СТЕРЕОЗРЕНИЕ ДЛЯ РОБОТОТЕХНИКИ

Стереозрение — это восстановление глубины сцены по двум изображениям с калиброванной пары камер. Для робота это не академический трюк, а сенсор опережающего торможения: без карты глубины платформа не избегает препятствий, манипулятор не знает, где остановиться перед поверхностью, дрон не может летать в помещении.

Последние восемь лет дали нам целую линию архитектур — от тяжёлых трёхмерных свёрточных сетей с полным cost volume до итеративных и гибридных методов, которые не только точнее, но и переносятся между сценами и освещениями. Параллельно развивались монокулярные модели глубины: они не дают метрическую глубину сами по себе, но обеспечивают очень устойчивые priors о структуре сцены. И сегодня эти два мира официально слились во что-то вроде DEFOM-Stereo (CVPR 2025), где стереомодель усиливают «фундаментальной» моделью глубины.

Ниже я соберу ключевые модели стереозрения (GC-Net, PSMNet, GA-Net, AANet, RAFT-Stereo, IGEV-Stereo), покажу роль монокулярных моделей (Monodepth2, AdaBins, MiDaS), и объясню, почему DEFOM-Stereo — это логичное направление следующего шага. В конце дам ссылки на оригинальные статьи.

1. ЭВОЛЮЦИЯ СТЕРЕОСЕТЕЙ: ОТ ОБУЧАЕМОГО COST VOLUME К ИТЕРАТИВНОМУ УТОЧНЕНИЮ

Современная история стереозрения стартовала с GC-Net (ICCV 2017) [1]. До неё стерео выглядело как каскад из отдельных блоков: выделили признаки, посчитали соответствия, сгладили эвристикой. GC-Net впервые обучила всё это сквозным образом end-to-end. Сеть строит трёхмерный cost

volume (соответствие «пиксель слева ↔ возможные смещения справа»), затем прогоняет его через 3D-CNN для регуляризации и восстанавливает карту диспаритета через дифференцируемый soft-argmin. Это дало субпиксельную точность без ручных правил и фактически узаконило идею: стерео — это одна нейросеть, а не пайплайн из костылей [1]. GC-Net стала отправной точкой почти для всех последующих работ.

Дальше пришла PSMNet (CVPR 2018) [2]. Её ключевая мысль — стереомодель должна понимать сцену глобально, а не только локальными патчами. PSMNet вводит Pyramid Spatial Pooling: многоуровневое агрегирование признаков, которое даёт сети контекст целого объекта и всей сцены. Это критично в реальных сценах, где половина предметов однотонные и блестящие (дверца шкафа, белый стол, пластиковый корпус). PSMNet также использует каскадные hourglass-блоки в 3D, постепенно уточняя диспаритет. В итоге она стала эталоном на KITTI и надолго закрепила идею «стерео должно видеть не только текстуру, но и структуру» [2].

Ранние модели были точными, но прожорливыми. Полный 3D cost volume + глубокий 3D-CNN — это тонна памяти и миллисекунды, которые превращаются в сотни миллисекунд. GA-Net (CVPR 2019) [3] предложила шаг к практичности. Она встроила в нейросетевой контур то, что раньше считалось «классикой» стерео: полу-глобальную агрегацию (в духе Semi-Global Matching) и guided-фильтрацию по исходному изображению. Эти механизмы стали дифференцируемыми слоями Guided Aggregation. Получилась сеть, которая ведёт себя почти как огромная 3D-CNN, но вычислительно дешевле [3]. Это был первый серьёзный признак зрелости: мы

больше не просто накачиваем модель параметрами, мы начинаем оптимизировать архитектурную логику.

AANet (CVPR 2020) [4] закрепила этот разворот в сторону скорости. Авторы сформулировали задачу почти в индустриальных терминах: стереомодель должна работать в реальном времени. AANet уходит от тяжёлой обработки полного 3D cost volume и использует адаптивную агрегацию признаков на разных масштабах плюс деформируемые свёртки. Деформируемые свёртки дают возможность «подогнуть» фильтр под геометрию локального объекта и не размыть границу. В результате глубина получается достаточно точной при латентности порядка десятков миллисекунд на изображениях масштаба KITTI, то есть близко к 15+ FPS на стандартном GPU [4]. Впервые можно честно сказать: да, это реалистично для бортового робота, а не только для оффлайн-анализа.

