

SEGMENT ANYTHING MODEL

Fine-Tuning

Kaiky Braga
Larissa Afonso
Luciano Sampaio

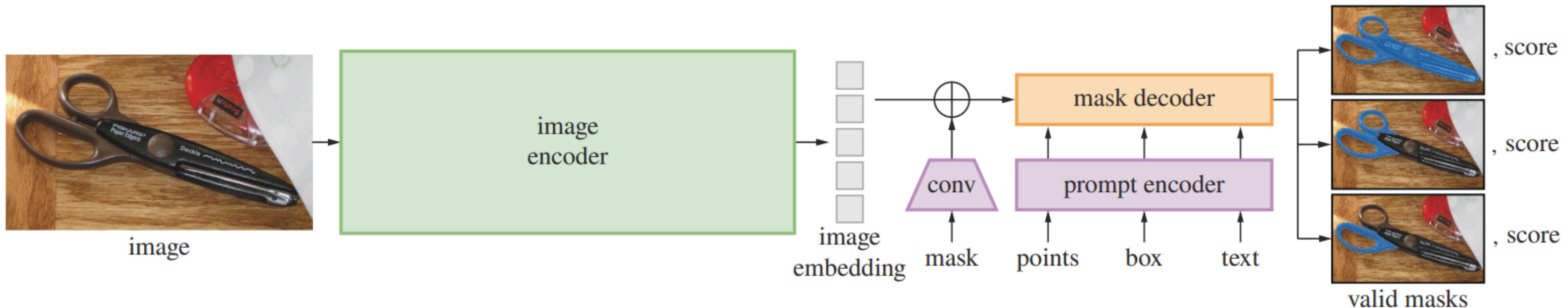
ARQUITETURA DO MODELO

O SAM é um modelo de segmentação generalista capaz de segmentar qualquer objeto em qualquer imagem usando apenas prompts fornecidos pelo usuário.

Características principais:

- Treinado em 1 bilhão de máscaras e 11 milhões de imagens (dataset SA-1B).
- Funciona via prompts:
 - Pontos (positivos/negativos)
 - Caixas delimitadoras
 - Máscaras aproximadas
- Alta generalização para domínios novos (zero-shot).
- Capaz de segmentar múltiplos objetos sem retrain.
- Rápido, escalável e adaptável via fine-tuning.

ARQUITETURA DO MODELO



ARQUITETURA DO MODELO

1. Image Encoder

Backbone Vision Transformer (ViT-H/L/B).

Converte a imagem em um mapa de embeddings de alta dimensão.

2. Prompt Encoder

Converte diferentes tipos de prompts (pontos, caixas, máscaras) em embeddings compatíveis com o modelo.

Tipos de prompts suportados:

- Pontos: codificados como embeddings espaciais.
- Caixas: codificadas como pares de coordenadas.
- Máscaras: passadas como mapas binários processados por CNN leve.

