INTERACTIVE SEPARATION NETWORK FOR IMAGE INPAINTING

Siyuan Li[⋆], *Lu Lu*[⋆], *Zhiqiang Zhang*[⋆], *Xin Cheng*[⋆], *Kepeng Xu*[⋆], ✉*Wenxin Yu*[⋆]
Gang He[†], *Jinjia Zhou*[§], *Zhuo Yang*[‡]

[⋆] Southwest University of Science and Technology, Mianyang, Sichuan, China
[†] Xidian University, Xi'an, Shaanxi, China
[§] Hosei University, Tokyo, Janpanese
[‡] Guangdong University of Technology, Guangzhou, Guangdong, China
✉yuwenxin@swust.edu.cn

ABSTRACT

Image inpainting, also known as image completion, is the process of filling in the missing region of an incomplete image to make the repaired image visually plausible. Strided convolutional layer learns high-level representations while reducing the computational complexity, but fails to preserve existing detail from the original images (eg, texture, sharp transients), therefore it degrades the generative model in image inpainting task. To reduce the erosion of high-resolution components of images meanwhile maintaining the semantic representation, this paper designs a brand-new network called Interactive Separation Network that progressively decomposites the features into two streams and fuses them. Besides, the rationality of network design and the efficiency of proposed network is demonstrated in the ablation study. To the best of our knowledge, the experimental results of proposed method are superior to state-of-the-art inpainting approaches.

***Index Terms*—** Image Inpainting, Deep Learning, Feature Representation, Multi-Scale Feature, Feature Fusion

1. INTRODUCTION

Image inpainting, the process of reconstructing the corrupted parts of an incomplete image, plays a vital role in various computer vision applications. Since Convolutional Neural Network (CNN) has been introduced into the research field of image inpainting, learning-based image inpainting approaches have been rapidly developed. Although the strided convolution and pooling layer in CNN inevitably drop out some high-resolution spatial signals, they have pivotal roles in compressing the features into a high-level representation.

Due to this advantage of CNN, most of learning-based methods have the ability to fill in the damaged area by hallucinating (imagining) some novel objects similar to the counterpart existing in the real world. The one of typical work benefitted from CNN is the Context Encoder [1], which firstly applies deep learning to image inpainting and demonstrates that the learning-based method can greatly outperform the traditional methods. Context Encoder embeds the corrupted image into the high-level feature maps with low-dimension, which the decoder then uses to reconstruct the predicted image. However, due to its monotonic loss function and sequential network structure, the resulting images are visually obscured and contain many inconsistent artifacts. Affected by this fundamental contribution, the encoder-decoder architecture has been commonly used as a generative model in the learning-based approaches for image inpainting.

Since Goodfellow *et al.* proposed Generative Adversarial Network (GAN) [2], it has become popular to use adversarial learning to enhance generated image quality. Iizuka [3] proposed two discriminators to reinforce the global and local consistency of restored images. But the training process in Iizuka's work is fragile and hard to converge due to the use of raw discriminator without optimization measurements, and their outcomes largely rely on post-processing to eliminate style inconsistencies around the boundaries of filled region.

In order to rule out invalid placeholders, Liu *et al.* [4] renormalizes the traditional convolution into Partial Convolution which features calculation is based on the mask maps. The Partial Convolution only calculates the valid pixels of features, while the validity of pixels is determined by the corresponding binary mask. On account of the perceptual loss are introduced into this work, their model is able to eliminate the color inconsistencies around the borders. Although their result is visually plausible, their structural information in the filled region mismatch the neighborhoods. Yu *et al.* [5] also proposed Gated Convolution and use hand-written sketches to guide the inpainting process.

Kamyar *et al.* [6] proposed two-stage inpainting schemes which firstly restores the damaged edge map and then colors the image with the recovered edge map. However, both Kam-

This research is supported by Sichuan Science and Technology Program (No. 2020YFS0307, No. 2019YFS0146, No. 2019YFS0155), National Natural Science Foundation of China (No. 61907009), Science and Technology Planning Project of Guangdong Province (No. 2019B010150002), Natural Science Foundation of Guangdong Province (No. 2018A030313802).

