Управление процессом аугментации данных для решения задачи сегментации снимков УЗИ щитовидной железы

Д.В. Холод, Е.В. Боброва, К.С. Зайцев, А.А. Трухин, Е.А.Трошина

Аннотация. Целью данной статьи было исследование влияния пропорций используемых реальных и аугментированных данных на решение задачи сегментации узлов щитовидной железы на ультразвуковых изображениях. Для этого был введён параметр R - отношение числа аугментированных изображений к числу оригинальных - и проанализировано его влияние на качество сегментации при использовании различных архитектур энкодеров. В работе рассматривалось формирование обучающих выборок с различной долей аугментированных данных, после чего проводилось обучение моделей и оценка результатов сегментации на датасете DDTI. Оценивание качества выполнялось с использованием стандартных метрик сегментации (Dice и IoU). В результате установлено, что зависимость качества сегментации от параметра R носит немонотонный характер: существует оптимальное значение Rmax, при котором достигается наилучшее качество, а также область Rmin, в которой наблюдается ухудшение результатов. Показано, что архитектура ViT-Base демонстрирует наибольшую чувствительность к избыточной аугментации, тогда как ConvNeXt-Tiny и Swin Transformer-Tiny характеризуются большей устойчивостью к изменению доли синтетических данных.

Ключевые слова: *сегментация, ультразвуковые изображения, щитовидная железа, аугментация данных, коэффициент Dice, ConvNeXt, Swin Transformer, Vision Transformer*.

I. ВВЕДЕНИЕ

Ультразвуковая диагностика является методом выявления узлов щитовидной железы благодаря своей доступности и отсутствию ионизирующего излучения. Качество интерпретации ультразвуковых изображений зависит от опыта врача и подвержена субъективным ошибкам, особенно при оценке эхогенности ткани узла и морфологических характеристик. В связи с этим автоматическая сегментация узлов щитовидной железы на УЗИ-изображениях с использованием методов компьютерного зрения и глубокого обучения является актуальной научно-практической задачей, решение которой позволит автоматизировать описание характеристик узловых образований щитовидной железы.

В последние годы методы глубокого обучения, в частности свёрточные нейронные сети (CNN), продемонстрировали высокую эффективность в задачах медицинской сегментации изображений. Архитектуры типа U-Net и их модификации стали де-факто стандартом для обработки медицинских данных [1]. Однако успешное обучение таких моделей требует значительных объёмов размеченных данных, получение которых в медицинской области затруднено из-за высокой стоимости работы врачей-специалистов по экспертной разметки.

Одним из ключевых подходов к решению проблемы ограниченного объёма данных является аугментация — искусственное расширение обучающей выборки путём применения различных преобразований к исходным изображениям. В ряде работ показано, что корректно подобранные аугментации позволяют повысить обобщающую способность модели и снизить переобучение [2, 3]. Вместе с тем, в литературе отсутствует единое мнение о том, насколько интенсивной должна быть аугментация и как её влияние зависит от архитектуры модели; некорректная или чрезмерная аугментация может приводить к ухудшению качества [3, 4].

Дополнительную сложность вносит активное развитие архитектур глубокого обучения. Помимо классических свёрточных сетей, в задачах сегментации всё чаще применяются трансформерные и гибридные модели, такие как Vision Transformer (ViT) [5], Swin Transformer [7] и ConvNeXt [8]. Эти архитектуры по-разному обрабатывают пространственную и контекстную информацию, что может приводить к различной зависимости качества сегментации от применяемых стратегий аугментации; в частности, трансформеры часто требуют аккуратного подбора регуляризации и аугментаций при ограниченном объёме данных [5, 6].

В данной работе исследуется влияние соотношения оригинальных и аугментированных данных на качество сегментации узлов щитовидной железы на УЗИ-изображениях. Основной акцент сделан на анализе зависимости метрик качества от параметра R — числа аугментированных версий, формируемых из одного оригинального изображения с использованием фиксированного набора аугментаций — для различных архитектур энкодеров. Целью исследования является проверка гипотезы о существовании критического значения R, после которого дальнейшее увеличение доли аугментированных данных приводит к ухудшению качества сегментации.

