MatteFormer: Transformer-Based Image Matting via Prior-Tokens Supplementary Material

GyuTae Park[1,2], SungJoon Son[1,2], JaeYoung Yoo[2], SeHo Kim[2], Nojun Kwak[1]
[1] Seoul National University, South Korea
[2] NAVER WEBTOON AI, South Korea
{gyutae.park, sjson718, yoojy31, seho.kim}@webtoonscorp.com, nojunk@snu.ac.kr

1. Experiments

1.1. Ablation Study

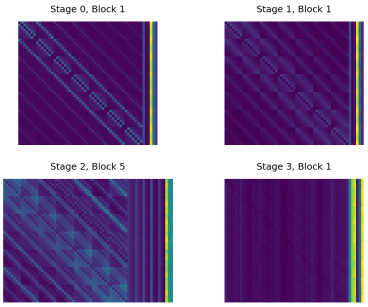
Visualization of attention maps in PA-WSA layers with prior-memory

In this subsection, we conduct an additional ablation study on the full model of the MatteFormer and show that not only prior-tokens of the current block but also prior-tokens of the previous blocks conveyed by prior-memory participate in the self-attention mechanism.

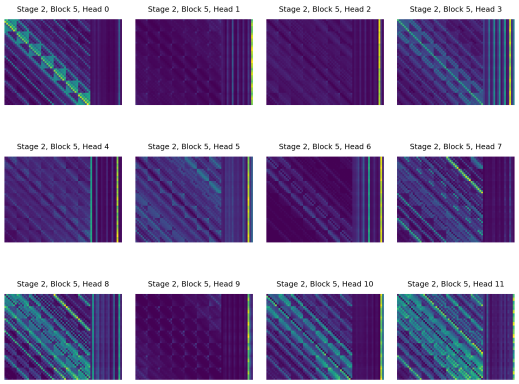
Note that the b-th block in each of the four encoder stages has $3 \times b$ prior-tokens in the corresponding prior-memory. Thus, in Fig. 1a, we visualize mean attention maps on the last PAST (Prior-Attentive Swin Transformer) block of each encoder stage to effectively show the aspect of attention on prior-tokens from the previous blocks. As window size is 7, the y-axis represents 49 query spatial-tokens in a local window and the x-axis denotes 49 spatial-tokens with appended prior-tokens. The last three columns are prior-tokens (unknown, foreground and background token in order) of the current block. The figure shows that prior-tokens stored in prior-memory also participate in the current self-attention layer as global priors.

Fig. 1b shows each multi-head attention map on the PAST block of stage index 2 and block index 5. Each head has a different property of attention in terms of prior-tokens usage. For example, head 3 attends on all unknown prior-tokens in the prior-memory and head 8 attends on all foreground prior-tokens in the prior-memory. Interestingly, in head 3, a token tends to attend more on the unknown prior-token of the current block than others of the previous blocks. Head 2 and 9 mainly focus on the unknown prior-token of the current block. On the contrast, head 1 focuses on the background prior-token of the current block. Meanwhile, head 0 and 7 do not much attend to the prior-tokens, rather to the spatial-tokens.

With the ablation study, we found that MatteFormer not only makes use of 3 prior-tokens (of foreground, background and unknown) from the current block but also uses prior-tokens generated from previous blocks through prior-memory. It suggests that introducing prior-tokens and prior-memory contributes to performance improvement.



(a) Mean attention maps are averaged on multi-heads. We plot the attention maps on the last block of each encoder stage.



(b) Multi-head attention maps in the PA-WSA. We plot only the last block (block index 5) of stage index 2 to show examples in a simple. The self-attention layer has 12 multi-heads.

1.2. Results on Image Matting Datasets

More qualitative results on Composition-1k and Distinctions-646 are shown in Fig. 2 and Fig. 3, respectively.

