

УДК 004.932.2

Нейросетевая реализация монокамерного измерения глубины сцены

Лысенко Олег Дмитриевич

olegkeatzin@yandex.ru

Корлякова Мария Олеговна

mkorlyakova@yandex.ru

КФ МГТУ им. Н.Э. Баумана, Калуга, Россия

Монокамерная глубина сцены — это техника, при которой глубинная информация о сцене оценивается с использованием изображений, сделанных одной камерой. В отличие от стереозрения, где для оценки глубины используются два или более изображений с разных углов, монокамерная глубина извлекает информацию о расстояниях на основе одного изображения. Монокамерные системы могут работать на любом устройстве с одной камерой, что делает их доступными для широкого спектра задач, от смартфонов до дронов. Описаны проблемы определения монокамерной глубины сцены. Выбрана нейросетевая модель для оценки глубины сцены. Построена и обучена U-Net модель.

Ключевые слова: монокамерная глубина сцены, сверточная нейронная сеть, U-Net, Depth estimation

Карты глубины играют важную роль в современных технологиях, обеспечивая трехмерное восприятие и анализ различных сцен и объектов. Их использование охватывает множество областей: автономные автомобили, дополненная реальность, фотография и видеосъемка, архитектура и многие другие. Они позволяют улучшить точность навигации, безопасность, качество изображений и видео, а также создавать более реалистичные и интерактивные виртуальные среды.

Для решения задачи оценки глубины на основе изображений часто используются специализированные устройства, такие как LiDAR и ультразвуковые датчики. Однако интересен подход, при котором глубина оценивается исключительно на основании изображения, без необходимости использования дополнительных сенсоров. Эта задача может быть решена с использованием нейросетей.

В качестве основной архитектуры была выбрана модель U-Net. Эта архитектура обладает способностью к эффективному извлечению признаков из изображения благодаря своей специфичной структуре, состоящей из двух частей: кодировщика (encoder) и декодера (decoder). Кодировщик выполняет сжатие входных данных, выделяя ключевые особенности изображения, тогда как декодер восстанавливает исходный размер изображения, одновременно добавляя контекстную информацию, необходимую для точной сегментации или предсказания глубины. Такая структура позволяет модели U-Net успешно справляться с задачами, требующими высокой точности, включая оценку глубины на изображениях.

В U-Net есть skip connections, соединяющие энкодер и декодер. Это позволяет сохранять и восстанавливать важную пространственную информацию, улучшая качество и точность обработки изображений.

Для обучения U-Net модели был выбран датасет NYU Depth V2 [2]. Он состоит из кадров помещений и карт глубины к ним.

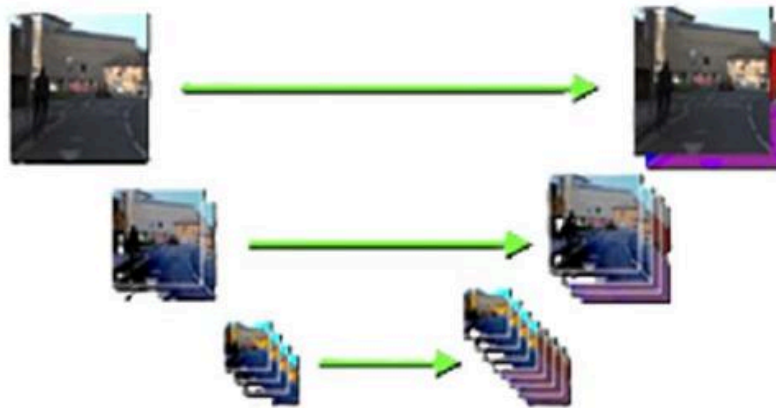


Рис. 1. Архитектура U-Net сети [1]