II. ПОСТАНОВКА ЗАДАЧИ

Рассматривается задача бинарной сегментации узлов щитовидной железы на ультразвуковых изображениях. Пусть задано конечное множество

оригинальных ультразвуковых изображений X и соответствующих им вручную размеченных масок узлов Y:

$$X = \{x_1, x_2, \dots, x_N\}, Y = \{y_1, y_2, \dots, y_N\}, \quad (1)$$

где N — число оригинальных снимков, $x_i \in \mathbb{R}^{H\times W\times C}$ — УЗИ-изображение, а $y_i \in \{0,1\}^{H\times W}$ — бинарная маска, определяющая область узла, где H и W — пространственные размеры изображения, C — число цветовых каналов. Предполагается, что пространственные размеры изображения и маски совпадают.

В условиях ограниченного объёма исходных медицинских данных для повышения обобщающей способности модели применяется аугментация. Аугментацию будем трактовать как отображение:

$$A: X \to \tilde{X}, \quad (2)$$

которое каждому оригинальному изображению x_i сопоставляет конечное множество его аугментированных версий:

$$A(x_i) = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_j, \dots, \tilde{x}_M\}, \quad (3)$$

где M — число аугментированных изображений, получаемых из одного оригинального снимка.

Для каждой аугментированной версии $\tilde{x}_j$ соответствующая ей маска $\tilde{y}_j$ преобразуется тем же геометрическим отображением. Таким образом, формируется множество аугментированных изображений:

$$\tilde{X} = \bigcup_{i=1}^{N} A(x_i), \quad |\tilde{X}| = M, \quad (4)$$

И множество всех аугментированных масок:

$$\tilde{Y} = \bigcup_{i=1}^{N} A(y_i), \quad |\tilde{Y}| = M \quad (5)$$

Совокупная обучающая выборка состоит из оригинальных и аугментированных данных:

$$X^{train} = X \cup \tilde{X}, \quad |X^{train}| = N + M, \quad (6)$$

$$Y^{train} = X \cup \tilde{Y}, \quad |Y^{train}| = N + M, \quad (7)$$

Введём параметр R:

$$R = \frac{M}{N}, \quad (8)$$

который характеризует число аугментированных изображений, приходящихся на один реальный снимок (то есть соотношение между аугментированными и оригинальными данными на уровне одного объекта). Тогда размер обучающей выборки имеет следующую размерность:

$$|X^{train}| = N + N \cdot R = N \cdot (1 + R), \quad (9)$$

$$|Y^{train}| = N + N \cdot R = N \cdot (1 + R), \quad (10)$$

Пусть f_θ — модель сегментации с параметрами θ, обучаемая на выборке X^{train}. Качество сегментации оценивается на фиксированной тестовой выборке X^{test} с помощью выбранной метрики ошибки $\mathcal{L}$ (или, эквивалентно, метрики качества Q).

Задача исследования формулируется следующим образом: найти такое значение $R = R^*$, при котором ошибка сегментации $\mathcal{L}(R)$ минимальна (или, эквивалентно, метрика качества $Q(R)$ максимальна), и проверить гипотезу о том, что при $R > R^*$ дальнейшее увеличение числа аугментированных данных приводит к ухудшению качества сегментации вследствие смещения распределения обучающей выборки.

III. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

A. Аугментация данных

Цель аугментации — повысить обобщающую способность модели, заставляя её быть инвариантной к несущественным преобразованиям входных данных. На практике из одного исходного УЗИ-снимка можно сформировать несколько новых, изменяя геометрию, фотометрические характеристики или внося шумовые искажения; при этом для модели такие примеры выступают как различные наблюдения одного и того же объекта.

В проведённых экспериментах использовался единая последовательность шагов аугментации, включающий набор преобразований, сгруппированных по следующим категориям.

Геометрические преобразования:

- *Случайные повороты:* поворот изображения на 90° или на произвольный угол до 359°.
- *Аффинные преобразования:* случайное смещение, масштабирование и вращение, реализуемые как единое преобразование.
- *Случайное кадрирование и изменение размера:* вырезка случайной области фиксированного размера 224×224 пикселя с последующим масштабированием до 256×256 пикселей.

