Орду Музаффер Али, магистрант
МГТУ им. Н.Э. Баумана
Ordu Muzaffer Ali
Bauman Moscow State Technical University

НЕЙРОСЕТЕВЫЕ МЕТОДЫ ОЦЕНКИ ГЛУБИНЫ СЦЕНЫ ПО МОНОКУЛЯРНОМУ ИЗОБРАЖЕНИЮ В ЗАДАЧАХ АНАЛИЗА ДОРОЖНОЙ ОБСТАНОВКИ
NEURAL NETWORK METHODS FOR MONOCULAR SCENE DEPTH ESTIMATION IN ROAD SCENE ANALYSIS

Аннотация. Рассмотрены нейросетевые методы оценки глубины сцены по одному изображению для определения расстояния до объектов дорожной обстановки. Предложена классификация методов: обучение с учителем, самообучение, модели относительной глубины и универсальные модели метрической глубины. Методы сравнены по опубликованным результатам на наборе данных KITTI, даны рекомендации по их применению.

Abstract. Neural network methods for single-image depth estimation are reviewed for measuring the distance to road scene objects. The methods are classified into supervised, self-supervised, relative depth and universal metric depth models, compared using published results on the KITTI dataset, and recommendations for their use are given.

Ключевые слова: Монокулярная оценка глубины, нейронные сети, карта глубины, расстояние до объектов, KITTI, дорожная обстановка

Keywords: Monocular depth estimation, neural networks, depth map, object distance, KITTI, road scene

Оценка расстояния до транспортных средств, пешеходов и других объектов дорожной обстановки необходима системам помощи водителю и автономным транспортным средствам. Для прямого измерения расстояния применяются лидары, радары и стереокамеры, однако они увеличивают стоимость оборудования и требуют точной калибровки [1]. Монокулярная камера наиболее дешева и распространена, но при проецировании сцены на плоскость изображения информация о глубине теряется, и её восстановление по одному кадру является некорректной задачей.

Классические подходы – стереозрение, восстановление структуры по движению, анализ дефокусировки и затенения – требуют нескольких камер или движения камеры либо чувствительны к освещению. Нейросетевые методы решают задачу по одному кадру за счёт априорных знаний о сцене, извлекаемых из обучающих выборок: размеров объектов, перспективы, перекрытий [2]; монокулярная оценка глубины применяется и в отечественных работах, например для картирования моделируемой среды [3]. Целью работы является классификация нейросетевых методов монокулярной оценки глубины и сравнение их применимости для анализа дорожной обстановки.

Постановка задачи

Требуется построить по изображению карту глубины, ставящую в соответствие каждому пикселю расстояние до точки сцены вдоль оптической оси камеры. Различают *метрическую* глубину, выраженную в метрах, и *относительную*, определённую с точностью до масштаба; для последней расстояние в метрах требует дополнительной информации, например высоты установки камеры над дорогой. Основным эталонным набором данных для дорожных сцен является KITTI [4] с изображениями камеры автомобиля и синхронными измерениями лидара. Качество оценивают средней относительной ошибкой AbsRel, среднеквадратичной ошибкой RMSE и долей пикселей $\delta < 1{,}25$, у которых предсказанная глубина отличается от истинной не более чем в 1,25 раза [2].

Классификация нейросетевых методов

Методы обучения с учителем используют пары «изображение – карта глубины», полученные лидаром или RGB-D камерой. Первой работой направления стала

многомасштабная сверточная сеть Д. Эйгена [2]; DenseDepth [5] применяет архитектуру «кодировщик – декодировщик» с предобученным кодировщиком, AdaBins [6] сводит оценку глубины к адаптивной классификации по интервалам глубины. Методы дают метрическую глубину, но хорошо работают лишь на данных, близких к обучающим, а плотная разметка дорогостояща.

Методы самообучения обучаются без разметки по стереопарам или монокулярному видео: глубина и движение камеры предсказываются совместно, а функцией потерь служит ошибка реконструкции соседнего кадра. Наиболее известный представитель – Monodepth2 [7]. Масштаб глубины при этом не определён, а на независимо движущихся объектах возникают ошибки.

Модели относительной глубины обучаются на смеси разнородных наборов данных с функцией потерь, инвариантной к масштабу и сдвигу, и обобщаются на новые сцены без дообучения (*zero-shot*). К ним относятся MiDaS [8], её трансформерная версия DPT и модели Depth Anything [9, 10], обученные на десятках миллионов изображений, размеченных моделью-учителем. На рисунке 1 приведены карты глубины для кадра городской сцены в дождь, сформированного в симуляторе CARLA [11]: Depth Anything V2 заметно лучше облегчённой версии MiDaS передаёт границы автобуса и пешехода.



Рис. 1. Оценка глубины по монокулярному изображению: а – исходный кадр; б – MiDaS; в – Depth Anything V2 (светлее – ближе)

Универсальные модели метрической глубины сочетают обобщающую способность моделей относительной глубины с метрическим масштабом. ZoeDepth [12] дополняет модель относительной глубины метрическими «головами», Metric3D v2 [13] приводит изображения разных камер к каноническому пространству с учётом фокусного расстояния, UniDepth [14] дополнительно оценивает параметры камеры по изображению. Эти модели наиболее перспективны для определения расстояния, но вычислительно наиболее затратны.