3. Mask Decoder

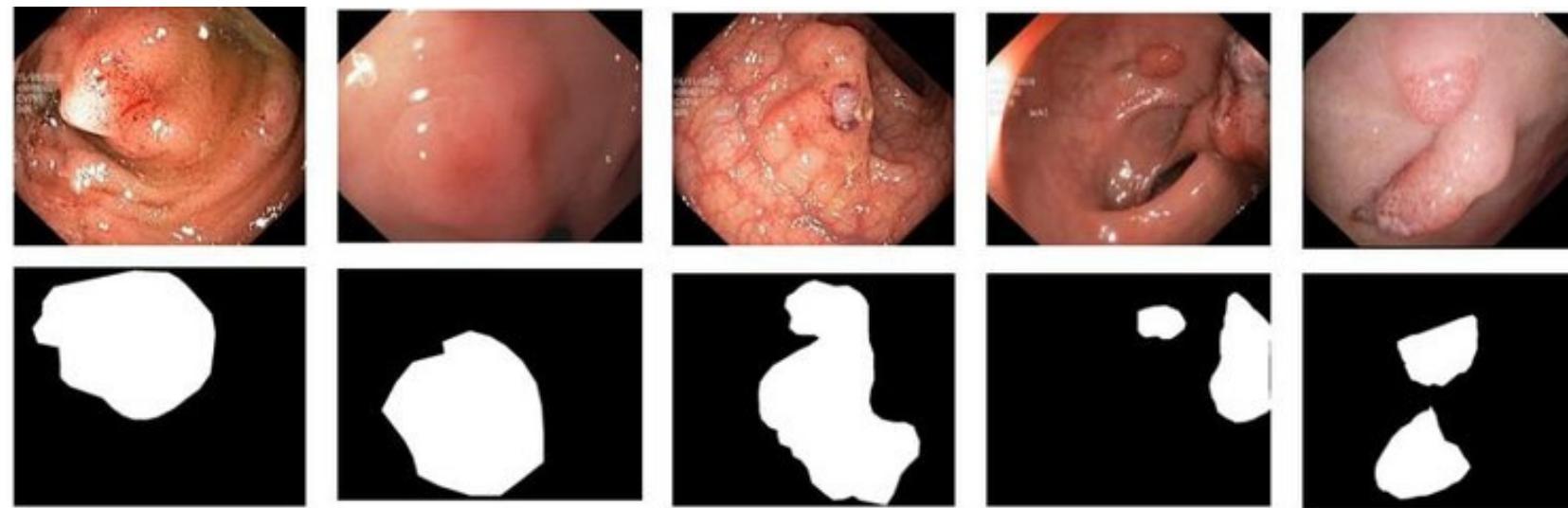
Combina:

- Embeddings da imagem
- Embeddings do prompt

Produz:

- 3 máscaras com níveis diferentes de confiança
- Scoring para selecionar a melhor máscara

Dataset Kvasir-SEG



Kvasir-SEG é um dataset público criado para tarefas de segmentação de pólipos em exames de colonoscopia.

Presença de:

- Pólipos pequenos, médios e grandes
- Variações de iluminação
- Alta diversidade anatômica

1.000 imagens

Cada imagem possui máscara anotada manualmente por especialistas

PREPARAÇÃO DO DATASET

- **Pareamento de imagens e máscaras:** pareadas pelo nome do arquivo;
- **Splits:** $train/val/test = 70\% / 15\% / 15\%$;
- **Pré-processamento de imagens:**
 - Conversão BGR→RGB;
 - Padding para obter dimensões fixas (ex.: 1024×1024);
 - Normalização com médias/variâncias.
- **Pré-processamento de máscaras:**
 - Padding igual ao da imagem
- **Extração de anotações:**
 - Centro da máscara (para pontos positivos);
 - Fallback no centro da imagem quando não há lesão;
 - Cálculo da bounding box a partir da máscara segmentada.
- **Data augmentation:** flips, rotações, mudanças de brilho/contraste e Alumentations

ADAPTAÇÃO DO SAM

Congelamento:

- Mask Encoder
- Prompt Encoder

Treinamento:

- Decoder

Batch Size: 4

Num. Epochs: 60

Loss Utilizada:

- Focal Loss + Dice Loss

Otimizador AdamW:

- Learning Rate: 0.00001
- Weight Decay: 0.0001

PROMPTS USADOS

Consideramos:

