Endoscopic Depth-of-Field Expansion via Cascaded Network with Two-streamed Multi-scale Fusion

Xiang Deng[1,†], Xing Liu[2,†], Tian Xu[1], Xiaoyue Liu[1], Tianyuan Gan[1], Chen Lu[3], Congcong Zhou[3], Peng Wang[4], Yong Lei[5], and Xuesong Ye[1,*]

[1] Biosensor National Special Laboratory, College of Biomedical Engineering and Instrument Science, Zhejiang University, China.
[2] Polytechnic Institute, Zhejiang University, China
[3] Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, China
[4] Hangzhou Xianao Technology Inc, China
[5] School of Mechanical Engineering, Zhejiang University, China
yexuesong@zju.edu.cn

Abstract. The application of ultrahigh definition endoscopy systems in minimally invasive surgeries has become increasingly widespread. However, their high resolution results in a reduced depth of field (DOF), making it difficult to achieve clear imaging across the entire frame. Unlike improvements in optical structures, we address this issue using a deep learning-based multi-focus image fusion (MFIF) approach. Traditional MFIF methods are less effective in endoscopic scenarios due to their inadequate design for extracting information from complex organ structures. To address these limitations, this work proposes a two-streamed cascaded encoder-decoder network that incorporates multi-scale feature extraction and fusion mechanisms validated in medical image segmentation. The network includes novel multi-scale fusion module with cross-axial attention that hierarchically integrates features using attention-guided weights and hybrid operations, effectively preserving intra-domain textures while modeling cross-domain dependencies. The framework is rigorously validated using novel real-world endoscopic datasets collected from imaging experimental platform. The experimental results demonstrate that the proposed method outperforms traditional approaches in benchmark tests. Code available at: https://github.com/luoyu5023/CTMFusion.

Keywords: Multi-focus image fusion · Cascaded Network · Multi-scale feature.

1 Introduction

Endoscopes serve as pivotal diagnostic instruments in contemporary medicine, enabling direct visualization of target organs and tissues through real-time morphological characterization. This imaging modality provides clinicians with crit-

† The authors contribute equally to this work. * Corresponding Author.

ical pathological evidence, thus facilitating rapid and accurate diagnostic evaluations. Minimally invasive endoscopic surgery, recognized for its procedural precision and reduced iatrogenic trauma, is considered a critical advance in next-generation clinical interventions [3].

However, due to the inherent limitations of optical imaging, the pursuit of high resolution and magnification in endoscopy inevitably results in reduced DOF [23]. Although autofocus technology has widespread implementation, substantial depth variations across anatomical structures cannot be simultaneously brought into sharp focus during single-frame acquisition. Unlike physical improvement in optical imaging systems, the simultaneous capture and fusion of images from multiple imaging planes present a feasible method for extending the DOF in endoscopic imaging.

Traditional methods [7][9][10] typically rely on handcrafted feature extraction, which frequently results in edge blurring and detail loss due to limited representational capacity. In recent years, deep learning-based approaches have emerged as the predominant methodology, evolving from CNNs [24] to architectures that incorporate Swin Transformers [8]. A primary focus of these advances lies in optimizing the extraction of both local and long-range features to enhance fusion performance. In particular, FusionDiff [6], as a diffusion model tailored for MFF tasks, demonstrates the potential of using limited training samples to approximate real-data distributions and generate high-quality fused output.

Previous studies, whether they employ CNNs or Transformers, predominantly decouple feature extraction and feature fusion into sequential stages. This paradigm allocates computational resources disproportionately to the feature extraction module, thereby delaying the integration of complementary information. Unsupervised ZMFF [4] utilizes two U-Net-based hourglass-structured networks but exhibits insufficient cross-level feature interaction, failing to take advantage of high-level semantic guidance to refine low-level feature learning.

By uniformly distributing computational resources across cascaded encoding-decoding stages and iteratively executing feature fusion at each stage [20], our network enables mutual guidance between preliminary features and fused outputs. This architecture facilitates a deeper exploration of multi-focus image interdependencies, where early fusion results inform subsequent feature extraction and refinement. It improves hierarchical feature utilization and the model's ability to capture latent correlations between focal planes.