Преобразования цвета и освещения:

- *Коррекция яркости и контраста:* случайное изменение яркости и контрастности изображения.
- *Преобразования в HSV:* случайная коррекция оттенка, насыщенности и значения (яркости) в цветовой модели HSV.
- *Сдвиг каналов RGB:* независимое изменение интенсивности красного, зелёного и синего каналов.

Добавление шумов:

- *Гауссов шум:* добавление случайного аддитивного шума.
- *Шум, имитирующий ISO:* добавление шума, характерного для повышенных значений светочувствительности.

Имитация размытия и артефактов сжатия:

- *Размытие:* размытие в движении и гауссово размытие, имитирующие смазывание из-за движения камеры/объекта или снижение резкости.

- *Сжатие изображения:* имитация потери качества при JPEG-сжатии

Структурные искажения:

- *Coarse Dropout:* случайное “вырезание” прямоугольных областей изображения и заполнение их нулевым значением (чёрным цветом), что снижает зависимость модели от локальных пиксельных паттернов и стимулирует использование более устойчивых признаков.

Отражения:

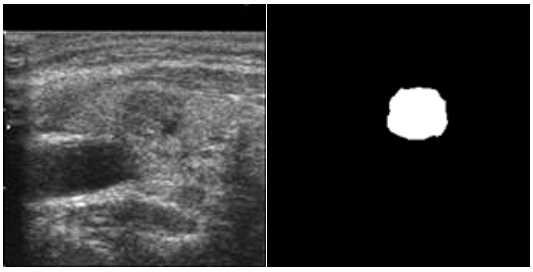
- *Горизонтальное и вертикальное отражение:* зеркальное отображение изображения по горизонтали и/или вертикали.

B. Обучающие данные

Для обучения и тестирования использован открытый датасет DDTI (Digital Database for Thyroid Ultrasound Images) [10] — база данных ультразвуковых изображений щитовидной железы, созданная в рамках совместного проекта Национального университета Колумбии, исследовательской группы CIM-LAB и Института медицинской диагностики (IDIME). Она содержит 637 УЗИ-изображений узлов щитовидной железы с размеченными полями. DDTI включает изображения с верифицированными диагнозами (злокачественные и доброкачественные узлы), подтверждёнными гистологическим исследованием образцов узлов, полученных с помощью тонкоигольной аспирационной биопсии (ТАБ). Снимки сопровождаются дополнительной информацией (метаданные пациента, параметры исследования).

Изображения имеют высокое разрешение, однако в них встречаются участки без диагностической информации (например, нерелевантные фоны), которые были предварительно удалены (пороговая фильтрация, свёртка). Подготовка данных включала также унификацию размера до 256×256 пикселей и нормализацию интенсивностей.

Набор предоставлен в формате пар «изображение — маска». Для экспериментов он был разделён на три части: 70% — обучающая выборка (445 снимков), 15% — валидационная и 15% — тестовая (96 в каждой). Такое разбиение позволило проводить обучение с контролем переобучения и объективно оценивать итоговую точность моделей.



а) б)

Рисунок 1 – Пример изображения щитовидной железы (а) и соответствующей маски узла (б)

C. Модели обучения

Для исследования влияния аугментации на различные архитектуры сегментации была реализована модульная модель, состоящая из двух функциональных компонентов: энкодера и декодера. Энкодер отвечает за извлечение признаков из входного изображения, постепенно сжимая пространственную информацию и формируя компактное представление объекта. Декодер восстанавливает пространственную структуру изображения из признакового представления, формируя карту сегментации с тем же разрешением, что и входное изображение. Такая декомпозиция позволяет варьировать архитектуру энкодера (например, ResNet, EfficientNet, или собственные блоки), не изменяя общую схему обработки данных и процедуру обучения, что важно для корректного сравнения влияния аугментации на разные архитектуры.

В качестве энкодера использовались предобученные модели из библиотеки timm (Torch Image Models), отвечающие за извлечение признаков из входного изображения. Энкодер преобразует вход 256×256 в компактное признаковое представление с пониженным пространственным разрешением и увеличенным числом каналов. В работе рассмотрены три архитектурно различающихся энкодера:

1. convnext_tiny — свёрточная архитектура ConvNeXt-Tiny [8], ориентированная на эффективное извлечение локальных признаков;
2. swin_tiny_patch4_window7_224 — иерархический трансформер Swin Transformer-Tiny [7] с локальным механизмом внимания;
3. vit_base_patch16_224 — Vision Transformer (ViT-Base) [5], обрабатывающий изображение как последовательность патчей.

