LEARNING TO FUSE HETEROGENEOUS FEATURES FOR LOW-LIGHT IMAGE ENHANCEMENT

Zhenyu Tang[⋆], *Long Ma*[‡], *Xiaoke Shang*[⋆⋆†], *Xin Fan*[⋆]

[⋆]DUT-RU International School of Information Science & Engineering, Dalian University of Technology
[‡]School of Software Technology, Dalian University of Technology
[†]College of Biomedical Engineering and Instrument Sciences, Zhejiang University

ABSTRACT

To see clearly in low-light scenarios, a series of learning-based techniques have been developed to improve visual quality. However, due to the absence of semantic-level features, the existing methods are perhaps less effective on semantic-oriented visual analysis tasks (e.g., saliency detection). To break down the limitation, we propose a new classification-driven enhancement method with heterogeneous feature fusion. Specifically, we construct a new low-light image enhancement network by integrating features acquired from the pre-trained classification network. Then, to better exploit the semantic-level information, we establish a Heterogeneous Feature Fusion (HF2) operation with channel-and-spatial attention to strength the effects of cross-domain features. HF2 acts on not only the fusion between classification and encoded features but also the fusion between encoded and decoded features. Extensive experiments are conducted to indicate our superiority against other state-of-the-art methods. The application on saliency detection further reveals our effectiveness in settling the semantic-oriented visual tasks.

***Index Terms*—** Low-light image enhancement · Classification network · Feature fusion · Saliency detection.

1. INTRODUCTION

The low-light images captured often suffer from low visibility and high noise, which affect severely the performance of high-level computer vision tasks such as saliency detection and visual tracking. In the following, we will briefly introduce the existing work and describe our main contributions.

1.1. Related Work

Over the past few decades, many approaches have been developed to address low-light image enhancement. They can be basically divided into two categories: model-based and learning-based methods.

This work is partially supported by the scientific research project of Zhijiang Laboratory (2019KB0AC02). * indicates the corresponding author. E-mail: sxk-1212@zju.edu.cn

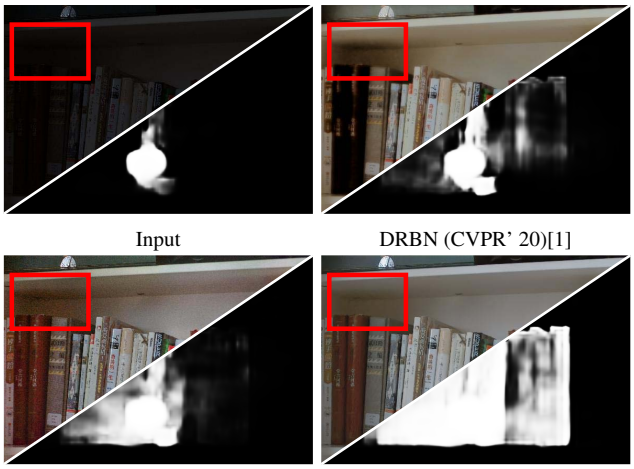


Fig. 1. The regions above and below diagonal are results of enhancement and saliency detection on the LOL dataset [3]. As for compared methods, there is severe color distortion or large noise as shown in the marked red rectangle. Our method realizes the best visual effects and obtains complete semantic-oriented structures while existing methods fail to realize.

Model-based Methods. Retinex theory [4, 5] describes the relationship between low and normal light image. Various traditional methods were proposed based on this theory and introduced different prior regularization terms to characterize desired layers [6, 7, 8, 9, 10]. However, limited to model capacity, these hand-crafted constraints are not adaptive enough so that their results may obtain intensive noises, inappropriate exposure, insufficient details and unsaturated colors.