The scarcity of task-specific training data remains a critical bottleneck that limits algorithmic precision. Although general-purpose fusion networks (e.g., those supporting MRI/CT fusion and multi-exposure fusion) have been developed, empirical evidence consistently shows that algorithms tailored to specific scenarios or tasks outperform generic solutions [22]. This observation further underscores that achieving optimal performance in endoscopic imaging requires both explicitly adapted algorithm designs for endoscopic modalities and the incorporation of real endoscopic datasets to address domain-specific challenges.

In this paper, our contributions are as follows. 1) We establish a two-streamed and cascaded encoding-decoding network, achieving DOF expansion through the generation of decision maps and MFIF. 2) We construct multi-scale feature inter-domain and cross-domain fusion modules, systematically integrating multi-input multi-scale features. 3) The method achieves state-of-the-art image fusion performance on EVP-MFI(a real laparoscopic dataset), with minimal blur artifacts and exceptional detail and texture preservation.

2 Methods

An overview of the proposed framework is illustrated in Fig. 1.The optical system simultaneously inputs far-focus and near-focus image of the same size. RGB inputs are first converted into the YCbCr color space. Next, the Y (luminance) channel is employed as the input of the fusion model since the structural details and intensity information are mainly concentrated in this channel. After initial feature extraction via the Stem module, multi-scale feature encoding-decoding and feature fusion are performed in the cascade network(num_stages = 3).

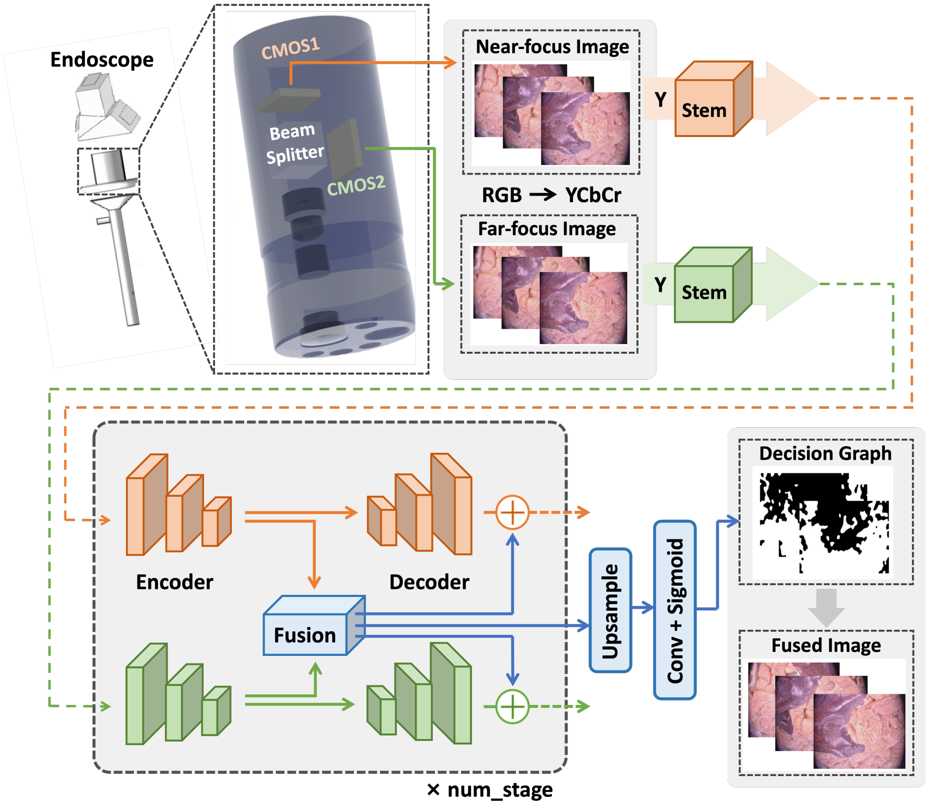

Fig. 1. The framework of the DOF expansion system.

The input image with a spatial size of $H \times W$, either from a far-focus or near-focus perspective, is processed through the stem module to extract high-resolution features of size $\frac{H}{4} \times \frac{W}{4}$. Each stem module incorporates two sequentially arranged 3×3 convolutional layers (each configured with a stride of 2) for early visual processing.

Sec.2.1 describes the cascade encoder and decoder framework, Sec.2.2 details the cross-domain and inter-feature fusion of multi-scale features, and Sec.2.3 outlines the loss function composition.