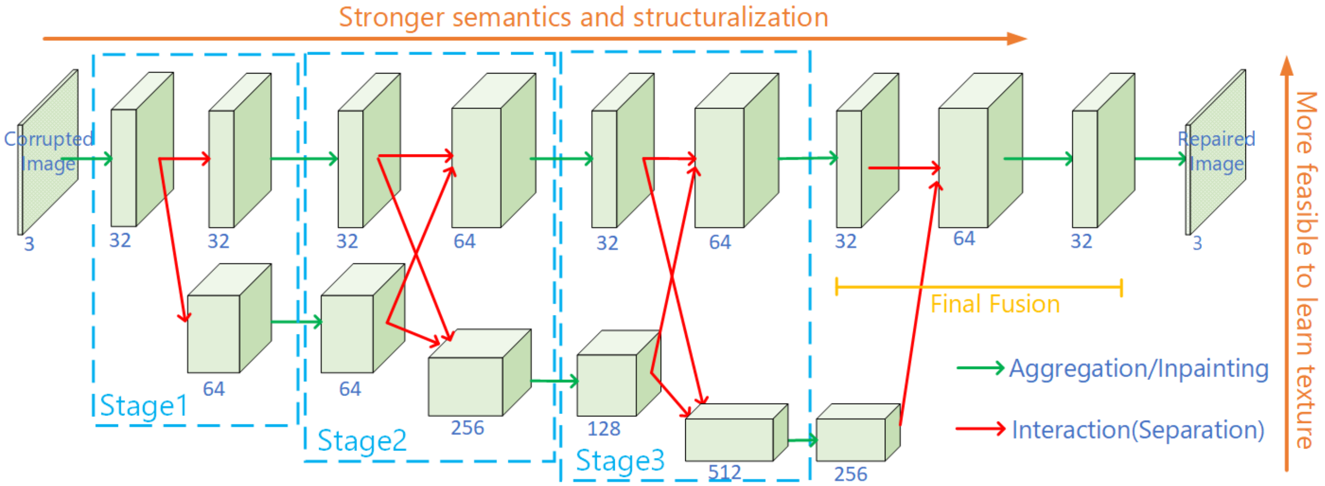

Fig. 1. The architecture of Interactive Separation Network. (The 4th stage and discriminator of ISNet are omitted for clarity.)

yar [6] and Yu [5, 7] exploit a two-stage inpainting strategy, they build a different network in each stage, thus it relies on intensive computation and memory, and their performance of their second network will suffer if the first stage network has poor inpainting prediction. Is there a network efficiently capturing both the structural information and texture information of images?

In other deep learning studies, Lin *et al.* proposed a pyramid network [8] for object detection in different sizes, which proves that the large scale features are more feasible to detect small objects. Sun *et al.* proposed a network with high-resolution representation for gesture estimation [9], which repeatedly conduct multi-scale fusions. The other valuable work [10] demonstrates the feature maps in different scales play different roles, this inspires us in image inpainting tasks to use high-resolution features to capture the high-frequency details and low-resolution features maps for processing low-frequency signals of the image.

This work design a novel inpainting network, named Interactive Separation Network (ISNet), which not only maintains the high-resolution features but also learns the high-level semantics by deep features. In the proposed ISNet, there are three main operations to manipulate the feature representation - Inpainting, Interaction, and Aggregation, which form one whole stage in ISNet, and the step-by-step connection of the stages constitute the main body of ISNet. The short version of the ISNet framework is shown in Figure 1. This paper also studies the efficiency of ISNet and different network settings in ablation experiments, and the integrated architecture of ISNet with GAN is demonstrated that achieve the state-of-the-art inpainting results. The code and pre-trained models can be accessed at https://github.com/GuardSkill/Large-Scale-Feature-Inpainting.

2. ISNET

As shown in Figure 1, ISNet consists of 4 consecutive stages (for clarity, stage 4 is omitted from the diagram). Different from the HRNet [9], it only has two parallel branches to deal with feature maps at two different scale levels. At each new stage, the channel number of smaller resolution features will increase, while the resolution of the smaller resolution features decreases. From the left to right, as the depth of network increases, the semantics of feature becomes stronger. In another aspect, from the bottom up, the feature is more feasible to learn high-resolution information of the image, such as texture, local gradience.

2.1. Inpainting, Interaction and Aggregation

As illustrated in Figure 2, each stage in ISNet has two branches to handle feature maps in different scales and consists of three processes, which named Inpainting, Interaction & Aggregation. After the Aggregation operation of the previous stage, 4 residual blocks [11] are adopted in Inpainting process to extract the high-level representation and repair feature maps in both branches, which reduces the degeneration caused by the network depth increasing. And the reason for the number of residual blocks is set to 4 will be explained in Section 3.1.

