Семантическая реконструкция 3D-сцен в реальном времени с использованием гауссовского сплаттинга

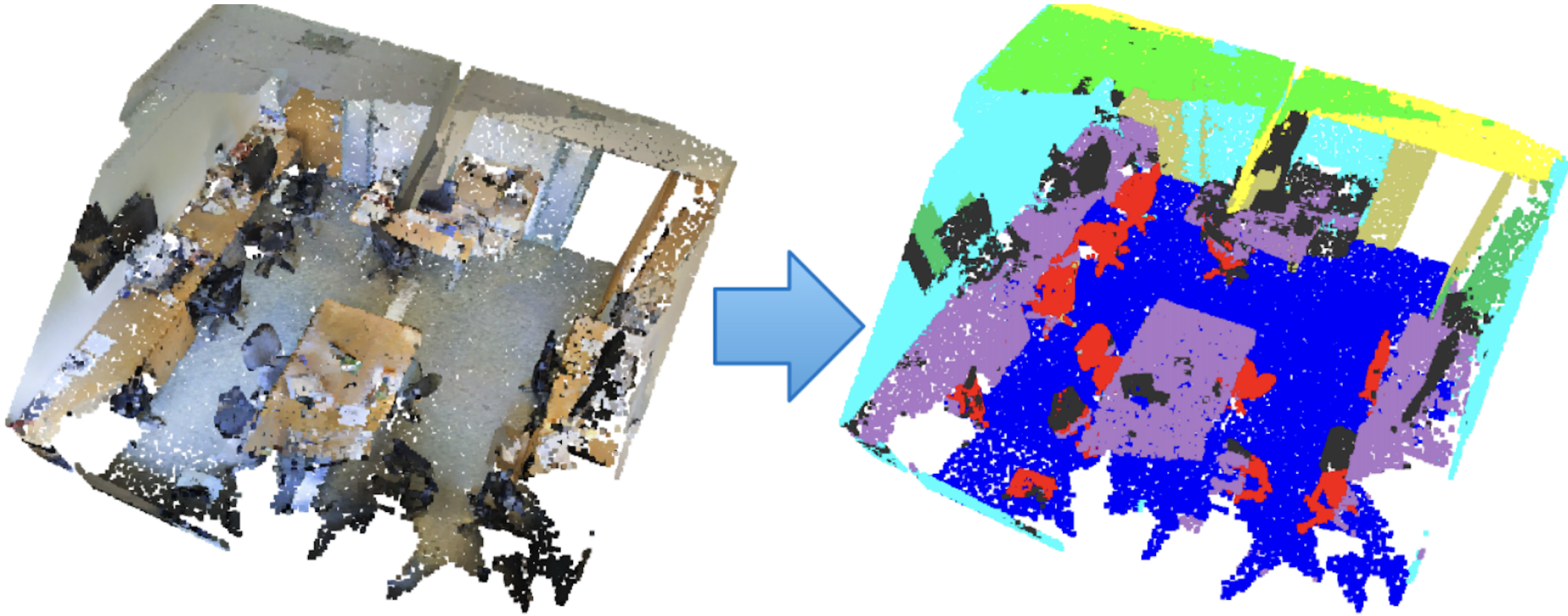
Титков Р.Е.
titkov.re@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

Мотивация

- Современные задачи в области робототехники требуют быстрой и точной семантической 3D-реконструкции сцен.
- Существующие методы обладают рядом ограничений:
  - высокая вычислительная сложность;
  - ограниченный набор семантических классов (closed-vocabulary);
  - низкая точность при использовании открытого словаря (open-vocabulary).



Цель работы

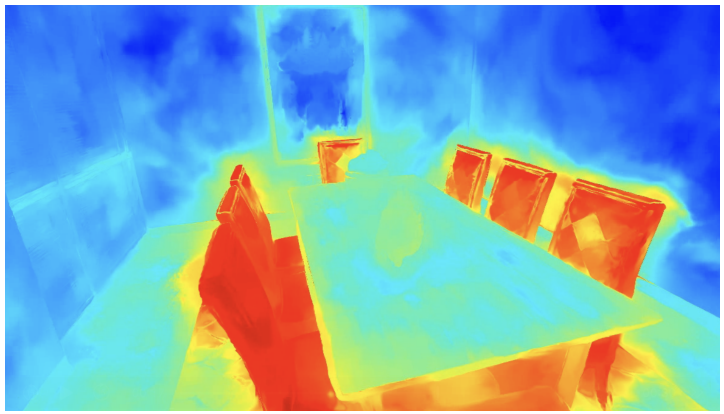
Разработка метода семантической 3D-реконструкции сцен в реальном времени на основе гауссовского сплаттинга с поддержкой open-vocabulary сегментации.