Learning-based Methods. With deep learning developing, it shows great superiority in many computer vision fields. Chen *et al.* [3] established a new dataset (i.e., LOL) obtained by changing the exposure time, and this work developed a RetinexNet based on Retinex theory. KinD [11] was then designed by slightly changing the architecture of RetinexNet and adding some effective loss functions. Yang *et al.* [1] proposed a semi-supervised learning framework based on a deep

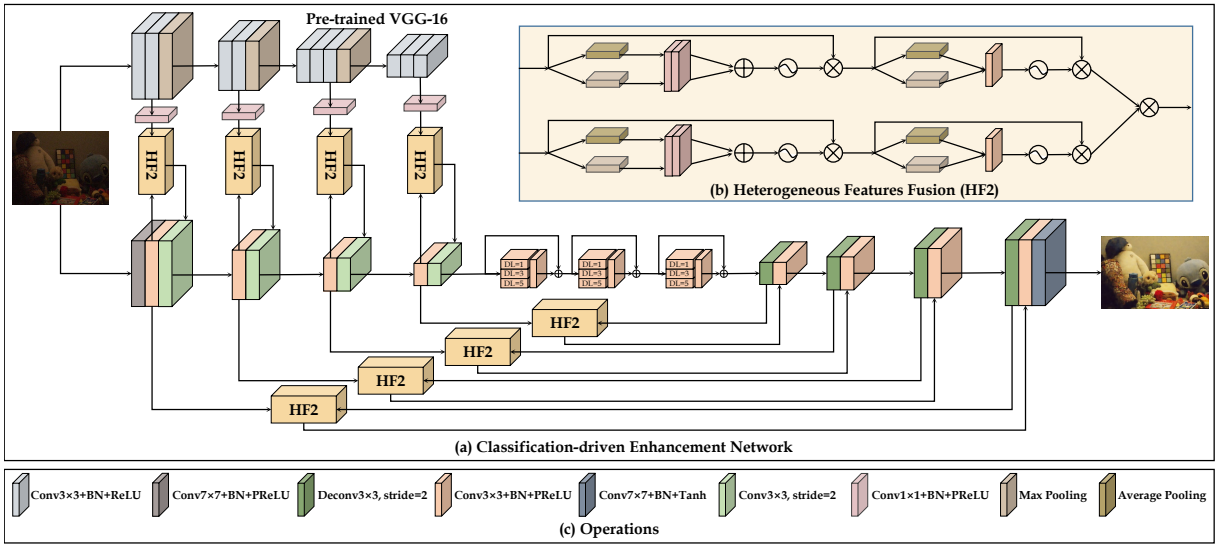


Fig. 2. The overall architecture of our proposed method. (a) Classification-driven enhancement network, which consists of the encoder, decoder and pre-trained classification network VGG16. (b) Heterogeneous Features Fusion (HF2), in which cross-domain features are through the attention mechanism and fused by using multiplication. (c) Operations, used in the figure.

recursive band network. Inspired by the adversarial mechanism, some recent works were proposed to handle LLIE by using GAN such as RetinexGAN [12] which proposed a generator and EnlightenGAN [2] which built generative adversarial network. Unfortunately, these methods usually lack the adequate utilization of features towards semantic-level tasks.

1.2. Our Contributions

To improve the semantic-friendliness of the enhanced results (ignored by most existing works), we propose a classification-driven enhancement network with heterogeneous features fusion. As shown in Fig. 1, our method performs better visual results and higher quality for the semantic-related tasks. We conclude our contribution as the following three-folds.

- We propose a novel classification-driven enhancement network that combines the semantic-level features derived from the classification network and possesses the ability of semantic perception, rather than only focusing on the visual quality as done in existing works.
- We establish a new-designed feature fusion mechanism-Heterogeneous Feature Fusion (HF2) to realize the valid information integration. HF2 acts on not only the fusion for classification and encoded features, but also the encoder and decoder features.
- Extensive evaluations indicate that our algorithm is remarkably superior to recently-proposed methods. We further apply our method to saliency detection to verify the ability of semantic perception which is usually less acquired in existing works.