2.1 Cascaded Network

This study uses SegNeXt [2] as the multi-scale feature encoder. Unlike Transformer-based models, SegNeXt introduces an efficient attention mechanism and employs cheaper and larger kernel convolutions. The encoder extracts features from the backbone network at three hierarchical levels with spatial resolutions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, and $\frac{H}{16} \times \frac{W}{16}$, respectively, which are subsequently channeled into the decoder for further processing.

The Decoder module performs a multi-scale feature aggregation through a down-top pathway with channel reduction and layer normalization. Low-resolution features are integrated with high-resolution features through unsample operations and element-wise addition. The final output retains the original spatial resolution, taking guidance from the current stage's fusion result, and serves as the input to the encoder in the next stage.

Under the rigorous requirements of minimally invasive surgical applications, this study adopts the decision map generation paradigm to preserve complete image detail integrity while circumventing erroneous texture synthesis [16]. To eliminate subjective bias, the proposed network architecture ensures spatial continuity of in-focus regions within input images through direct mask operations between the generated decision map and source images, bypassing conventional post-processing steps.

Thus, the output from the last layer undergoes upsampling, convolution, and sigmoid activation, producing the decision map and then directly generating the final fused image combined with origin input. This study employs CARAFE [15] for upsampling, which effectively aggregates contextual information within a large receptive field while being lightweight and fast to compute.

2.2 Feature fusion module

To establish long-range dependencies across domains, the Cross SwinTransformer module is used to integrate features of the same scale from two image streams. For the query of one image stream, the value and key of the other image stream are used to perform an attention-guided combination of cross-domain information, along with residual connections to preserve information within the domain.

A hierarchical hybrid fusion strategy is implemented: In shallow layers, channel concatenation preserves spatial details from dual-stream input, maintaining texture fidelity while enhancing local feature representation for subtle biological structures. Deeper layers adopt max fusion to process high-level semantics (e.g. organ contours), where competitive selection enhances salient features and suppresses blur artifacts, aligning with model lightweighting requirements [19]. We further incorporate multi-scale cross-axial attention features [13], proven effective in medical segmentation, to model long-range tissue dependencies and mitigate segmentation errors at ambiguous boundaries.

The fused features are ultimately generated at $\frac{H}{4} \times \frac{W}{4}$ resolution through feature concatenation followed by depth-wise and point-wise convolutions.

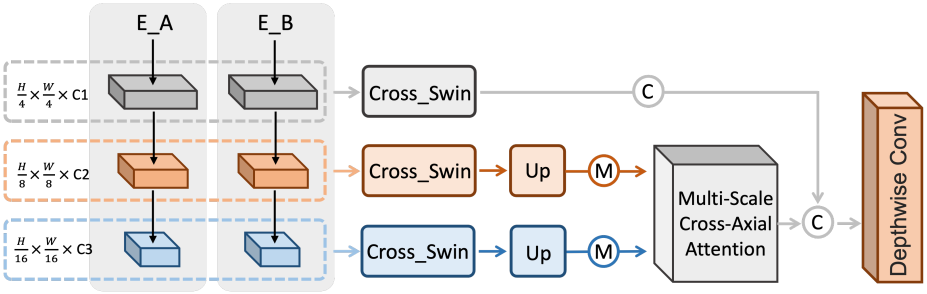
E_A
E_B
$\frac{H}{4} \times \frac{W}{4} \times C1$
$\frac{H}{8} \times \frac{W}{8} \times C2$
$\frac{H}{16} \times \frac{W}{16} \times C3$
Cross_Swin
Cross_Swin
Cross_Swin
Up
Up
M
M
C
C
Multi-Scale Cross-Axial Attention
Depthwise Conv

Fig. 2. Feature Fusion Module. The architecture processes multi-scale features from near/far-focus inputs, where **C** denotes channel-wise concatenation and **M** represents channel-wise max pooling operation.

2.3 Loss Function

For training stage , the total loss $\mathcal{L}_{total}$ is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{int} + \lambda_2 \mathcal{L}_{text} + \lambda_3 \mathcal{L}_{ssim}, \tag{1}$$

$$\mathcal{L}_{ssim} = 0.5 \times (1 - SSIM(F, A)) + 0.5 \times (1 - SSIM(F, B)) \tag{2}$$

$$\mathcal{L}_{text} = \frac{1}{HW} \parallel |\nabla F| - \max(|\nabla A|, |\nabla B|) \parallel_1 \tag{3}$$

$$\mathcal{L}_{int} = \frac{1}{HW} \parallel F - 0.5 \times (A + B) \parallel_1 \tag{4}$$

where F represents the fused image, A / B denotes the near or far focus image with the size of $H \times W$. $\|\cdot\|_1$ denotes the $l_1 - norm$ and ∇ indicates Sobel gradient operator.

