LIGHTWEIGHT PORTRAIT SEGMENTATION VIA EDGE-OPTIMIZED ATTENTION

Xinyue Zhang [1] Guodong Wang* [1] Lijuan Yang [2] Chenglizhao Chen* [1]

[1]College of Computer Science and Technology, Qingdao University, Qingdao, 266071 P.R. China

[2]Hisense Visual Technology Co., Ltd, Qingdao, 266555 P.R. China

Overview

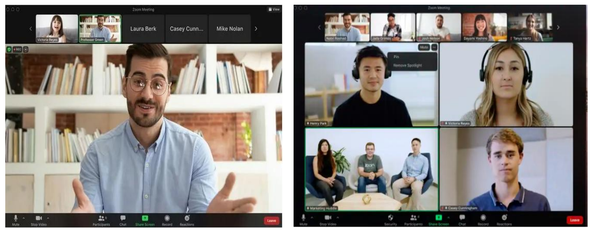
★We built a **lightweight** architecture with 0.06G FLOPs and 0.02M parameters. Our network achieves an FPS of 39.02 on CPU, which is more than three times faster than other networks.

★The **edge-optimized attention mechanism (EOAM)** is designed to collect specific edge areas for the bottom features and the high-level features in the process of feature fusion.

Motivation

With the outbreak of COVID-19 around the world, the frequency of video conferencing at home is increasing. Therefore, a segmentation architecture that can quickly carry out close-range portrait segmentation has become a current need. However, the current portrait segmentation architectures cannot meet the requirements of lightweight and edge-friendly.

Application scenario: The background of a user can be changed in a video conference to protect user privacy.



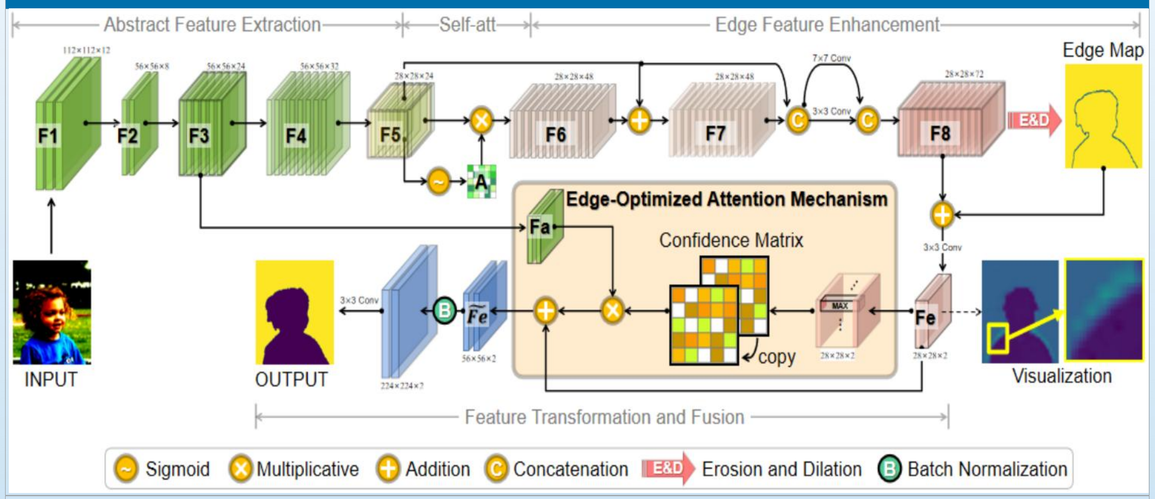
Key Idea: Lightweight architecture design and an edge-optimized attention mechanism.

Authors' Institution

QINGDAO UNIVERSITY

Hisense

Algorithm Framework



Edge-optimized Attention Mechanism

Edge-optimized attention mechanism plays an important role in the process of feature fusion. It needs input from two parts. One part is from the low-level feature Fa in abstract feature extraction, and the other part is the feature obtained by the Batchnormlization calculation of Fe . The probability value of incomplete storage to each pixel region is obtained by a calculation similar to the following equation and stored in the confidence matrix($\widehat{CM}$). "Incomplete Storage" means that these areas need to be further supplemented with relevant information.

$$\widehat{CM} = 1 - MaxChannel(\frac{e^{g_i}}{\sum_k^n e^{g_k}}),$$

where i represents a category in k (portrait, edge area, or background), and g i represents the value of this feature region. And n is the total number of categories. MaxChannel means calculating each channel's maximum value based on the channel hierarchy. The final feature output $\widehat{F_e}$ from this attention mechanism follows the following calculation. Calculating the above equation will give the incomplete region a larger value reserve in $\widehat{CM}$. Then, with the help of these reserve values in $\widehat{CM}$, the following equation is implemented to strengthen the attention and supplement the characteristics of the "Incomplete Storage".

