

Шпаргалка: задержка и точность по уровням

Сопоставьте каждую аналитику с уровнем под её бюджет производительности: как быстро приходит детекция и как часто она верна. Представительные цифры на 2026.

А. Бюджет задержки детекции (четыре этапа)

- Захват и кодирование: ~15-33 мс, одинаково на всех уровнях (происходит внутри камеры).
- Кадр к вычислениям: ~0 на камере, 1-5 мс по LAN, 40-200 мс через интернет.
- Инференс: 8-30 мс малая модель на NPU камеры; 10-25 мс крупнее на GPU.
- Доставка результата: 2-10 мс к локальному реле или 40-200 мс обратно через интернет.

В. Сквозная задержка и правило размещения

- На камере ~25-100 мс; локальный сервер ~30-120 мс; облако ~300-800 мс (сотовая >1 200 мс).
- Облако медленно из-за сетевого обмена, а не из-за вычислений.
- До 200 мс - только на краю; 200 мс-1 с = гибрид; несколько секунд = облако подходит.
- За 800 мс (обмен с облаком) человек проходит ~1,1 м, а за 60 мс (на камере) - ~8 см.

С. Потолок точности по уровням

- Размер модели задаёт потолок: nano ~40% mAP (камера) до xlarge ~55% mAP (облако) на COCO.
- Квантизация под камеру стоит 5-8% mAP; меньше 3% при калибровке на реальном видео.
- Реальный плюс облака - рассуждение по камерам (ре-ид, VLM), а не детекция в одном кадре.
- Точность - диапазон precision/recall: LPR 90-98%, лица зависят от условий. Никогда не 100%.

Д. Когда миллисекунды важны (а когда нет)

- Край (до 200 мс): периметр, вторжение, защита от потерь на кассе, стоп-линия, падения.
- Локальный сервер: тяжёлая детекция в реалтайме по многим камерам в одной локальной сети.
- Облако (секунды - норма): подсчёт людей, dwell-время, теплокарты, трендовая аналитика.
- Рассуждение в облаке: ре-идентификация по камерам, описание сцен VLM, поиск по архиву.

Размещайте детекцию у камеры, где важны рефлекс, и оставляйте облаку рассуждение по камерам, где важно суждение, а секунды допустимы. Указывайте точность как диапазон precision/recall, измеренный на своём видео, а не цифру из даташита. Инженерное руководство, не юридический совет - распознавание лиц и номеров биометрично и ограничено законом (EU AI Act - высокий риск; Illinois BIPA - согласие) независимо от уровня. Цифры - представительные для оборудования 2026 (NPU камеры, локальный GPU-сервер, облачный GPU-регион) и зависят от размера модели, разрешения, FPS и сети. Источники: NIST FRVT; ONVIF Profile M; IETF RFC 3550; бенчмарки Ultralytics/COCO YOLO; NVIDIA DeepStream; актуальные разборы задержки edge-vs-cloud.