The loss function consists of three components: 1) Intensity Loss ($\mathcal{L}_{int}$), which guides the fusion model to capture appropriate intensity information based on the global intensity of the source images. 2) Texture loss ($\mathcal{L}_{text}$), designed to preserve texture details from source images, uses a maximum selection strategy to aggregate these details effectively. 3) SSIM Loss ($\mathcal{L}_{ssim}$), which constrains the structural similarity between the fused image and the source images.

3 Experiments

3.1 Datasets and Evaluation Metrics

The datasets employed in this research are summarized in Table. 1. We first pretrain model using MFF-WHU [21] and NYD [14], which synthesize multi-focus images based on depth information from wild images.

Table 1. Dataset settings.

Purpose	Dataset	Number	Image Size
Pre-training	MFI-WHU [21]	120	Not fixed
	NYD [14]	100	600×440
Fine-tuning	SCARED-Depth	110	1280×1024
Test	EVP-MFI	115	960×540

To enhance the network perception of endoscopic images, the model is fine-tuned on medical data after pre-training on natural images. We processed the SCARED dataset [1], which contains a large number of intraoperative laparoscopic images, using Depth Anything V2 [18] and Otsu binarization to generate depth masks. These masks, combined with Gaussian noise applied to the original images, are used to create laparoscopic multi-focus image pairs. Ultimately, 110 pairs were selected for fine-tuning.

Addressing the current lack of practical multi-focus medical imaging datasets in clinical research, we developed a multi-focus experimental platform to systematically acquire images as test dataset named EVP-MFI, with the schematic diagram illustrated in Fig.1. We constructed a simulated abdominal cavity environment using fresh ex vivo porcine organs (liver, stomach, and intestines), subsequently acquiring 115 paired multi-focus images across multiple poses.

The effectiveness of MFIF was evaluated using six quantitative metrics: Feature Mutual Information (FMI), Peak Signal-to-Noise Ratio (PSNR), Nonlinear Correlation Information Entropy ($\mathrm{Q_{NICB}}$), Edge-Based Similarity Measurement ($\mathrm{Q^{AB/F}}$), SSIM-based fusion quality metric ($\mathrm{Q^{Y}}$) for structural consistency verification, and Human visual perception($\mathrm{Q_{CB}}$).Higher metrics indicate that the fusion image is better [22].

3.2 Implementation details

The batch size is set to 12 by default, with 1000 training epoches in total. At each step, images from the training set are randomly cropped into 128 × 128 patches, and data augmentation is applied by flipping to enhance image diversity. The Adam optimizer is used to update the network weights with an initial learning rate of 2e-4, while MultiStepLR controls the decay. The hyperparameters for the loss function are empirically set as $\lambda_1 = 20$, $\lambda_2 = 20$, $\lambda_3 = 50$.

The proposed network was implemented in PyTorch and training was carried out for 1000 epochs on four Nvidia A100 GPUs, which took approximately 5 hours to complete.

3.3 Quantitative Comparison with SOTA

Table 2 presents a quantitative comparison between our method and multiple MFIF approaches, where the baseline methods were implemented by their best model according to official codes. Our method achieves significant improvements over previous work. Although marginally underperforming generative networks in PSNR, it still exceeds previous decision map-based models. The superior FMI scores(+**0.019**) indicate that the fused images preserve more comprehensive feature information from source images, validating the enhanced capability of our feature fusion module. The higher $\mathrm{Q}^{\mathrm{AB/F}}$(+**0.101**) and Q^{Y}(+**0.167**) metrics reveal that our cascaded network architecture effectively extracts deep-level edge and structural information from source images. Notably, the performance exhibits a marginal improvement through network fine-tuning, whereas networks without prior medical image training may underperform compared to traditional methods.

Table 2. Quantitative results on EVP-MFI dataset. **Bold** and Underline show the best and second-best results, respectively.

