УДК 004.932.2 **DOI** 10.24412/2413-7383-2025-4-39-71-83

Б. В. Павленко
Федеральное государственное бюджетное научное учреждение
«Институт проблем искусственного интеллекта», г. Донецк
283048, г. Донецк, ул. Артема, 118 б

# ПОДХОД К МУЛЬТИМОДАЛЬНОМУ ОБЪЕДИНЕНИЮ ДАННЫХ В ЗАДАЧАХ РЕГРЕССИИ И КЛАССИФИКАЦИИ НА СНИМКАХ С БПЛА

B.V. Pavlenko
Federal State Budgetary Scientific Institution «Institute of Artificial Intelligence Problems»
283048, Donetsk, Artema str, 118-b

# AN APPROACH TO MULTIMODAL DATA FUSION IN TASKS OF REGRESSION AND CLASSIFICATION IN DRONE IMAGES

В данной работе предлагается метод мультимодального объединения признаков с использованием изображений с беспилотной воздушной платформы (БПЛА) и закодированных метаданных. Рассматривается применимость предлагаемого подхода к задачам регрессии значений положения БПЛА и классификации объектов на снимках с минимальным набором известных метаданных. Для регрессии введён механизм интерполяции значений высоты, позволяющий более точно связать разницу визуальных представлений снимка с его высотой и обеспечить большую устойчивость к изменениям высоты и углов поворота камеры БПЛА. Для регулирования вклада текстовой модальности использован метод, обеспечивающий случайное игнорирование текстовых признаков и уменьшающий переобучение на текстовую модальность дополнительно с регуляризацией ошибкой выравнивания признаков и коэффициентом ее влияния на общую ошибку модели. Для классификации применен тот же подход с сравнением нескольких вариантов текстовых шаблонов. Проведены эксперименты на выборках аэрофотосъёмки с вариацией высоты и углов ориентации камеры, которые показали, что предложенный подход превосходит простую конкатенацию признаков, подтверждены выдвинутые гипотезы об улучшении обобщающей способности модели с применением рассматриваемых методов.

**Ключевые слова:** мультимодальные нейронные сети, регрессия, классификация, объединение признаков, позиционирование, БПЛА.

In this paper, a method of multimodal feature combination is proposed using images from an unmanned aerial vehicle (UAV) and encoded metadata. The applicability of the proposed approach to the tasks of UAV position regression and object classification in images with a minimal set of known metadata is considered. For regression, a mechanism for interpolating altitude values is introduced, which allows for a more accurate association between the visual representation of an image and its altitude, and provides greater stability to changes in altitude and camera rotation angles of the UAV. To regulate the contribution of textual modality, a method is used that ensures random ignoring of textual features and reduces overfitting on textual modality in addition to regularization by the feature alignment error and its coefficient of influence on the overall model error. The same approach is used for classification, with the comparison of several variants of textual templates. Experiments were conducted on the following datasets.

**Key words:** multimodal neural networks, regression, classification, feature fusion, positioning, UAV.

## Введение

Одним из ключевых вызовов в разработке автономных систем навигации для беспилотных летательных аппаратов (БПЛА) является обеспечение высокой точности оценки их положения при ограниченных вычислительных ресурсах [1], [2]. Классические методы навигации, основанные на GPS и инерциальных системах, показывают себя менее надежными в условиях сценариев ухудшения или полной потери сигнала, в том числе при полетах в городских каньонах или в условиях радиоэлектронного противодействия. Это стимулирует развитие альтернативных подходов, опирающихся на сенсорные данные с бортовых камер, методы компьютерного зрения и глубокого обучения.

В последние годы большое внимание уделяется исследованию архитектур, способных эффективно объединять визуальные признаки изображения с дополнительными источниками информации, семантическими описаниями сцены, геометками, метеоданными и прочими контекстными метаданными. Ряд работ [3-6] демонстрируют, что включение дополнительных семантических сигналов в модель приводит к росту точности предсказаний и улучшению устойчивости модели к шуму и вариативности входных данных, которая может быть выражена различиями в содержании изображаемой сцены или различиями перспективы. При этом данные результаты наблюдались не только в экспериментах в сфере компьютерного зрения для беспилотных устройств, перекрестной геолокализации, а в том числе в визуально-языковом обнаружении объектов и визуальном понимании сцены [5], [6].

В контексте использования семантических дополнений возникает задача обогащения открытых наборов данных [7], в случаях отсутствия таких метаданных, как описание местности, высота снимка, информация о положении камеры. При сборе синтетических наборов данных с использованием специализированных сред [8] полнота метаданных зависит от используемой среды. При поиске реальных наборов данных проблема неполноты метаданных может оставаться нерешенной. Использование таких адаптированных для мультимодального обучения наборов данных является важной частью экспериментов в сфере решения задач повышения автономности беспилотной навигации [9].

В данной работе демонстрируются результаты экспериментов с применением метода объединения признаков для регрессии значений положения БПЛА, и классификации объектов на снимках с дополнительной семантической и числовой информацией, передаваемой через текстовое описание. Целью исследования являлся поиск **легковесного мультимодального решения**, пригодного для использования в реальном времени на бортовых платформах с ограниченными вычислительными ресурсами.

Мы рассматриваем несколько вариантов объединения признаков для анализа их влияние на итоговую ошибку модели на целевых задачах, дополнительные методы улучшения процесса обучения, а также вариации текстовых шаблонов. Для оценки предложенного подхода проведены эксперименты с различными схемами объединения и методами регуляризации, а также исследовано влияние текстовой семантики на устойчивость модели в условиях, когда на режиме работы доступен только визуальный сигнал.