2. THE PROPOSED APPROACH

2.1. Classification-driven Enhancement Network

Different from existing networks for low-light image enhancement, we utilize the semantic features extracted from the classification network to construct the network. To realize it, we adopt the U-Net architecture consisting of encoder and decoder. Because the encoder possesses a similar structure with the classification network, we incorporate the extracted features from the classification network into the encoder.

To be specific, we utilize the pre-trained VGG-16 [13] as the classification network. On the one hand, the pre-trained network model is acquired from the ImageNet [14], with strong generalization ability for different scenarios. On the other hand, we cannot retrain the network in low-light scenarios, because of lacking class labels in the paired dataset.

As shown in Fig. 2, our network architecture is composed two main backbones, i.e., pre-trained VGG-16 and U-Net with feature boost. The architecture of the U-Net with feature boost contains the encoder and decoder connected by HF2 in which the features are through channel attention [15] and spatial attention [15] and fused by using multiplication.

2.2. Heterogeneous Features Fusion Operation (HF2)

In fact, we can directly fuse the cross-domain features (e.g., the classification and encoded features) by using the commonly-used operator, such as concatenation, element-wise operation (addition/multiplication). However, these operators usually embody an implicit assumption, i.e., fused features are from the same domain or source, which leads to the ignorance on some important contextual information.

Table 1. Quantitative comparison on the LOL dataset. The best result is in **red** whereas the second one is in **blue**.

Metrics	SSIENet	RetinexNet	MBLLEN	EnGAN	KinD	FIDE	DRBN	ZeroDCE	Ours
PSNR	14.176	13.096	16.491	15.644	16.244	**17.609**	15.324	15.512	**19.924**
SSIM	0.573	0.424	0.630	0.565	0.667	0.688	**0.702**	0.519	**0.802**

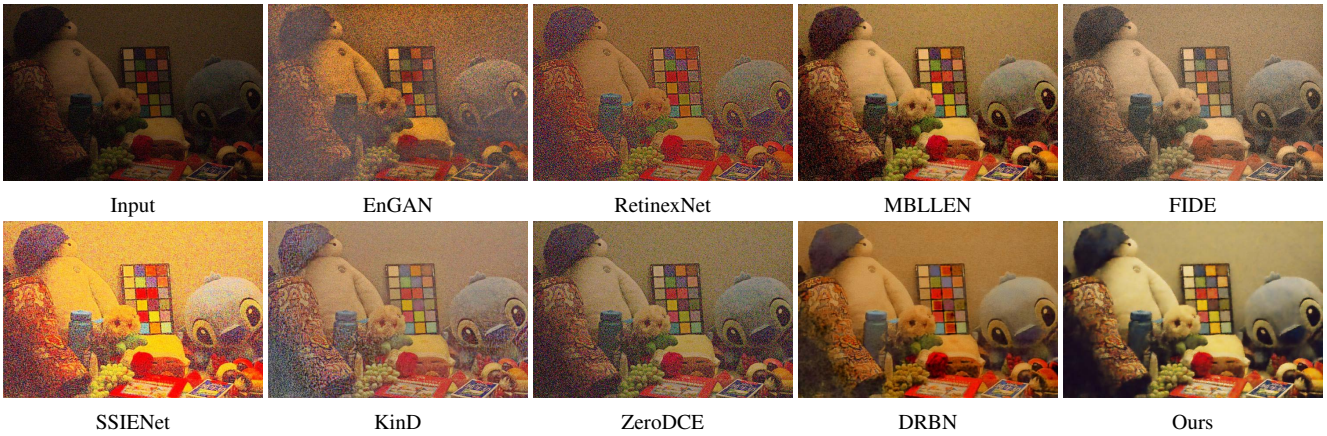


Fig. 3. Visual comparison on LOL dataset among state-of-the-art low-light image enhancement approaches.

