Сегментация объектов сложной формы по текстовому запросу

Ветошкин Л.Н.
vetoshkin.ln@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар



План доклада

- Мотивация
- Постановка задачи
- Описание метода
- Проделанная работа
- Дальнейшие исследования

Мотивация

Сегментация объектов по запросу — это перспективное направление, которое применяется в различных областях, таких как робототехника, беспилотные автомобили и разметка данных. Данный тип сегментации способен сегментировать объекты различных классов.

Проблемой на данный момент является то, что современные архитектуры показывают низкое качество при сегментации объектов сложной формы.



.

Задача

Если в структуре объекта будут присутствовать тонкие линии, такие объекты будем называть - *объектами сложной формы*. Например: провода, велосипед, клетка и т.п.

Задача состоит в разработке метода, который будет максимизировать mIoU для объектов сложной формы. Формально задачу можно описать следующим образом:

$$\mathcal{C} = \{c_0, c_1, ...\} - \text{классы объектов сложной формы}$$

$$\text{mIoU}(\mathbf{M}^{\mathcal{C}}, \Phi(\mathbf{I}^{\mathcal{C}}, \mathbf{T})) \to max$$

где $\mathbf{I}^{\mathcal{C}}$ - изображения, содержащие объекты $c \in \mathcal{C}$, $\mathbf{T}$ - текстовый запрос, $\mathbf{M}^{\mathcal{C}}$ - истинные маски объектов $c \in \mathcal{C}$ находящихся на изображениях $\mathbf{I}^{\mathcal{C}}$, Φ - разработанный метод.

Проделанная работа

- Использован датасет ThinObject-5K[1] для сегментации тонких объектов
- Обучена модель Talk2DINO на этом датасете
- С помощью Talk2DINO получены маски и карты внимания
- Эти данные использованы для дообучения модели SAM
- Проведена валидация и визуализация результатов