## Наборы данных

Для экспериментов необходимы наборы данных, содержащих не только разнообразные снимки местности со спутника и БПЛА, но также и дополнительную семантику в виде метаданных позиционирования летательного аппарата.

GTA-UAV [10] – синтетический набор данных, созданный на основе игрового движка Grand Theft Auto V с использованием сценариев автоматического управления камерой и БПЛА. Набор данных включает в себя изображения высокого разрешения, снятые с различных высот, углов наклона и направлений полета, точные метки положения и ориентации (позиции в координатах, углы Эйлера и кватернионы), синхронизированные с каждым кадром, семантические маски и классы объектов (здания, дороги, растительность, транспорт) – что делает набор пригодным для задач регрессии, сегментации и обнаружения объектов.

VisLoc [11] – набор данных для задач визуальной локализации и регрессии положения. Набор содержит 6742 снимка с реальных бортовых камер в разнообразных сценариях полета – городские кварталы, пригороды, сельские территории, и учитывает вариации освещенности, шум сенсора, блики и другие артефакты. Высота полета представлена в диапазоне 400 – 2000 м.

UC Merced Land Use Dataset [12] – набор данных, созданный для задач классификации и анализа типов землепользования. Он состоит из 2100 изображений. Каждый снимок вручную аннотирован и отнесён к одному из 21 класса, таких как «жилая зона», «аэропорт», «ферма», «лес», «река», «дорога» и др. Изображения в наборе сбалансированы по классам (по 100 снимков каждого типа) и характеризуются разнообразием текстур, структуры и контекста сцены.

## Методология

Благодаря способности трансформеров кодировать входные данные практически любой модальности, некоторые модели реализуют подход single-stream multimodal transformer или unified encoder. При таком подходе у модели отсутствуют отдельные базовые сети извлечения признаков (backbone) или отдельные энкодеры, данные кодируются одним энкодером, как например в ViLT [13]. Примерами иного подхода являются LXMERT [14], CLIP [15], использующие отдельные ветви кодирования для разных модальностей, с последующим применением перекрестного внимания (Cross-Attention) или контрастивного обучения.

В данной работе был использован подход с применением экстракторов признаков, и разделением сети на две ветви – обучающую и валидационную с двумя регрессорами. Изначально была выдвинута гипотеза о том, что внедрение дополнительных числовых данных повысит численные показатели модели в задаче регрессии, т.к. различия высоты снимков и углов наклона камеры явно отражаются в визуальном контексте изображения, а значит, можно подкрепить модель дополнительным контекстом. Отсюда следовала вторая гипотеза о применимости интерполяции значения высоты для повышения стабильности предсказаний для задачи регрессии. Она состоит из округления высоты до ближайшего шага (1) и последующим ограничением диапазона (2):

$$h_q = round\left(\frac{h}{d_{step}}\right) \cdot d_{step}, \tag{1}$$

где $h$ – исходная высота, $d_{step}$ – шаг интерполяции. Итоговая формула выглядит следующим образом:

$$h_{out} = max\big(h_{minDs}, \min\left(h_{dsMax}, h_q\right)\big), \tag{2}$$

где $h_{minDs}$ и $h_{maxDs}$ – минимальное и максимальное значение высота в наборе данных.

Перед подачей в текстовый энкодер при обучении метаданные внедряются в текстовые шаблоны. Для регрессии, значения высоты снимка внедряются в текстовый шаблон вида «This is a drone aerial image taken from a height of *h* meters». Для классификации были разработаны следующие шаблоны: «The target is a <class> in an aerial photo», «An aerial photo contains a <class>», «An aerial view showing a <class>».

Для снижения затрат на обучение был введен механизм text-dropout, при котором с заданной вероятностью $p_{textDrop}$ текстовые признаки игнорировались и модель получала на вход только изображение. В соответствии с этим адаптировалась итоговая функция потерь (3) т.к. при подаче обеих модальностей к ошибке регрессии $L_{reg}$ прибавляется ошибка выравнивания признаков $L_{align}$ – L2-регуляризация между векторными представлениями изображения $v_{img}$ и текста $v_{text}$ умноженная на коэффициент влияния $\lambda$ (4).

$$L = \begin{cases} L_{reg}, & p_{textDrop} = true \\ L_{reg} + \lambda \cdot L_{align}, & p_{textDrop} = false \end{cases}, \quad (3)$$

где $L_{reg}$ – комбинация Hubert Loss и Focal Loss, которая обеспечивает меньшую чувствительность к выбросам и дисбалансу примеров, а $\lambda \cdot L_{align}$:

$$L_{align} = 1 - \frac{\langle v_{img},\ v_{text} \rangle}{\|v_{img}\| \cdot \|v_{text}\|}, \quad (4)$$

## Архитектура

Модели состоят из двух ветвей обработки входных данных, которые работают в разных режимах в зависимости от режима обучения или теста. Каждая ветвь имеет свой экстрактор признаков для своей модальности: сверточную сеть (CNN) StripNet-small [16] для изображений и языковую модель MobileClip2-s0 [17] с частичной разморозкой, за счет чего обучаемыми остаются только слои текстовой проекции и последний блок трансформера. После экстрактора следует блок информативного усиления признаков – шея – включающая в себя блок, вдохновленный архитектурой UAV-YOLOv12 [18] для выделения релевантных признаков при сохранении низкой вычислительной стоимости (рисунок 2). Multi-Pool агрегирует разномасштабные признаки перед объединением. Такое мульти-адаптивное сжатие позволило избежать проблемы дорогих вычислений, возникшей при подаче на разные ветки и шею сырых тензоров из CNN с огромными ядрами.