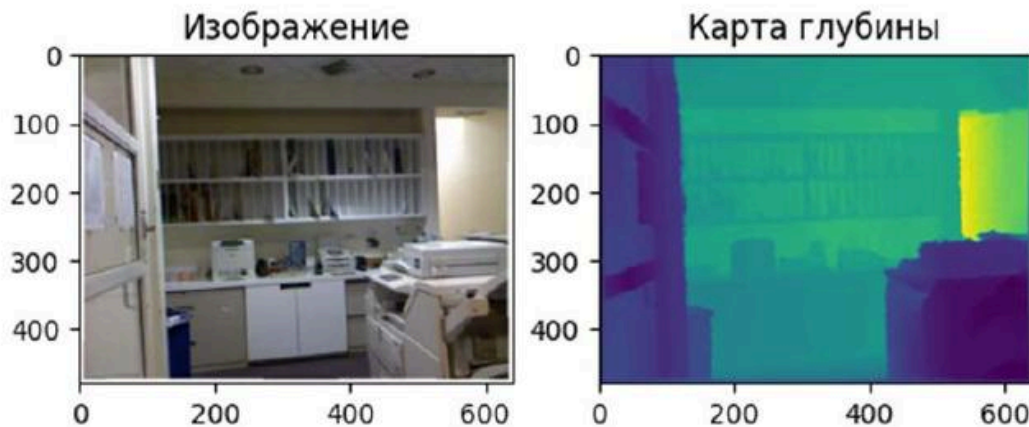


Рис. 2. Пример изображения и его карты глубины из NYU Depth V2

Датасет из 1449 изображений был поделен в соотношении: 80 % на обучение, 10 % на тест и 10 % на валидацию. Модель имеет 5 encoder слоев и 5 decoder слоев. В качестве функции потерь использовалась MSEloss и logMSEloss. Модели обучались 50 эпох. Для оценки готовой модели использовалась метрика MSE — средняя квадратичная ошибка [3]:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

где y_i — истинное значение; $\hat{y}_i$ — предсказанное значение.

Сравнение обученных моделей

Модель	Потери	MSE
U-Net	MSE	1,3266
U-Net	logMSE	0,5529

Из таблицы видно, что модель использующая logMSE лучше конкурента. Логарифмирование позволило лучше справляться с предсказаниями, имеющими широкий диапазон значений, минимизируя влияние больших ошибок и обеспечивая более устойчивую и точную оптимизацию модели [4].