2. ИТЕРАЦИОННЫЕ МЕТОДЫ

Следующий скачок — отказ от идеи «сеть предсказывает глубину один раз и всё». RAFT-Stereo (3DV 2021) [5] перенесла в стереозрение принцип, придуманный для оптического потока в RAFT: итеративное уточнение. Вместо одного жёсткого прогноза сеть строит плотную карту соответствий между левым и правым изображением, а затем многократно, шаг за шагом, обновляет оценку диспаритета с помощью рекуррентных ConvGRU-блоков. Это как маленький внутренний оптимизатор, который раз за разом подчищает ошибки на границах, в бликах, в зонах с плохой текстурой. Важный факт: RAFT-Stereo стала лидером на Middlebury и ETH3D — наборах, которые считаются особо сложными из-за близких дистанций, тонких деталей и нелинейных отражений [5]. То есть модель перестала быть

«только для дороги днём» и стала реально сильной там, где живут манипуляторы, складские тележки и сервисные роботы: стол, полка, рука оператора, блестящая упаковка.

IGEV-Stereo (CVPR 2023) [6] пошла ещё дальше. Авторы отметили, что одной итеративности мало: нужно явно кодировать геометрию сцены, иначе сеть может путаться в областях с окклюзиями и слабой текстурой. IGEV-Stereo вводит так называемый Iterative Geometry Encoding Volume — объём признаков, который уже на старте несёт геометрию сцены. Сеть сначала получает разумную начальную карту глубины из этого геометрического объёма, а потом дорабатывает её итеративно в стиле RAFT-Stereo. Получается гибрид: у нас есть «хороший черновик» и есть механизм его дошлифовки.

Практический результат оказался показательный. IGEV-Stereo заняла первое место среди опубликованных методов на KITTI 2012 и KITTI 2015, но не только там: она демонстрирует устойчивую переносимость между доменами (уличные сцены KITTI → сложные indoor-сцены Middlebury и ETH3D) при адекватной скорости инференса [6]. Это уже звучит как то, что хочется поставить на робота без тотальной переобучаемости под каждую комнату.

3. МОНОКУЛЯРНАЯ ГЛУБИНА

Интересно, что параллельно с эволюцией стерео росла другая ветка — монокулярная оценка глубины из одного RGB кадра. Эти модели по определению не дают абсолютную метрическую глубину «в метрах» без доп.калибровки, но они невероятно хорошо понимают относительную

структуру сцены и обобщаются между доменами. Сегодня они не конкуренты стерео, а его союзники.

Monodepth2 (ICCV 2019) [7] стала «золотым стандартом» для self-supervised монокулярной глубины. Идея проста и мощна: мы учим модель предсказывать глубину так, чтобы с помощью этой глубины и оценки относительного движения камеры можно было пересинтезировать соседние кадры и минимизировать фотометрическую ошибку. Авторы добавили несколько практичных улучшений — маскирование движущихся объектов, многомасштабный лосс, аккуратное обращение с окклюзиями — и добились того, что одно изображение даёт очень резкую, детализированную карту глубины без GT лидара [7]. Это взорвало область, потому что снимает зависимость от дорогой плотной разметки.

AdaBins (CVPR 2021) [8] показала, что можно ещё поднять потолок точности, если относиться к глубине не только как к регрессии пиксель-за-пикселем. Модель делит диапазон глубины на набор «карманов» (bins), но не фиксированных — она адаптивно предсказывает центры этих бинов под конкретную сцену с помощью трансформерного блока. Затем итоговая карта глубины получается, как взвешенная комбинация этих центров. Такой дискретно-непрерывный взгляд оказался очень эффективным на датасетах с точной аннотацией (NYU-Depth-v2, KITTI), и долгое время AdaBins считалась эталоном supervised-точности для одного кадра [8].