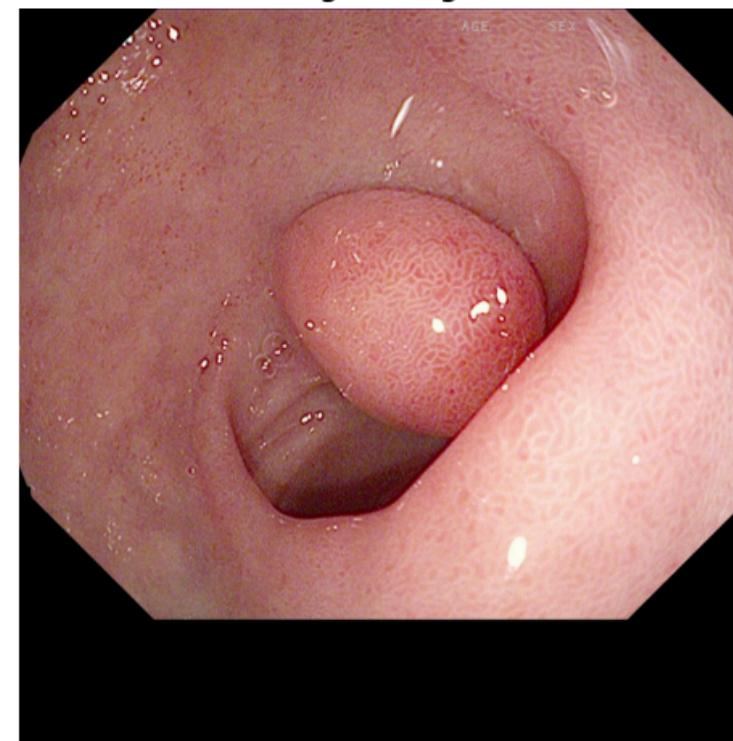
- Usar somente pontos
- Usar somente caixas
- Usar os dois alternativamente

Decisão final:

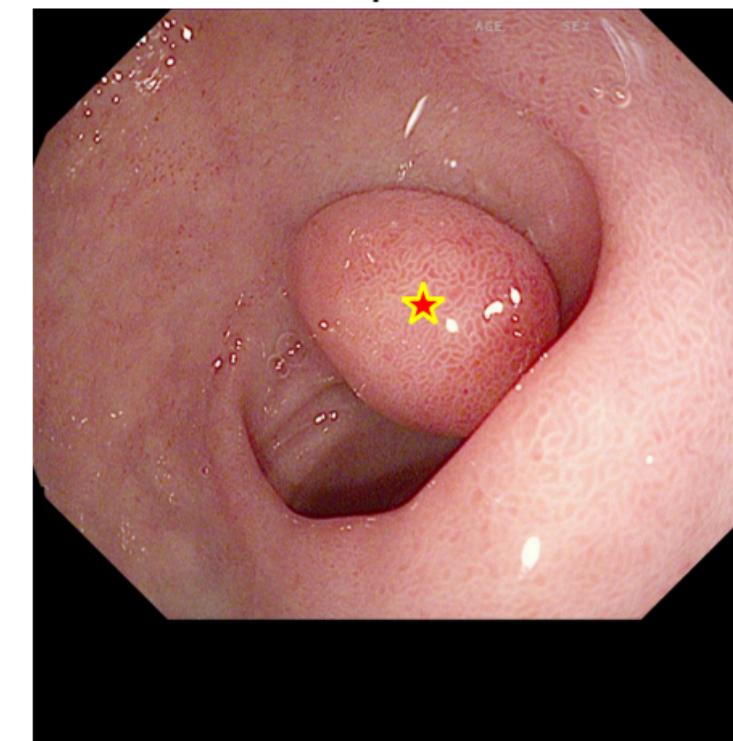
- Usar ambas as formas de prompt simultaneamente, contribuindo com contexto