RGB-реконструкция

$\Rightarrow$

Запрос:
“chair”



Семантическая карта

Рис.: open-vocabulary сегментация по текстовому запросу.

Метрики оценки качества реконструкции

Визуальные метрики:

- **PSNR:** $10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right)$ — измеряет степень искажения реконструированного изображения.
- **SSIM:** оценивает структурное сходство изображений по яркости, контрасту и локальной структуре.

Семантические метрики:

- **Accuracy:** $\frac{\text{Correct}}{\text{Total}}$ — доля правильно классифицированных точек.
- **IoU:** $\frac{TP}{TP + FP + FN}$, **mIoU** — среднее по классам, мера перекрытия предсказанных и истинных масок.

Метрики рассчитываются по каждому кадру с последующим усреднением.

Постановка проблемы

Дано:
Последовательность RGB-D кадров $\{F_1, \ldots, F_n\}$
Требуется:
Построить 3D-реконструкцию сцены $\mathcal{R}$, которая обеспечивает:

- Высокое качество визуальной реконструкции;
- Возможность семантической сегментации по произвольным текстовым запросам (оценка на тестовых классах).

Задача оптимизации:

$$\max_{\mathcal{R}} \left[Q_{\text{visual}}(F_i, \mathcal{R}) + Q_{\text{sem}}(F_i, \mathcal{R}, \text{Classes}_{\text{test}}) \right]$$

Где:

- Q_{visual} — качество визуальной реконструкции: PSNR, SSIM;
- Q_{sem} — точность семантической разметки: mIoU, Accuracy.

Текущий прогресс работы

Завершённые этапы:

✓ Анализ моделей OpenSeg и DINOv2 для извлечения признаков; финальный выбор — DINOv2
✓ Разработка модели для сжатия признаков (PCA)
✓ Проведены эксперименты с размерностью сжатых признаков
✓ Интеграция DINOv2 и Compressor в основной пайплайн
✓ Интеграция Talk2DINO для сегментации по текстовому запросу
✓ Реализация полного пайплайна визуально-семантической 3D-реконструкции

Планируемые задачи:

☐ Внедрение обновлённого растеризатора для улучшения качества рендеринга
☐ Расширенные эксперименты с альтернативными методами сжатия признаков

Используемые методы

Базовый метод:

- **Photo-SLAM** — SLAM-система на основе Gaussian Splatting, выполняющая 3D-реконструкцию сцен с использованием RGB-D последовательностей.

Базовый метод был расширен с помощью:

- **DINOv2** — извлечение визуальных признаков из изображений.
- **Talk2DINO** — преобразование текстовых запросов в пространство визуальных признаков.
- **PCA** — сжатие признаков до размерности K для повышения вычислительной эффективности.

Предложенный подход

Ключевые особенности предлагаемого решения:

+ Использование семантики в 3D-реконструкции
+ Единая функция потерь, учитывающая цвет, глубину и семантику
+ Реконструкция в реальном времени
– Сжатие эмбеддингов может снижать точность сегментации