MiDaS (Ranftl et al., 2019–2020) [9] решила другую боль: робастность. Авторы показали, что можно обучать одну модель на смеси разных датасетов с несовместимой разметкой глубины (разные диапазоны, разные форматы — лидар, стерео, структурированный свет) и при этом добиться потрясающей

zero-shot переносимости. MiDaS не всегда даёт метры, но даёт очень хорошую относительную глубинную структуру для практически любых изображений «из мира», от улицы до офиса, без дообучения [9]. Это был первый сигнал того, что большие универсальные модели глубины могут стать чем-то вроде «фундамента», на который потом можно навесить более точную геометрию.

4. НОВАЯ ВОЛНА: DEFOM-STEREO И СЛИЯНИЕ ФУНДАМЕНТАЛЬНЫХ МОДЕЛЕЙ С СОПРЯЖЁННОЙ ГЕОМЕТРИЕЙ

Логичное продолжение этой истории — DEFOM-Stereo (CVPR 2025) [10]. Авторы исходят из наблюдения: современные монокулярные модели глубины (например, Depth Anything V2 на базе больших vision transformer-архитектур) демонстрируют фантастическую переносимость между доменами, но не знают абсолютного масштаба; стереомодели наоборот дают метрическую глубину, но могут «поплыть» в текстурно бедных или блестящих областях, особенно вне своего тренировочного домена.

DEFOM-Stereo берёт рекуррентный стереофреймворк в духе RAFT-Stereo и встраивает в него мощную предобученную монокулярную модель глубины как источник начальной геометрии. Сначала эта «фундаментальная» модель (depth foundation model) даёт хорошую относительную карту глубины и богатые признаки глобальной сцены. Эти признаки сливаются с обычными CNN-признаками, а полученная карта используется как инициализация диспаритета. Далее идёт итеративное уточнение диспаритета, но уже не с нуля, а начиная с осмысленной инициализации. Дополнительно добавлен модуль масштабной коррекции,

который правит метрику — то есть помогает превратить «я знаю форму сцены» в «я знаю расстояния в корректной шкале» [10].

Результат впечатляющий: DEFOM-Stereo показывает улучшенную zero-shot робастность (то есть перенос без дообучения) и выходит на топовые места сразу на четырёх классических бенчмарках стереозрения — KITTI 2012, KITTI 2015, Middlebury и ETH3D — а также сильно выступает в общем зачёте Robust Vision Challenge [10]. Это уже не просто ещё одна SOTA-кривая. Это архитектурный сдвиг: стерео перестаёт быть чисто бинокулярным зрением и превращается в гибрид, который опирается как на геометрию двух камер, так и на высокоуровневые представления сцены из больших «foundation» моделей глубины.

5. КОНТЕКСТ ДАННЫХ: INSTEREO2K КАК БАЗОВЫЙ ДОМЕН ДЛЯ РОБОТА В ПОМЕЩЕНИИ

Важный практический момент — на каком типе данных всё это должно учиться и проверяться, если нам нужен робот в помещении, а не беспилотник на трассе. Для такой задачи критичен ближний диапазон (десятки сантиметров — пара метров), сложные отражающие поверхности, полки, провода, человеческие руки в кадре. Именно под это был собран датасет InStereo2K [11]. Он содержит около 2050 пар реальных стереоизображений с очень точной картой диспаритета, снятых в бытовых и офисных помещениях: из них ~2000 пар идут на обучение и ~50 — на тест. Авторы показали, что дообучение классических стереомоделей на InStereo2K резко улучшает их качество на других indoor-бенчмарках вроде Middlebury, то есть модель действительно становится лучше в «жизни на столе», а не только «на дороге» [11].

Для задач робототехники в помещении (манипулятор, складской помощник, сервисный робот) это значит следующее. Модели уровня RAFT-Stereo [5], IGEV-Stereo [6] и теперь DEFOM-Stereo [10] — это не просто красивые цифры в таблице KITTI. Они уже умеют работать на ближних дистанциях и переноситься между доменами, а наличие такого набора как InStereo2K позволяет целево дообучить их под Indoor-сценарий без потери общей устойчивости. Это ровно то, что нужно, если робот завтра должен не умереть, столкнувшись с блестящей алюминиевой деталью на складе.

