# Automatic trimap generation by a multimodal neural network

Masaki Taniguchi†, Taro Tezuka‡

†University of Tsukuba, Graduate School of Library, Information and Media Studies
‡University of Tsukuba, Faculty of Library, Information and Media Science

Paper #1900

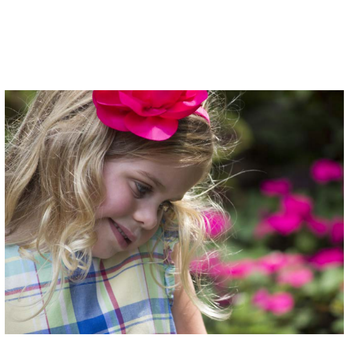
Code: 11.6 – Computational Imaging Methods and Models

# What is trimap?

- Alpha matte: map for composing FG and BG image

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \qquad \alpha_i \in [0, 1]$$

- Task contains 2 stege
  - Roughly main object detection (creating trimap)
  - Fine scale matting (creating alpha matte)

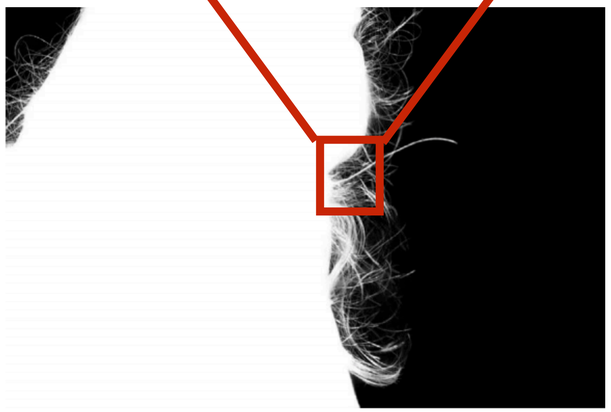


**Photo**

object detection stage

**trimap**

fine matting stage

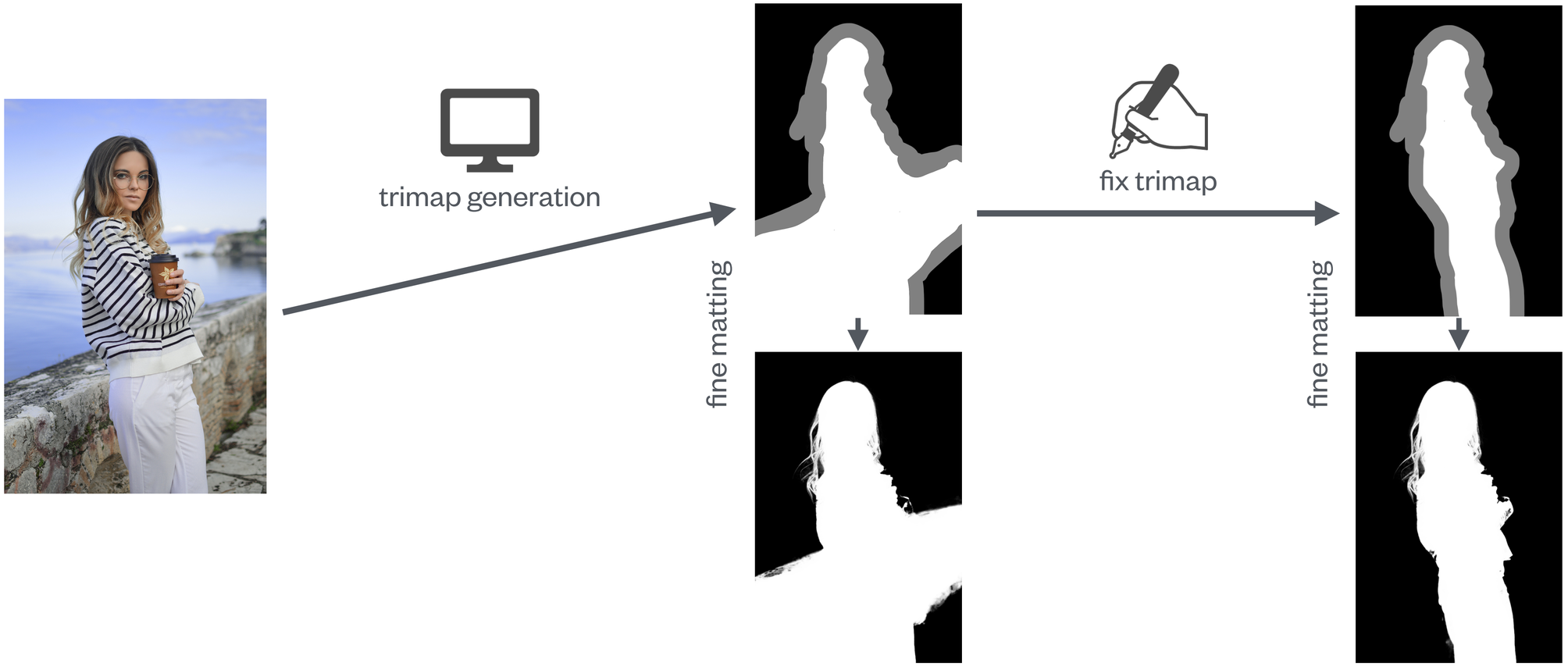