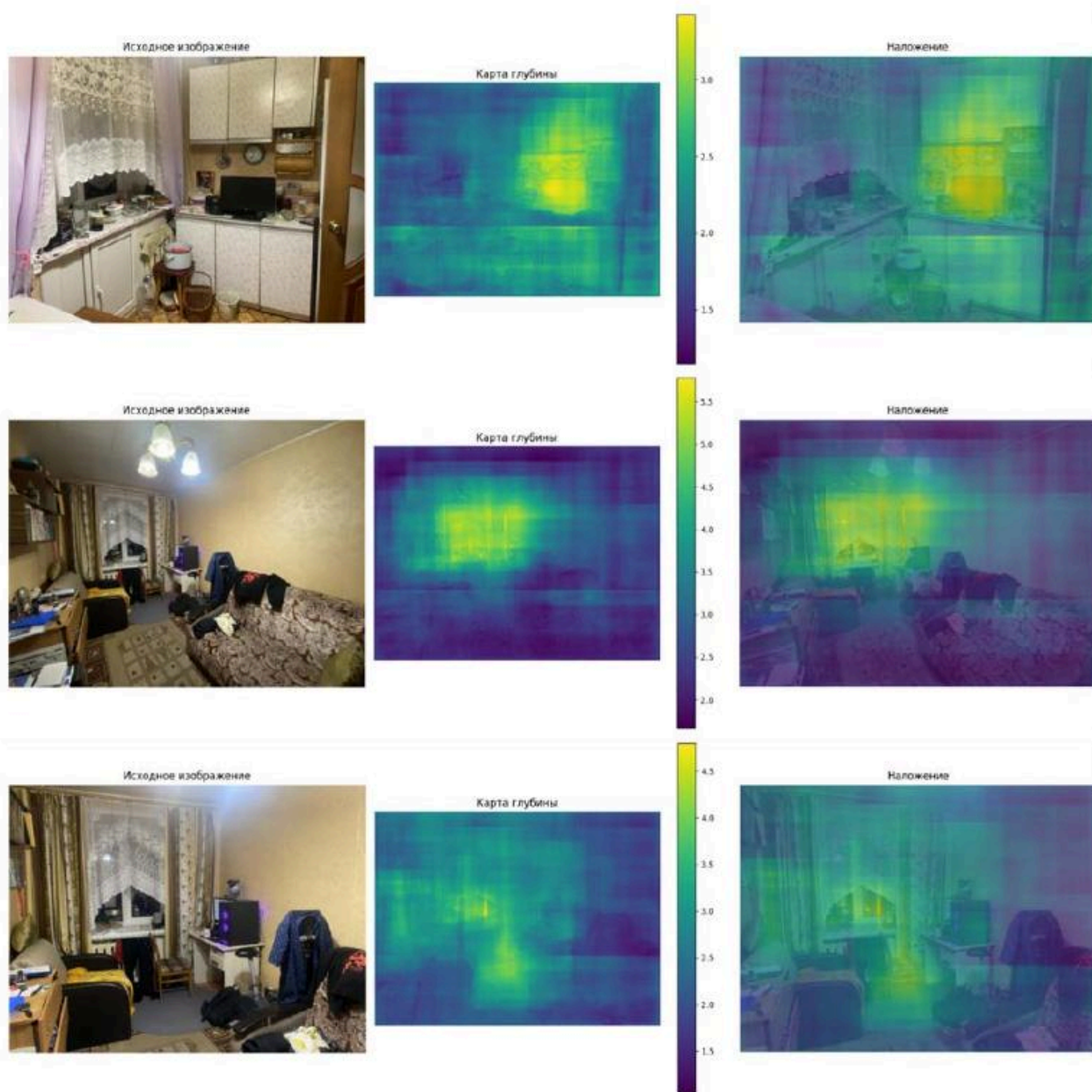


Рис. 3. Тестирование модели (logMSE)