В конце модель разветвляется на две головы в зависимости от режима обучения или теста. В режиме обучения используются текстовый экстрактор и объединение, результат которого подается в голову для объединенных признаков. В тестовом режиме SLM не работает и текстовые признаки не извлекаются, что существенно экономит ресурсы и дает ощутимый прирост в скорости. Общая архитектура представлена на рисунке 1.

Шея (рисунок 2) состоит из сверток, C3k2-блоков и блоков A2C2F-Attention (рисунок 3), комбинирующих пространственное и канальное внимание, позволяя модели фокусироваться на важных областях и подавлять шум в фоновых регионах. C3k2-блоки представляют собой модифицированные C3-блоки (Cross Stage Partial) в виде бутылочных горлышек с двумя свертками, где используются два сверточных ядра (k=2) для баланса между глубиной сети и эффективностью вычислений. Они помогают лучше интегрировать контекст и сохранить детальные признаки для мелких объектов, что критично для обнаружения на UAV-снимках.

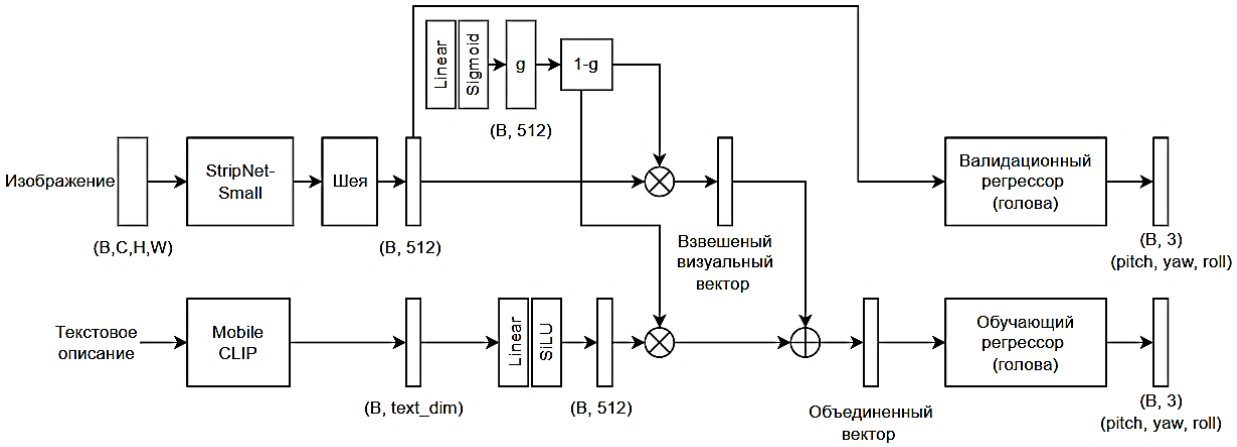


Рисунок 1 – Архитектура регрессионной модели с раздельными ветвями обработки объединенных и визуальных признаков

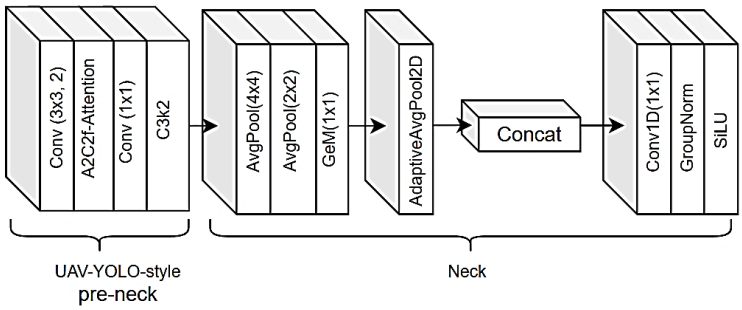


Рисунок 2 – Архитектура шеи с блоком предобработки признаков с блоками YOLO и адаптивным усредняющим сжатием разноразмерных карт признаков

Голова AngleHead (рисунок 3) принимает на вход размерности (B, D) результат объединения или признаки изображения после шеи. Так как для объединения не использовалась конкатенация, размерность выходного вектора не удваивалась, оставаясь равной 512 в независимости от режима модели. Выход головы имеет размерность равную 3 по числу регрессионных параметров.

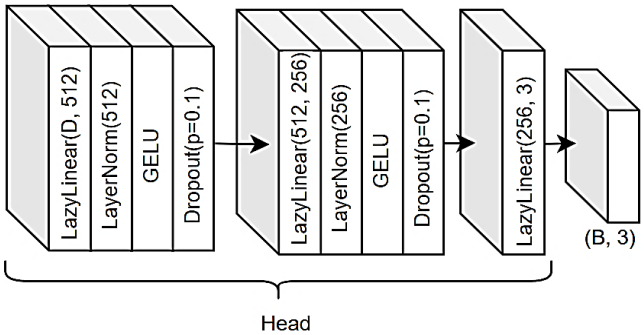


Рисунок 3 – Подробная схема головы-регрессора значений углов

## Обучение

При обучении сравнивалось объединение через конкатенацию (5) или gate-сумму с коэффициентом $g$ (6), которое происходит после шеи, что позволяет работать с компактными и семантически информативными признаками:

$$fused = concat(v_{img}, v_{text}), \quad (5)$$

$$fused = v_{img}g + v_{text}(1 - g), \quad (6)$$

Обучение модели проводилось в течение 20 эпох с batch-size=32, что позволило достичь стабильной сходимости без переобучения. В качестве оптимизатора использовался Ranger [19], объединяющий преимущества Rectified Adam (RAdam) [20] и Lookahead [21], что обеспечивает более плавную и надёжную оптимизацию за счёт адаптивной скорости обучения и усреднения весов по «медленным» и «быстрым» шагам. Для управления скоростью обучения применялся планировщик скорости обучения WarmupReduceLROnPlateauScheduler, выполняющий постепенный «прогрев» скорости обучения (learning rate) на начальных итерациях, а затем автоматически снижал его при стабилизации валидационной метрики, способствуя лучшему обучению и повышению финальных метрик модели. Для управления влиянием ошибки выравнивания признаков в режиме обучения использовался коэффициент $\lambda$ при различных значениях $p_{textDropout}$. Для более стабильной до-настройки (fine-tuning) текстового энкодера было принято решение использовать «частичную разморозку». Для сохранения чувствительности к домену оставлен последний блок трансформера, а для адаптации векторных представлений оставлены слои проекции. Реализованная таким образом до-настройка позволяет не терять предобученные представления модели.