In the Interaction process, we adopt one convolution layer with stride 2 to scale down the feature maps in the second branch and n (where n equal to stage index) convolution layers with stride 2 to downsample the feature in the first branch to same targeted resolution, then we concatenate these feature maps as the input of second branch in Aggregation process. Similarly, for generating the features of the first branch, we put the second branch feature into $n-1$ Sub-Pixel [12] convolutional layers to get high-resolution features, which take

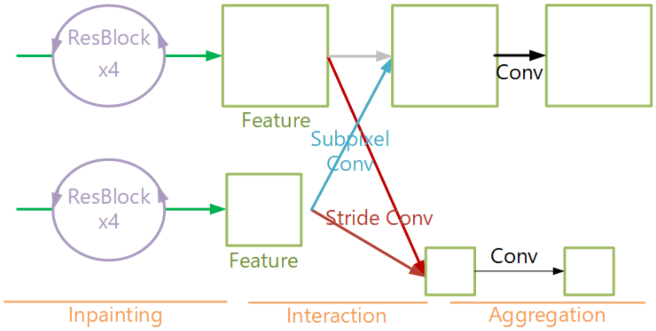

Fig. 2. The detailed design of each stage.

advantage of convolution to learn an adaptive upsample strategy, then concatenate them with the original feature of the first branch to updated feature. In the 1st stage, it is worth noting that the Inpainting and Interaction process only inputs the first branch because there only are larger-scale feature maps.

In the Aggregation operation, we exploit 3x3 convolution to compress the channels of feature to specific numbers, which depends on feature scales. When the resolution is reduced, the ISNet distributes 32, 64, 128, 256, 512 specific channels to different scales.

2.2. Network and Loss Function

At the head of ISNet, the damaged image is embedded into 32-dimensional feature maps with the same resolution, then these features maps are sequentially fed into 4 stages, finally produce two kinds of features with different scales. Although it's feasible to directly output high-resolution feature in the end, we apply a Final Fusion (FF) process which decodes the two kinds of features to 32-dimensional feature maps.

At the end of ISNet, we also build a discriminator model similar to EdgeConnect [6], which facilitates the inpainting result by adversarial learning. In spite of spectral normalization [13] was originally used only in discriminators, Odena [14] has recently shown that spectral normalization can keep generators away from dramatic changes of parameters and gradient. Therefore, we apply spectral normalization [13] to each layer of ISNet.

For the convenience of representation, this paper denotes the function mapping of generator and discriminator in ISNet as G and D respectively. If the $\mathbf{I}_{gt}$, $\mathbf{M}$, $\mathbf{I}_{pred}$ respectively refer to ground truth image, the binary mask that labels corrupted region as 0, the image ISNet predicts, the function of generator can be described as equation 1.

$$\mathbf{I}_{pred} = G\left(\mathbf{I}_{gt} \odot \mathbf{M}\right) \tag{1}$$

Where $\odot$ represents element-wise multiplication. ISNet also exploits the loss funcion similar to those state-of-art work [4–6], the total loss function is formaluted as equation 2.

$$\mathcal{L}_G = \mathcal{L}_{\ell_1} + 0.1(-\mathcal{L}_D) + 0.1\mathcal{L}_{perc} + 250\mathcal{L}_{style} + 10\mathcal{L}_{FM} \tag{2}$$

where $\mathcal{L}_{perc}$ and $\mathcal{L}_{style}$ refer to perceptual loss [15], $\mathcal{L}_{FM}$ is the feature matching loss that represent the difference of the activation maps in the intermediate layers of the discriminator. $\mathcal{L}_{\ell_1}$ is the L1 distance between repaired image and ground truth. $\mathcal{L}_D$ means adverarial loss, this paper adopts the hinge loss (Equation 3) as the objective function of discriminator, which train the discriminator more strictly.

$$\begin{aligned}\mathcal{L}_D = \mathbb{E}_{gt}\left[\psi(1 - D\left(\mathbf{I}_{gt}\right))\right] \\ + \mathbb{E}_{pred}\left[\psi\left(1 + D\left(\mathbf{I}_{pred} \odot (1 - \mathbf{M})\right)\right)\right]\end{aligned} \tag{3}$$

3. EXPERIMENTS

This Section explains some network settings of ISNet through ablation experiments, and evaluates the comparative performance of ISNet on two public datasets - Places2 [16] and CelebA [17]. The binary mask dataset used in all the experiments is made by Liu *et al.* [4].