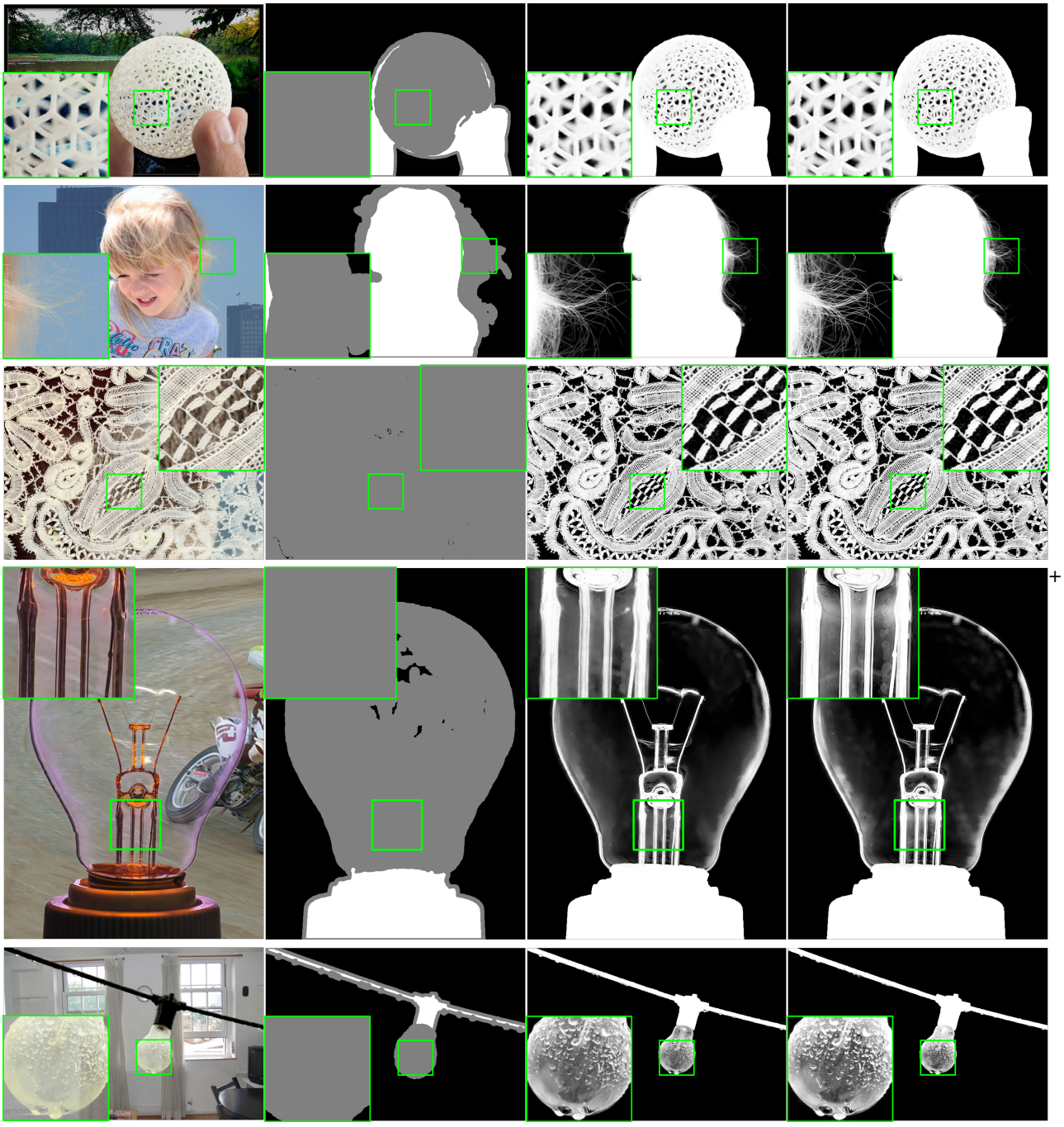


Figure 2. The qualitative results on Composition-1k. Best viewed by zooming in.

