УДК 004.852
DOI 10.17513/snt.39945

СЕМАНТИЧЕСКАЯ СЕГМЕНТАЦИЯ ТЕКСТОВЫХ ПОЛЕЙ И ТАБЛИЦ В ДОКУМЕНТЕ НА ОСНОВЕ ПРИМЕНЕНИЯ АРХИТЕКТУРЫ UNETFORMER

[1, 2]Климов А.М., [1, 2]Котюжанский Л.А., [2]Четверкин Н.В.

[1]ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», Екатеринбург, e-mail: ak12wirexia122@gmail.com;
[2]ООО «Нексус», Екатеринбург, e-mail: nexus077@gmail.com

Аннотация. В статье рассматривается применение архитектуры UNetFormer для решения задачи семантической сегментации текстовых строк и таблиц в документах. Цель исследования – решение задачи семантической сегментации для документов, имеющих особенности, которые могут встречаться на одной странице документа: различные ориентации текста, таблицы, шумы и инородные объекты (печати, подписи). В качестве решения поставленной задачи была выбрана архитектура нейронной сети для семантической сегментации – UNetFormer, которая показывает высокую эффективность в других задачах: семантической сегментации спутниковых и медицинских снимков. Также для более эффективного обучения авторы предлагают использование метода аугментации данных в реальном времени с помощью генерации и преобразования реальных данных. Для определения ориентации текста в обучающих данных использовались карты, соответствующие различным ориентациям текста, а также карты для детекции таблиц (их ребер и узлов) и ядер строк для более точного вырезания текстовых прямоугольников с последующей обработкой моделью распознавания текста. Полученные результаты демонстрируют высокий показатель среднего значения индекса Жаккара (mIoU = 0,833) на датасете из 1230 размеченных документов, собранном авторами.

Ключевые слова: UNet, UNetFormer, детекция текста, семантическая сегментация документов, оптическое распознавание текста

SEMANTIC SEGMENTATION OF TEXT FIELDS AND TABLES IN A DOCUMENT BASED ON THE UNETFORMER ARCHITECTURE

[1, 2]Klimov A.M., [1, 2]Kotyuzhanskiy L.A., [2]Chetverkin N.V.

[1]Ural Federal University named after the First President of Russia B.N. Yeltsin, Yekaterinburg, e-mail: ak12wirexia122@gmail.com;
[2]OOO «Nexus», Yekaterinburg, e-mail: nexus077@gmail.com

Annotation. The article discusses the application of the UNetFormer architecture to solve the problem of semantic segmentation of text strings and tables in documents. The purpose of the study is to solve the problem of semantic segmentation for documents that have features that can occur on one page of a document: different orientations of text, tables, noises and foreign objects (seals, signatures). As a solution to this problem, the neural network architecture for semantic segmentation, UNetFormer, was chosen, which shows high efficiency in other tasks: semantic segmentation of satellite and medical images. Also, for more effective training, the authors suggest using the real-time data augmentation method by generating and converting real data. To determine the orientation of the text in the training data, maps corresponding to different orientations of the text were used, as well as maps for detecting tables (their edges and nodes) and string kernels for more accurate cutting of text rectangles with subsequent processing by a text recognition model. The results obtained demonstrate a high indicator of the average value of the Jacquard index (mIoU = 0.833) on a dataset of 1230 marked-up documents collected by authors.

Keywords: UNet, UNetFormer, text detection, documents semantic segmentation, optical text recognition

В настоящее время системы оптического распознавания текста обладают широким спектром функциональных применений, включая возможность интеграции в системы автоматизированного документооборота. Кроме того, данные системы могут использоваться для автоматизации создания форм документов в случаях, когда отсутствует исходный файл, а также для автоматизированного извлечения необходимых данных из фотографий или отсканированных изображений документов [1]. Обычно система оптического распознавания текста включает в себя два основных модуля [2]: сегментации текстовых строк в документе и распознавания детектированных строк, то есть преобразования изображения в текстовую строку.

Задача сегментации текстовых строк в документе является ключевым этапом в области оптического распознавания текста. Разработка эффективных методов для решения этой задачи представляет собой обязательное условие для обеспечения высокой точности и производительности системы распознавания текста.

