Seamless Object Transfer Between Images

Batuhan Solmaz
Bogazici University

Introduction

Objective: Transferring objects between images using prompt-based input.
Methods: SAM for segmentation extract precise object masks based on user prompts and Deep Image Blending for integration ensuring realistic alignment

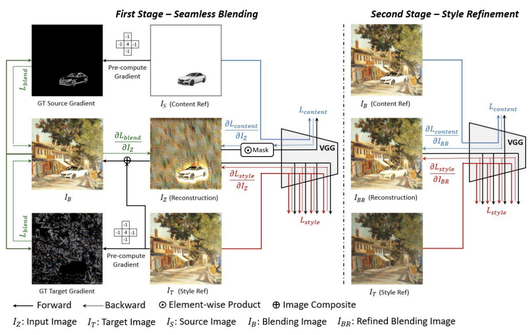
Deep Image Blending

Description: Task is image harmonization. Stage 1: Seamless blending with gradient, content, and style losses.
The blending image is computed as:

$$I_B = I_Z \odot M + I_T \odot (1 - M)$$

Loss Functions:

- **Gradient Loss** (L_{blend}):
  - Ensures the gradients of I_B match the gradients of I_S and I_T in the blending region.
- **Content Loss** ($L_{content}$):
  - Compares deep features of I_Z (from VGG) with I_S to preserve structural information of the source object.
  - Helps ensure that the blended image retains the recognizable shape, layout, and meaningful content of the object.
- **Style Loss** (L_{style}):
  - Matches the style of I_Z with I_T using Gram matrices computed from VGG layers.
- **Histogram Loss** (L_{hist}):
  - Matches the marginal distributions of feature maps in I_B with those in I_T for each filter at each VGG layer. Stabilizes the style transfer (brightness, contrast, and texture intensity).
- **Total Variation Loss** (L_{tv}):
  - Promotes spatial smoothness in I_B by penalizing large intensity differences between neighboring pixels.



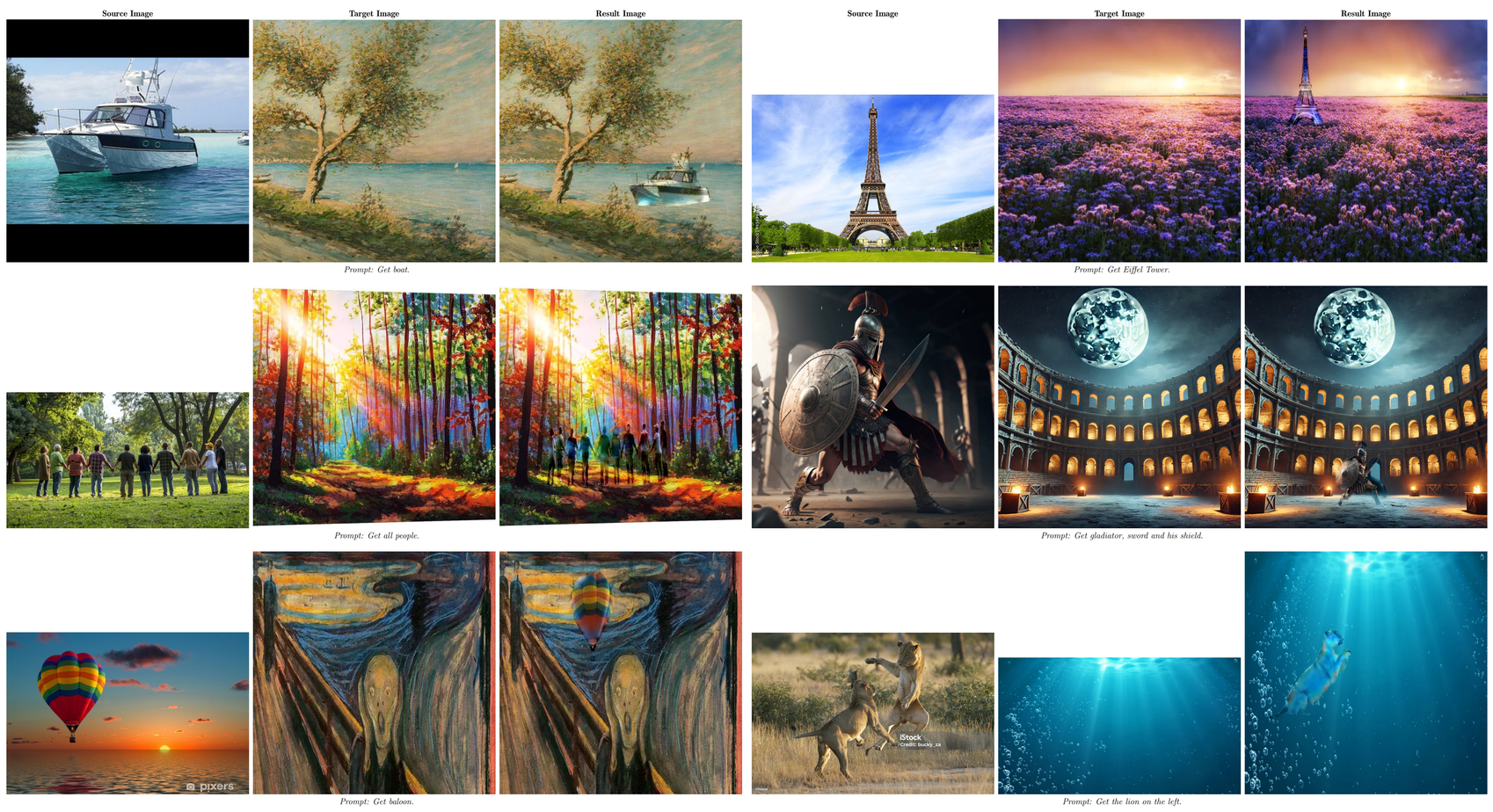
Stage 2: Re-optimizing the blending image (I_B) with a focus on style losses. Same logic with the first stage but with different weights on different loss funcitons
A finer style refinement process is applied.