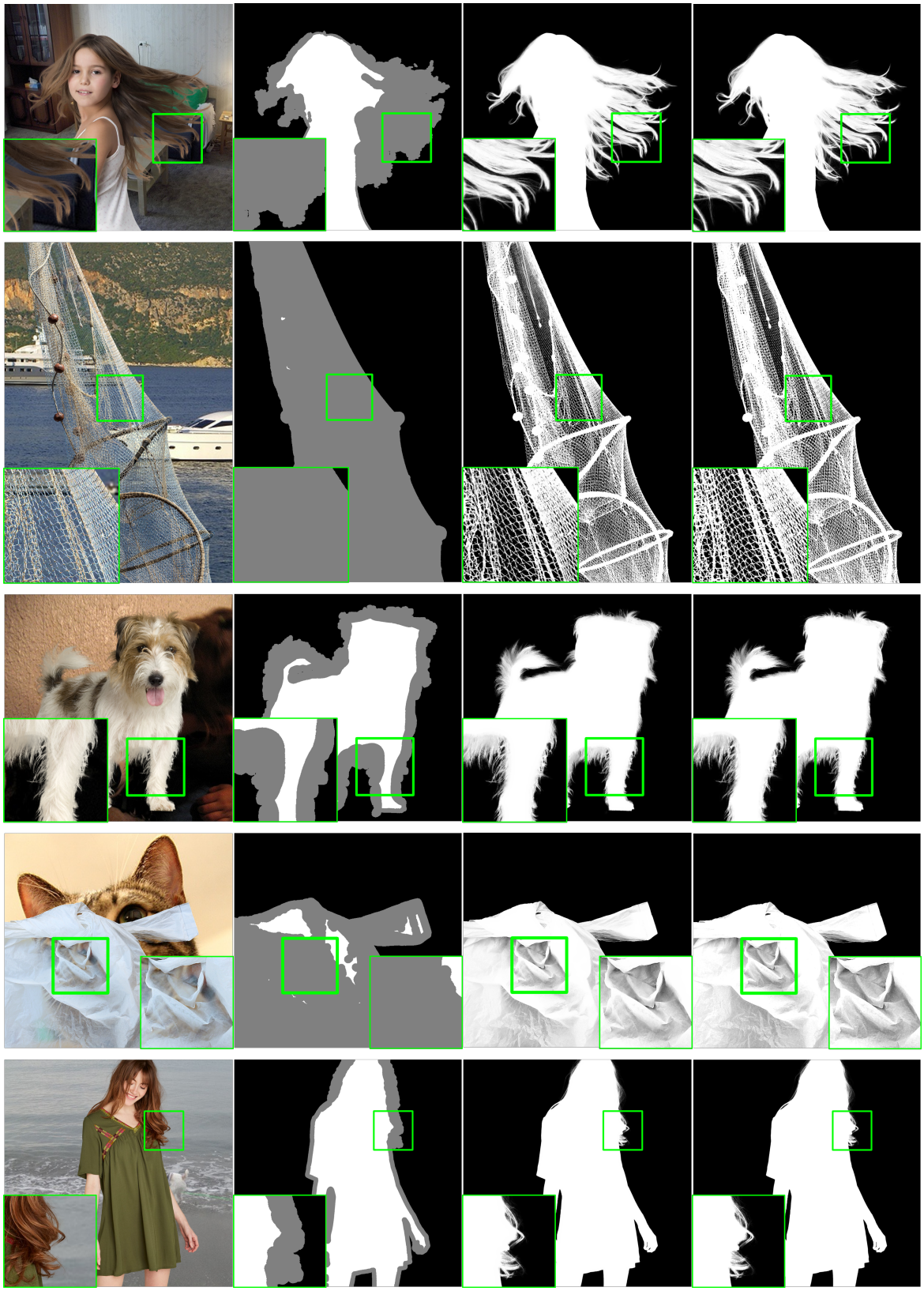


Figure 3. The qualitative results on Distinctions-646. Best viewed by zooming in.