# Camouflaged Object Detection with Feature Decomposition and Edge Reconstruction

Chunming He[1] , Kai Li[2*], Yachao Zhang[1] , Longxiang Tang[1] , Yulun Zhang[3] , Zhenhua Guo[4] , and Xiu Li[1*]
[1]Shenzhen International Graduate School, Tsinghua University,
[2]NEC Laboratories America, [3]ETH Zürich, [4]Tianyi Traffic Technology

## Contents



## A. Supplementary Instructions for GFA

### A.1. Derivation of the Linear Coefficients

In the high frequency (HF) bands, the linear coefficients $\{\sigma_w, \mu_w\}$ can be acquired by minimizing the following objective function $L_h$:

$$L_h = \sum_{i \in s_w} \left[ \left(p_k^h\right)_i^2 \left(\left(f_{k-1}^{dh}\right)_i - \left(\left(f_k^r\right)_{HF}\right)_i\right)^2 + \epsilon\sigma_w^2 \right], \quad (1)$$

where

$$\left(f_{k-1}^{dh}\right)_i = \sigma_w down\left(\left(f_{k-1}^r\right)_{HF}\right)_i + \mu_w, \forall i \in s_w. \quad (2)$$

We define $A_H = \left(p_k^h\right)_i (down((f_{k-1}^r)_{HF}))_i$, $A_L = \left(p_k^h\right)_i \left((f_k^r)_{HF}\right)_i$, and $p_n$ as the number of the pixels in $s_w$. By assigning the partial derivatives of the optimization function $L$ to $\sigma_w$ and $\mu_w$ and locating the zero points, we can calculate the optimized results of $\{\sigma_w, \mu_w\}$:

$$\begin{aligned}
&\frac{\partial L_h}{\partial \sigma_w} = \sum_{i \in s_w} [2\epsilon\sigma_w + 2\ down\left(\left(f_{k-1}^r\right)_{HF}\right)_i \left(p_k^h\right)_i^2 \\
&\left(\sigma_w down\left(\left(f_{k-1}^r\right)_{HF}\right)_i + \mu_w - \left(\left(f_k^r\right)_{HF}\right)_i\right)] = 0, \\
&\Rightarrow \left(\overline{A_H^2} - p_n \times \overline{\overline{\left(p_k^h\right)_i} A_H} \times \overline{A_H} + \epsilon\right) \sigma_w = \\
&\left(\overline{A_H A_L} - p_n \times \overline{\overline{\left(p_k^h\right)_i} A_H} \times \overline{A_L}\right), \\
&\Rightarrow \sigma_w = \frac{\overline{A_H A_L} - p_n \times \overline{\overline{\left(p_k^h\right)_i} A_H} \times \overline{A_L}}{\overline{A_H^2} - p_n \times \overline{\overline{\left(p_k^h\right)_i} A_H} \times \overline{A_H} + \epsilon},
\end{aligned} \quad (3)$$

$$\begin{aligned}
&\frac{\partial L_h}{\partial \mu_w} = \sum_{i \in s_w} [\left(p_k^h\right)_i^2 \left(\sigma_w down\left(\left(f_{k-1}^r\right)_{HF}\right)_i\right. \\
&\quad \left. + \mu_w - ((f_k^r)_{HF})_i\right)] = 0, \\
&\Rightarrow \overline{\left(p_k^h\right)_i} \mu_w = \overline{A_L} - \sigma_w \times \overline{A_H}, \\
&\Rightarrow \mu_w = \left(\overline{A_L} - \sigma_w \times \overline{A_H}\right) / \left(\overline{\left(p_k^h\right)_i}\right),
\end{aligned} \quad (4)$$

where $\overline{(\cdot)}$ denote the average operation.

### A.2. Low Frequency Feature Aggregation with GFA

In low frequency (LF) components, given $(f_{k-1}^r)_{LF}$, $(f_k^r)_{LF}$, $k \in \{2, 3, 4\}$, the aggregated feature map $f_{k-1}^l$ can be acquired by minimizing the following optimization function with the assistance of the attention map $p_k^l$:

$$L_l = \sum_{i \in s_w} \left[ \left(p_k^l\right)_i^2 \left(\left(f_{k-1}^{dl}\right)_i - \left(\left(f_k^r\right)_{LF}\right)_i\right)^2 + \epsilon\sigma_w^2 \right], \quad (5)$$

where

$$\left(f_{k-1}^{dl}\right)_i = \sigma_w down\left(\left(f_{k-1}^r\right)_{LF}\right)_i + \mu_w, \forall i \in s_w, \quad (6)$$