TREINAMENTO

Imagen Original



Prompt: Ponto



Prompt: Bounding Box

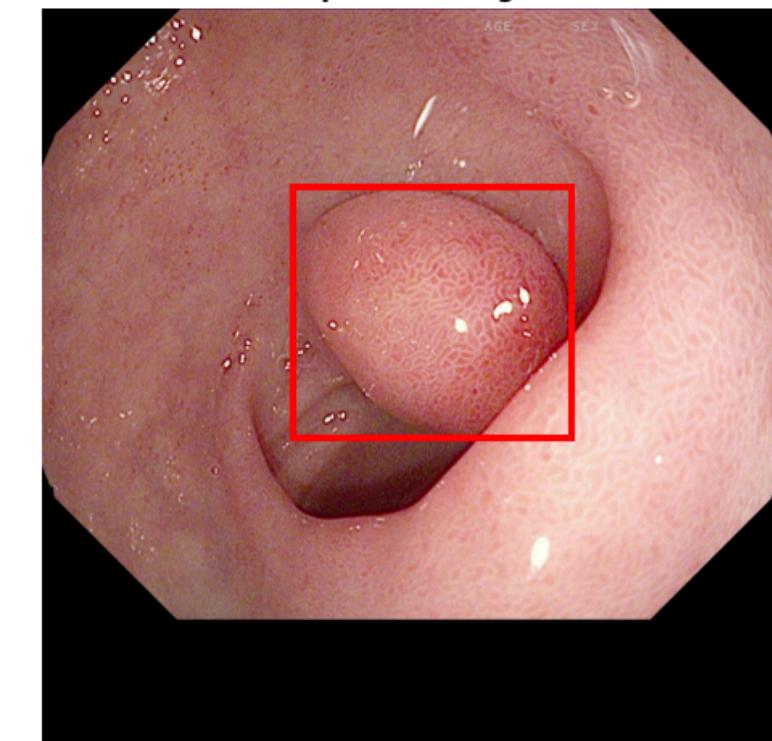