Рис. 1. На верхнем изображении – пример работы MultiplexedOCR [5], красные рамки – найденные алгоритмом текстовые строки, зеленые – текст, распознанный системой. На нижнем изображении – пример работы алгоритма, предлагаемого авторами, зеленые рамки – найденные текстовые строки, красные рамки – разделение текстовых строк на отдельные слова

Существует множество подходов для решения данной задачи, основанных как на алгоритмах компьютерного зрения, например smearing [3], так и методах с использованием нейронных сетей [4]: например, в работе Ц. Хуана [5] и в статье М. Буста, Л. Неймана и Дж. Мэйтаса [6] для детекции текста на изображении используется подход с использованием нейронных сетей. Эти решения не обладают достаточным качеством сегментации строк на собранных авторами данных (далее датасет Nexus): возникают ложноположительные срабатывания на дефекты печати и сканирования, ложноотрицательные срабатывания на ориентациях, отличных от нормальной. К системе предъявляются высокие требования качества распознавания, поэтому подобные ошибки имеющихся решений неприемлемы. Существующие системы имеют еще один существенный недостаток: они не способны определять ориентацию строк (например, текст, повернутый на 90 градусов), что ведет к ошибке в OCR. Сравнение работы на датасете Nexus алгоритма MultiplexedOCR [5] и предлагаемого авторами можно увидеть на рис. 1. Собранный датасет имеет следующую специфику: строки в пределах одного документа могут иметь разную ориентацию, имеются наложения рукописного текста и печатей поверх таблиц и печатного текста.

Эффективное определение позиции и размера текстовых областей непосредственно влияет на качество всей системы оптического распознавания текста. Неверное определение текстовой области приводит к падению точности распознавания текста, к нему можно отнести:

1) недовыделение текста;

2) выделение как текста объекта, не являющегося текстом;

3) «срастание» выделенных областей текста.

Поэтому повышение эффективности методов семантической сегментации документов играет решающую роль в обеспечении успешного функционирования системы оптического распознавания текста.

В работе предлагается метод решения вышеизложенных проблем, основанный на подходе с использованием архитектуры UNetFormer. Эта архитектура направлена на повышение эффективности и точности в задачах семантической сегментации спутниковых снимков [7] и медицинских изображений [8]. Целью этого исследования является решение задачи семантической сегментации документов на основе архитектуры UNetFormer.

Материалы и методы исследования

Архитектура решения

Для решения задачи используется архитектура UNetFormer на базе предобученного ResNet-18 в качестве энкодера и декодера с блоками Global Local Attention. Известные архитектуры нейронных сетей для решения задачи семантической сегментации, основанные на сверточных слоях, такие как SegNet [9], UNet [10], имеют ограниченное рецептивное поле, поэтому они извлекают только признаки ближнего контекста, игнорируя или слабо реагируя на признаки, находящиеся вне зоны рецептивного поля. Из-за этого существенного недостатка могут пропадать недостаточно контрастные строки и линии таблиц. Блоки Global Local Attention помогают модели из-

влекать также признаки дальнего контекста, что приводит к существенному росту качества семантической сегментации, что и показывают в статье, посвященной архитектуре UNetFormer [7]. Модель возвращает N карт сегментации по количеству классов. То есть выходной тензор имеет размерность

$$(B, N, H, W),$$

где B – размер батча, N – число семантических классов, H, W – высота и ширина входного изображения соответственно. Каждая карта размерности (B, 1, H, W) – батч бинарных изображений, соответствующих положению объектов присущего этой карте класса на исходных изображениях. В нашем случае число классов $N = 7$:

1) текст, ориентированный горизонтально (0 градусов);

2) текст, повернутый на 90 градусов по часовой стрелке;

3) текст, повернутый на 90 градусов против часовой стрелки;

4) текст, повернутый на 180 градусов;

5) узлы таблиц;

6) ребра таблиц;

7) ядра строк – тонкие линии, проведенные через середину строки вдоль направления текста (рис. 4).

В отличие от статьи [7] карта фона (специальный класс для пикселей, не принадлежащих ни одному из семантических классов) не используется.

Подобное разбиение на классы позволяет определить ориентацию каждого текстового поля на этапе сегментации и повернуть вырезанный прямоугольник с текстом на необходимый угол при подаче на вход системе распознавания текста. Также данное разбиение позволяет определить позицию и структуру таблиц, а также ориентацию страницы документа в целом.