## Эксперименты

Эксперименты проводились с использованием StripNet в качестве CNN-экстрактора на основании результатов из таблицы 1:

Таблица 1 – Сравнение показателей моделей StripNet-small и InceptionNext-tiny на наборах данных GTA-UAV и VisLoc.

| Backbone | Набор данных | test loss | train loss | test MAE | train MAE |
|---|---|---|---|---|---|
| StripNet-small | GTA-UAV | 0.05 | 0.05 | 0.1 | 0.11 |
| InceptionNext-tiny | GTA-UAV | 0.1 | 0.08 | 0.21 | 0.17 |
| StripNet-small | VisLoc | 0.11 | 0.10 | 0.22 | 0.22 |
| InceptionNext-tiny | VisLoc | 0.11 | 0.11 | 0.21 | 0.24 |

Основные эксперименты состояли в сравнении методов объединения конкатенацией (3) и суммированием с gate-коэффициентом $g$ (4) при различных значениях $p_{textDropout}$ и интерполяции высоты с шагом равным 5 или 10. Результаты показали, что $p_{textDropout} > 0.5$ перестает давать улучшения показателей.

В первом наборе экспериментов модель обучалась на 50 эпохах для VisLoc с объединением конкатенацией и ошибками регрессии (3) (значение $\lambda$ в описании к изображениям указывает на прибавление $L_{align}$ к ошибке).

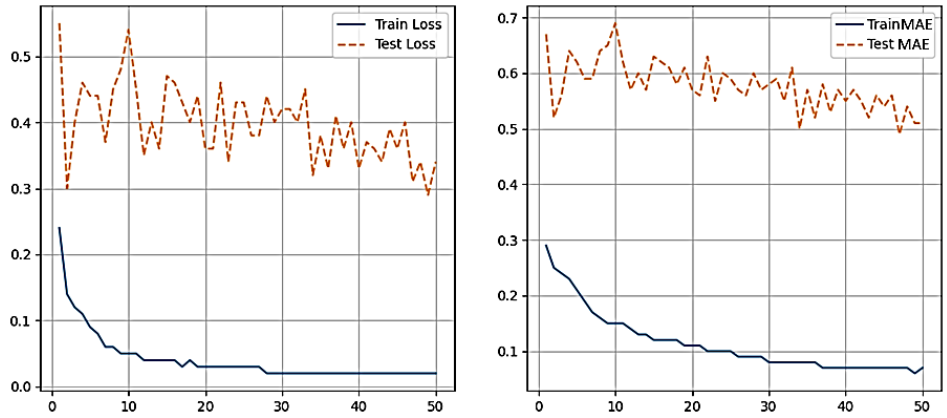


Рисунок 4 – Изменение ошибки регрессии $L_{reg}$ и MAE без применения text-dropout в течении обучения

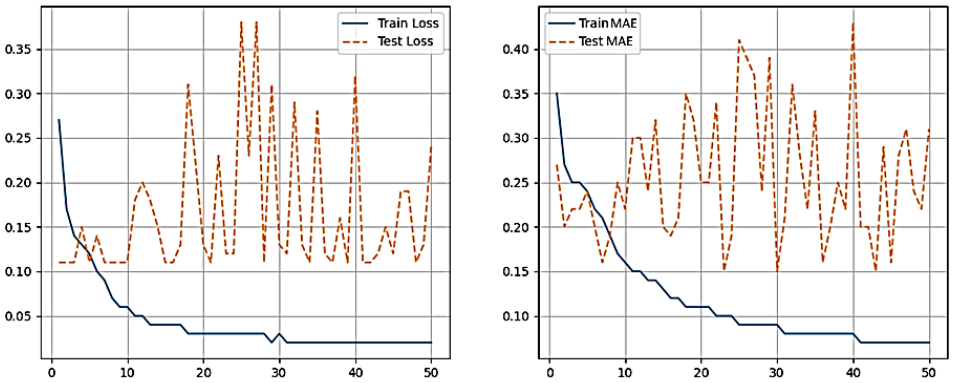


Рисунок 5 – Изменение ошибки регрессии $L_{reg}$ и MAE с применением text-dropout с вероятностью $p_{textDropout} = 0.5$

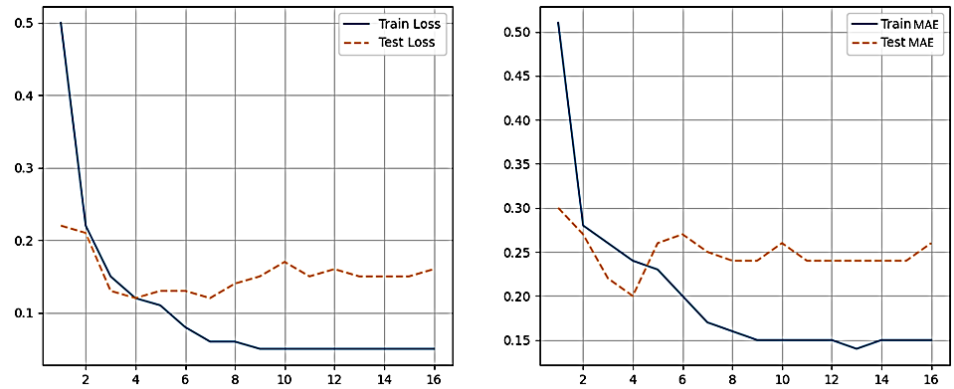


Рисунок 6 – Изменение ошибки регрессии $L_{reg}$ и MAE с применением text-dropout с вероятностью $p_{textDropout} = 0.3$, коэффициентом влияния ошибки выравнивания признаков $\lambda = 0.1,$ шагом интерполяции высоты $d_{step} = 5$