$$\widehat{F_e} = F_{bcbr}((F_a \otimes \widehat{CM} \oplus F_e), 2),$$

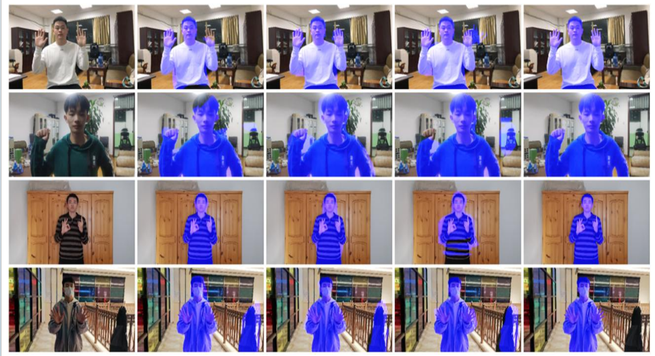
Finally, Batchnormalization, upsampling, and3×3convolutionareperformed on $\widehat{F_e}$ in turn to obtain the preliminary mask. After preliminary conversion into masks, we will optimize them again and save them as the final results.

Input | Original Result | Region to be Filled | Final Mask

In the optimization process, the regions in masks are divided into several sub-connected regions according to the outermost contour, and color filling is carried out in the largest sub-connected region. This is because the image transmitted into the network may suffer from the phenomenon that objects block the portrait area. Therefore, this color filling redefined the image area that is not divided in the portrait due to object occlusion and other reasons in the prediction process as part of the portrait mask.

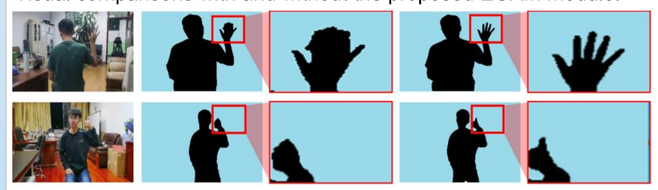
Visual Comparisons

Visual comparisons for the segmentation results.



Input | SiNet | CoNet | MoNet | Ours

Visual comparisons with and without the proposed EOAM module.



Input | Without EOAM | With EOAM

Comparisons Based on Different Measures

Computation cost comparisons. Notice that the FPS of our method can outperform recent architectures by more than three times, and the mIOU value can reach new SOTA on the EG1800 set.

A: Enet B: BiseNet C: ESPNet D: PoNet E: DSNet F: DSNet+ G: ENetV2+ H: ENetV2++ I: CoNet J: MoNet K: SiNet L: SiNet+

Methods	A	B	C	D	E	F	G	H	I	J	K	L	Ours
FPS	8.060	2.990	6.990	3.300	7.750	9.260	4.950	3.650	1.550	10.87	12.35	12.35	**39.02**
Para (M)	0.355	0.124	0.345	2.08	0.143	0.064	0.458	0.778	0.838	0.458	0.087	0.087	**0.021**
FLOPs(G)	0.346	2.310	0.328	0.325	0.199	0.139	0.137	0.231	1.870	0.066	0.064	0.064	**0.061**
mIOU	95.16	94.91	94.65	95.99	94.10	93.58	94.00	94.71	95.71	94.19	94.81	95.29	**96.01**

Effectiveness of the proposed EOAM.

A: Enet B: BiseNet C: ESPNet D: DSNet E: DSNet+ F: CoNet G: PoNet H: ENetV2+ I: ENetV2++ J: SiNet K: SiNet+ L: MoNet

mIOU	A 2016	B 2018	C 2018	D 2018	E 2018	F 2018	G 2019	H 2019	I 2019	J 2020	K 2020	L 2022
Baseline	95.16	94.91	94.65	94.10	93.58	95.71	95.99	94.00	94.71	94.81	95.29	94.19
+EOAM	95.72	95.49	95.27	94.61	93.93	95.94	96.27	94.68	95.06	95.49	95.98	94.75
Gain	**+0.56**	**+0.58**	**+0.62**	**+0.51**	**+0.35**	**+0.23**	**+0.28**	**+0.68**	**+0.35**	**+0.68**	**+0.69**	**+0.56**

Email

Guodong Wang: doctorwgd@gmail.com

Xinyue Zhang: 2020025793@qdu.edu.cn