Supplementary Materials for Highly Efficient Natural Image Matting

Yijie Zhong
dun.haski@gmail.com
Bo Li[1]
libraboli@tencent.com
Lv Tang
luckybird1994@gmail.com
Hao Tang[2]
hao.tang@vision.ee.ethz.ch
Shouhong Ding[1]
ericshding@tencent.com

[1] Youtu Lab,
Tencent,
Shanghai, China

[2] Computer Vision Lab,
ETH Zurich

1 Introduction

This supplemental material contains 6 parts:

- Section 2 gives more details about our proposed lightweight backbone.
- Section 3 gives more analysis of the proposed ENA.

We hope this supplemental material can help you get a better understanding of our work.

2 More Details of Our Network

2.1 Details of Network Architecture

When OCBolck has only one scale input, OCBlock will be built like Fig.1(b) to obtain multi-scale output. When OCBlock has only one scale output, it will be built like Fig. 1(c) to fuse the multi-scale input. The general OCBlock is shown in Fig. 1(a).

Specifically, in stage1 of the proposed network, we use another OCBlock like Fig.1(b) as En1-0 to decompose the input image into features at two resolutions (256 and 512). So the dimension of the input of our backbone is $[256, 512]$. The dimension output of each level is $[256, 512], [256], [128], [64]$ respectively. We only use the 512 features from En1-1 as described in main text. Thus, the first OCBlock in En1-3 and En1-4 uses Fig. 1(b) structure. The last OCBlock in En1-2, En1-3 and En1-4 uses Fig. 1(c) structure. MRN uses a similar structure.

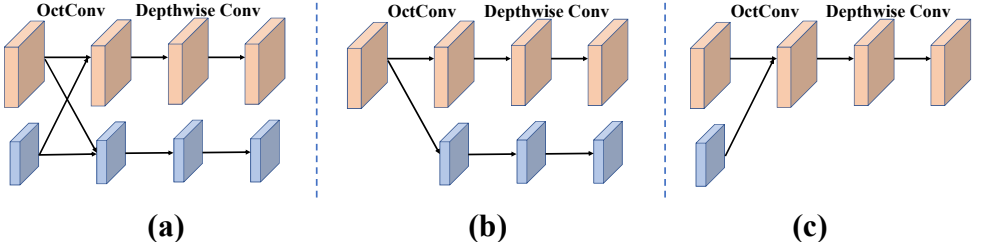

Figure 1: Different types of OCBlock.

3 More analysis of ENA

3.1 The Efficiency of ENA

Given an input feature map of size $H \times W \times C$, we analyze the computation cost of both the common non-local attention mechanism in the whole image and our proposed ENA.

For common non-local attention mechanism, three 1×1 convolution layers acting on Q, K, V to change them from $\mathbb{R}^{C \times H \times W}$ to $\mathbb{R}^{\frac{C}{2} \times H \times W}$ and one 1×1 convolution layers acting on the output feature to change it from $\mathbb{R}^{\frac{C}{2} \times H \times W}$ to $\mathbb{R}^{C \times H \times W}$. The complexity of this part is $\mathcal{O}(2C^2HW)$. And, the complexity of multiplying between features is $\mathcal{O}(\frac{3}{2}C(HW)^2)$. So, the complexity of self-attention mechanism is

$$C_1 = \mathcal{O}(2C^2HW + \frac{3}{2}C(HW)^2). \tag{1}$$

Considering our approach, we divide the height and width dimension to $\sqrt{k}$ groups in calculating long-range relations, and $\frac{H}{\sqrt{k}}$ and $\frac{W}{\sqrt{k}}$ groups in calculating short-range relations. So the complexity of our approach is

$$C_2 = \mathcal{O}(4C^2HW + \frac{3}{2}C(HW)^2(\frac{1}{k} + \frac{k}{HW})). \tag{2}$$

Since the channel number C is a smaller value in our approach, it is much smaller than HW. The increase in the first term is negligible compared to the decrease in the second term. The complexity of our approach can be minimized to

$$min(C_2) = \mathcal{O}(4C^2HW + 3C(HW)^{\frac{3}{2}}). \tag{3}$$

when $k = \sqrt{HW}$. Although $\sqrt{k}$ will take a small value in the implementation (e.g. 4) and will not be able to minimize C_2, our approach is efficient enough compared to original self-attention.

For the numerical complexity, we consider the feature dimensions of 80 and $\sqrt{k} = 4$. When the input size is 64×64, the computation cost of the self-attention mechanism is 2.06 GFLOPs and ours ENA is 0.11 GFLOPs. When the input size is 128×128, it becomes 32.42 GFLOPs and 0.45 GFLOPs. So, our proposed efficient non-local attention module not only guarantees a small amount of computation cost when the input size is small but also guarantees a slow growth when the input size increases.

3.2 The location choice of ENA in our network

CFM wants to avoid the gradual dilution of the high-level semantic information during decoding and propagate it to the low-level. Multiple CFMs will lead to duplication of operations and produce redundant computations, so we only build CFM between En1-2 corresponding to the last decoder and En1-4. For the ENA module, our experiments show that putting it at En/De2-2 does not improve performance while increasing complexity.