Во втором наборе экспериментов модель обучалась на 50 эпохах с объединением суммированием с весовым gate-коэффициентом и ошибкой регрессии $L_{reg} + \lambda \cdot L_{align}$ (1). Также использовался механизм интерполяции высоты со значением $d_{step}$ равным 5 и 10.

Таблица 3 – Минимальные значения ошибки и MAE при gate-суммировании

| $p_{textDropout}$ | Train Loss | Train MAE | Test Loss | Test MAE | $\lambda$ | $d_{step}$ |
|---|---|---|---|---|---|---|
| 0.2 | 0.03 | 0.10 | 0.11 | 0.19 | 0.5 | 10 |
| 0.2 | 0.03 | 0.10 | 0.13 | 0.15 | 0.5 | 10 |
| 0.3 | 0.02 | 0.08 | 0.11 | 0.17 | 0.1 | 5 |
| 0.3 | 0.04 | 0.13 | 0.12 | 0.20 | 0.5 | 5 |

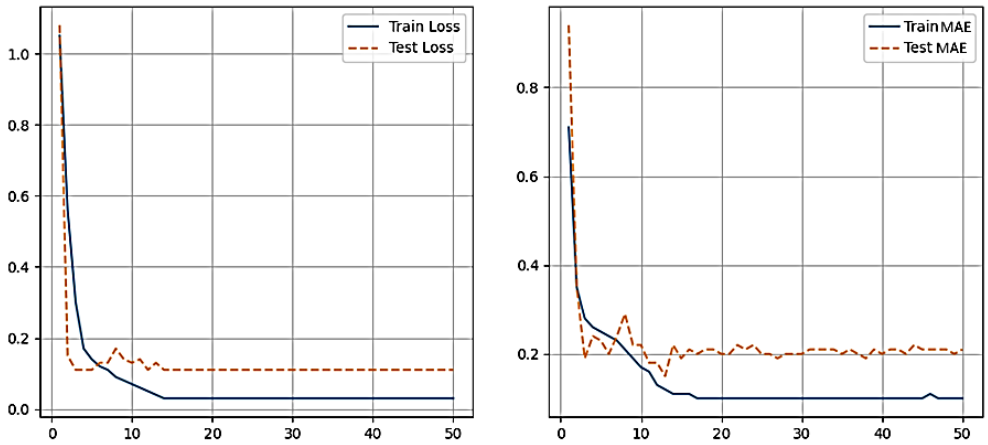


Рисунок 7 – Изменение ошибки регрессии $L_{reg}$ и MAE с применением text-dropout с вероятностью $p_{textDropout} = 0.2$, коэффициентом влияния ошибки выравнивания признаков $\lambda = 0.5$, шагом интерполяции высоты $d_{step} = 10$

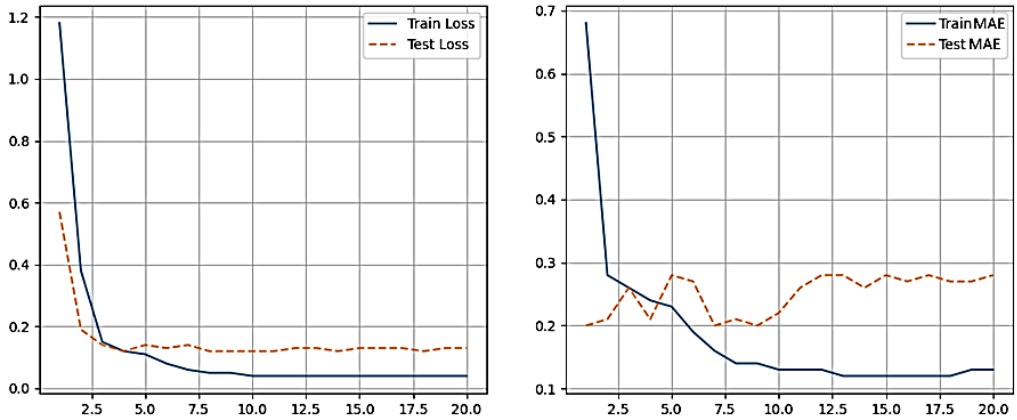


Рисунок 8 – Изменение ошибки регрессии $L_{reg}$ и MAE с применением text-dropout с вероятностью $p_{textDropout} = 0.3$, коэффициентом влияния ошибки выравнивания признаков $\lambda = 0.5$, шагом интерполяции высоты $d_{step} = 5$