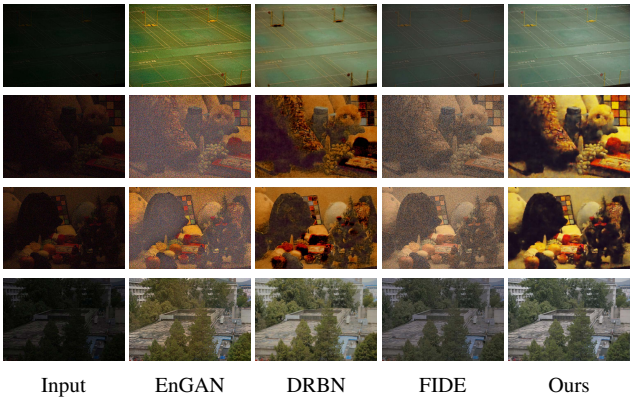


Fig. 4. More visual results on LOL dataset.

As described in the above section, the fused features usually originate from different tasks. Thus, we try to build a new aggregation mechanism to fully exploit various features.

We name our fusion mechanism as Heterogeneous Features Fusion operation, abbreviated as HF2. Specifically, this operation aims to automatically combine different features. Fortunately, the widely-used attention behavior can exactly caters to our appeal. Considering the inherent spatial and channel dimensions of the given feature, we cascade the channel and spatial attention of HF2. As shown in the right top corner of Fig. 2, HF2 can be formulated as

$$\mathcal{F}_{\mathtt{f}}^{i} = \mathcal{A}_{\mathtt{s}}(\mathcal{A}_{\mathtt{c}}(\mathcal{F}_{\mathtt{a}}^{i})) \otimes \mathcal{A}_{\mathtt{s}}(\mathcal{A}_{\mathtt{c}}(\mathcal{F}_{\mathtt{b}}^{i})), \tag{1}$$

where $\mathcal{F}_{\mathtt{a}}^{i}$ and $\mathcal{F}_{\mathtt{b}}^{i}$ represent the extracted features from two different sub-network. In this work, we define the $\mathtt{a}$ as the encoder, and $\mathtt{b}$ as the pre-trained VGG-16/decoder. $\mathcal{F}_{\mathtt{f}}^{i}$ denotes the fusion feature which is outputted to the next layer. $\mathcal{A}_{\mathtt{s}}$ and $\mathcal{A}_{\mathtt{c}}$ are the spatial and channel attention, respectively. $\otimes$ denotes multiplication as the fusing operator of HF2.

By performing the channel and spatial attention to acquire the valuable features, our HF2 can aggregate the information from different domains to great extent[1].

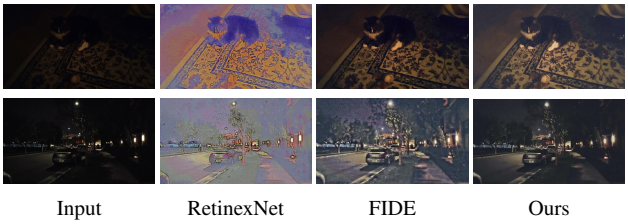


Fig. 5. More visual comparison on LoLi-Phone dataset.

2.3. Training Loss

Considering the importance of image details and textures, we use multi-scale structure similarity (MS-SSIM) loss function $\mathcal{L}_{MS-SSIM}$ [16] to optimize our network and increase the sensitivity towards contrast and color of images. Besides, we choose to adopt $\mathcal{L}_1$ loss function as a pixel-wise objective function. Therefore, this loss function is defined as:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{MS-SSIM} + \lambda_2 \cdot \mathcal{L}_1. \tag{2}$$

where λ_1, λ_2 are respectively set to 0.8 and 0.2.

3. EXPERIMENTS

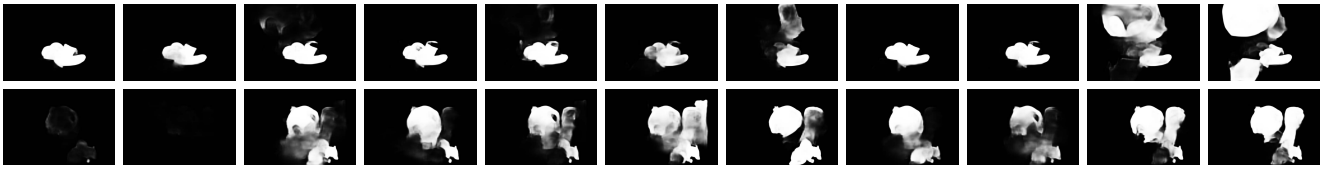
3.1. Implementation Details

We use LOL dataset for training and evaluation and utilize 689 samples for training and the rest 100 samples for eval-

[1]Experimental evaluation can be found in Sec. 3.4.

