ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 3. С. 49–53
DOI: https://doi.org/10.5281/zenodo.24540654
Сайт журнала: https://cajournal.kz

УДК 004.932:004.85

ВКЛАД КАНАЛА ГЛУБИНЫ В СЕМАНТИЧЕСКУЮ СЕГМЕНТАЦИЮ ПОМЕЩЕНИЙ: ФАКТОРНЫЙ ЭКСПЕРИМЕНТ С КЛАСТЕРНОЙ СТАТИСТИЧЕСКОЙ ОЦЕНКОЙ

Е.Б. Маратов
Международный университет Астана, Астана, Казахстан

Маратов Ерасыл Балканович — магистрант 2 курса образовательной программы, «7М06101 – Вычислительная техника и программное обеспечение», Международный университет Астана, Астана, Казахстан.

Научный руководитель: Жукабаева Тамара Кокеновна — PhD, профессор, Международный университет Астана, Астана, Казахстан.

© Е.Б. Маратов, 2026

Аннотация: В статье оценивается, насколько добавление карты глубины к цветному изображению улучшает семантическую сегментацию помещений, если архитектура сети и режим обучения остаются неизменными. На наборе NYU Depth V2 (13 классов) проведён факторный эксперимент 2×2 с сетью U-Net: вход RGB или RGB-D, кодировщик со случайной инициализацией или предобученный на ImageNet, по три запуска на вариант; доверительные интервалы построены методом кластерного бутстрепа по 215 сценам. Глубина повышает mIoU на 9,1 п.п. при обучении с нуля и на 3,0 п.п. с предобученным кодировщиком ($p < 0{,}001$), причём выигрыш сосредоточен в классах, определяемых геометрией: пол, кровать, стул, потолок.

Ключевые слова: семантическая сегментация, RGB-D, карта глубины, U-Net, NYU Depth V2, предобучение, бутстреп, компьютерное зрение

Введение

Семантическая сегментация — присвоение класса каждому пикселю изображения — является базовой задачей понимания сцен для мобильных роботов, систем дополненной реальности и построения семантических 3D-карт [1]. Полностью свёрточные сети [2] и архитектуры типа «кодировщик–декодировщик» [3, 4, 5] решают её по цветному изображению. В помещениях, однако, многие объекты похожи по цвету и текстуре (белая стена, белый потолок, светлый пол), но различаются положением и формой поверхности. Потребительские RGB-D камеры дают для каждого пикселя расстояние до сцены, и эту информацию естественно использовать как дополнительный источник признаков [6, 7].

Предложено несколько способов объединить цвет и глубину: глубина как дополнительный входной канал (раннее слияние) [6], геоцентрическое кодирование HHA [8], два кодировщика со слиянием признаков [9, 10], кросс-модальное внимание в трансформерах [11]. Сравнивая такие методы между собой, сложно отделить вклад глубины от вклада архитектуры, предобучения и режима обучения: новые методы, как правило, отличаются от базовых сразу по нескольким факторам. Кроме того, результаты обычно приводятся по одному запуску, хотя разброс между запусками с разным начальным состоянием генератора

случайных чисел может быть сопоставим с различием методов [12], а кадры тестовой выборки, снятые в одной комнате, не являются независимыми наблюдениями.

Цель работы — количественно оценить вклад канала глубины в семантическую сегментацию помещений при фиксированной архитектуре и проверить, сохраняется ли этот вклад при использовании кодировщика, предобученного на обычных фотографиях. Для этого решены задачи: (1) построить факторный план эксперимента, в котором вход и инициализация варьируются независимо; (2) оценить неопределённость результатов с учётом разброса между запусками и зависимости кадров одной сцены; (3) определить, для каких классов глубина полезна, а для каких нет.

Материалы и методы

Данные.