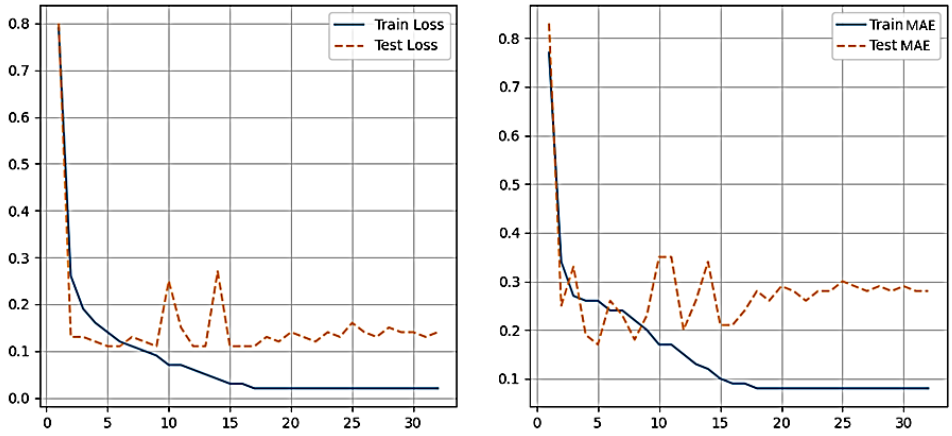


Рисунок 9 – Изменение ошибки регрессии $L_{reg}$ и MAE с применением text-dropout с вероятностью $p_{textDropout} = 0.3$, коэффициентом влияния ошибки выравнивания признаков $\lambda = 0.1$, шагом интерполяции высоты $d_{step} = 5$

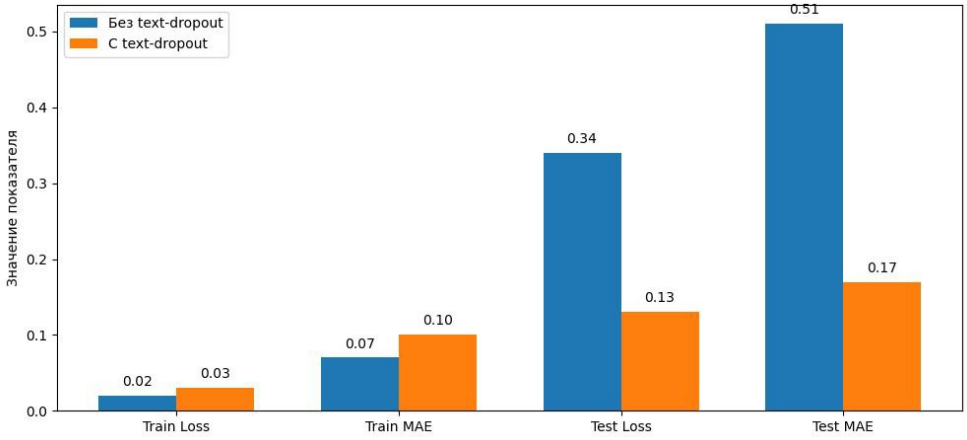


Рисунок 10 – Сравнение показателей ошибки обучения и валидации при обычном обучении и с применением метода text-dropout

Эксперименты показали, что внедрение числовых метаданных в текстовые шаблоны позволяет связать различия визуального контекста в зависимости от высоты с ее значением, что положительно отразилось на значениях ошибки при обучении. Однако, из-за отсутствия ветки текстовых признаков в тестовом режиме, ошибка была крайне нестабильна.

Для достижения хороших показателей были применены такие методики, как случайное обнуление текстовых признаков при обучении, интерполяция высоты с разным значением шага, а также введение дополнительной ошибки выравнивания признаков с коэффициентом влияния на общую ошибку.

На графиках можно отчетливо видеть, что при использовании конкатенации с различными конфигурациями $p_{textDropout}$, $\lambda$, $d_{step}$ снижение ошибки на обучении прекращается приблизительно к 30 эпохе и далее уходит на плато. Также наблюдается высокая нестабильность показателей с одинаковой динамикой на тестовых данных на всем диапазоне значений $p_{textDropout}$. Более лучшую сходимость при обучении показал метод использования gate-суммирования, при значениях $p_{textDropout} = 0.2$, $\lambda = 0.5$, $d_{step} = 10$ и $p_{textDropout} = 0.3$, $\lambda = 0.5$, $d_{step} = 5$.

Полученные результаты наталкивают на гипотезу о том, что увеличение $p_{textDropout}$ с 0.2 до 0.3 коррелирует с уменьшением шага интерполяции высоты с 10 до 5. Объединение суммированием показало себя лучше, в том числе с точки зрения вычислений, т.к. при суммировании размерность объединенного вектора признаков не увеличивается. Использование $\lambda$=0.5 для регулирования вклада ошибки выравнивания признаков дало лучшую стабильность и динамику общей ошибки. Гипотетически, наиболее оптимальные значения $\lambda$ находятся в окрестности 0.5.

## Заключение

В настоящей работе предложен метод мультимодальной регрессии положения БПЛА, основанный на совместной обработке визуальных признаков и текстовых закодированных метаданных. Предложенный подход учитывает высоту съёмки как важный фактор для точности предсказания и интегрирует её в текстовый шаблон, что позволило повысить обобщающую способность модели и уменьшить чувствительность к изменению ракурсов.

Применение регуляризации через функцию выравнивания признаков и случайное обнуление текстовых признаков способствовало стабилизации процесса обучения и снижению риска переобучения на текстовую модальность. Эксперименты показали, что методы адаптивного объединения признаков с использованием механизма *gate*-суммирования демонстрируют преимущество над простой конкатенацией как по точности, так и по вычислительной эффективности. Анализ параметров модели выявил оптимальные настройки коэффициента выравнивания (~0.5) и шага интерполяции высоты, которые обеспечили наилучший баланс между точностью регрессии и стабильностью обучения.

Результаты подтверждают гипотезу о важности согласования числовых и визуальных признаков при решении задач позиционирования. В перспективе планируется расширить подход за счёт учёта дополнительных сенсорных данных и изучения влияния различных стратегий регуляризации внимания на устойчивость модели в условиях шумных измерений.