Methods	FMI	$\mathrm{Q}^{\mathrm{AB/F}}$	PSNR	Q^{Y}	$\mathrm{Q}_{\mathrm{NICB}}$	Q_{CB}
DSIFT [10]	0.8302	0.3376	33.81	0.5630	0.8121	0.6671
NSCT [9]	0.8352	0.4170	33.36	0.6686	0.8114	0.7038
GFF [7]	0.8320	0.3641	33.15	0.6548	0.8120	0.6816
IFCNN [24]	0.8291	0.4033	33.30	0.6689	0.8111	0.6753
SESF [11]	0.8320	0.3641	33.15	0.6548	0.8120	0.6816
DRPL [5]	0.8439	0.4917	33.25	0.8336	0.8136	0.6617
U2Fusion [17]	0.8369	0.3961	33.05	0.6344	0.8125	0.6717
DeFusion [8]	0.8291	0.3262	33.56	0.5428	0.8119	0.6594
SwinFusion [12]	0.8444	0.4792	**33.89**	0.8110	0.8135	0.6941
SwinMFF [16]	0.8359	0.4429	31.33	0.7518	0.8115	0.6549
ZMFF [4]	0.8317	0.4163	32.89	0.6764	0.8104	0.6735
FusionDiff [6]	0.8320	0.3769	30.19	0.6585	0.8118	0.6842
ours	0.8638	0.5926	33.36	0.9780	0.8252	0.7193
ours(fine-tuned)	**0.8672**	**0.5977**	33.34	**0.9848**	**0.8279**	**0.7228**

3.4 Qualitative comparison with SOTA

Fig. 3 highlights the qualitative visualization superiority of our framework over existing deep learning-based baselines in EVP-MFI. Our method preserves intricate textural details from far-focused regions (see blue insets) while maintaining

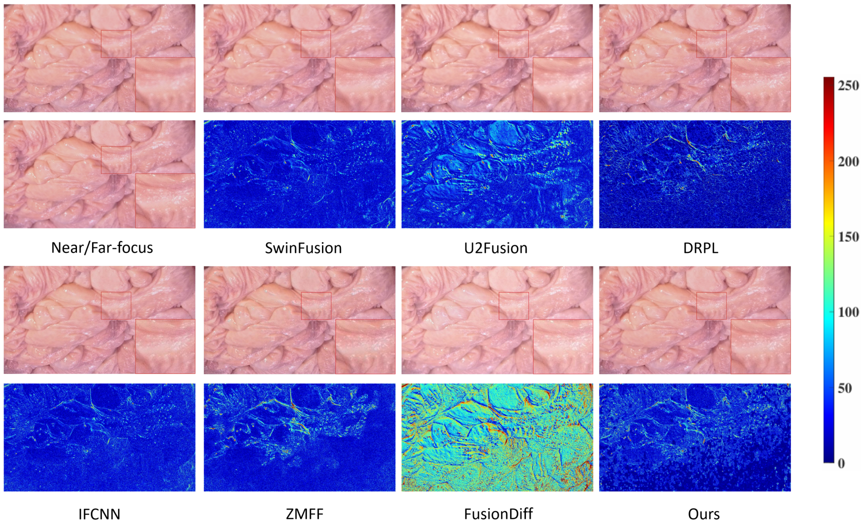

Fig. 3. Qualitative results of EVP-MFI. From left to right, top to bottom: near/far-focus image, the fused results and difference maps. The difference maps represent the difference between the near-focus image and fused results, manifesting as obvious texture details in the upper region.

sharper focus transition boundaries (red boxed area). The proposed framework generates fewer blurring artifacts at the fusion boundaries while circumventing the luminance distortion inherent in the FusionDiff method. These results support the superior applicability of our model in clinical settings.

3.5 Ablation Study

To further investigate the impact of the proposed fusion module and the number of cascade stages on performance, ablation experiments are performed in Table. 3. The ablation of either the Decoder or hybrid strategy in the fusion module leads to a consistent degradation of performance metrics. Performing feature fusion at an earlier stage effectively assists the network in learning deeper features($+0.004$, $+0.029$, $+0.053$ in FMI, $\mathrm{Q}^{\mathrm{AB/F}}$ and Q^{Y}). Cross-Swin proves its strong capability in extracting edge and structural information across diverse domains. And we validate that three stages yield the optimal balance in performance.