Функция потерь и оптимизатор

В качестве функции потерь была выбрана комбинированная функция:

$$L = \alpha L_{dice} + L_{SoftLogitsBCE} + Þ, \qquad (1)$$

основанная на работе [7], адаптированная под задачу добавлением слагаемого $Þ$, где $\alpha = 0.4$, а составляющие функции потерь определяются как

$$L_{dice} = 1 - \frac{2}{B}\sum_{i=1}^{B}\sum_{j=1}^{N}\frac{\hat{y}_j^i \wedge y_j^i}{\hat{y}_j^i \vee y_j^i}, \qquad (2)$$

где $\hat{y}_j^i$ – предсказанная моделью карта для i-го сэмпла и j-го класса, y_j^i – маска i-го сэмпла и j-го, класса, $\wedge$ – операция поэлементного логического «и», $\vee$ – операция поэлементного логического «или».

$$L_{SoftLogitsBCE} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N} y_j^i \cdot \ln\left(\sigma\left(\tilde{y}_j^i\right)\right) + \left(1 - y_j^i\right) \cdot \ln\left(1 - \tilde{y}_j^i\right), \qquad (3)$$

$$\tilde{y}_j^i = \beta \cdot \left(1 - \hat{y}_j^i\right) + \left(1 - \beta\right) \cdot \hat{y}_j^i, \qquad (4)$$

где β – сглаживающая константа.

С целью уменьшения «срастания» строк было добавлено слагаемое в функцию потерь:

$$Þ = \gamma \sum_{i=1}^{B}\sum_{j=1}^{N} \hat{y}_j^i \oplus y_j^i,$$

где $\oplus$ – операция поэлементного XOR, а $\gamma = 14$. Это слагаемое позволяет дополнительно штрафовать модель за расхождение с целевым тензором. Задача минимизации $L \rightarrow min$ решалась оптимизатором Adam [11] с различными скоростями обучения для энкодера и декодера: $lr_{encoder} = 6 \cdot 10^{-5}$, $lr_{decoder} = 10^{-3}$.

Подход к обучению и аугментации данных

На вход модели подавались трехканальные изображения разрешением 1024x1024, то есть тензор размерности (B, 3, 1024, 1024), тензор ожидаемых масок имеет размерность (B, 7, 1024, 1024). Схема подготовки обучающих данных представлена на рис. 2.

Все эти действия происходят в реальном времени, то есть генерация изображений происходит непосредственно во время обучения модели. Предложенный подход позволяет создать разнообразие данных, подаваемых модели и увеличить ее обобщающую способность, что положительно сказывается на качестве обучения и предотвращает переобучение. Пример входных данных и масок для них по классам можно увидеть на рис. 3.

Решение проблемы пересечения строк

При использовании обученной модели обнаружилась проблема «срастания» («срастание» или пересечение строк – вид ошибки, при котором n строк определяются алгоритмом сегментации строк как одно целое) находящихся близко строк, возникающая из-за ошибок в разметке датасета, а также за счет ошибок самой модели сегментации.