**alpha matte**

Images from: Deep Image Matting [1]

# Why trimap is important for alpha matting

- Separate trimap generation stage and fine matting stage
  → Can fix main object detection result when it missed



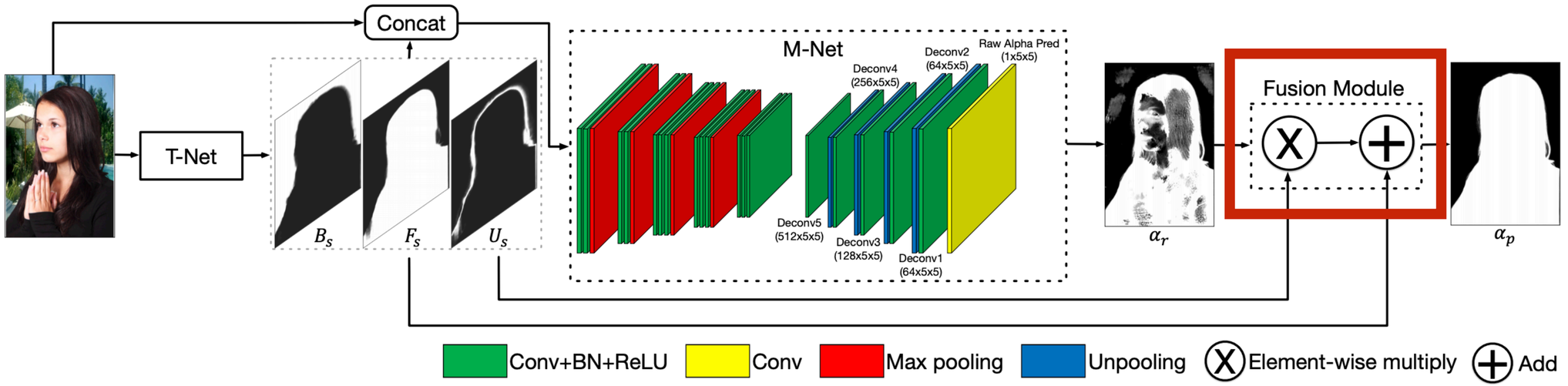
## What is the difficulty of trimap generation?

- No GT trimap data
    - difficult to define GT trimap
- Less public alpha matting dataset
    - Adobe Image Matting[1]: 493 unique training image
    - Distinctions-646[2]: 646 unique training image
    - Semantic Human Matting[3] (not public): 35,513 unique images

Existing trimap generation method

# Semantic Human Matting[3]

- End-to-end alpha matting training
  → output trimap as intermediate representation (Fusion module)
- Using huge dataset
  → 35,513 unique human images (not public)