3.1. Ablation Study

Places2 has more than one million training images and 328,500 test images. In order to confidently demonstrate the design of ISNet, this paper selects the bigger Places2 as the evaluation dataset for the ablation study. Structural similarity index (SSIM) [18] with a window size of 11, peak signal-to-noise ratio (PSNR) and L1 distances are evaluated on the test dataset of Places2 for different models. It's worth to note that all these measurements are based on the computation between composite image $\mathbf{I}_{comp}$ and ground truth image $\mathbf{I}_{gt}$, where $\mathbf{I}_{comp} = \mathbf{I}_{pred} \odot (1 - M) + \mathbf{I}_{gt} \odot M$.

We start with the base model of the missing final fusion (FF) process, and all the upsample layers are traditional bilinear interpolation instead of Sub-Pixel layers. It's noticeable that FF process will increase the learnable weights. To demonstrate the improvement does not only benefits from the increase of weight number, this experiment manually reduces the number of channel in some intermediate layers of base model when adding the FF operation. (Row 2, Table 1) To prove the efficiency of ISNets-like models, we also introduced a UNet-like [19] generator model from [6], and make its magnitude of parameters similar to ISNet via adding additional layers both in encoder and decoder. Table 1 shows the performance of models with different components on the Places2 test dataset. All the models are trained until convergence, and the training environments are constant(eg, random seed, learning rate). It can be concluded that the ISNets-style models have better quantitative performance than UNet-like mode, regardless of the number of parameters.

This paper also explores the impact of the number setting of residual blocks in each stage of ISNet, it records the performance and parameter number of ISNet in different blocks

Table 1. The evaluation of models with different components, here the ISNet refers to the BaseModel+FF+Sub-pixel.

Models	Parameters	PSNR	L1(%)	SSIM
BaseModel	15.75M	27.9095	2.12	0.8800
BaseModel+FF	13.93M	28.2757	2.06	**0.8860**
ISNet	23.63M	**28.5304**	**1.91**	0.8854
UNet-like Net	23.27M	25.9900	2.51	0.8508

Table 2. The performance of ISNet with different numbers of residual blocks setting in each stage.

Blocks	Parameters	PSNR	L1(%)	SSIM
4	23.63M	**28.5304**	**1.91**	**0.8854**
3	22.08M	28.4679	1.97	0.8843
2	20.53M	28.4707	1.97	0.8841
1	18.99M	28.0165	2.05	0.8816

setting (Table 2). In addition to proving effectiveness of residual blocks, it provides a reference for ISNet application that requires different memory.

3.2. Comparison and Discussion

This work also evaluates the ISNet using the metrics involved in Section 3.1 in the different proportions of the corrupted region and compares these results with the state-of-art inpainting approaches. The quantitative results on places2 are shown in Table 3[1] , and the comparison of CelebA dataset is displayed on Table 4[1]. These data show that regardless of the proportion of damaged areas, ISNet can achieve better qualitative (Figure 3) and quantitative performance in all aspects of the indicators. It demonstrates that high-resolution components and deeper features play an important role in addressing the issue of texture inconsistent in image inpainting.

Table 3. The performance of ISNet on Places2 test dataset, these data[1] are taken from literature [6].

Mask Ratio		CA [7]	GLCIC [3]	PConv [4]	EdgeCnt [6]	ISNet
10-20%	PSNR	24.36	23.49	28.02	27.95	**31.41**
	SSIM	0.893	0.862	0.869	0.920	**0.960**
	L1(%)	2.41	2.66	1.14	1.50	**0.63**
20-30%	PSNR	21.19	20.45	24.9	24.92	**26.88**
	SSIM	0.815	0.771	0.777	0.861	**0.912**
	L1(%)	4.23	4.7	1.98	2.59	**1.38**
30-40%	PSNR	19.13	18.5	22.45	22.84	**23.90**
	SSIM	0.739	0.686	0.685	0.799	**0.854**
	L1(%)	6.15	6.78	3.02	3.77	**2.34**
40-50%	PSNR	17.75	17.17	20.86	21.16	**21.34**
	SSIM	0.662	0.603	0.589	0.731	**0.778**
	L1(%)	8.03	8.85	4.11	5.14	**3.74**

4. CONCLUSION

Due to ISNet plays a critical role in the maintenance of high-resolution feature and semantic information in deeper layers, it achieves promising inpainting results. Additionally, the additional residual block can further facilitate inpainting quality in ISNet. However, evidence suggests that the improvement of repaired images in large damaged regions is limited, we can focus on how to enhance the robustness of inpainting model to large damaged region in the future work.

