TE-CTC для распознавания рукописных полей карт скорой медицинской помощи. Наилучшие результаты были получены после предметно-ориентированного дообучения модели на целевых изображениях: общий CER составил 0,0922, WER — 0,3475, а Digit CER для поля цифровых показателей — 0,0776. Полученные данные показывают, что адаптация к структуре документа, характеру медицинского почерка и особенностям заполнения бланков является необходимым условием создания прикладной HTR-системы.

Вместе с тем пилотное исследование выявило ряд ограничений, включая сравнительно более высокую ошибку в многострочных текстовых полях, возможное влияние элементов соседних строк, ограниченность объёма и разнообразия предметной выборки, а также отсутствие языкового медицинского постобработки. Дальнейшее развитие MedForm-HTR предполагает расширение и внешнюю валидацию датасета, архитектурное сравнение, изоляцию целевых строк, применение медицинского словаря и предметной языковой модели, а также оценку достоверности и ручную проверку критичных результатов.

Обсуждение

Таким образом, научный вклад работы заключается в аналитическом обосновании и экспериментальной проверке адаптации ResNet18 в составе гибридной архитектуры (ResNet18-TE-CTC) к задаче распознавания рукописного кириллического текста, что позволило снизить CER на 70% по сравнению с базовой конфигурацией.

В отличие от специализированных CNN-CTC архитектур, где параметры сжатия учитываются при проектировании, для ResNet18 требуется дополнительная настройка. Контролируемое исследование показало, что изменение шага свёртки в слоях Layer 3 и Layer 4 с 2 на 1 улучшает качество распознавания за счёт оптимизации выходной последовательности признаков.

Пилотная проверка на фрагментах карт СМП подтвердила эффективность предложенного подхода.

Заключение

Полученные результаты дают все основания считать модифицированную гибридную архитектуру ResNet18-TE-CTC как перспективную модель для HTR-модуля разрабатываемой интеллектуальной системы MedForm-HTR. Анализ полученных результатов исследований показывает, что для успешной реализации данной системы требуется дальнейшая оптимизация и развития как самой модели HTR-модуля, так и всего процесса обработки документа, включая приём скана, предобработку, определение типа страницы, выравнивание, выделение полей, разбиение строк, их классификацию, HTR-распознавание, OCR, OMR и верификацию.

В соответствии с этим основные направления совершенствования системы будут включать:

- расширение набора медицинских документов с разнообразием почерков и типов записей;
- исследование подходов к оценке качества распознавания: как для конкретных специалистов, так и для общей применимости;

- улучшение выделения текстовых строк в многострочных полях;
- оптимизация обработки границ фрагментов и отступов для повышения точности распознавания;
- разработка методов подавления помех от посторонних элементов;
- внедрение специальных преобразований изображений, учитывающих особенности медицинских документов;
- совершенствование алгоритмов распознавания с использованием медицинского словаря и правил проверки;
- создание системы оценки достоверности результатов с передачей сложных случаев на проверку оператору.

Таким образом, планируется подготовка системы MedForm-HTR до полноценного решения задач по обработке медицинских документов (карты вызовов СМП и др.), включающих распознавание текста, анализ структуры документа и контроль качества получаемых результатов.

Благодарности

Благодарим коллектив редакции.

Конфликт интересов

Не указан.

Рецензия

Нуриев М.Г., Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань Российская Федерация
DOI: <https://doi.org/10.60797/IRJ.2026.171.59.15>

Acknowledgement

The authors express their gratitude to the editorial team.

Conflict of Interest

None declared.

Review

Nuriev M.G., Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan Russian Federation
DOI: <https://doi.org/10.60797/IRJ.2026.171.59.15>

Список литературы на английском языке / References in English

1. AlKendi W. Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey / W. AlKendi, F. Gechter, L. Heyberger [et al.] // Journal of Imaging. — 2024. — Vol. 10, No. 1. — Art. 18. — DOI: 10.3390/jimaging10010018.



2. Bluche T. Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention / T. Bluche, J. Louradour, R. Messina // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). — Kyoto, Japan, 2017. — P. 1050–1055. — DOI: 10.1109/ICDAR.2017.174.
3. Coquenat D. DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition / D. Coquenat, C. Chatelain, T. Paquet // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2023. — Vol. 45, No. 7. — P. 8227–8243. — DOI: 10.1109/TPAMI.2023.3235826.
4. Shi B. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition / B. Shi, X. Bai, C. Yao // IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). — 2017. — Vol. 39, No. 11. — P. 2221–2235. — DOI: 10.1109/TPAMI.2016.2646371.
5. Li Y. HTR-VT: Handwritten Text Recognition with Vision Transformer / Y. Li, D. Chen, T. Tang [et al.] // Pattern Recognition. — 2025. — Vol. 158. — Art. 110967. — DOI: 10.1016/j.patcog.2024.110967.
6. Wick C. Transformer for Handwritten Text Recognition Using Bidirectional Post-Decoding / C. Wick, J. Zöllner, T. Grüning // Document Analysis and Recognition — ICDAR 2021. Lecture Notes in Computer Science. — Vol. 12823. — Springer, Cham, 2021. — P. 112–126. — DOI: 10.1007/978-3-030-86334-0_8.
7. Li M. TrOCR: Transformer-based optical character recognition with pre-trained models / M. Li, T. Lv, J. Chen [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. — 2023. — Vol. 37, No. 11. — P. 13094–13102. — DOI: 10.1609/aaai.v37i11.26538.
8. Coquenat D. End-to-end handwritten text line recognition with attention-based recurrent neural networks / D. Coquenat, C. Chatelain, T. Paquet // Pattern Recognition. — 2023. — Vol. 138. — Art. 109389. — DOI: 10.1016/j.patcog.2023.109389.
9. Kang L. Pay Attention to What You Read: Non-Recurrent Handwritten Text-Line Recognition / L. Kang, P. Riba, M. Rusinol [et al.] // Pattern Recognition. — 2022. — Vol. 129. — Art. 108766. — DOI: 10.1016/j.patcog.2022.108766.
10. Coquenat D. Recurrence-Free Unconstrained Handwritten Text Recognition Using Gated Fully Convolutional Network / D. Coquenat, C. Chatelain, T. Paquet // 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). — Dortmund, Germany, 2020. — P. 19–24. — DOI: 10.1109/ICFHR2020.2020.00015.
11. Vaswani A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems 30 (NeurIPS, 2017). — P. 5998–6008.
12. Atienza R. Vision Transformer for fast and efficient scene text recognition / R. Atienza // Document Analysis and Recognition — ICDAR 2021. — Springer, Cham, 2021. — P. 319–334. — DOI: 10.1007/978-3-030-86330-2_21.
13. Michael J. Evaluating sequence-to-sequence models for handwritten text recognition / J. Michael, R. Labahn, J. Zöllner [et al.] // 2019 International Conference on Document Analysis and Recognition (ICDAR). — IEEE, 2019. — P. 1286–1293. — DOI: 10.1109/ICDAR.2019.00208.
14. Yousef M. OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by Learning to Unfold / M. Yousef, T.E. Bishop // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — Seattle, WA, USA, 2020. — P. 14710–14719.
15. Puigcerver J. Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? / J. Puigcerver // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). — Kyoto, Japan, 2017. — P. 67–72. — DOI: 10.1109/ICDAR.2017.20.
16. Tabzaoui A. Validated and Decoding-Aware CNN-Transformer-CTC Framework for Multi-Font Printed Arabic Word Recognition / A. Tabzaoui, L. Chakir // Applied Sciences. — 2026. — Vol. 16. — Art. 4071. — DOI: 10.3390/app16094071.