## Список литературы

1. Пикалёв Я. С. Обнаружение ключевых объектов и перекрёстная геолокализация: Анализ наборов данных и методологические перспективы //Проблемы искусственного интеллекта.2024. Т. 35. № 4. С. 25-37.
2. Пикалёв Я. С., Ермоленко Т. В. О нейронных архитектурах извлечения признаков для задачи распознавания объектов на устройствах с ограниченной вычислительной мощностью // Проблемы искусственного интеллекта. 2023. №. 3 (30). С. 44-54.
3. Sheng K., Zhan H., Xie Z., Chen H., Xu Y., Tang L., Luo B. GeoText-1652: A Benchmark for Language-Aware Cross-View Geo-Localization // European Conference on Computer Vision (ECCV). 2024.
4. Xu Y., Zhang Z., He W., Wang J., Luo B. A Survey of Open-Vocabulary Object Detection for UAV Imagery // Drones. 2025. Vol. 9, No. 1. P. 12. DOI: 10.3390/drones9010012.
5. Sun X. et al. Spatial-LLaVA: Enhancing Large Language Models with Spatial Referring Expressions for Visual Understanding //arXiv preprint arXiv:2505.12194. – 2025.
6. Ye J. et al. Where am I? Cross-View Geo-localization with Natural Language Descriptions //arXiv preprint arXiv:2412.17007. 2024.
7. Павленко, Б. В. Методика создания набора аэрофотоснимков для задачи перекрёстной геолокализации / Б.В. Павленко, Я.С. Пикалёв // Проблемы искусственного интеллекта. 2024. №4(35). С. 101-112. DOI:10.24412/2413-7383-2024-4-101-112.

8. Пикалёв Я. С. Обнаружение ключевых объектов и перекрёстная геолокализация: Анализ наборов данных и методологические перспективы //Проблемы искусственного интеллекта. 2024. Т. 35. №. 4. С. 25-37.
9. Зуев В. М., Иванова С. Б. Оценка собственного местоположения аппарата на основе анализа видеоизображения //Проблемы искусственного интеллекта. 2024. Т. 33. №. 2. С. 21-28.
10. Ji Y. et al. Game4loc: A uav geo-localization benchmark from game data //Proceedings of the AAAI Conference on Artificial Intelligence. 2025. Т. 39. №. 4. С. 3913-3921.
11. Xu W. et al. Uav-visloc: A large-scale dataset for uav visual localization //arXiv preprint arXiv:2405.11936. 2024.19:06
12. Yang Y., Newsam S. Bag-of-visual-words and spatial extensions for land-use classification //Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. 2010. – С. 270-279. Kim W., Son B., Kim I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision // arXiv preprint, arXiv:2102.03334, 2021.
13. Huang S. et al. Language is not all you need: aligning perception with language models. arXiv //Preprint posted online February. 2023. Т. 27.
14. Tan H., Bansal M. LXMERT: Learning cross-modality encoder representations from transformers [Электронный ресурс] // arXiv preprint arXiv:1908.07490. 2019. URL: https://arxiv.org/abs/1908.07490 (дата обращения: 18.09.2025).
15. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision // arXiv preprint arXiv:2103.00020, 2021. URL: https://arxiv.org/abs/2103.00020
16. Qu G. et al. Stripnet: Towards topology consistent strip structure segmentation //Proceedings of the 26th ACM international conference on Multimedia. 2018. С. 283-291.
17. Faghri F. et al. MobileCLIP2: Improving Multi-Modal Reinforced Training //arXiv preprint arXiv:2508.20691. 2025.
18. Cui B., Liu Z., Yang Q. UAV-YOLO12: A Multi-Scale Road Segmentation Model for UAV Remote Sensing Imagery // Drones. 2025. Т. 9. №. 8. С. 533.
19. Less Wright, Nestor Demeure et al. Ranger: synergistic combination of RAdam + Lookahead for the best of both [Электронный ресурс] // GitHub / arXiv preprint. 2019. URL: https://arxiv.org/abs/2004.01461 (дата обращения: 18.09.2025).
20. Liu L., Jiang H., He P., Chen W., Liu X., Gao J., Han J. On the Variance of the Adaptive Learning Rate and Beyond [Электронный ресурс] // arXiv preprint. – 2019. – № arXiv:1908.03265. – URL: https://arxiv.org/abs/1908.03265. (дата обращения: 18.06.2025).
21. Zhang M. R., Lucas J., Hinton G., Ba J. Lookahead Optimizer: k steps forward, 1 step back [Электронный ресурс] // Advances in Neural Information Processing Systems, NeurIPS 2019. – 2019. – Vol. 32. – P. 9593-9604. – URL: https://arxiv.org/abs/1907.08610. (дата обращения: 18.06.2025).

# References

1. Pikalyov, Ya. S. Key Object Detection and Cross-Geolocalization: Dataset Analysis and Methodological Prospects // Problems of Artificial Intelligence. - 2024. - Vol. 35. - No. 4. - Pp. 25-37.
2. Pikalyov, Ya. S., Ermoolenko, T.V. On Neural Feature Extraction Architectures for Object Recognition on Devices with Limited Computing Power // Problems of Artificial Intelligence. - 2023. - No. 3 (30). - Pp. 44-45
3. Sheng K., Zhan H., Xie Z., Chen H., Xu Y., Tang L., Luo B. GeoText-1652: A Benchmark for Language-Aware Cross-View Geo-Localization // European Conference on Computer Vision (ECCV). – 2024.
4. Xu Y., Zhang Z., He W., Wang J., Luo B. A Survey of Open-Vocabulary Object Detection for UAV Imagery // Drones. – 2025. – Vol. 9, No. 1. – P. 12. – DOI: 10.3390/drones9010012.
5. Sun X. et al. Spatial-LLaVA: Enhancing Large Language Models with Spatial Referring Expressions for Visual Understanding //arXiv preprint arXiv:2505.12194. – 2025.
6. Ye J. et al. Where am I? Cross-View Geo-localization with Natural Language Descriptions //arXiv preprint arXiv:2412.17007. – 2024.
7. Pavlenko, B.V. Methodology for Creating a Set of Aerial Photographs for the Task of Cross-Geolocalization / B. V. Pavlenko, Ya. S. Pikalev // Problems of Artificial Intelligence. - 2024. - No. 4 (35). - Pp. 101-112. - DOI: 10.24412/2413-7383-2024-4-101-112.
8. Pikalyov, Ya. S. Detection of Key Objects and Cross-Geolocalization: Analysis of Datasets and Methodological Prospects // Problems of Artificial Intelligence. - 2024. - Vol. 35. - No. 4. - Pp. 25-37.

