Journal of Image Processing Theory and Applications (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2026.090102
ISSN 2560-6239 Vol. 9 Num. 1

U-Net Handwriting Removal Method and Dataset with ResNetV2 Fusion

Runqing Yan[1], Jianye An[1,*]

[1]Tianjin University of Commerce, Tianjin, China
[*]Corresponding Author

Keywords: ResNetV2 Fusion, U-Net, Handwriting Erase, Removal of Handwritten Text

Abstract: With the growing demand for paperless offices and digital archiving, many paper documents are scanned into image formats for management and dissemination. These images often contain handwritten annotations, signatures, or markings, which interfere with accurate understanding and automatic analysis, especially in educational scenarios where exams and assignments include extensive handwritten content. This highlights the need for effective handwritten text removal techniques. This work proposes an end-to-end handwritten text removal method based on a U-Net enhanced with ResNetV2 modules. The model leverages multi-scale feature extraction, residual learning, and skip connections to remove handwritten marks while preserving printed text and document layout. In addition, a high-quality, large-scale handwritten text removal dataset is constructed and publicly released to provide a standardized benchmark for evaluation and reproducibility. Experimental results show that the proposed approach efficiently removes handwritten traces while maintaining document structure and visual consistency, improving the usability of digital documents. The study contributes to research on handwritten text removal and provides technical support for educational resource digitization, smart learning, and document management.

1. Introduction

With the continuous growth of paperless office practices, digital archiving, and the electronic management of documents, a large number of paper documents are scanned and converted into image formats. These images are then stored in cloud databases for unified management or distributed efficiently through networks, enabling document retrieval, sharing, and long-term preservation. However, document images often contain various types of unstructured information, such as handwritten annotations, signatures, and notes. These handwritten elements not only interfere with the accurate understanding and automated analysis of the original document content but also pose significant challenges for tasks such as information extraction and data mining. Therefore, how to accurately remove handwritten text from document images has become an important research topic with both academic value and practical application potential.

In educational scenarios, the application of handwritten text removal is particularly representative. Exam papers and assignments serve as important carriers of students' learning and teachers' instruction, but they usually contain a large number of handwritten answers, grading marks, and

comments. These handwritten traces can cause interference when the documents are reused. For example, when teachers or students wish to reorganize an exam paper into a “blank version” for review and practice, handwritten marks can significantly affect the user experience. This supports students in building personalized collections of incorrect questions and conducting targeted review, while also facilitating the reuse and sharing of questions by teachers. Such an approach not only improves the efficiency of teaching and learning but also provides strong support for the development of smart education environments.

At present, several companies in the industry have launched application products that integrate handwritten text removal functions (such as CamScanner developed by INTSIG). This indirectly demonstrates that handwritten text removal has clear demand in real-world applications and promising commercial potential. In contrast, systematic research on this topic in academia remains relatively limited. Publicly available datasets are scarce, and reproducible experimental benchmarks as well as well-established lines of research are almost nonexistent, making it difficult to conduct, compare, and reproduce related studies. In the absence of mature research frameworks and standardized data support, the technical challenges of handwritten text removal become even more pronounced.

This task not only requires the complete removal of handwritten traces in document images but also demands the maximum preservation of printed content and layout structures (such as question text, table lines, and legends). Handwritten content is often randomly distributed and highly variable in form, sometimes directly overlapping with printed elements. If handwritten and printed strokes cannot be precisely distinguished at the stroke level, two typical problems may occur: residual handwritten traces that are not fully removed, or the unintended deletion of printed content.

Meanwhile, in scenarios such as exam papers and homework assignments, a single page may contain dozens or even hundreds of densely distributed handwritten regions. The increase in the number of handwritten marks amplifies local detection and generation errors, which can lead to unstable results and visual inconsistencies, ultimately affecting document usability. Therefore, ensuring both the completeness and stability of handwritten text removal under large-scale and multi-target conditions has become one of the key challenges of this task.

2. Related Work

Relevant studies treat the text removal task as a type of Image-to-Image Translation problem. Such methods typically do not include an explicit text detection stage; instead, they directly map the input image from the original image domain to the target image domain. In the text removal task, the original image domain corresponds to input images containing text, while the target image domain corresponds to images in which the text has been removed.