4 Conclusion

This study proposes a high-resolution DOF expansion method for endoscopic systems, which effectively fuses multi-focus endoscopic images. Our method outperforms existing networks on a novel real-world endoscopic test set(EVP-MFI),

Table 3. Ablation experiments results with EVP-MFI. **Bolded** values indicate the best performance.(w/o stage1-2 indicates single stage, and so forth)

Configurations	FMI	$Q^{AB/F}$	Q^{Y}
w/o stage1-2	0.8613	0.5832	0.9629
w/o stage1	0.8616	0.5914	0.9727
w/o Decoder	0.8620	0.5763	0.9716
w/o fusion from stage1-2	0.8599	0.5635	0.9250
w/o Cross-Swin	**0.8640**	0.5790	0.9559
hybrid strategy → cat	0.8602	0.5821	0.9443
CARAFE→bilinear	0.8621	0.5813	0.9727
ours	0.8638	**0.5926**	**0.9780**

with gains of 0.019, 0.101 and 0.167 in FMI, $Q^{AB/F}$ and Q^{Y}. The cascaded network architecture developed in this research enables a more efficient learning of complementary information and the deep features of the image pairs. In future work, we plan to extend the application domain of this fusion framework to other clinical endoscopic systems.

Acknowledgments This work was supported by the National Key Research and Development Program of China (2022YFB4700800).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
2. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in neural information processing systems **35**, 1140–1156 (2022)
3. Hamad, G.G., Curet, M.: Minimally invasive surgery (2010)
4. Hu, X., Jiang, J., Liu, X., Ma, J.: Zmff: Zero-shot multi-focus image fusion. Information Fusion **92**, 127–138 (2023)
5. Li, J., Guo, X., Lu, G., Zhang, B., Xu, Y., Wu, F., Zhang, D.: Drpl: Deep regression pair learning for multi-focus image fusion. IEEE Transactions on Image Processing **29**, 4816–4831 (2020)
6. Li, M., Pei, R., Zheng, T., Zhang, Y., Fu, W.: Fusiondiff: Multi-focus image fusion using denoising diffusion probabilistic models. Expert systems with applications **238**, 121664 (2024)
7. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Transactions on Image processing **22**(7), 2864–2875 (2013)
8. Liang, P., Jiang, J., Liu, X., Ma, J.: Fusion from decomposition: A self-supervised decomposition approach for image fusion. In: European Conference on Computer Vision. pp. 719–735. Springer (2022)

9. Liu, Y., Liu, S., Wang, Z.: A general framework for image fusion based on multi-scale transform and sparse representation. Information fusion **24**, 147–164 (2015)
10. Liu, Y., Liu, S., Wang, Z.: Multi-focus image fusion with dense sift. Information Fusion **23**, 139–155 (2015)
11. Ma, B., Zhu, Y., Yin, X., Ban, X., Huang, H., Mukeshimana, M.: Sesf-fuse: An unsupervised deep model for multi-focus image fusion. Neural Computing and Applications **33**, 5793–5804 (2021)
12. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022)
13. Shao, H., Zeng, Q., Hou, Q., Yang, J.: Mcanet: Medical image segmentation with multi-scale cross-axis attention. arXiv preprint arXiv:2312.08866 (2023)
14. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. pp. 746–760. Springer (2012)
15. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe: Content-aware reassembly of features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3007–3016 (2019)
16. Xie, X., Guo, B., Li, P., He, S., Zhou, S.: Swinmff: toward high-fidelity end-to-end multi-focus image fusion via swin transformer-based network. The Visual Computer pp. 1–24 (2024)
17. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE transactions on pattern analysis and machine intelligence **44**(1), 502–518 (2020)
18. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2025)
19. Yang, Y., Xu, W., Huang, S., Wan, W.: Low-light image enhancement network based on multi-scale feature complementation **37**(3), 3214–3221 (2023)
20. Zhang, G., Li, Z., Tang, C., Li, J., Hu, X.: Cednet: A cascade encoder–decoder network for dense prediction. Pattern Recognition **158**, 111072 (2025)
21. Zhang, H., Le, Z., Shao, Z., Xu, H., Ma, J.: Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Information Fusion **66**, 40–53 (2021)
22. Zhang, X.: Deep learning-based multi-focus image fusion: A survey and a comparative study. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4819–4838 (2021)
23. Zhang, Y.: Deep learning-enhanced microscopy with extended depth-of-field. Light: Science & Applications **12**(1), 284 (2023)
24. Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: Ifcnn: A general image fusion framework based on convolutional neural network. Information Fusion **54**, 99–118 (2020)