[1] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In WACV, 2021

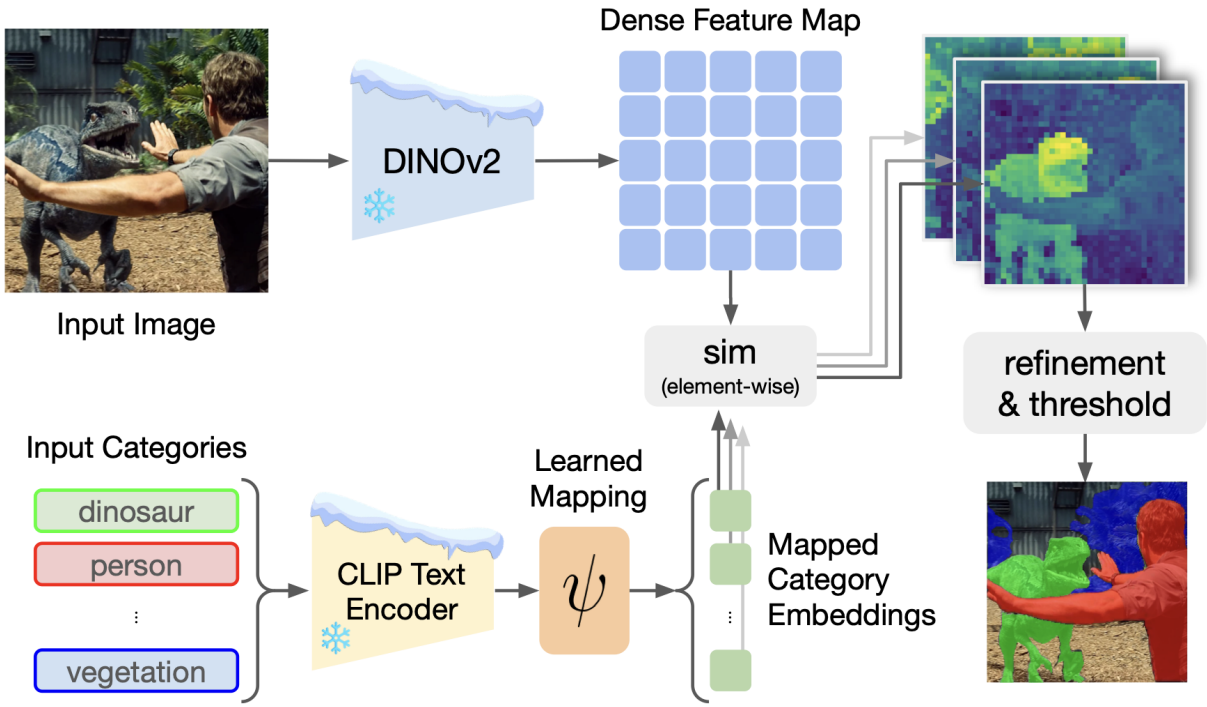
Датасет ThinObject-5K

- Датасет для задач сегментации **тонких объектов**, где стандартные модели часто теряют детали
- Содержит ∼5,000 изображений с ручной разметкой масок
- Включает категории как проволока, верёвки, антенны, зонтики, велосипеды и другие тонкоструктурные объекты
- **Имена файлов** имеют формат вида `[object]_PNG[N].jpg`, что позволяет удобно извлекать класс объекта



Talk2DINO

- Выполняет выравнивание текстовых эмбеддингов **CLIP** с визуальными признаками **DINOv2**
- Использует поканальные признаки DINO и обучаемую проекцию для сопоставления с языковыми запросами
- Позволяет обнаруживать и локализовать объекты по произвольному текстовому описанию



Дообучение Talk2DINO на ThinObject-5K

- Talk2DINO была дообучена на датасете **ThinObject-5K** для лучшего распознавания тонких объектов
- Улучшена способность модели локализовать и интерпретировать тонкие и узкие структуры
- Метрика **mIoU** улучшилась с **0.68** до **0.72**

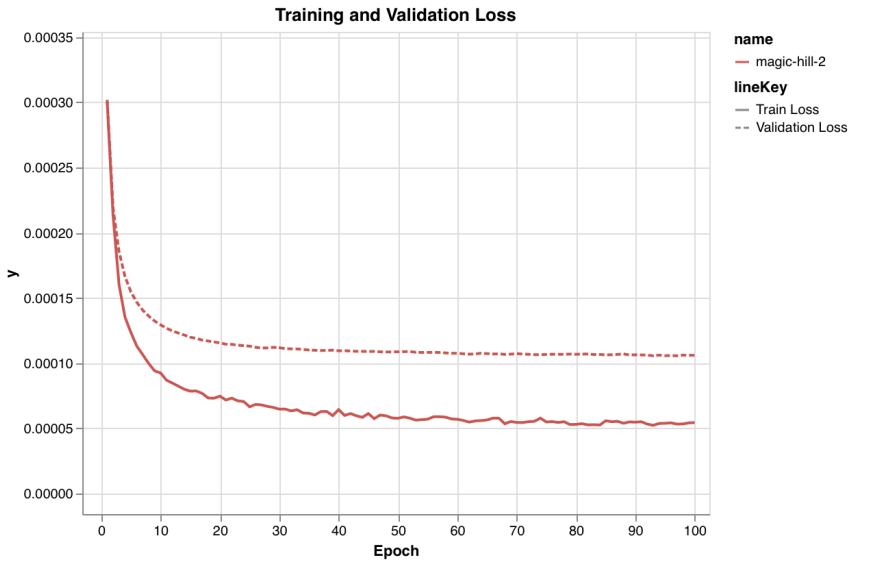
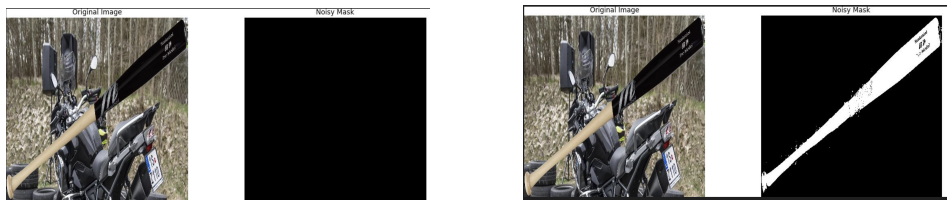