Использован размеченный набор NYU Depth V2 [7]: 1449 пар «цветное изображение – карта глубины», снятых камерой Kinect в жилых и офисных помещениях, с официальным разбиением на 795 обучающих и 654 тестовых изображений из 215 сцен. Исходные классы приведены к 40 классам по отображению [8] и далее к 13 классам [6]. Изображения уменьшены до 240×320 пикселей (разметка — по ближайшему соседу). Неразмеченные пиксели исключаются из функции потерь и метрик.

Модель и обучение.

Во всех вариантах используется одна архитектура — U-Net [3] с кодировщиком ResNet-18 [13] (14,3 млн параметров) из библиотеки Segmentation Models PyTorch [14, 15]. Варьируются два фактора: вход — RGB (3 канала) или RGB-D (4 канала, глубина подаётся как дополнительный канал), и инициализация кодировщика — случайная или предобученная на ImageNet [16]. Для предобученного кодировщика веса первого слоя для канала глубины копируются из красного канала, а все четыре набора весов умножаются на 3/4 (стандартная процедура библиотеки). Глубина нормализуется средним и стандартным отклонением обучающей выборки.

Обучение одинаково для всех четырёх вариантов: 40 эпох, оптимизатор AdamW [17] (скорость обучения $3 \cdot 10^{-4}$, коэффициент затухания весов 10^{-4}) с расписанием OneCycle, размер батча 8, перекрёстная энтропия, аугментация — горизонтальное отражение. Каждый вариант обучен трижды с разными начальными значениями генератора случайных чисел (seed 0, 1, 2). Тестовая выборка оценивается один раз после последней эпохи, без выбора лучшей эпохи по тесту. Вычисления выполнены на ноутбуке Apple M-series (встроенный GPU, MPS); обучение одной модели занимает 13–15 минут.

Метрики и статистическая оценка.

Качество оценивается средним по классам индексом Жаккара:

$$\mathrm{mIoU} = (1/K) \cdot \Sigma_k \, TP_k / (TP_k + FP_k + FN_k), \qquad (1)$$

где $K = 13$ — число классов; TP_k, FP_k, FN_k — число верно отнесённых, ложно отнесённых и пропущенных пикселей класса k, накопленные по всей тестовой выборке.

Поскольку кадры одной сцены похожи, доверительные интервалы строятся методом кластерного бутстрепа [18, 19]: на каждой из $B = 2000$ итераций с возвращением выбираются целые сцены, их матрицы ошибок суммируются, и mIoU вычисляется заново. 95%-й интервал задаётся перцентилями бутстреп-распределения. Для сравнения RGB-D и RGB обе модели оцениваются на одной и той же выборке сцен (парный бутстреп), а двусторонний p-уровень равен

$$p = 2 \cdot \min(P^*(\Delta \le 0), P^*(\Delta \ge 0)), \qquad (2)$$

где $\Delta = \mathrm{mIoU(RGB\text{-}D)} - \mathrm{mIoU(RGB)}$, P^* — доля бутстреп-итераций, удовлетворяющих условию.

Результаты и обсуждение

Таблица 1 – Качество сегментации на тестовой выборке NYU Depth V2 (13 классов)

Вход	Кодировщик	mIoU по запускам, %	mIoU, %	95% ДИ	Pixel acc, %

RGB	с нуля	29,2 / 28,5 / 29,9	**29,2**	[27,8; 30,5]	57,0
RGB-D	с нуля	39,7 / 37,9 / 37,1	**38,2**	[36,7; 39,6]	65,6
RGB	ImageNet	50,2 / 49,5 / 49,6	**49,8**	[47,3; 52,0]	72,8
RGB-D	ImageNet	52,1 / 52,7 / 53,5	**52,8**	[50,4; 54,8]	74,9

Таблица 2 – Прирост mIoU от добавления глубины (парный кластерный бутстреп)

Сравнение	Δ mIoU по запускам, п.п.	Среднее Δ, п.п.	95% ДИ	p
RGB-D − RGB, с нуля	+10,6 / +9,4 / +7,2	**+9,1**	[8,1; 10,1]	< 0,001
RGB-D − RGB, ImageNet	+1,9 / +3,2 / +3,9	**+3,0**	[2,2; 3,8]	< 0,001

Результаты четырёх вариантов приведены в таблице 1, прирост от глубины — в таблице 2. Добавление глубины статистически значимо улучшает сегментацию в обоих режимах и в каждом из трёх запусков: на 9,1 п.п. mIoU (95% ДИ [8,1; 10,1]) при обучении с нуля и на 3,0 п.п. ([2,2; 3,8]) с предобученным кодировщиком. Время обучения при этом не растёт: четвёртый канал затрагивает только первый свёрточный слой.