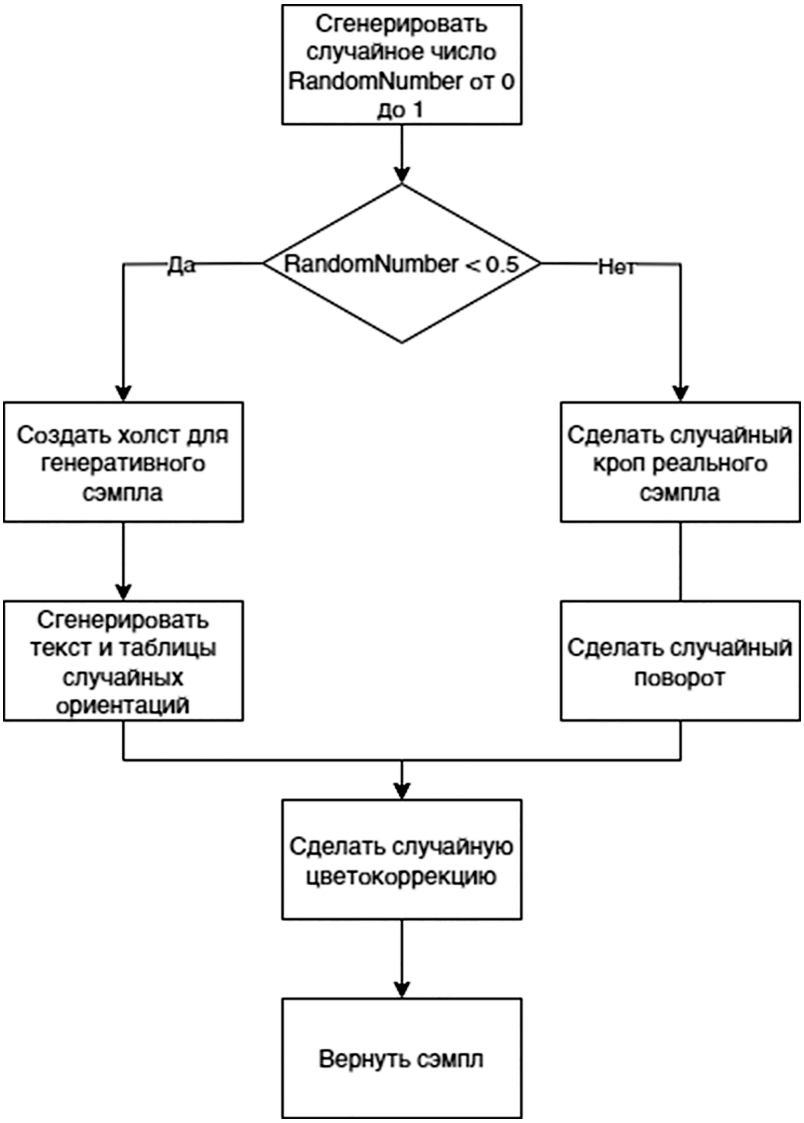

Рис. 2. Схема генерации сэмплов для обучения и валидации модели

«Срастание» строк приводит к неправильной вырезке текстовых прямоугольников, а далее к неверному распознаванию текста в вырезанных областях. Для решения этой проблемы используются два подхода: дополнительная карта, о которой написано выше – карта ядер строк и дополнительном слагаемом в функции потерь. Комбинированный подход позволяет избежать большинства проблем при распознавании текста, связанных с такой особенностью работы модели сегментации. Пример разделения пересекающихся строк можно увидеть на рис. 4.

Результаты исследования и их обсуждение

Для вычисления метрик качества обученной модели используются не попавшие в тренировочную выборку, а также генерируемые в реальном времени данные. Ключевая метрика качества модели для нашего исследования – mIoU (средний индекс Жаккара). Также важным является индекс Жаккара для каждого класса отдельно. Значения для этих метрик приведены в таблице.

Высокие значения этих метрик обусловлены самой архитектурой выбранной модели, поскольку UNetFormer учитывает как локальный (ближний), так и дальний контекст. Не последнюю роль играет количество и разнообразие данных. Предложенный подход к обучению позволяет постоянно подавать модели новые данные для обучения, что положительно сказывается на работе модели в реальных условиях на неизвестных ранее данных.

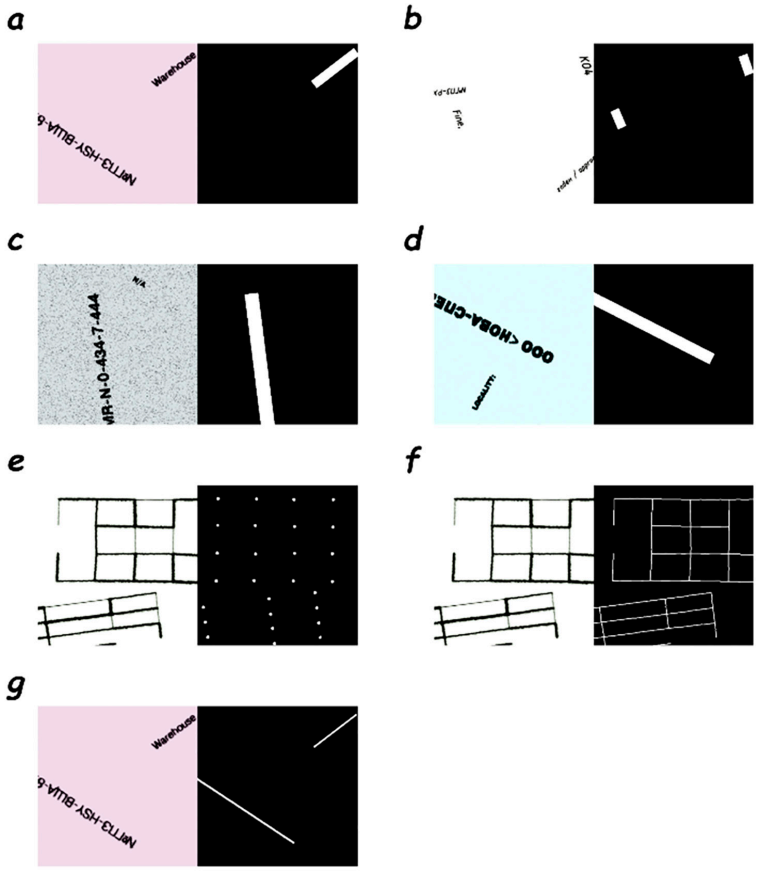

Рис. 3. Сгенерированные изображения и маски для обучения. Пары входное изображение – маска: a – строк горизонтальной ориентации текста (0 градусов), b – строк, повернутых на 90 градусов по часовой стрелке, c – строк, повернутых на 90 градусов против часовой стрелки, d – строк, повернутых на 180 градусов, e – узлов таблиц, f – ребер таблиц, g – ядер строк