Table 4. The performance of ISNet over CelebA dataset.

Mask Ratio		CA [7]	GLCIC [3]	EdgeCnt [6]	ISNet
10-20%	PSNR	25.32	24.09	33.51	**37.31**
	SSIM	0.888	0.865	0.961	**0.980**
	L1(%)	2.48	2.53	0.76	**0.33**
20-30%	PSNR	22.09	20.71	30.02	**32.13**
	SSIM	0.819	0.773	0.928	**0.954**
	L1(%)	3.98	4.67	1.38	**0.75**
30-40%	PSNR	19.94	18.50	27.39	**28.69**
	SSIM	0.750	0.689	0.890	**0.921**
	L1(%)	5.64	6.95	2.13	**1.3**
40-50%	PSNR	18.41	17.09	25.28	**25.37**
	SSIM	0.678	0.609	0.846	**0.869**
	L1(%)	7.35	9.18	3.03	**2.26**

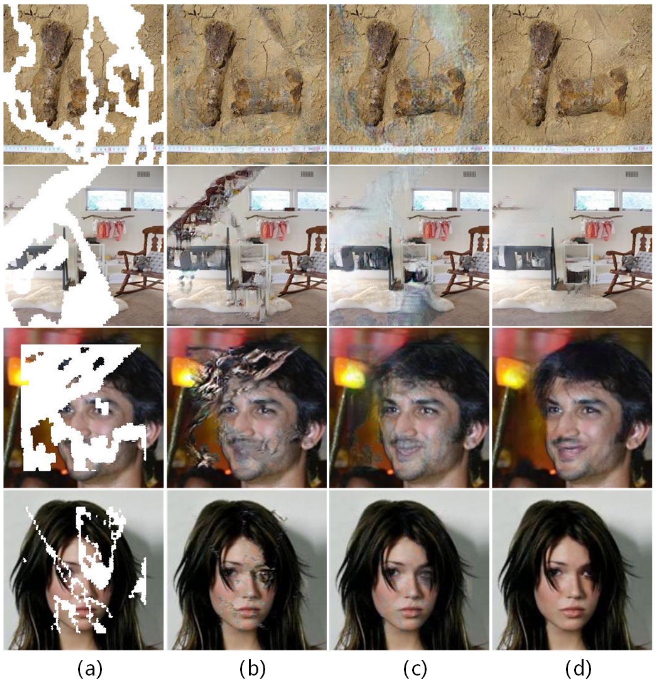

Fig. 3. Comparison of qualitative results over Places2 (first 2 rows) and CelebA (last 2 rows) with official pre-trained models: (a) Damaged Image. (b) CA [7]. (c) EdgeCnt [6]. (d) ISNet.

[1] According to the code of [6], the formulation of L1 measurement in [6] is not standard average Manhattan distance (L1 norm).

5. REFERENCES

[1] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, “Context encoders: Feature learning by inpainting,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[3] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, “Globally and locally consistent image completion,” *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 107, 2017.

[4] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[5] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, “Free-form image inpainting with gated convolution,” *arXiv preprint arXiv:1806.03589*, 2018.

[6] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi, “Edgeconnect: Generative image inpainting with adversarial edge learning,” 2019.

[7] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang, “Generative image inpainting with contextual attention,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[9] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, “Deep high-resolution representation learning for human pose estimation,” *arXiv preprint arXiv:1902.09212*, 2019.

[10] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng, “Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3435–3444.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[13] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, “Spectral normalization for generative adversarial networks,” *arXiv preprint arXiv:1802.05957*, 2018.

[14] Augustus Odena, Jacob Buckman, Catherine Olsson, Tom B Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow, “Is generator conditioning causally related to gan performance?,” *arXiv preprint arXiv:1802.08768*, 2018.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[16] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, “Places: A 10 million image database for scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.