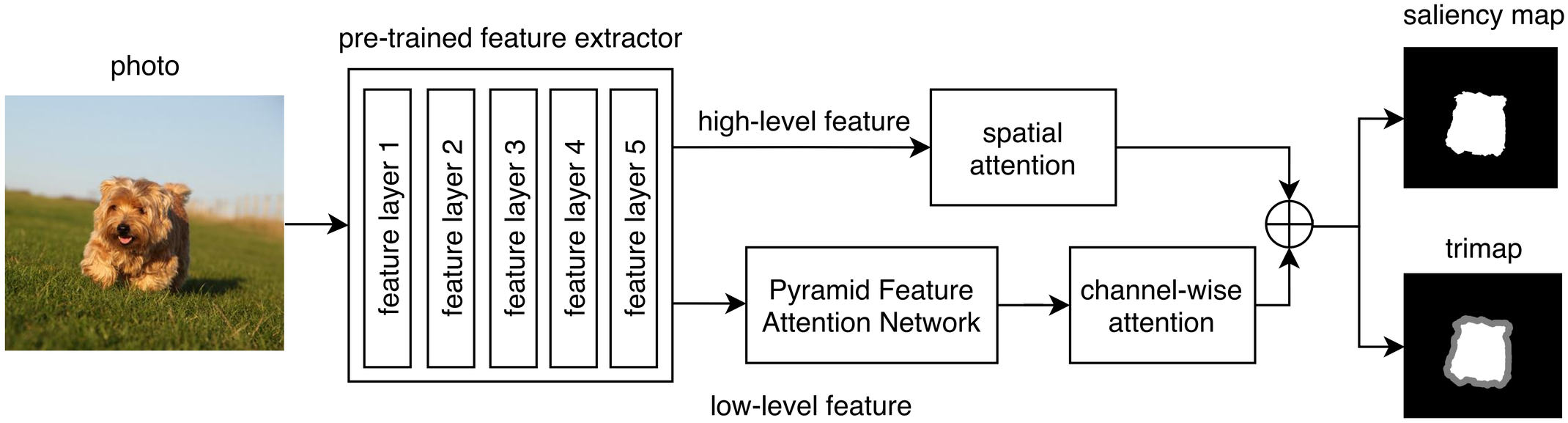

Images from: Semantic Human Matting [3]

# Our method

- Multimodal training by pseudo trimap and saliency map dataset
    - Increase number of GT trimap image
- Additional end-to-end training by GT alpha matte dataset
    - Fine-tune trimaps by alpha matting loss

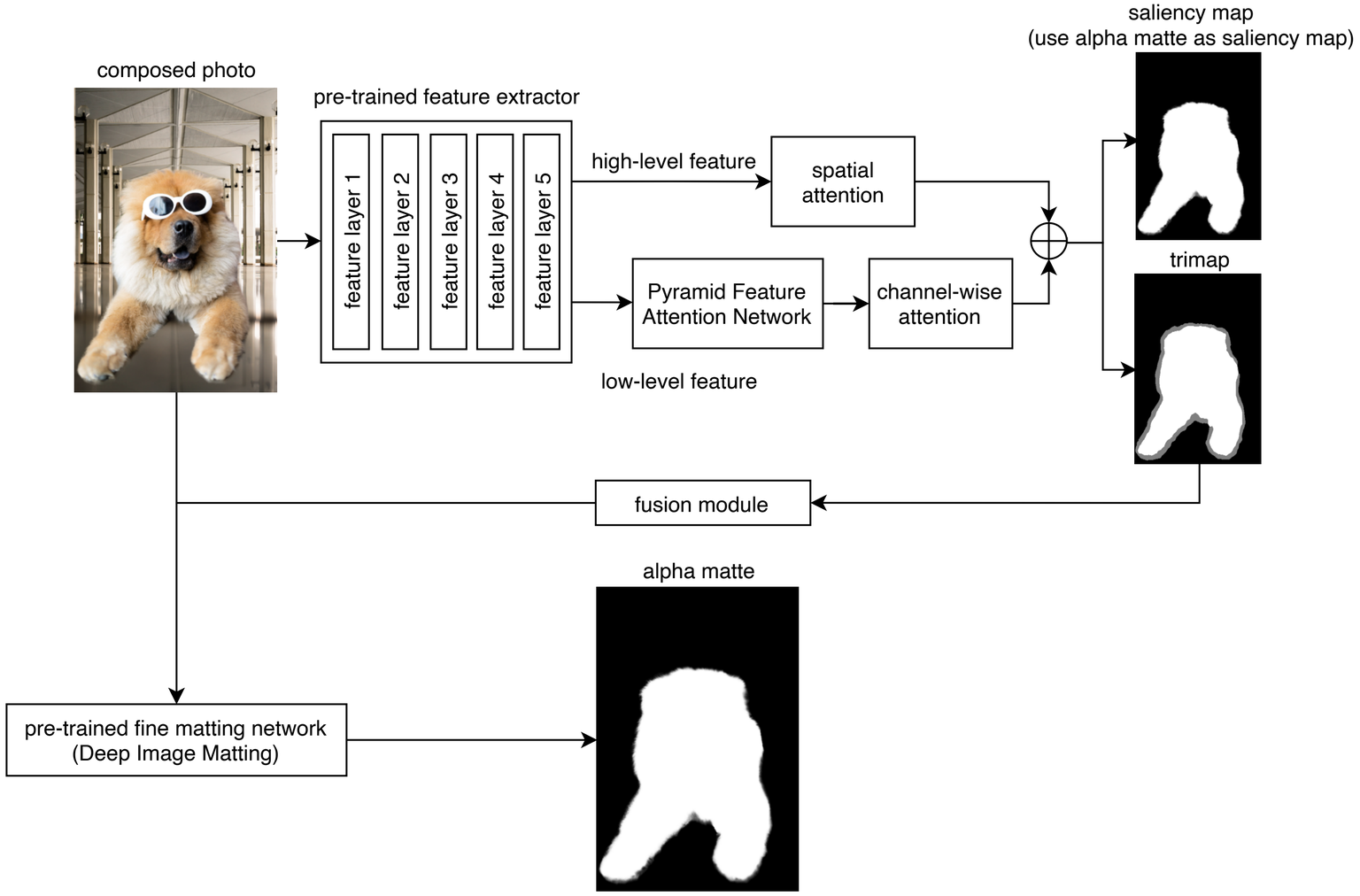
# STEP1 - Train main object detection stage