Nakamura[1] et al. proposed a text removal method based on a convolution–deconvolution architecture. In this approach, a four-layer convolutional network first acts as an encoder to extract and represent features from the input image. The encoded features are then passed to a decoder composed of four deconvolution layers to reconstruct an image with the text removed. Unlike methods that process the entire image directly, this approach adopts a sliding window of size 64×64 to process the image region by region. As a result, removing text from a single image requires multiple inference operations. In addition, because the network architecture is relatively simple and shallow, and the training process constrains the generated images using only the L2 loss function, the generated results tend to exhibit blurred textures or artifacts in background regions, which negatively affects the overall visual quality.

In subsequent studies, some researchers introduced Generative Adversarial Networks (GANs) into the text removal task and employed deeper neural network architectures to process the entire image

in an end-to-end manner. Through the adversarial training mechanism between the generator and the discriminator, as well as the stronger feature representation capability of deeper networks, these methods are able to generate richer and more natural background textures, thereby significantly improving the quality of images after text removal.

Isola[2] et al. proposed a general image-to-image translation framework called Pix2Pix, which can also be applied to text removal in images. Pix2Pix is built on the Conditional Generative Adversarial Network (cGAN) framework. It uses a generator $G(x, z) \rightarrow y_G$ to map the input image to the target image domain, while a discriminator $D(x, y_G)$ determines whether the input image pair $\{x, y_G\}$comes from real data or is generated by the model. Here, x, z, y, andy_G denote the original image, random noise, the real target-domain image, and the generated target-domain image, respectively. Through adversarial training between the generator and discriminator, the model progressively improves the quality of the generated outputs, producing images with finer textures and more realistic visual effects.

Inspired by Pix2Pix, Zhang[3] et al. proposed the EnsNet model for text removal in images. The model consists of two components: a generator and a discriminator, responsible for image reconstruction and authenticity discrimination, respectively.

In terms of generator design, EnsNet adopts an encoder–decoder architecture. The encoder is based on ResNet-18[4] to extract image features, while the decoder is composed of multiple deconvolution layers that gradually restore the image content. To effectively fuse information from different hierarchical levels, EnsNet introduces skip connections, which directly pass features from the encoding stage to the decoding stage. This allows high-level semantic information and low-level texture details to be combined effectively. During the decoding process, the model generates multi-scale images through deconvolution operations at different layers and applies pixel-level supervision at multiple scales to improve the detail quality of the generated images.

In addition, EnsNet employs VGG16[5] to extract features from both the generated images and reference images, and constrains them in a high-dimensional feature space, thereby further enhancing the structural consistency and visual quality of the generated results. For the discriminator, EnsNet adopts a Patch-based local discriminator structure, which evaluates the authenticity of corresponding image regions at different locations on the feature map. To encourage the model to focus more on text regions, adversarial loss is applied only within text areas during training.

3. Data Annotation

3.1 Data Collection

During the data collection stage, we first retrieved over 7,000 candidate images from the Rednote platform using the keyword query "{grade} handwritten {subject} exam paper," where "grade" covers both high school and middle school levels, and "subject" includes major disciplines such as mathematics, physics, chemistry, biology, geography, and Chinese. To ensure the representativeness and validity of the data, we established a series of collection criteria:

(1) Image quality and resolution: Exam paper images must have high clarity and resolution (no less than 96 dpi, with both width and height $\geq$ 640 pixels) to facilitate subsequent annotation and model training.

(2) Presence of genuine handwritten marks: Images should contain authentic handwriting, covering key areas such as questions and answers as much as possible.

(3) Diverse image styles: Images should exhibit variation in writing tools, writing density, shooting angles, and lighting conditions.

(4) Authenticity of source: Papers must originate from real scenarios, with any artificially modified or non-natural images excluded.

All collected images were cross-validated by two graduate students, and only those confirmed to meet the criteria by both were retained.

During the data annotation and cleaning stage, we leveraged the API service of the TAL Education Cloud (https://ai.100tal.com/product/hand-writing-clear) to perform automated cleaning. This process takes the original handwritten exam paper as input and outputs a corresponding clean version (Ground Truth image), automatically performing whitening, watermark removal, and other preprocessing steps. Subsequently, we manually refined any remaining handwritten traces using Adobe Photoshop, employing the clone-stamp tool to replace target areas with surrounding background pixels. This yielded visually natural and detail-consistent clean images.

It is noteworthy that no geometric corrections (e.g., image warping) were applied during the annotation and cleaning process, preserving the authentic geometric features of the captured images. This approach allows us to directly construct paired "original image–GT image" relationships, enabling the model to learn an end-to-end mapping from handwritten images to clean exam paper images.