Table 2. Saliency detection results on LOL dataset. The best result is in **red** whereas the second best one is in **blue**.

Metrics	RetinexNet	MBLLEN	GLADNet	SSIENet	KinD	EnGAN	FIDE	ZeroDCE	DRBN	Ours
S-measure ↑	0.601	0.664	0.690	0.659	**0.708**	0.656	0.667	0.657	0.685	**0.733**
E-measure ↑	0.577	0.741	0.761	0.732	**0.768**	0.680	0.704	0.705	0.732	**0.767**
F-measure ↑	0.262	0.409	0.438	0.402	**0.459**	0.377	0.393	0.410	0.443	**0.474**
MAE ↓	0.135	0.114	0.116	0.111	0.111	0.107	0.108	0.112	**0.105**	**0.089**



Input RetinexNet MBLLEN EnGAN SSIENet KIND DRBN FIDE ZeroDCE Ours GT

Fig. 6. Application examples of the proposed method and state-of-the-arts for salient object detection from LOL dataset.

Table 3. Ablation Study.

Method	VGG	W/O HF2	W/ HF2			PSNR	SSIM
			$\oplus$	$\bigcup$	$\otimes$		
M_0	✓	✓				17.615	0.742
M_1	✓		✓			19.794	0.790
M_2	✓			✓		19.688	0.799
M_3	✓				✓	**19.924**	**0.802**

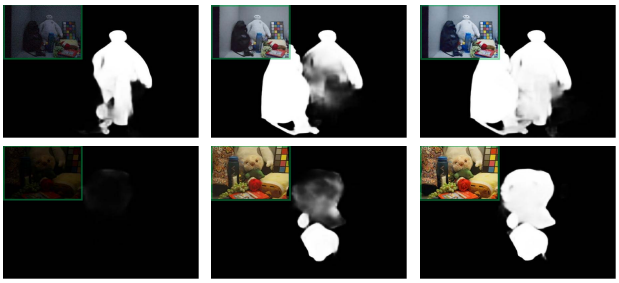
uation. Meanwhile, we mainly use two widely-used metrics (i.e., PSNR and SSIM) to measure the quality of images. We compare our method with eight state-of-the-art deep networks, including SSIENet [17], RetinexNet [3], MBLLEN [18], EnGAN [2], KinD [11], FIDE [19], DBRN [1]and ZeroDCE [20]. We set the batch size to 8 and use ADAM optimizer with default parameters($\beta_1 = 0.9$, and $\beta_2 = 0.999$). The initial learning rate is fixed as 5×10^{-4}, decreasing to 0.2 times every 7000 iterations during training.

3.2. Comparisons with State-of-the-Art

From the table 1, we discover our result is superior to other methods in terms of PSNR, SSIM. As shown in Fig. 3, we find that some existing methods such as EnGAN [2], RetinNex [3], MBLLEN [18], SSIENet [17] and KinD [11], will produce excessive noise and inappropriate exposure, while DRBN [1] and FIDE [19] cause severe contrast reduction and color distortion. Almost all compared methods fail to make full use of texture information. In contrast, our network can obtain richer texture information and form favorable contrast and color of images. Fig. 4 provides more comparisons. In Fig. 5, images of LoLi-Phone dataset are enhanced by models of existing methods and ours which are trained on LOL dataset.