- **Style Loss** (L_{style}) with a higher weight.

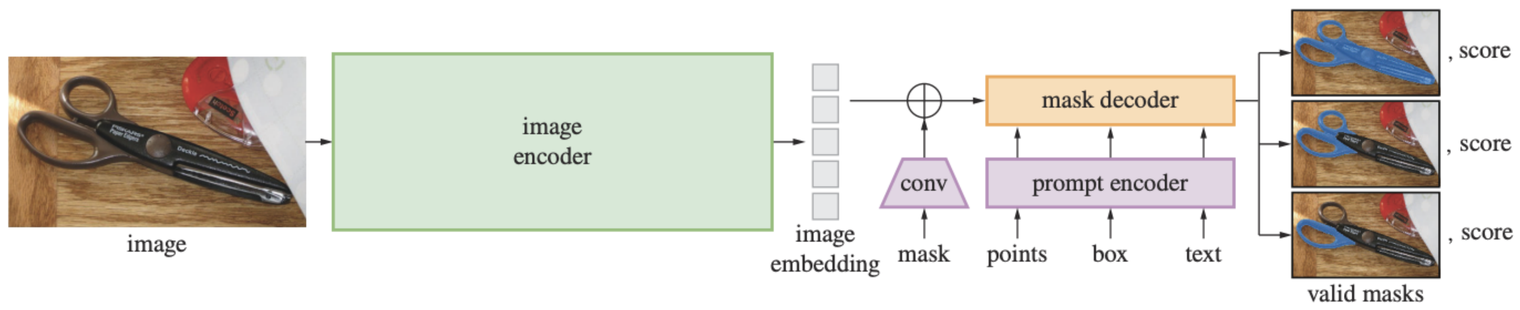
The total loss is computed as:

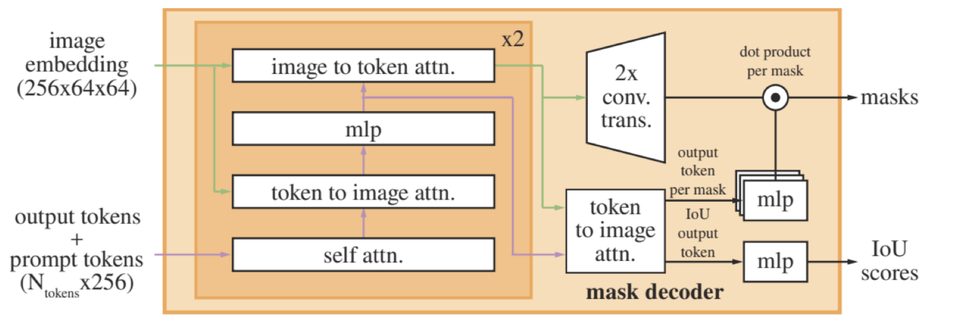
$$L_{total} = \lambda_{cont}L_{cont} + \lambda_{style}L_{style} + \lambda_{hist}L_{hist} + \lambda_{tv}L_{tv}$$

Visual Examples



Segment Anything Model(SAM)





Description: The image encoder is a masked autoencoder Vision Transformer (ViT-H) model, which is a language model pre-trained on a massive dataset. It utilizes self-attention mechanisms and feed-forward networks (FFNs). Images are divided into fixed-size patches, and positional encodings are added to these patches to provide spatial information to the model
Prompt Encoder: This component uses CLIP's encoder for text prompt.
Mask Decoder: The mask decoder generates the actual segmentation mask based on the encoded image and prompt information. It is trained on large datasets where objects have labeled masks. Similar to BERT's `[CLS]` tokens, the output token is provided to self-attention layers. Then, cross-attention is applied, followed by a MLP. Reversed cross-attention is also applied. The transformer produces two outputs: sequences of tokens and embeddings of the image. Each token has its own MLP layer. There are 2 loss functions focal loss and dice loss

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

$$D = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2}$$

Future Work and Conclusion

- Improve blending efficiency for real-time applications can be done to improve. Proposed method successfully integrates SAM and deep image blending to achieve seamless object transfer.

References

- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R. (2023). *Segment Anything.* arXiv. https://arxiv.org/abs/2304.02643.
- Zhang, L., Wen, T., Shi, J. (2019). *Deep Image Blending.* arXiv. https://arxiv.org/abs/1910.11495.