where $down$ denotes the down-sampling operator. $s_w$ is a squared window centered by pixel $w$. $i$ represents pixel $i$ in $s_w$. For uniformity, we still conduct the linear transformation with the coefficients $\{\sigma_w, \mu_w\}$, which can be calculated following Appendix A.1:

$$\begin{aligned}
\sigma_w &= \frac{\overline{A_H^l A_L^l} - p_n \times \overline{\overline{\left(p_k^l\right)_i} A_H^l} \times \overline{A_L^l}}{\overline{\left(A_H^l\right)^2} - p_n \times \overline{\overline{\left(p_k^l\right)_i} A_H^l} \times \overline{A_H^l} + \epsilon}, \\
\mu_w &= \left(\overline{A_L^l} - \sigma_w \times \overline{A_H^l}\right) / \left(\overline{\left(p_k^l\right)_i}\right),
\end{aligned} \quad (7)$$

where $A_H^l = \left(p_k^l\right)_i (down((f_{k-1}^r)_{LF}))_i$, $A_L^l = \left(p_k^l\right)_i \left((f_k^r)_{LF}\right)_i$. By averaging, matrixing, and upsampling the window-based coefficients, we get the final linear coefficients $\left(\boldsymbol{\sigma}_h^l, \boldsymbol{\mu}_h^l\right)$. Therefore, the aggregated feature map $f_{k-1}^l$ can be acquired as follows:

$$\begin{aligned}
f_{k-1}^l &= GFA\left(\left(f_k^r\right)_{LF}, \left(f_{k-1}^r\right)_{LF}, p_k^l\right), \\
&= \boldsymbol{\sigma}_h^l \odot \left(f_{k-1}^r\right)_{LF} + \boldsymbol{\mu}_h^l,
\end{aligned} \quad (8)$$

*Corresponding author.

Figure 1. Failure cases.

where $\odot$ is the Hadamard product. In this case, $f_3^l$ can be calculated as follows:

$$\begin{aligned} f_3^l &= GFA\left((f_4^r)_{LF}, (f_3^r)_{LF}, p_4^l\right), \\ &= \boldsymbol{\sigma}_h^l \odot (f_3^r)_{LF} + \boldsymbol{\mu}_h^l. \end{aligned} \tag{9}$$

## B. Limitations and Future Work

As shown in Fig. 1, FEDER generates inaccurate segmentation maps when the camouflaged object is heavily obscured by surrounding environments. This is mainly because such obscuration can cause unusual object shapes, which pose challenges to the OER module for accurate edge reconstruction, and further suppress the segmentation performance. To address this issue, we will consider designing some background-based detection strategies to identify occlusions from the perspective of background consistency, and thus precisely detect those camouflaged objects.

Additionally, we will consider incorporating more powerful backbones, e.g., ScalableViT [6], with more strategic pretrain methods into the encoder of the COD task, such as SimVTP [5]. Furthermore, it would be desirable to employ image quality assessment techniques [1, 2] to aid in screening out low-quality camouflaged images in the dataset that are severely affected by degradation because these degraded images are deemed to have the potential to seriously affect the quality of downstream tasks [3, 4].

## References


[1] Runze Hu, Yutao Liu, Ke Gu, Xiongkuo Min, and Guangtao Zhai. Toward a no-reference quality metric for camera-captured images. *IEEE T. Cybernetics*, 2021. 2

[2] Runze Hu, Yutao Liu, Zhanyu Wang, and Xiu Li. Blind quality assessment of night-time image. *Displays*, 69:102045, 2021. 2

[3] Mingye Ju, Can Ding, Charles A Guo, Wenqi Ren, and Dacheng Tao. Idrlp: image dehazing using region line prior. *IEEE Trans. Image Process.*, 30:9043–9057, 2021. 2

[4] Mingye Ju, Chunming He, Juping Liu, Bin Kang, Jian Su, and Dengyin Zhang. Ivf-net: An infrared and visible data fusion deep network for traffic object enhancement in intelligent transportation systems. *IEEE Transactions Intell Transp Syst*, 2022. 2

[5] Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*, 2022. 2

[6] Rui Yang, Hailong Ma, Jie Wu, Yansong Tang, Xuefeng Xiao, Min Zheng, and Xiu Li. Scalablevit: Rethinking the context-oriented generalization of vision transformer. In *ECCV*, volume 13684, pages 480–496, 2022. 2