3.3. Saliency Detection

To further verify the effectiveness and applicability of our proposed network, we also apply our method to the salient object detection task. Then we made salience predictions for enhanced images using F^3Net [21]. When measuring metrics, we choose the salient map of ground truth as the label. The quantitative results of the salience maps are displayed in the table 2. As summarized in it, our method performs against the state-of-the-art methods in four evaluation metrics. Fig. 6 shows that our method enables enhanced images to possess more complete structure and precise boundary, while existing methods even fail to capture roughly outline of the objects.



Input W/O VGG W/ VGG

Fig. 7. Effects of pre-trained VGG network.

3.4. Analysis and Discussions

In the table 3 and Fig. 7, we conduct the ablation analysis. $\oplus$, $\bigcup$ and $\otimes$ respectively represent addition, concatenation and multiplication as the fusing operators of HF2.

Effects of VGG network. From Fig. 7, we discover that with VGG network introduced, the results of saliency detection possess more complete structure, which proves the importance of the semantic features extracted by VGG network.

Effects of HF2. As the table 3 seen, with the module HF2, the quality of images are improved greatly. In the comparison of different operators, M_3 is separately superior to M_1 0.13dB and M_2 0.24dB. Therefore, we choose multiplication as the fusing operator of HF2.

4. CONCLUSIONS

In this paper, we propose our enhancement network by fusing cross-domain features acquired from VGG network to better exploit the semantic-level information. The abundant Experiments fully demonstrated the superiority of our algorithm in the semantic-oriented visual tasks.

5. REFERENCES

[1] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *CVPR*, 2020, pp. 3063–3072.

[2] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, and et al, "Enlightengan: Deep light enhancement without paired supervision," *IEEE TIP*, vol. 30, pp. 2340–2349, 2021.

[3] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC*, 2018.

[4] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *CVPR*, 2016, pp. 2782–2790.

[5] Edwin H Land and John J McCann, "Lightness and retinex theory," *JOSA*, vol. 61, no. 1, pp. 1–11, 1971.

[6] Xueyang Fu, Yinghao Liao, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding, "A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation," *IEEE TIP*, vol. 24, no. 12, pp. 4965–4977, 2015.

[7] Xiaojie Guo, Yu Li, and Haibin Ling, "Lime: Image enhancement via illumination map estimation," *IEEE TIP*, vol. 26, no. 2, pp. 982–993, 2017.

[8] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE TIP*, vol. 27, no. 6, pp. 2828–2841, 2018.

[9] Qing Zhang, Yongwei Nie, Lei Zhu, Chunxia Xiao, and Wei-Shi Zheng, "Enhancing underexposed photos using perceptually bidirectional similarity," *IEEE TMM*, vol. 23, pp. 189–202, 2021.

[10] Risheng Liu, Shichao Cheng, Long Ma, Xin Fan, and Zhongxuan Luo, "Deep proximal unrolling: Algorithmic framework, convergence analysis and applications," *IEEE TIP*, vol. 28, no. 10, pp. 5013–5026, 2019.

[11] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo, "Kindling the darkness: A practical low-light image enhancer," in *ACM MM*, 2019, pp. 1632–1640.

[12] Yangming Shi, Xiaopo Wu, and Ming Zhu, "Low-light image enhancement algorithm based on retinex and generative adversarial network," *arXiv*, p. 1906.06027, 2019.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, p. 1409.1556, 2015.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, September 2018.

[16] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE TCI*, vol. 3, no. 1, pp. 47–57, 2017.

[17] Yu Zhang, Xiaoguang Di, Bin Zhang, and Chunhui Wang, "Self-supervised image enhancement network: Training with low light images only," *arXiv*, p. 2002.11300, 2020.

[18] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim, "Mbllen: Low-light image/video enhancement using cnns.," in *BMVC*, 2018, p. 220.

[19] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *CVPR*, 2020, pp. 2281–2290.

[20] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, June 2020.

[21] Jun Wei, Shuhui Wang, and Qingming Huang, "F³net: Fusion, feedback and focus for salient object detection," *AAAI*, vol. 34, no. 07, pp. 12321–12328, Apr. 2020.