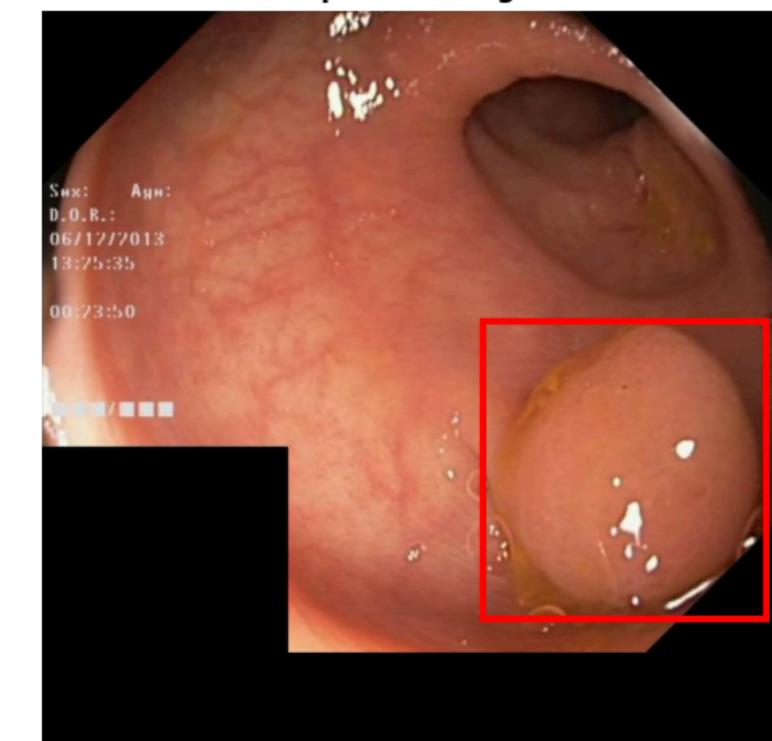
Imagen Original



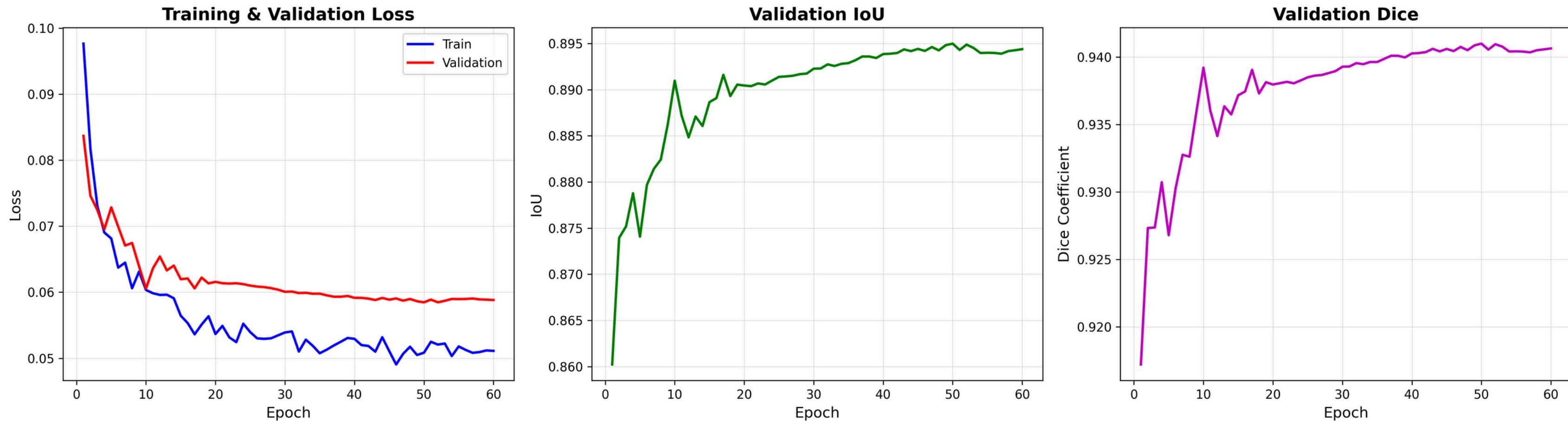
Prompt: Ponto