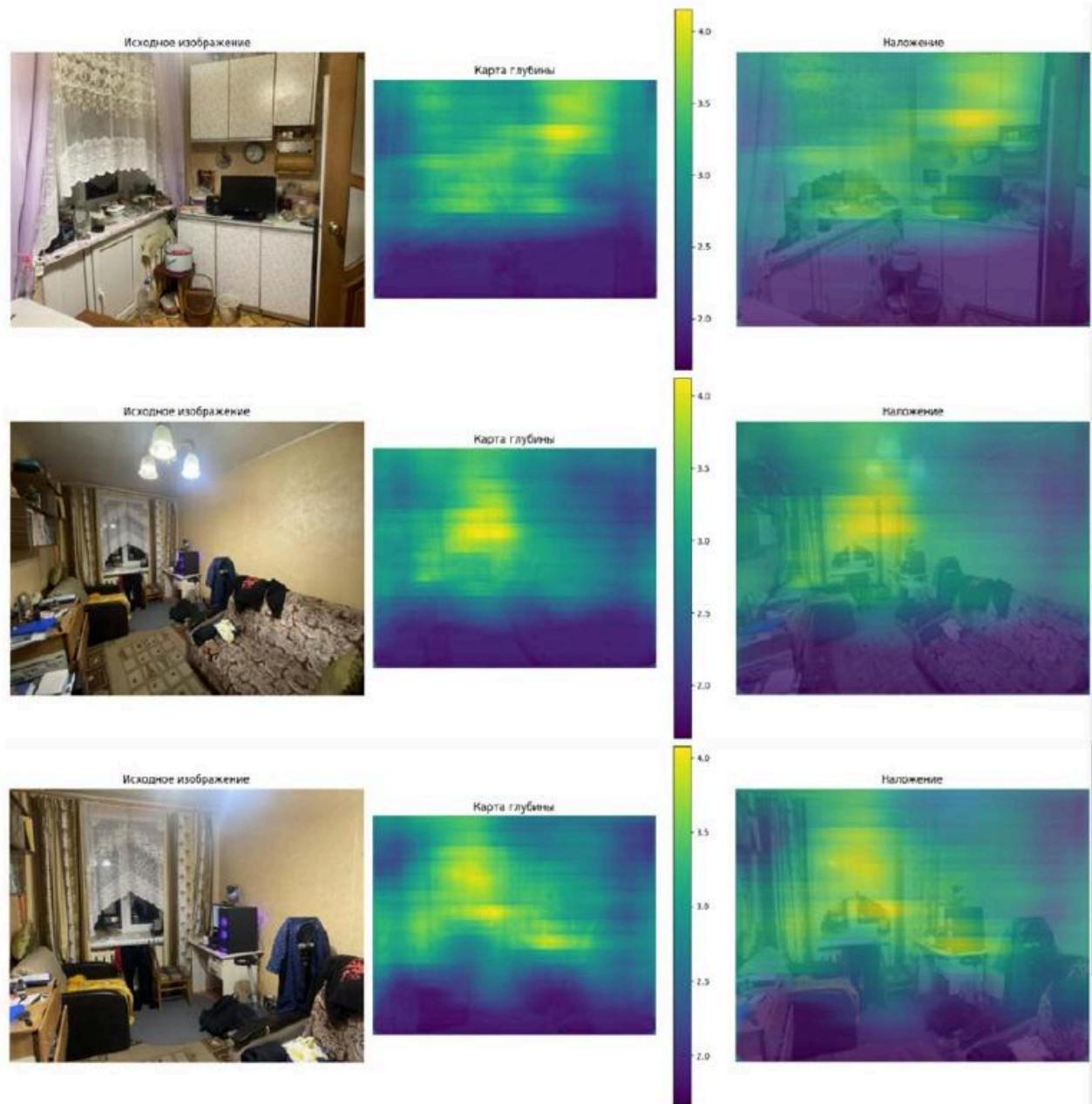


Рис. 4. Тестирование модели (MSE)

На тепловой карте светлые области соответствуют удаленным объектам, а темные — близким. В общем, модель демонстрирует удовлетворительное понимание динамики изменения глубины на изображении, однако точность ее работы оставляет желать лучшего. В дальнейшем планируется расширить набор данных посредством аугментации, а также исследовать другие типы сверточных архитектур, сравнивая их эффективность и вычислительную сложность.

Литература

- [1] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. URL: <https://arxiv.org/abs/1505.04597> (дата обращения 08.11.2024).
- [2] Silberman N., Kohli P., Hoiem D., Fergus R. Indoor Segmentation and Support Inference from RGBD Images ECCV. URL: https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html (дата обращения 08.11.2024).
- [3] MSE в машинном обучении простыми словами. URL: <https://dzen.ru/a/YGSuWtWwHiE0Ixfo> (дата обращения 08.11.2024).
- [4] Kapronczay M. Mean squared error (mse) vs. mean squared logarithmic error (msle): a guide. URL: <https://builtin.com/data-science/msle-vs-mse> (дата обращения 08.11.2024).

Neural network implementation of monocular depth estimation

Lysenko O.D. olegkeatzin@yandex.ru

Korlyakova M.O. mkorlyakova@yandex.ru

BMSTU (KB), Kaluga, Russia

Monocular depth of the scene is a technique in which in-depth information about a scene is evaluated using images taken with a single camera. Unlike stereo vision, where two or more images from different angles are used to estimate depth, single-camera depth extracts distance information based on a single image. Multi-camera systems can work on any device with a single camera, making them available for a wide range of tasks, from smartphones to drones. The problems of determining the multi-chamber depth of the scene are described. A neural network model has been selected to estimate the depth of the scene. A U-Net model has been built and trained.

Keywords: monocular scene depth, convolutional neural network, U-Net, Depth estimation