Во всех экспериментах декодер фиксировался и имел минималистичную структуру: свёрточный слой 1×1 для согласования числа каналов и билинейный апсэмплинг для восстановления пространственного разрешения до размера входного изображения. Это позволяет интерпретировать различия в итоговых метриках преимущественно как следствие свойств выбранного энкодера, а не изменений в декодирующей части сети.

D. Параметры эксперимента

Применялась стратегия ранней остановки по метрике IoU. Параметр чувствительности (delta) был установлен на уровне 0.01, а параметр patience — 10 эпох. Обучение автоматически останавливалось в случае, если улучшение значения метрики IoU не превышало заданный порог в течение десяти последовательных эпох. Для обновления весов нейронной сети использовался адаптивный оптимизатор Adam с начальной скоростью обучения $lr = 10^{-4}$. Обучение проводилось с размером батча

8, максимальным числом эпох 50 и размером входного изображения 256×256.

E. Методы оценки качества сегментации

Для оценки качества сегментации использовались коэффициент Dice [12] и коэффициент Жаккарда (IoU) [11]. Пусть $P \subset \Omega$ — множество пикселей, отнесённых моделью к узлу, а $G \subset \Omega$ — множество истинных пикселей узла. Тогда индекс IoU определяется как:

$$IoU(P, G) = \frac{|P \cap G|}{|P \cup G|}, \qquad (11)$$

А коэффициент Dice (F1-score для масок) вычисляется по формуле:

$$\mathrm{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}, \qquad (12)$$

Концептуально обе метрики измеряют степень перекрытия предсказанной и истинной масок. IoU показывает, какую долю объединённой области составляет их пересечение. Dice оценивает, насколько хорошо совпадают две маски относительно их среднего размера. Таким образом, IoU и Dice показывают, насколько сильно предсказанная область совпадает с реальной.

Помимо метрик перекрытия, дополнительно рассчитывались ROC-AUC и PR-AUC, позволяющие оценить качество сегментации как задачи бинарной классификации пикселей. Пусть каждому пикселю $x \in \Omega$ модель сопоставляет вещественное значение $s(x) \in [0,1]$, интерпретируемое как вероятность принадлежности к узлу. Тогда, изменяя порог $t \in [0,1]$, получаем бинарное предсказание:

$$P_t = \{x \in \Omega \mid s(x) \geq t\}$$

Для каждого значения порога вычисляются:

$$TPR = \frac{TP}{TP + TN}, \qquad FPR(t) = \frac{FP}{FP + TN}, \qquad (13)$$

где TP — число истинно положительных пикселей, FP— число ложно положительных, FN— число пропущенных пикселей узла, TN— число правильно классифицированных фоновых пикселей.

Площадь под ROC-кривой определяется как:

$$ROC - AUC = \int_0^1 TPR(FPR)dFPR\,, \qquad (14)$$

ROC-AUC показывает, насколько хорошо модель в целом отделяет пиксели объекта от пикселей фона. Однако при сильном дисбалансе классов, когда фон существенно преобладает над узлом ROC-AUC может быть избыточно оптимистичной. Поэтому дополнительно рассчитывалась PR-кривая. Для каждого порога вычисляются:

$$\mathrm{Precision}(t) = \frac{TP}{TP + FP}, \qquad (15)$$

$$\mathrm{Recall}(t) = \frac{TP}{TP + FN}, \qquad (16)$$

Площадь под PR-кривой (Average Precision, PR-AUC) вычисляется как:

$$PR - AUC = \int_0^1 \mathrm{Precision}(\mathrm{Recall})d\mathrm{Recall}\,, \qquad (17)$$

PR-AUC отражает, насколько хорошо модель выделяет пиксели объекта, не захватывая лишние пиксели фона, при различных значениях порога классификации.