Prompt: Bounding Box



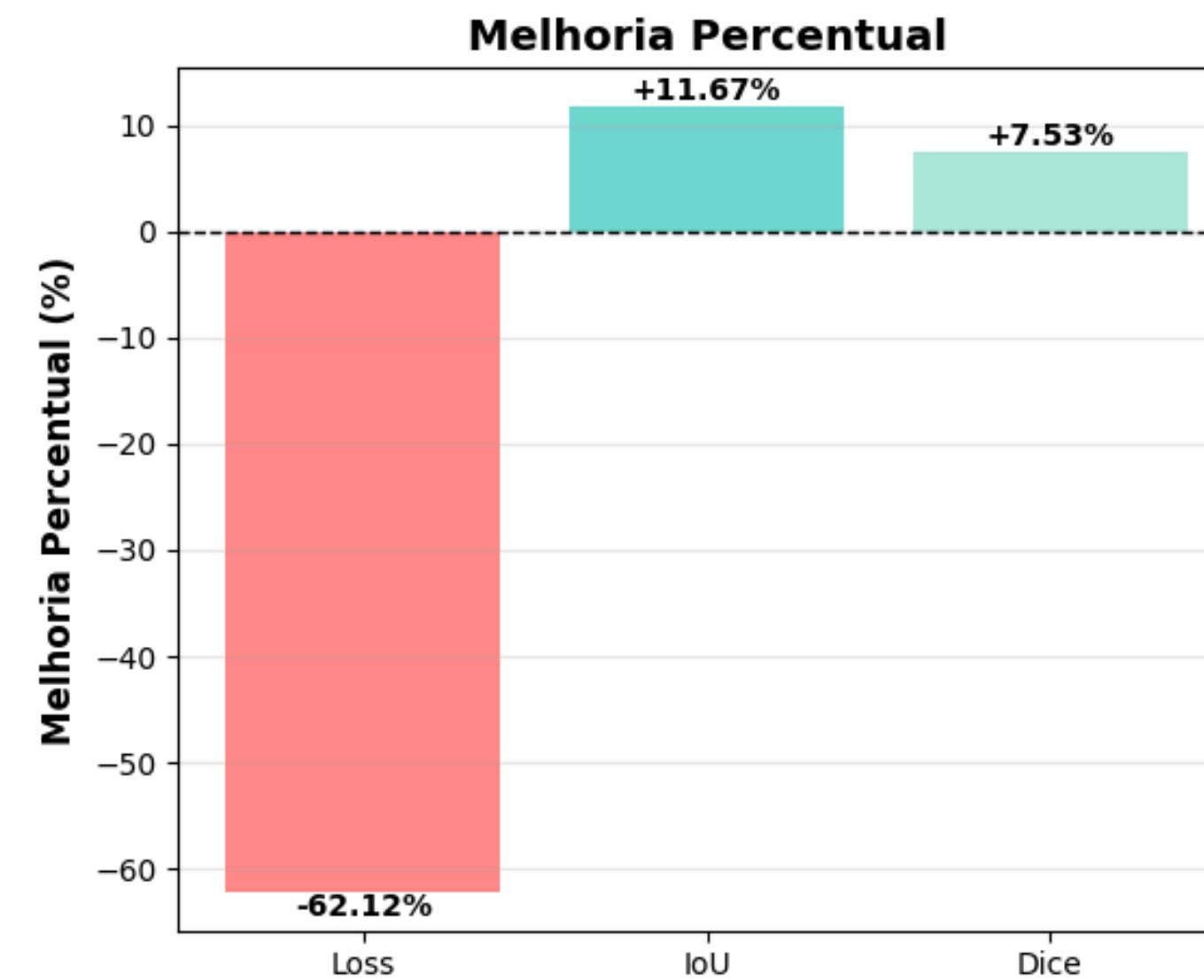
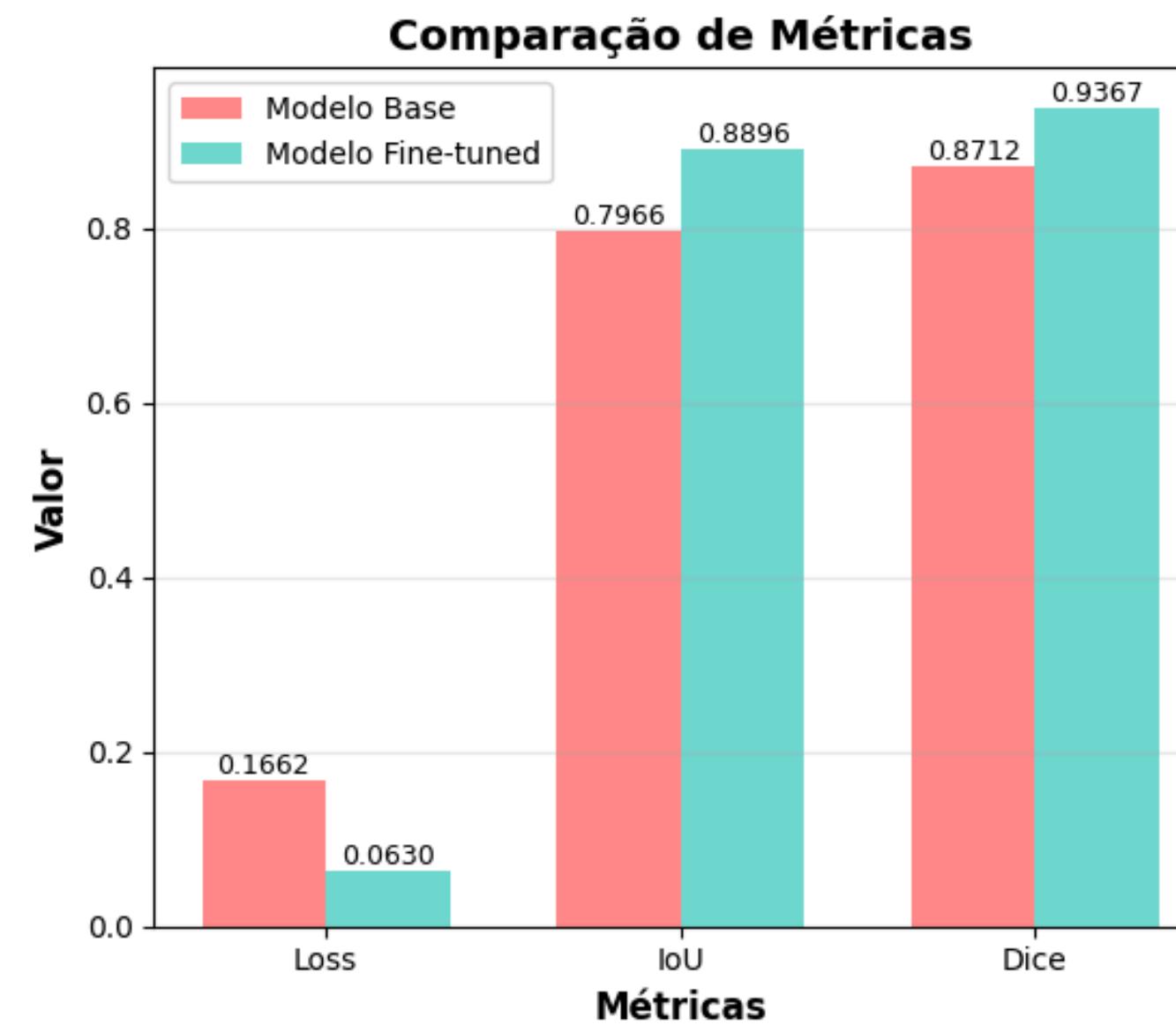
TREINAMENTO