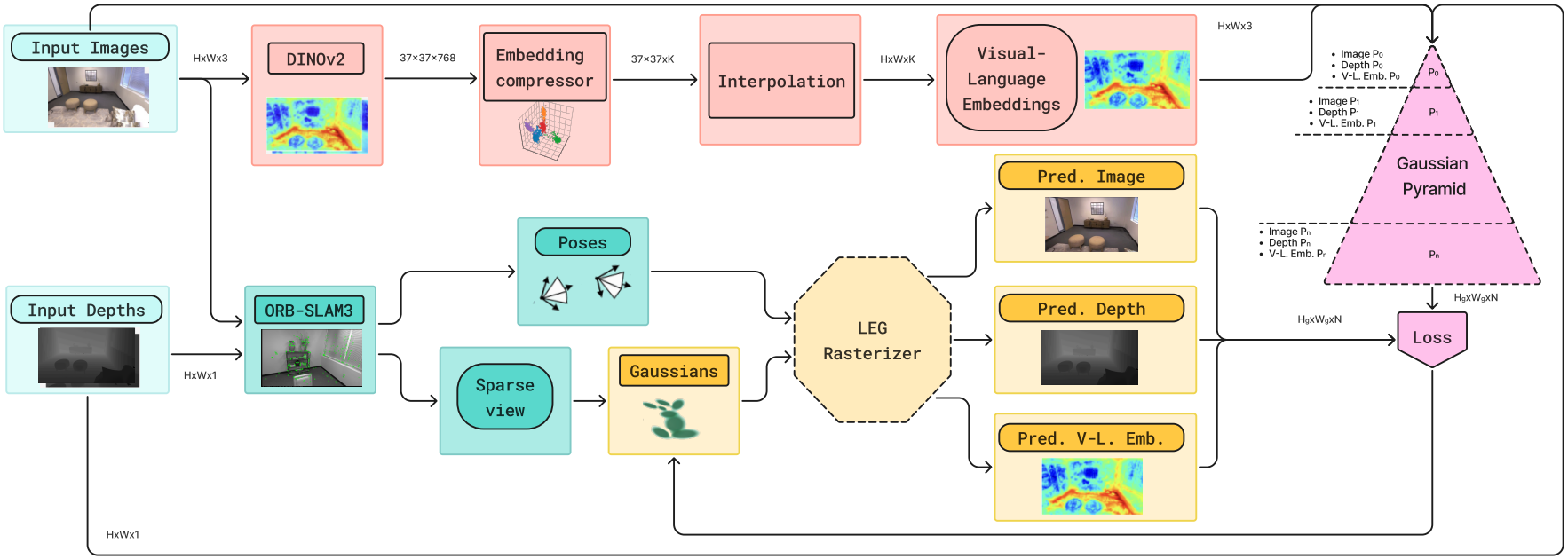
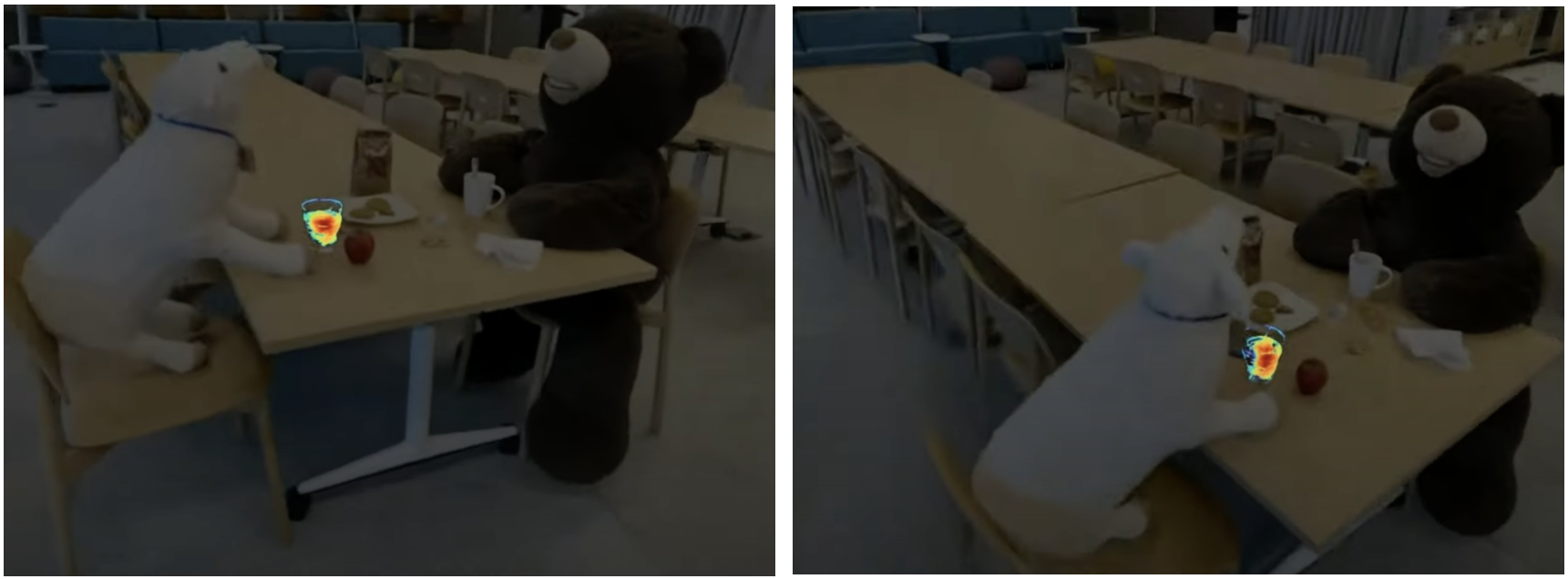