В качестве функции потерь использовалась сумма Dice Loss и бинарной кросс-энтропии (Binary Cross Entropy, BCE):

$$\mathcal{L}(P, G) = \mathcal{L}_{\mathrm{Dice}}(P, G) + \mathcal{L}_{\mathrm{BCE}}(P, G), \qquad (18)$$

Такое сочетание позволяет одновременно учитывать как ошибки на уровне совпадения масок (Dice Loss), так и ошибки бинарной классификации отдельных пикселей (BCE).

IV. АНАЛИЗ РЕЗУЛЬТАТОВ

В настоящей работе представлены результаты экспериментов по оценке влияния R — числа аугментированных изображений, приходящихся на один реальный снимок — на качество сегментации для трёх архитектур энкодеров. На графиках ниже показана зависимость метрик качества от параметра R.

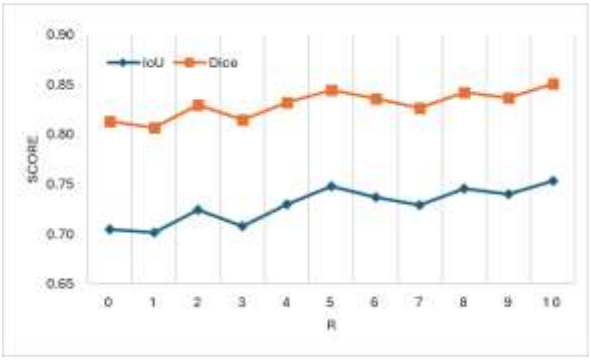


Рисунок 2 – Зависимость метрик Dice и IoU от параметра R для ConvNeXt-Tiny.

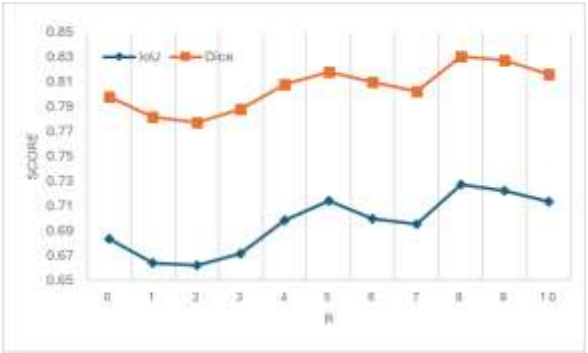


Рисунок 3 – Зависимость метрик Dice и IoU от параметра R для Swin Transformer-Tiny.

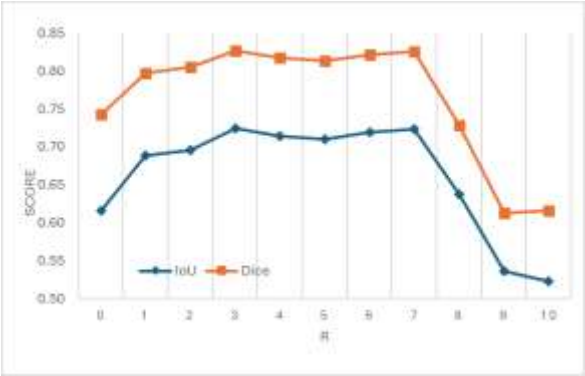


Рисунок 4 – Зависимость метрик Dice и IoU от параметра R для ViT-Base.

Табл.1. Метрики качества моделей при отсутствии аугментации (R = 0)

Метрика	*ConvNeXt-Tiny*	*Swin-Tiny*	*Vit-Base*
IoU	0.705	0.683	0.616
Dice	0.813	0.798	0.743
ROC-AUC	0.984	0.982	0.961
PR-AUC	0.905	0.898	0.880

Табл.2. Метрики качества моделей при наилудшем соотношении аугментации (R = R_{max})

Метрика	*ConvNeXt-Tiny*	*Swin-Tiny*	*Vit-Base*
R_{max}	10	8	3
IoU	0.754	0.727	0.725
Dice	0.851	0.831	0.827
ROC-AUC	0.990	0.989	0.982
PR-AUC	0.937	0.936	0.913

Табл.3. Метрики качества моделей при наихудшем соотношении аугментации (R = R_{min})

Метрика	*ConvNeXt-Tiny*	*Swin-Tiny*	*Vit-Base*
R_{min}	1	2	10
IoU	0.702	0.662	0.523
Dice	0.807	0.777	0.615
ROC-AUC	0.976	0.979	0.831
PR-AUC	0.893	0.890	0.781

Результаты, представленные в Таблицах 1—3, показывает, что влияние аугментации данных на качество сегментации носит выраженно немонотонный характер и существенно зависит от архитектуры используемой модели. Для всех рассмотренных моделей наблюдается наличие оптимального значения коэффициента аугментации R_{max}, при котором достигается максимум метрик IoU и Dice, а также области значений R_{min}, при которых качество сегментации деградирует.