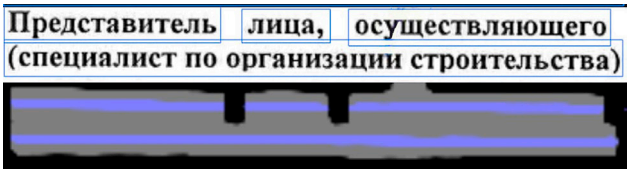

Рис. 4. Пример разделения пересекающихся строк на карте сегментации с помощью карты ядер строк. Серым цветом изображен результат сегментации текстовых строк нормальной ориентации, фиолетовым – результат сегментации ядер строк. Синие рамки вокруг текста – определенные алгоритмом текстовые строки

Метрики обученной модели на тестовых данных

Метрика	Значение
mIoU	0,833
IoU для карты нормальной ориентации	0,895
IoU для карты ориентации 90 градусов по часовой стрелке	0,886
IoU для карты ориентации 90 градусов против часовой стрелки	0,887
IoU для карты ориентации 180 градусов	0,894
IoU для карты узлов таблиц	0,798
IoU для карты ребер таблиц	0,741
IoU для карты ядер строк	0,730

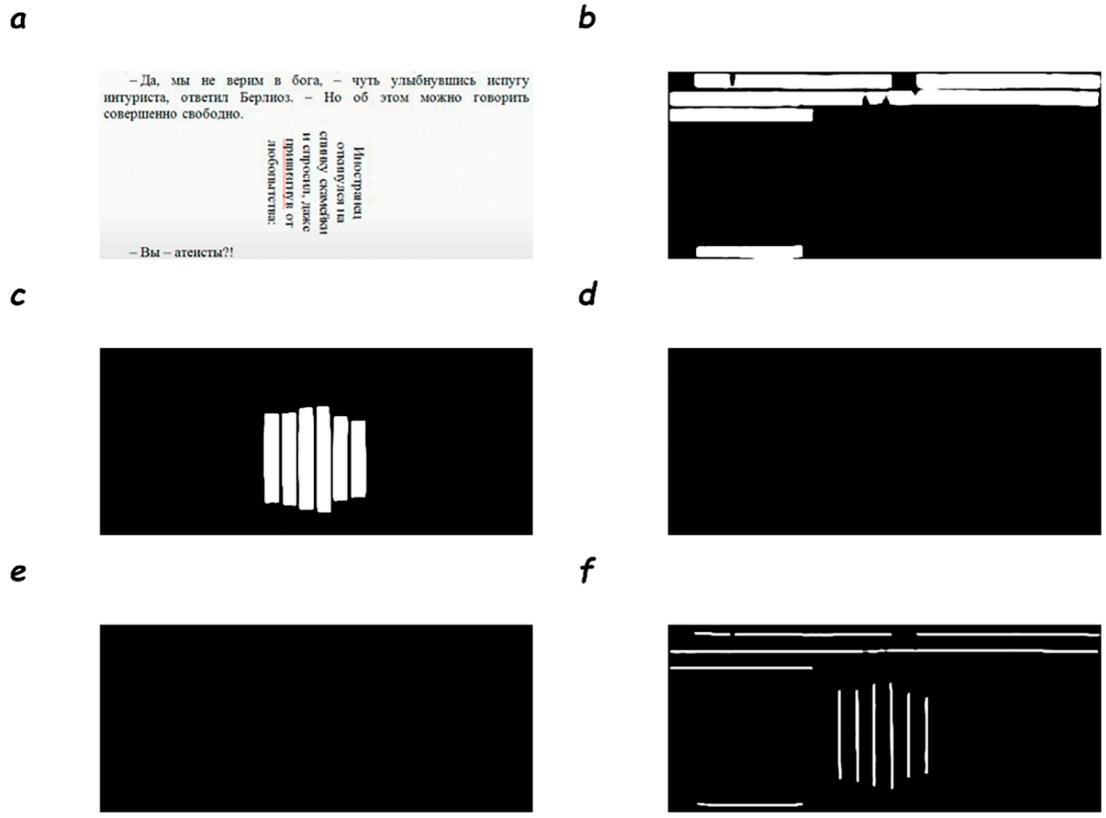

Рис. 5. Результат сегментации моделью; a – входное изображение, b – карта горизонтальной ориентации текста, c – карта текста, повернутого на 90 градусов по часовой стрелке, d – карта текста, повернутого на 90 градусов против часовой стрелки, e – карта текста, повернутого на 180 градусов, f – карта ядер строк