Рис.: Архитектура предложенного подхода

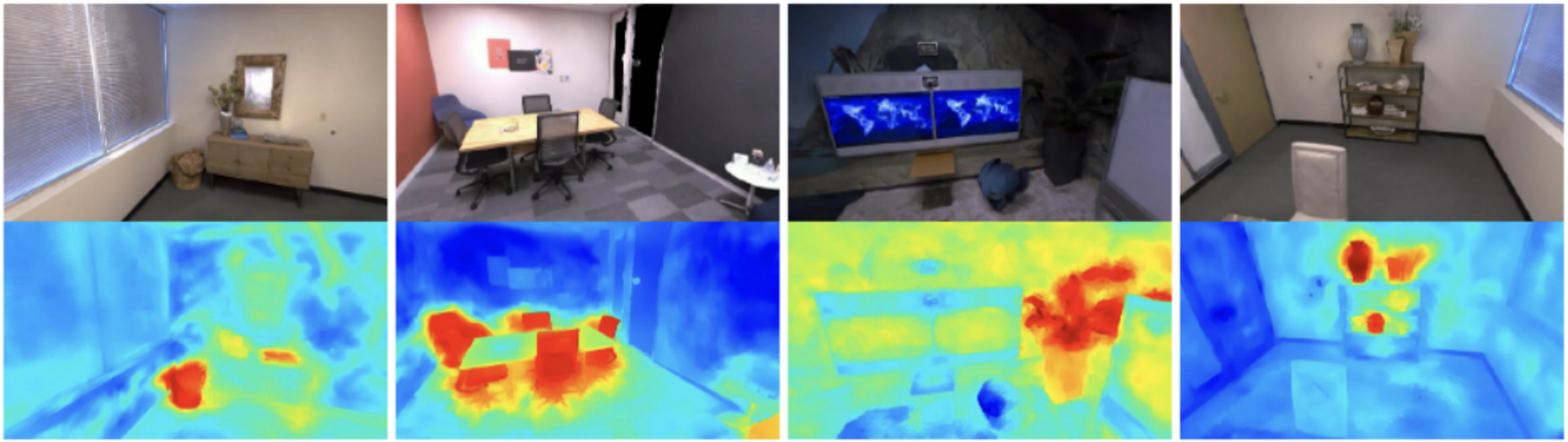
Семантическая сегментация объектов

1. Семантическая карта признаков формируется путём проекции гауссиан из 3D-сцены на плоскость текущей камеры.
2. Текстовый запрос преобразуется в эмбеддинг с помощью модели Talk2DINO.
3. Объекты локализуются на изображении на основе сравнения эмбеддингов сцены и текстового описания.



Локализация объекта по текстовому запросу “tea mug”

Семантическая сегментация объектов



(a) Text query: "basket" (b) Text query: "chair" (c) Text query: "plant" (d) Text query: "vase"