С точки зрения теории машинного обучения данный эффект можно интерпретировать в рамках компромисса bias–variance. Умеренная аугментация увеличивает разнообразие обучающей выборки и снижает дисперсию модели, улучшая её обобщающую способность. Однако при чрезмерной аугментации происходит смещение распределения данных (*data distribution shift*), в результате чего модель начинает оптимизироваться под синтетические паттерны, плохо соответствующие реальным ультразвуковым изображениям, что приводит к росту ошибки на тестовой выборке.

Для архитектуры ConvNeXt-Tiny оптимальное значение $R_{max} = 10$ обеспечивает стабильный прирост качества (IoU +6%, Dice +5%), при этом деградация при неблагоприятных значениях R минимальна (около 1%). Такая устойчивость может быть объяснена сильным индуктивным смещением, присущим свёрточным нейронным сетям: локальностью, трансляционной эквивалентностью и иерархическим извлечением признаков. Кроме того, использование современных механизмов нормализации, таких как Global Response Normalization, способствует стабилизации обучения при наличии шумов и геометрических искажений. Аналогичные выводы о высокой робастности свёрточных архитектур в медицинской сегментации приводятся в ряде работ, посвящённых анализу аугментаций и обучению моделей на медицинских данных [3, 9].

Модель Swin Transformer-Tiny демонстрирует промежуточное поведение. С одной стороны, она выигрывает от аугментации (IoU +7%, Dice +4% при $R_{max} = 8$), с другой — показывает более заметную деградацию при неблагоприятных значениях R по сравнению с ConvNeXt. Это согласуется с архитектурными особенностями Swin Transformer, который сочетает трансформерный механизм внимания с локальностью и иерархией за счёт оконного внимания [7]. Подобный компромисс снижает чувствительность к небольшим искажениям, однако не устраняет полностью зависимость от распределения данных, характерную для трансформерных моделей; примеры применения Swin-подобных подходов в медицинской сегментации описаны, в частности, в работе о сочетании Swin Transformer и UperNet [16].

Анализ графика зависимости метрик от коэффициента аугментации показывает наличие трёх локальных падений IoU и Dice на Swin Transformer-Tiny. Эти снижения отражают архитектурно-зависимое влияние аугментации: при слишком большой аугментации происходит переобучение на синтетических паттернах, когда модель подстраивается под комбинации трансформаций, редко встречающиеся в реальных УЗИ-изображениях, что ухудшает результаты на тестовой выборке. Архитектурные особенности Swin

Transformer-Tiny позволяют частично компенсировать локальные искажения через оконное внимание, но модель слабее устойчива к глобальным трансформациям по сравнению с ConvNeXt, что проявляется как промежуточное поведение: сильный эффект аугментации при умеренных значениях R, но чувствительность к неблагоприятным значениям R, вызывающая падения метрик.

Наиболее выраженная зависимость от коэффициента аугментации наблюдается для Vision Transformer (ViT-Base). При оптимальном значении $R_{\max} = 3$ достигается максимальный прирост качества среди всех моделей (IoU $+16\%$, Dice $+12\%$), что указывает на высокую эффективность умеренной аугментации для трансформеров в условиях ограниченного объёма данных. Однако при агрессивной аугментации ($R_{\min} = 10$) качество резко снижается (-16% по обеим метрикам), причём деградация превосходит эффект отсутствия аугментации вовсе.

Данный результат согласуется с выводами работ, показывающих, что ViT обладает слабым встроенным индуктивным смещением и требует либо больших объёмов данных, либо тщательно подобранных стратегий регуляризации и аугментации [5, 6].