Факторный план позволяет сравнить вклад двух факторов. Предобучение на ImageNet даёт прирост от +14,5 до +20,6 п.п., т.е. существенно больше, чем глубина. При этом эффекты не складываются: с предобученным кодировщиком выигрыш от глубины втрое меньше. Признаки, выученные на большом наборе фотографий, частично передают ту же информацию о форме и расположении поверхностей, которую несёт глубина, но не заменяют её полностью. Если бы RGB-модель с нуля сравнивалась с предобученной RGB-D моделью, разница в 23,6 п.п. была бы ошибочно приписана глубине.

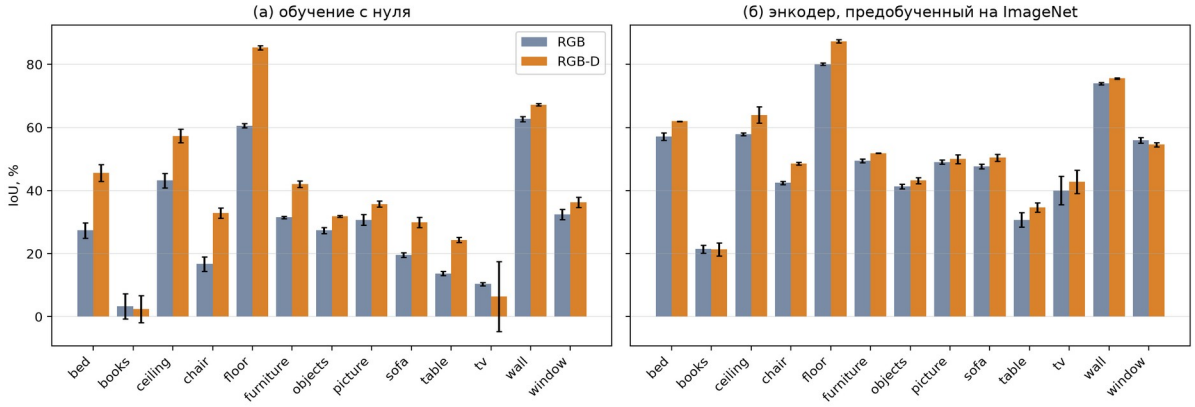


Рисунок 1 - IoU по классам, среднее по трём запускам (отрезки — стандартное отклонение): (а) обучение с нуля, (б) кодировщик ImageNet

Анализ по классам (рисунок 1) показывает, что глубина помогает прежде всего классам, которые определяются формой и положением в пространстве: пол (+24,7 п.п. с нуля, +7,2 с ImageNet), кровать (+18,2), стул (+16,2), потолок (+14,1), стол (+10,7). Пол и потолок различаются прежде всего высотой и ориентацией поверхности, а не цветом, — именно эту информацию и даёт карта глубины. Для классов, которые отличаются текстурой, а не геометрией, — книги (-0,1), картины (+0,9), окна (-1,4 п.п. с ImageNet) — прироста нет: картины и окна лежат в плоскости стены и отличаются от неё только цветом. Класс «телевизор» при обучении с нуля нестабилен между запусками, что связано с его малой долей в данных.