Рис.: Примеры различных запросов на сценах из датасета Replica

Эксперименты с разной степенью сжатия эмбеддинга

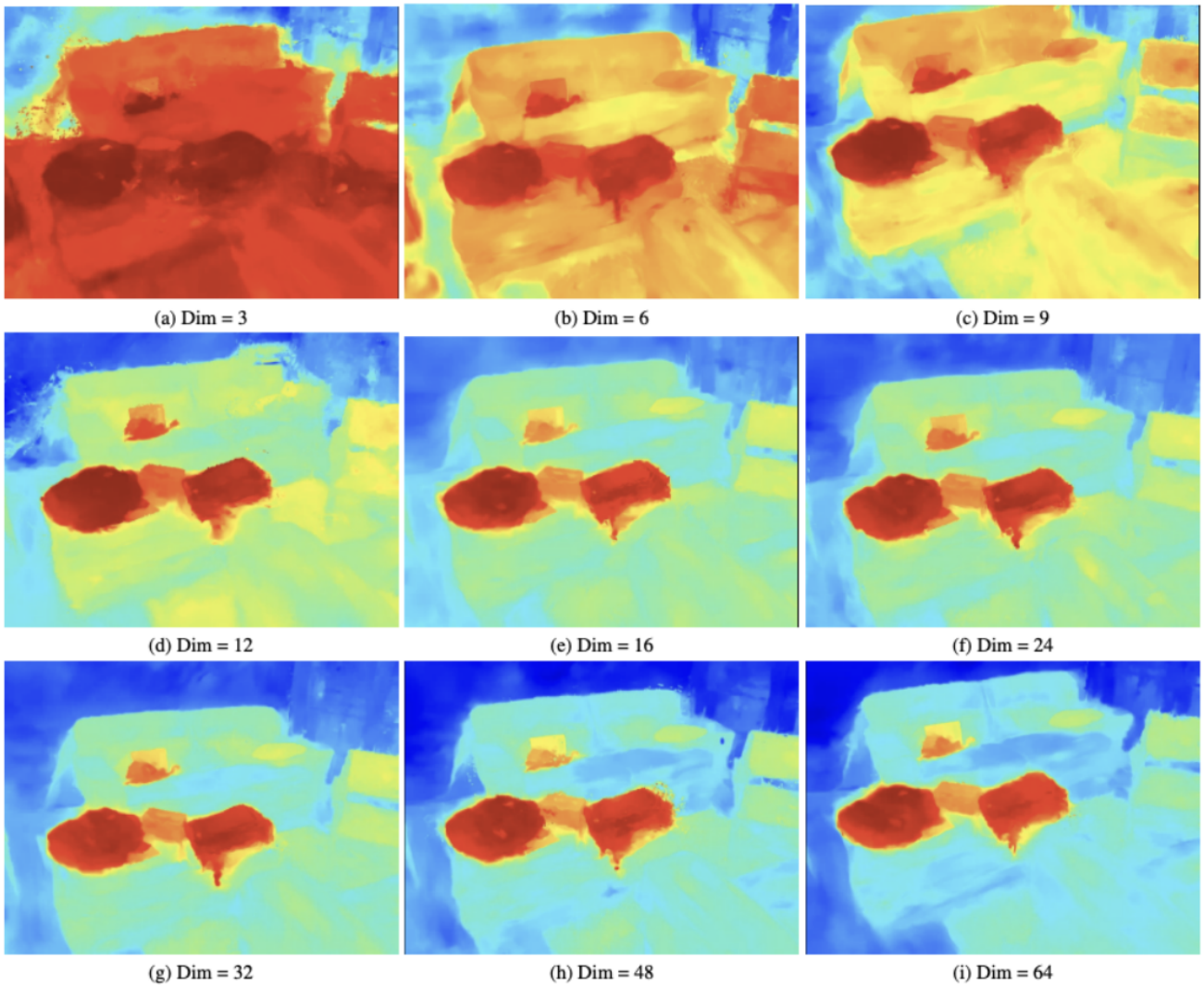


Рис.: Сравнение качества сегментации по запросу **backpack** для разных рахмерностей сжатого эмбеддинга

Сравнение качества сегментации с SOTA-методами

Method	**Backbone**	**mIoU, %**	**mAcc, %**	**Training Time**
LERF	NeRF+CLIP	31.2	61.7	45 min
PVLFF	NeRF+LSeg	52.9	67.0	65 min
LangSplat	3DGS+CLIP	24.7	42.0	180 min
Feature3DGS	3DGS+LSeg	59.2	75.1	150 min
Semantic Gaussians	3DGS+LSeg	**62.0**	**77.0**	90 min
LEG-SLAM	3DGS+DINOv2	41.4	74.3	**1.5 min**

Таблица: Сравнение качества Open-Vocabulary сешментации на 12 сценах датасета ScanNet (в среднем 1800 кадров на сцену)