Также многочисленные гибридные архитектуры, использующие трансформеры как энкодеры в сегментационных сетях (например, TransUNet), демонстрируют конкурентоспособные результаты на медицинских данных при корректной настройке обучающего пайплайна [15].

В целом полученные результаты подтверждают, что аугментация данных не является универсальным способом повышения качества сегментации и должна рассматриваться как архитектурно-зависимый инструмент. С практической точки зрения это означает, что использование максимально возможного числа аугментированных изображений может быть не только неэффективным, но и вредным, особенно для трансформерных моделей. Выбор стратегии аугментации должен учитывать баланс между увеличением разнообразия данных и сохранением статистических свойств реального распределения медицинских изображений.

ЗАКЛЮЧЕНИЕ

Результаты демонстрируют, что различные энкодеры нейронной сети по-разному реагируют на увеличение доли аугментированных данных. Свёрточная модель ConvNeXt и гибридная архитектура Swin Transformer показывают наибольшую устойчивость к росту R, тогда как Vision Transformer оказывается более чувствительным к переизбыточной аугментации. Оптимальные значения коэффициента для приведенного в публикации датасета составили $R_{\max} = 10$ для ConvNeXt-Tiny, $R_{\max} = 8$ для Swin Transformer-Tiny и $R_{\max} = 3$ для ViT-Base, наблюдалось повышение качества вплоть до 16%; при этом при неблагоприятных значениях $R_{\min}$ наблюдалось снижение качества вплоть до -16% для ViT-Base.

Таким образом, полученные результаты подтверждают гипотезу о существовании оптимального значения R и подчёркивают необходимость подбора стратегии аугментации с учётом архитектуры модели и объёма исходного датасета. Практическая рекомендация по итогам работы заключается в том, что увеличение доли аугментированных данных должно выполняться контролируемо: для CNN и гибридных моделей допустим более широкий диапазон R, тогда как для ViT-подобных предпочтительна умеренная аугментация и тщательная настройка обучающего пайплайна.

БЛАГОДАРНОСТИ

Авторы благодарят Высшую инжиниринговую школу Национального исследовательского ядерного университета МИФИ за поддержку в публикации материалов этого исследования.

БИБЛИОГРАФИЯ

[1] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation // MICCAI. 2015. P. 234–241.

[2] Shorten C., Khoshgoftaar T. A survey on Image Data Augmentation for Deep Learning // Journal of Big Data. 2019. Vol. 6. Art. 60.

[3] Kim M., Bae H.-J. Data Augmentation Techniques for Deep Learning-Based Medical Image Analyses // J Korean Soc Radiol. 2020. Vol. 81(6). P. 1290–1304.

[4] García S., et al. Data augmentation for medical imaging: A systematic literature review. 2022.

[5] Dosovitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929. 2020.

[6] Touvron H., Cord M., Douze M., Massa F., Sablayrolles A., Jégou H. Training data-efficient image transformers & distillation through attention // ICML. 2021. P. 10347–10357.

[7] Liu Z., Lin Y., Cao Y., et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows // ICCV. 2021. P. 10012–10022.

[8] Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S. A ConvNet for the 2020s. arXiv:2201.03545. 2022.

[9] Lee L. H., Gao Y., Noble J. A. Principled Ultrasound Data Augmentation for Classification of Standard Planes. arXiv:2103.07895. 2021.

[10] Pedraza L., Vargas C., Narváez F., Durán O., Muñoz E., Romero E. An Open Access Thyroid Ultrasound Image Database // Proc. SPIE. 2015. Vol. 9287. Art. 92870W. DOI: 10.1117/12.2073532.

[11] Kim K., Lee H. S. Probabilistic Anchor Assignment with IoU Prediction for Object Detection // European Conference on Computer Vision. Cham: Springer International Publishing, 2020. P. 355–371.

[12] Shamir R. R., et al. Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations. arXiv:1906.11031. 2019.

[13] M. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review // J. Med. Syst. 2024. Vol. 48. Article 84.

[14] Advantages of transformer and its application for medical image segmentation: a survey // BioMed Eng. Online. 2024.

[15] Chen J., et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv. 2021.

[16] Medical image segmentation by combining feature enhancement Swin Transformer and UperNet // Sci. Rep. 2025.