Ограничения работы. Функция потерь на обучающей выборке к концу обучения снижается до ≈0,32 (с нуля) и ≈0,12 (с ImageNet), что при точности на тесте 57–75 % указывает на переобучение на 795 изображениях. Отдельной валидационной выборки нет, поэтому гиперпараметры не подбирались и одинаковы для всех вариантов. Рассмотрено только раннее слияние; более сложные схемы — HHA-кодирование и отдельный

кодировщик глубины [8, 9, 10] — могут дать больший выигрыш и будут исследованы при том же протоколе оценки.

Заключение

В факторном эксперименте с фиксированной архитектурой U-Net показано, что карта глубины статистически значимо улучшает семантическую сегментацию помещений: на 9,1 п.п. mIoU при обучении с нуля и на 3,0 п.п. с кодировщиком, предобученным на ImageNet. Выигрыш сосредоточен в классах, определяемых геометрией (пол, потолок, мебель), и отсутствует для «плоских» объектов. Предобучение даёт больший эффект, чем глубина, и частично его поглощает, поэтому сравнивать RGB и RGB-D методы следует при одинаковой инициализации. Оценка по нескольким запускам с кластерным бутстрепом по сценам позволяет отличить реальный эффект от случайного разброса. Полученная модель используется как модуль сегментации в прототипе семантической 3D-реконструкции помещений, разрабатываемом в рамках магистерской диссертации.

Список литературы:

1. McCormac J., Handa A., Davison A., Leutenegger S. SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks // IEEE International Conference on Robotics and Automation (ICRA). – 2017. – P. 4628–4635.

2. Long J., Shelhamer E., Darrell T. Fully Convolutional Networks for Semantic Segmentation // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 3431–3440.

3. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention (MICCAI). – 2015. – P. 234–241.

4. Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39, No. 12. – P. 2481–2495.

5. Chen L.-C., Zhu Y., Papandreou G., Schroff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation // European Conference on Computer Vision (ECCV). – 2018. – P. 801–818.

6. Couprie C., Farabet C., Najman L., LeCun Y. Indoor Semantic Segmentation using Depth Information // International Conference on Learning Representations (ICLR). – 2013.

7. Silberman N., Hoiem D., Kohli P., Fergus R. Indoor Segmentation and Support Inference from RGBD Images // European Conference on Computer Vision (ECCV). – 2012. – P. 746–760.

8. Gupta S., Girshick R., Arbeláez P., Malik J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation // European Conference on Computer Vision (ECCV). – 2014. – P. 345–360.

9. Hazirbas C., Ma L., Domokos C., Cremers D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture // Asian Conference on Computer Vision (ACCV). – 2016. – P. 213–228.

10. Seichter D., Köhler M., Lewandowski B., Wengefeld T., Gross H.-M. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis // IEEE International Conference on Robotics and Automation (ICRA). – 2021. – P. 13525–13531.

11. Zhang J., Liu H., Yang K., Hu X., Liu R., Stiefelhagen R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers // IEEE Transactions on Intelligent Transportation Systems. – 2023. – Vol. 24, No. 12. – P. 14679–14694.

12. Bouthillier X., Delaunay P., Bronzi M. et al. Accounting for Variance in Machine Learning Benchmarks // Proceedings of Machine Learning and Systems (MLSys). – 2021. – Vol. 3. – P. 747–769.

13. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770–778.

14. Iakubovskii P. Segmentation Models PyTorch [Электронный ресурс]. – 2019. – URL: https://github.com/qubvel/segmentation_models.pytorch (дата обращения: 20.11.2026).

15. Paszke A., Gross S., Massa F. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems (NeurIPS). – 2019. – Vol. 32.

16. Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2009. – P. 248–255.

17. Loshchilov I., Hutter F. Decoupled Weight Decay Regularization // International Conference on Learning Representations (ICLR). – 2019.

18. Efron B., Tibshirani R. J. An Introduction to the Bootstrap. – New York: Chapman & Hall/CRC, 1993. – 436 p.

19. Field C. A., Welsh A. H. Bootstrapping Clustered Data // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2007. – Vol. 69, No. 3. – P. 369–390.

© Central Asian Scientific Journal, 2026