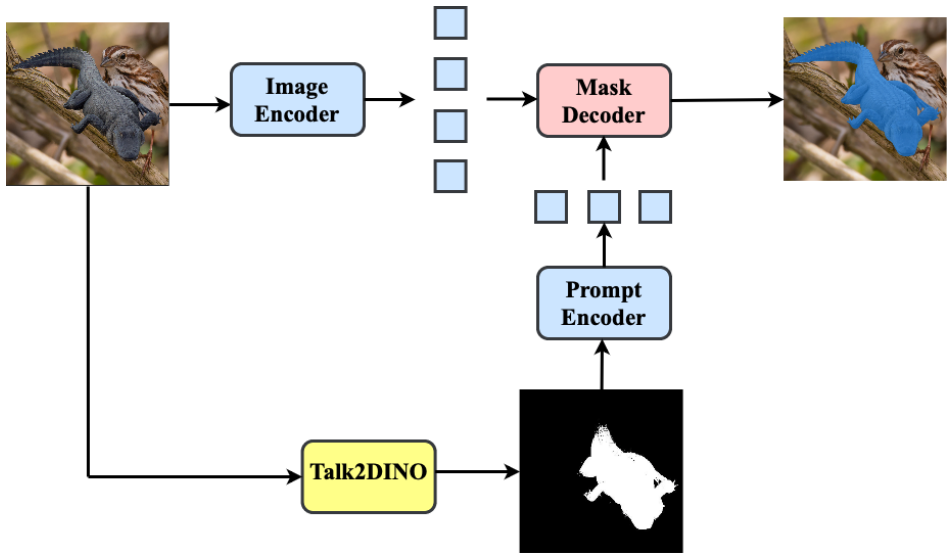
График обучения



Маска до / после обучения

SAM + Talk2DINO

- Для улучшения качества сегментации использованы маски, сгенерированные моделью **Talk2DINO**
- Дообучался только **Mask Decoder** модели **SAM**, остальные компоненты были заморожены
- Улучшение метрики **mIoU**:
  - Без Talk2DINO: **0.905**
  - С Talk2DINO: **0.919**



Архитектура SAM. Дообучался только блок **Mask Decoder** (обозначен цветом)

SAM+Talk2IDNO

- Визуальное сравнение предсказанных масок:
  - Слева — SAM
  - Справа — SAM+Talk2DINO
- Видно улучшение выделения тонких объектов и границ



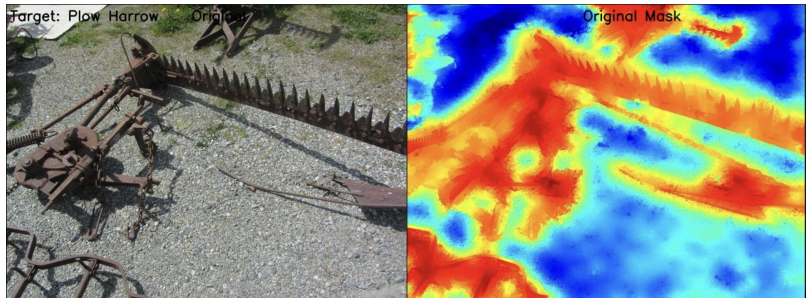
Без Talk2DINO С Talk2DINO

Улучшение сегментации краёв объектов

- Реализован модуль **AttentionFusion** для объединения признаков SAM и карты внимания DINOv2
- Цель — усилить границы объектов путём контекстного внимания между признаками
- **mIoU** улучшилось с **0.905** до **0.908**

Математическая формализация:

$$F'_{\text{dino}} = \text{Interpolate}(F_{\text{dino}}, H, W)$$
$$F^{256}_{\text{dino}} = \text{Conv}_{1\times1}(F'_{\text{dino}})$$
$$Q = \text{Flatten}(F^{256}_{\text{dino}}), \quad K, V = \text{Flatten}(F_{\text{sam}})$$
$$F_{\text{attn}} = \text{MultiHeadAttn}(Q, K, V)$$
$$F_{\text{fused}} = \text{LayerNorm}(F_{\text{attn}} + F_{\text{sam}})$$



Карта внимания DINOv2

Дальнейшие исследования

- **Доработка модуля уточнения границ объектов**
  - Улучшение механизма фьюзинга признаков для более точной локализации
  - Исследование альтернатив attention-механизмов
- **Тестирование модели на других датасетах**
  - Проверка обобщающей способности на других типах тонких объектов
  - Адаптация модели под более сложные и разнообразные сцены

Спасибо за внимание!

cogmodel.mipt.ru

vetoshkin.ln@phystech.edu