Сравнение качества сегментации с SOTA-методами

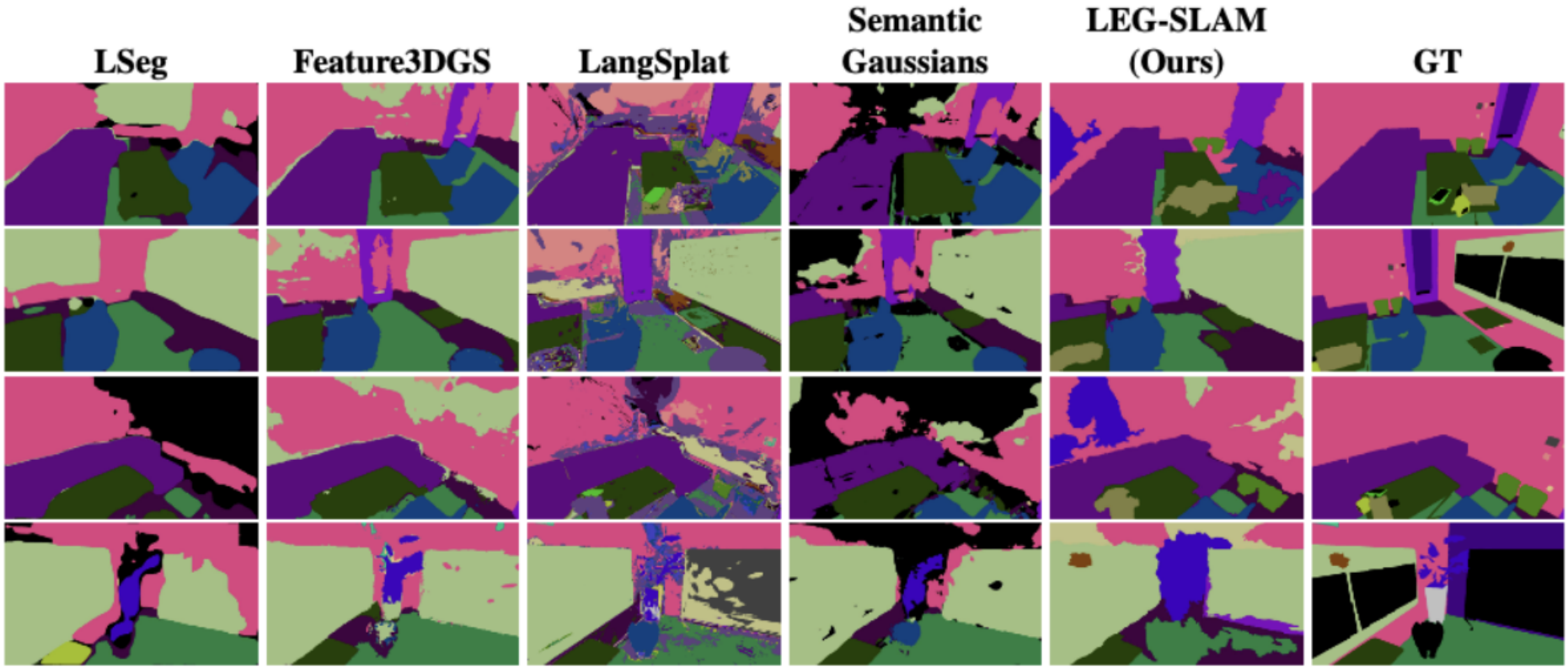


Рис.: Сравнение качества Open-Vocabulary сегментации на датасете Replica

Сравнение с современными SLAM-подходами

Method	**ATE RMSE ↓**	**FPS ↑**	**PSNR (dB) ↑**	**SSIM ↑**	**LPIPS ↓**	**Semantics**
NICE-SLAM	2.51	0.198	24.42	0.809	0.233	-
Vox-Fusion	1.47	1.07	24.41	0.801	0.236	-
Co-SLAM	1.06	6.41	30.24	0.939	0.252	-
ESLAM	0.62	3.02	29.08	0.929	0.239	-
SplaTAM	**0.41**	0.21	**34.11**	**0.968**	**0.102**	-
Photo-SLAM	1.24	27.35	30.91	0.883	0.143	-
SNI-SLAM	0.46	2.15	29.43	0.935	0.235	Closed-Vocabulary
LEG-SLAM (Ours)	0.94	**10.56**	32.12	0.9151	0.1456	**Open-Vocabulary**

Таблица: Сравнительный анализ точности локализации, скорости и качества рендеринга на датасете Replica (8 сцен).

План публикаций

✓ **Принята к публикации:**

- **Beyond Bare Queries: Open-Vocabulary Object Grounding with 3D Scene Graph**
  Sergey Linok, Tatiana Zemskova, Svetlana Ladanova, Roman Titkov, Dmitry Yudin, Maxim Monastyrny, Aleksei Valenkov
  ICRA 2025 (Accept)

☐ **Подана на рецензию:**

- **LEG-SLAM: Real-Time Language-Enhanced Gaussian Splatting for SLAM**
  Roman Titkov, Egor Zubkov, Dmitry Yudin, Jaafar Mahmoud, Malik Mohrat, Gennady Sidorov
  ICCV 2025 (Under Review)

Дальнейшие работы

- Провести расширенное тестирование различных моделей для сжатия признаков.
- Внедрить улучшенный растеризатор для повышения качества рендеринга.

Спасибо за внимание!

titkov.re@phystech.edu