- Train saliency map and trimap multimodality
- Introduce PFAN[4] network (SoTA saliency map prediction network)
- DUTS image dataset[5]: (Train: 10,553, Test: 1,000)



# STEP2 - Fine-tune with end-to-end training

- Combine pre-trained fine matting stage
- Additional training by Distinctions-646[2]
- Adopt the thickness of unknown part to the object



# Result

- The main object in the image is accurately detected
- Unknown parts become thicker at the complex boundaries

original photos

trimaps

# Quantitative comparison: network structure

| Method | SAD | MSE ($\times 10^{-2}$) | Gradient ($\times 10^{3}$) | Connectivity |
|---|---|---|---|---|
| *ECSSD* | | | | |
| VGG16 + PSPNet | 11.499 | 6.098 | 2.247 | 1.236 |
| VGG16 + proposed | **9.219** | **4.865** | **1.812** | **1.127** |
| Resnet18 + PSPNet | 10.814 | 5.730 | 2.131 | 1.231 |
| Resnet18 + proposed | **8.069** | **4.075** | **1.478** | **1.161** |
| Densenet + PSPNet | 8.933 | 4.706 | 1.721 | **1.021** |
| Densenet + proposed | **7.265** | **3.596** | **1.287** | **1.021** |
| *Distinction-646* | | | | |
| VGG16 + PSPNet | 14.552 | 6.550 | 2.304 | 1.838 |
| VGG16 + proposed | **13.228** | **6.038** | **2.103** | **1.716** |
| Resnet18 + PSPNet | 14.918 | 6.875 | **2.419** | 1.815 |
| Resnet18 + proposed | **14.215** | **6.776** | 2.439 | **1.480** |
| Densenet + PSPNet | 21.106 | 10.789 | 4.013 | 1.663 |
| Densenet + proposed | **13.619** | **6.421** | **2.267** | **1.509** |

**Table 1**. Performance comparison.

# Quantitative comparison: with- and without-each components

| Method | SAD | MSE ($\times 10^{-2}$) | Gradient ($\times 10^{3}$) | Connectivity |
|---|---|---|---|---|
| no SA | 16.618 | 8.062 | 2.903 | 1.726 |
| no CA | 15.904 | 7.587 | 2.707 | 1.877 |
| no $L_S$ | 14.837 | 7.037 | 2.483 | 1.692 |
| no $L_\alpha$ | 16.332 | 8.154 | 3.024 | **1.364** |
| proposed | **14.215** | **6.776** | **2.439** | 1.480 |

**Table 2**. Performance with and without each component. $L_S$ and $L_\alpha$ are saliency map loss and alpha matte loss.