Важным достоинством модели является скорость предсказания: 8–12 мс при размере входного изображения 6 мегапикселей на графическом процессоре Nvidia RTX 3080Ti. Примеры работы модели сегментации на реальных документах можно увидеть на рис. 5.

Заключение

В статье показана возможность эффективного применения архитектуры UNetFormer в задаче детекции текстовых полей с учетом их ориентации и таблиц в документах, и подход к решению такой задачи. Для любой системы OCR детекция текста является одной из подзадач, которую требуется решить для построения системы. Документы, обрабатываемые такими системами, могут иметь особенности: таблицы, текст, имеющий ориентацию, отличную от нормальной, близко расположенный текст, наложения объектов друг на друга, тени, печати. Разработанная система способна распознавать все возможные ориентации текста и таблицы, что позволяет избежать проблем неверного распознавания текста из-за ошибки в определении ориентации, а также сохранять информацию о структуре документа (положении текстовых полей, таблицах).

UNetFormer показывает высокие значения метрики Жаккара в задаче семантической сегментации документов, значит, можно сделать вывод о том, что модель подходит для решения такой задачи.

Список литературы

1. Котюжанский Л.А., Четверкин Н.В., Протасевич А.А., Кочеров Р.В., Рыжкова Н.Г. Классификация сканированных документов с использованием сверточной нейросети // Современные наукоемкие технологии. 2021. № 6. С. 45–49.

2. Полюхин Д.А., Сальников И.И. Методы и этапы распознавания рукописного текста // Научное обозрение. Педагогические науки. 2019. № 3. С. 71–74.

3. Wong K.Y., Casey R.G., Wahl F.M. Document analysis system // IBM Journal of Research and Development. 1982. Vol. 26, Is. 6. P. 647–656. DOI: 10.1147/rd.266.0647.

4. Wang X., He Z., Wang K., Wang Y., Zou L., Wu Z. A survey of text detection and recognition algorithms based on deep learning technology // Neurocomputing. 2023. Vol. 556. DOI: 10.1016/j.neucom.2023.126702.

5. Huang J., Pang G., Kovvuri R., Toh M., Liang K.J., Krishnan P., Yin X., Hassner T. A multiplexed network for end-to-end, multilingual OCR // 2021 IEEE/CVF Conference on

Computer Vision and Pattern Recognition. 2021. P. 4547–4557. [Электронный ресурс]. URL: https://ieeexplore.ieee.org/document/9577633 (дата обращения: 17.01.2024).

6. Busta M., Neumann L., Matas J. Deep textspotter: An end-to-end trainable scene text localization and recognition framework // 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017. P. 2380–7504. [Электронный ресурс]. URL: https://ieeexplore.ieee.org/document/8237504 (дата обращения: 17.01.2024).

7. Wang L., Li R., Zhang C., Fang S., Duan C., Meng X., Atkinson P.M. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery // ISPRS Journal of Photogrammetry and Remote Sensing. 2022. Vol. 190. P. 196–214. DOI: 10.1016/j.isprsjprs.2022.06.008.

8. Hatamizadeh A., Xu Z., Yang D., Li W., Roth H., Xu D. Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation // Cornell University. 2022. [Электронный ресурс]. URL: https://arxiv.org/abs/2204.00631 (дата обращения: 17.01.2024).

9. Badrinarayanan V., Handa A., Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling // Cornell University. 2015. [Электронный ресурс]. URL: https://arxiv.org/abs/1505.07293 (дата обращения: 17.01.2024).

10. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation // MICCAI 2015. P. 234–241. [Электронный ресурс]. URL: https://arxiv.org/abs/1505.04597 (дата обращения: 17.01.2024).

11. Kingma D.P., Ba J. Adam: A method for stochastic optimization // Cornell University. 2017. [Электронный ресурс]. URL: https://arxiv.org/abs/1412.6980 (дата обращения: 17.01.2024).