9. Zuev, V.M., Ivanova S. B. Estimation of the Device's Own Location Based on Video Image Analysis // Problems of Artificial Intelligence. – 2024. – T. 33. – No. 2. – Pp. 21-28.
10. Ji Y. et al. Game4loc: A uav geo-localization benchmark from game data //Proceedings of the AAAI Conference on Artificial Intelligence. – 2025. – T. 39. – №. 4. – С. 3913-3921.
11. Xu W. et al. Uav-visloc: A large-scale dataset for uav visual localization //arXiv preprint arXiv:2405.11936. – 2024.19:06
12. Yang Y., Newsam S. Bag-of-visual-words and spatial extensions for land-use classification //Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. – 2010. – С. 270-279. Kim W., Son B., Kim I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision // arXiv preprint, arXiv:2102.03334, 2021.
13. Huang S. et al. Language is not all you need: aligning perception with language models. arXiv //Preprint posted online February. – 2023. – T. 27.
14. Tan H., Bansal M. LXMERT: Learning cross-modality encoder representations from transformers [Электронный ресурс] // arXiv preprint arXiv:1908.07490. – 2019. – URL: https://arxiv.org/abs/1908.07490 (дата обращения: 18.09.2025).
15. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision // arXiv preprint arXiv:2103.00020, 2021. URL: https://arxiv.org/abs/2103.00020
16. Qu G. et al. Stripnet: Towards topology consistent strip structure segmentation //Proceedings of the 26th ACM international conference on Multimedia. – 2018. – С. 283-291.
17. Faghri F. et al. MobileCLIP2: Improving Multi-Modal Reinforced Training //arXiv preprint arXiv:2508.20691. – 2025.
18. Cui B., Liu Z., Yang Q. UAV-YOLO12: A Multi-Scale Road Segmentation Model for UAV Remote Sensing Imagery //Drones. – 2025. – Т. 9. – №. 8. – С. 533.
19. Less Wright, Nestor Demeure et al. Ranger: synergistic combination of RAdam + Lookahead for the best of both [Электронный ресурс] // GitHub / arXiv preprint. – 2019. – URL: https://arxiv.org/abs/2004.01461 (дата обращения: 18.09.2025).
20. Liu L., Jiang H., He P., Chen W., Liu X., Gao J., Han J. On the Variance of the Adaptive Learning Rate and Beyond [Электронный ресурс] // arXiv preprint. – 2019. – № arXiv:1908.03265. – URL: https://arxiv.org/abs/1908.03265. (дата обращения: 18.06.2025).
21. Zhang M. R., Lucas J., Hinton G., Ba J. Lookahead Optimizer: k steps forward, 1 step back [Электронный ресурс] // Advances in Neural Information Processing Systems, NeurIPS 2019. – 2019. – Vol. 32. – P. 9593-9604. – URL: https://arxiv.org/abs/1907.08610. (дата обращения: 18.06.2025).

# RESUME

*B. V. Pavlenko*

*An approach to multimodal data fusion in the problem of UAV position value regression*

Research into multimodal learning models and methods is currently one of the most active areas. Along with the development and implementation of intelligent unmanned systems, the need for mobile and precise intelligent positioning systems is growing. The use of additional semantics in the form of metadata allows for improving the quantitative and qualitative performance of models. Since existing datasets do not always provide rich numerical metadata on UAV position and altitude, a solution can be found in a multimodal approach to regressing the corresponding values, predicted by clarifying the context and meaning of the numerical value through text descriptions. Experiments have demonstrated the suitability of this approach for determining missing UAV positioning values based on aerial photographs and altitude using interpolation during the training process.

# РЕЗЮМЕ

*Б.В. Павленко*
*Подход к мультимодальному объединению данных в задаче регрессии значений положений БПЛА*

Исследования моделей и методов мультимодального обучения являются одними из наиболее активных на сегодняшний день. Совместно с развитием и внедрением интеллектуальных беспилотных систем растет необходимость в мобильных и точных интеллектуальных системах позиционирования. Использование дополнительной семантики в виде метаданных позволяет повышать количественные и качественные показатели моделей. В силу того, что имеющиеся наборы данных не всегда располагают богатыми числовыми метаданными о положении и высоте БЛА, решением может быть мультимодальный подход к регрессии соответствующих значений, предсказываемых за счет уточнения контекста и смысла числового значения через текстовое описание. Эксперименты показали пригодность подхода к определению недостающих значений позиционирования БЛА по аэрофотоснимку и высоте с применением интерполяции в процессе обучения.

**Павленко Богдан Викторович** – младший научный сотрудник, аспирант, ФГБНУ «Институт проблем искусственного интеллекта», 283048, г. Донецк, ул. Артема, д. 118 б, телефон +7(949) 438-6450, bogdanpavl2000@mail.ru. *Область научных интересов:* системы распознавания образов, нейронные сети, мультимодальные модели.

Статья поступила в редакцию 10.09.2025.