Статья получена 17 марта 2026.

Холод Данила Витальевич, Национальный Исследовательский Ядерный Университет МИФИ, магистрант email:danila.kholod@gmail.com

Боброва Елизавета Витальевна, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, email:EVBobrova@mephi.ru

Зайцев Константин Сергеевич, Национальный Исследовательский Ядерный Университет МИФИ, профессор, email:KSZaytsev@mephi.ru

Трухин Алексей Андреевич, ФГБУ «НМИЦ эндокринологии» Минздрава России, медицинский физик, email:alexey.trukhin12@gmail.com

Трошина Екатерина Анатольевна, ФГБУ «НМИЦ эндокринологии» Минздрава России, чл. корр., директор Института клинической эндокринологии, email:troshina@inbox.ru

Adaptive Control of Data Augmentation for Thyroid Ultrasound Image Segmentation

D. V. Kholod, E. V. Bobrova, K. S. Zaitsev, A. A. Trukhin, E. A. Troshina

Abstract - The aim of this study was to investigate the influence of the proportion of real and augmented data on the performance of thyroid nodule segmentation in ultrasound images. For this purpose, a parameter R was introduced, defined as the ratio of the number of augmented images to the number of original images, and its effect on segmentation quality was analyzed for different encoder architectures. The study considered the formation of training datasets with varying shares of augmented data, followed by model training and evaluation of segmentation results on the DDTI dataset. Segmentation performance was assessed using standard evaluation metrics, namely the Dice coefficient and Intersection over Union (IoU). The results show that the dependence of segmentation quality on the parameter R is non-monotonic: there exists an optimal value R_{max} at which the best performance is achieved, as well as a region R_{min} where the segmentation quality deteriorates. It is demonstrated that the ViT-Base architecture is the most sensitive to excessive augmentation, whereas ConvNeXt-Tiny and Swin Transformer-Tiny exhibit greater robustness to changes in the proportion of synthetic data.

Keywords: *segmentation, ultrasound images, thyroid gland, data augmentation, Dice coefficient, ConvNeXt, Swin Transformer, Vision Transformer.*

REFERENCIES

[1] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation // MICCAI. 2015. P. 234–241.

[2] Shorten C., Khoshgoftaar T. A survey on Image Data Augmentation for Deep Learning // Journal of Big Data. 2019. Vol. 6. Art. 60.

[3] Kim M., Bae H.-J. Data Augmentation Techniques for Deep Learning-Based Medical Image Analyses // J Korean Soc Radiol. 2020. Vol. 81(6). P. 1290–1304.

[4] García S., et al. Data augmentation for medical imaging: A systematic literature review. 2022.

[5] Dosovitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929. 2020.

[6] Touvron H., Cord M., Douze M., Massa F., Sablayrolles A., Jégou H. Training data-efficient image transformers & distillation through attention // ICML. 2021. P. 10347–10357.

[7] Liu Z., Lin Y., Cao Y., et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows // ICCV. 2021. P. 10012–10022.

[8] Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S. A ConvNet for the 2020s. arXiv:2201.03545. 2022.

[9] Lee L. H., Gao Y., Noble J. A. Principled Ultrasound Data Augmentation for Classification of Standard Planes. arXiv:2103.07895. 2021.

[10] Pedraza L., Vargas C., Narváez F., Durán O., Muñoz E., Romero E. An Open Access Thyroid Ultrasound Image Database // Proc. SPIE. 2015. Vol. 9287. Art. 92870W. DOI: 10.1117/12.2073532.

[11] Kim K., Lee H. S. Probabilistic Anchor Assignment with IoU Prediction for Object Detection // European Conference on Computer Vision. Cham: Springer International Publishing, 2020. P. 355–371.

[12] Shamir R. R., et al. Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations. arXiv:1906.11031. 2019.

[13] M. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review // J. Med. Syst. 2024. Vol. 48. Article 84.

[14] Advantages of transformer and its application for medical image segmentation: a survey // BioMed Eng. Online. 2024.

[15] Chen J., et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv. 2021.

[16] Medical image segmentation by combining feature enhancement Swin Transformer and UperNet // Sci. Rep. 2025.