Сравнение методов

В таблице 1 приведены опубликованные результаты рассмотренных методов на тестовой выборке KITTI по разбиению Д. Эйгена [2]. Режим «KITTI» означает обучение или дообучение модели на этом наборе данных, режим «zero-shot» – оценку без использования KITTI при обучении. Значения для AdaBins, DPT и ZoeDepth приведены по сводным таблицам [9, 13].

Таблица 1

Результаты монокулярной оценки глубины на наборе данных KITTI

Метод	**Группа**	**Режим**	**AbsRel**	**RMSE, м**	**δ < 1,25**
Eigen и др [2]	с учителем	KITTI	0,190	7,156	0,692
DenseDepth [5]	с учителем	KITTI	0,093	4,170	0,886
AdaBins [6]	с учителем	KITTI	0,058	2,360	0,964
Monodepth2 [7]	самообучение	KITTI	0,115	4,863	0,877
DPT	относительная	KITTI	0,062	2,573	0,959
ZoeDepth [12]	метрическая	KITTI	0,054	2,281	0,971

Продолжение таблицы 1

Метод	**Группа**	**Режим**	**AbsRel**	**RMSE, м**	**δ < 1,25**
Depth Anything V2 [10]	метрическая	KITTI	0,045	1,861	0,983
Metric3D v2 [13]	метрическая	zero-shot	0,052	2,511	0,974
UniDepth [14]	метрическая	zero-shot	0,042	1,75	0,986

Из таблицы 1 видно, что за десять лет средняя относительная ошибка снизилась более чем в четыре раза: с 0,190 до 0,04–0,05. Наиболее важен для практики результат универсальных метрических моделей Metric3D v2 и UniDepth: без обучения на KITTI они достигают точности дообученных моделей (UniDepth оценивалась на уточнённой разметке, поэтому сравнение приближённое). Облегчённые модели пригодны для работы в реальном времени на встраиваемых ускорителях, а приведённые метрики получены в благоприятных условиях: в дождь или ночью качество снижается, что видно на рисунке 1.

Заключение

Предложена классификация нейросетевых методов монокулярной оценки глубины на четыре группы. Наилучшее сочетание точности, обобщающей способности и метрического масштаба обеспечивают универсальные метрические модели, однако ценой вычислительных затрат. Для определения расстояния до объектов дорожной обстановки рекомендуется использовать универсальные метрические модели либо модели относительной глубины с восстановлением масштаба по геометрии установки камеры.

Список литературы:

1. Колузов А. В., Корнеев К. С., Гой А. С. Определения расстояния до объекта с использованием технического зрения // Интеллектуальные системы в производстве. – 2025. – Т. 23, № 4. – С. 33–39.

2. Eigen D., Puhrsch C., Fergus R. Depth map prediction from a single image using a multi-scale deep network // Advances in Neural Information Processing Systems 27 (NIPS 2014). – 2014. – P. 2366–2374.

3. Бархум М., Пыркин А. А. Использование монокулярной оптики при оценке глубины объектов для двумерного картирования моделируемой среды // Научно-технический вестник информационных технологий, механики и оптики. – 2024. – Т. 24, № 1. – С. 118–123.

4. Geiger A., Lenz P., Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite // 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2012. – P. 3354–3361.

5. Alhashim I., Wonka P. High quality monocular depth estimation via transfer learning. – arXiv preprint arXiv:1812.11941. – 2018.

6. Bhat S. F., Alhashim I., Wonka P. AdaBins: depth estimation using adaptive bins // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2021. – P. 4009–4018.

7. Godard C., Mac Aodha O., Firman M., Brostow G. J. Digging into self-supervised monocular depth estimation // 2019 IEEE/CVF International Conference on Computer Vision (ICCV). – 2019. – P. 3827–3837.

8. Ranftl R., Lasinger K., Hafner D., Schindler K., Koltun V. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2022. – Vol. 44, № 3. – P. 1623–1637.

9. Yang L., Kang B., Huang Z., Xu X., Feng J., Zhao H. Depth Anything: unleashing the power of large-scale unlabeled data // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024. – P. 10371–10381.

10. Yang L., Kang B., Huang Z., Zhao Z., Xu X., Feng J., Zhao H. Depth Anything V2 // Advances in Neural Information Processing Systems 37 (NeurIPS 2024). – 2024. – P. 21875–21911.

11. Dosovitskiy A., Ros G., Codevilla F., Lopez A., Koltun V. CARLA: an open urban driving simulator // Proceedings of the 1st Annual Conference on Robot Learning (CoRL). – 2017. – P. 1–16.

12. Bhat S. F., Birkl R., Wofk D., Wonka P., Müller M. ZoeDepth: zero-shot transfer by combining relative and metric depth. – arXiv preprint arXiv:2302.12288. – 2023.

13. Hu M., Yin W., Zhang C., Cai Z., Long X., Chen H., Wang K., Yu G., Shen C., Shen S. Metric3D v2: a versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2024. – Vol. 46, № 12. – P. 10579–10596.

14. Piccinelli L., Yang Y.-H., Sakaridis C., Segu M., Li S., Van Gool L., Yu F. UniDepth: universal monocular metric depth estimation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024. – P. 10106–10116.