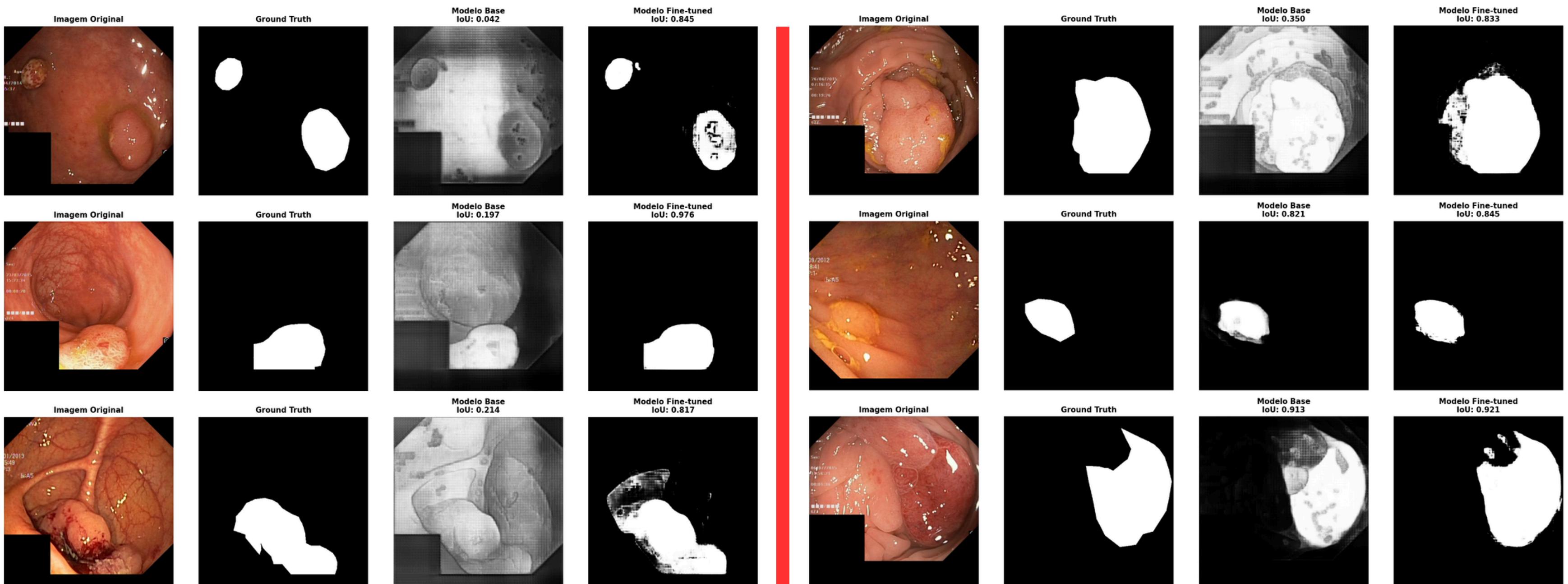
RESULTADOS

Comparação de Resultados			
Métrica	Original	Fine-Tuned	Melhoria
Loss (Dice 1/21 + Focal 20/21)	0.1662	0.0630	-0.1033
IoU	0.7966	0.8896	+0.0930
Dice	0.8712	0.9367	+0.0656

RESULTADOS



RESULTADOS



MELHORIAS FUTURAS

- Adaptação via LoRa (ou equivalentes)
- Treinar a partir de um checkpoint mais robusto
- Testar novas losses
- Data Augmentation mais robusto
- Testar com novas versões do SAM

OBRIGADO!!!