# Qualitative comparision: with- and without-step2 training

photo

trimap

alpha matte

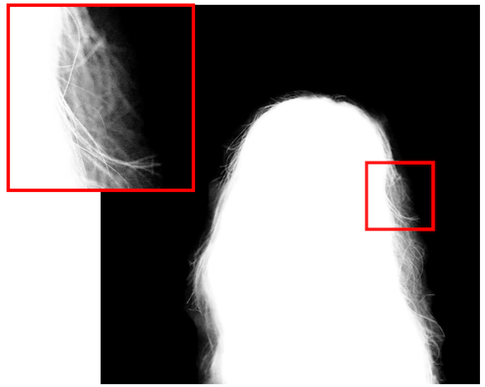
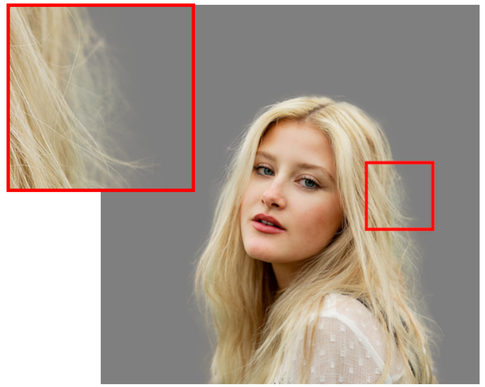
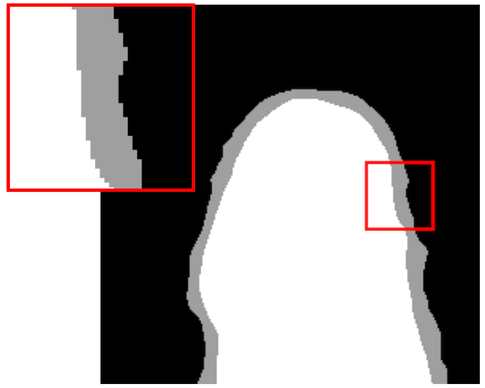
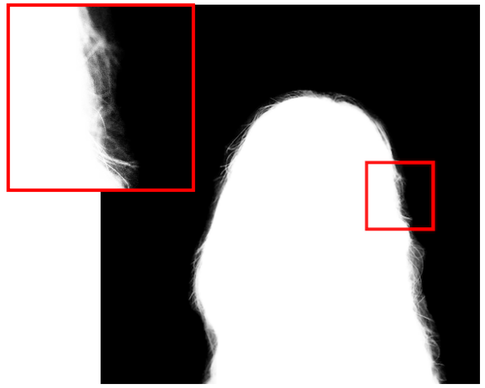


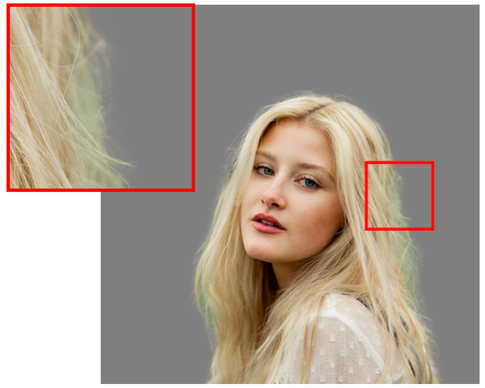
photo + alpha ch



with step2 training







Without-step2 training

# References


1. Xu, Ning, et al. "Deep image matting." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
2. Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei, “Attentionguided hierarchical structure aggregation for image matting,” in CVPR, 2020, pp. 13676–13685.
3. Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai, “Semantic human matting,” in ACM Multimedia, 2018, pp. 618–626.
4. Zhao, Ting, and Xiangqian Wu. "Pyramid feature attention network for saliency detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
5. Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan, “Learning to detect salient objects with image-level supervision,” in CVPR, 2017.

# Thank you for listening