6. КОРОТКО

GC-Net [1] и PSMNet [2] сделали стереозрение полностью обучаемым и научили сеть смотреть на сцену целиком. GA-Net [3] и AANet [4] показали, что глубину можно считать быстро и экономно, а не только дорого и очень точно. RAFT-Stereo [5] принесла идею итеративного уточнения, которая особенно хорошо работает в ближнем диапазоне — там, где живёт робот. IGEV-Stereo [6] добавила явную геометрию сцены и стала новым эталоном переносимой точности. Параллельно Monodepth2 [7], AdaBins [8] и MiDaS [9] научили модели понимать глубину по одному кадру и обобщаться между доменами. DEFOM-Stereo [10] официально скрестила эти две линии: стерео теперь использует знания крупной монокулярной «foundation» модели глубины, получая и метрическую точность, и робастность к новым условиям.

С практической точки зрения, для робототехники в помещении это означает, что стереозрение больше не выглядит как лабораторная роскошь. При правильном обучении (включая данные вроде InStereo2K [11]) современные модели действительно стали тем, что можно ставить на борт.

ССЫЛКИ

[1] A. Kendall, Y. Gal, R. Cipolla. "End-to-End Learning of Geometry and Context for Deep Stereo Regression." ICCV 2017. arXiv:1703.04309. https://arxiv.org/abs/1703.04309

[2] J.-R. Chang, Y.-S. Chen. "Pyramid Stereo Matching Network (PSMNet)." CVPR 2018. arXiv:1803.08669. https://arxiv.org/abs/1803.08669

[3] F. Zhang, V. Prisacariu, R. Yang, P. H. S. Torr. "GA-Net: Guided Aggregation Net for End-to-End Stereo Matching." CVPR 2019. arXiv:1904.06587. https://arxiv.org/abs/1904.06587

[4] H. Xu, J. Zhang. "AANet: Adaptive Aggregation Network for Efficient Stereo Matching." CVPR 2020. arXiv:2004.09548. https://arxiv.org/abs/2004.09548

[5] L. Lipson, Z. Teed, J. Deng. "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching." 3DV 2021. arXiv:2109.07547. https://arxiv.org/abs/2109.07547

[6] G. Xu, X. Wang, X. Ding, X. Yang. "Iterative Geometry Encoding Volume for Stereo Matching (IGEV-Stereo)." CVPR 2023. arXiv:2303.06615. https://arxiv.org/abs/2303.06615

[7] C. Godard, O. Mac Aodha, M. Firman, G. J. Brostow. "Digging Into Self-Supervised Monocular Depth Estimation (Monodepth2)." ICCV 2019. openaccess.thecvf.com/... / arXiv version. https://openaccess.thecvf.com/content_ICCV_2019/papers/Godard_Digging_Into_Self-Supervised_Monocular_Depth_Estimation_ICCV_2019_paper.pdf

[8] S. F. Bhat, I. Alhashim, P. Wonka. "AdaBins: Depth Estimation Using Adaptive Bins." CVPR 2021. arXiv:2011.14141. https://arxiv.org/abs/2011.14141

[9] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer (MiDaS)." arXiv:1907.01341, ICCV 2019 / TPAMI later. https://arxiv.org/abs/1907.01341

[10] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, R. Huang. "DEFOM-Stereo: Depth Foundation Model Based Stereo Matching." CVPR 2025. arXiv:2501.09466. https://arxiv.org/abs/2501.09466

[11] W. Bao, W. Wang, Y. Xu, Y. Guo, S. Hong, X. Zhang. "InStereo2K: A large real dataset for stereo matching in indoor scenes." Science China Information Sciences, 2020. DOI:10.1007/s11432-019-2803-x. https://link.springer.com/article/10.1007/s11432-019-2803-x