During the data integration stage, we introduced an additional layer of manual review to compensate for possible over-erasure or residual traces from automated cleaning. Specifically, each pair of original and GT images was cross-validated by two graduate students to ensure that the GT image is visually reasonable, printed text remains intact, and background consistency is maintained. Only high-quality samples passing this manual quality check were included in the final dataset, resulting in a total of 4,165 images.

3.2 Data Augmentation

To further enhance the model's generalization ability and its adaptability to structural features at different scales, this section employs a paired data augmentation strategy based on synchronized random cropping to expand the dataset.

Let the original paired training set be

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{1}$$

where $x_i \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$ denotes the noisy input image, and $y_i \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}}$ denotes the corresponding pixel-level clean annotation. Let the multi-scale cropping set be

$$\mathcal{S} = \{512,640,768,896,1024\} \tag{2}$$

(1) Randomly sample the crop size: For each pair (x_i, y_i), sample the crop side length from a uniform distribution over $\mathcal{S}$:

$$s \sim \text{Unifor m}(\mathcal{S}) \tag{3}$$

(2) Randomly sample the crop starting coordinates: For a given crop size s, uniformly sample the top-left corner coordinates within the feasible range of the original image:

$$(u, v) \sim \text{Uniform}(\{0,1,\dots,W-s\} \times \{0,1,\dots,H-s\}) \tag{4}$$

(If the image size is smaller than s, first resize or pad according to a fixed strategy before proceeding with the following steps.)

(3) Define the cropping operator $C_{s,u,v}(\cdot)$: For any image z of the same size as (x_i, y_i),

$$C_{s,u,v}(z) = z[u{:}\,u+s-1, v{:}\,v+s-1] \tag{5}$$

which extracts the top-left subimage of size $s \times s$ starting at (u, v).

(4) Synchronized pixel-level cropping (to ensure alignment): Apply the same cropping operator to both images in the pair:

$$\tilde{x}_i = C_{s,u,v}(x_i), \tilde{y}_i = C_{s,u,v}(y_i) \tag{6}$$

Thus, a new paired sample $(\tilde{x}_i, \tilde{y}_i)$ is obtained while preserving pixel-level alignment.

(5) Randomness and sample distribution: By repeating the above sampling procedure K_i times for each pair (x_i, y_i) (either fixed or adaptive per sample), the augmented sample set is obtained:

$$\tilde{\mathcal{D}} = \left\{\left(\tilde{x}_i^{(j)}, \tilde{y}_i^{(j)}\right)\right\}_{i=1,j=1}^{N,K_i} \tag{7}$$

with a total number of samples

$$N_{\text{aug}} = \sum_{i=1}^{N} K_i \tag{8}$$

This strategy can significantly increase the number of training samples without changing the data semantics, thereby enhancing the model's ability to learn local texture details, edge structures, and overall layout. Based on the original dataset, a total of 12,377 augmented samples were generated.

3.3 Data Comparison

Currently, the academic community still lacks large-scale, high-quality public datasets for handwritten exam mark removal. At present, only two representative resources are available. The first is the SCUT-EnsExam[6] dataset developed by the Jin Lianwen team at South China University of Technology. The second is the Handwritten Text Erasure Task (Task 2) from the 2022 AI competition jointly organized by Baidu Wangpan and Paddle (Baidu Wangpan AI Competition: Handwritten Text Erasure (Task 2), Paddle AI Studio Galaxy Community).

Compared with conventional datasets, the annotation method proposed in this work offers notable advantages in efficiency and data consistency. The SCUT-EnsExam dataset relies primarily on manual processing: researchers used Adobe Photoshop to erase handwritten marks individually, replacing the target region with surrounding background pixels via the clone stamp tool to produce visually plausible results. However, this approach generated only 545 images at resolutions around 1700×2500. For handwriting adjacent to printed text, extremely high precision is required to preserve the integrity of the underlying printed content. Consequently, this method is labor-intensive and prone to subjective variation.

In contrast, the dataset provided in the Baidu Wangpan AI Competition was synthetically constructed rather than collected from real-world scenarios. Clean document images were selected as backgrounds, and manually collected handwritten marks were overlaid to produce "source images," while the original clean documents served as the corresponding ground-truth (GT) images. This procedure enables the rapid generation of paired "handwritten–clean" samples with perfectly aligned layout and printed text. Nevertheless, synthetic datasets present inherent limitations. The realism of handwriting-background integration depends on the synthesis algorithm and cannot fully reproduce stroke dynamics, pressure variation, bleed-through, or lighting effects found in authentic writing. Furthermore, the distribution, color diversity, and overlap of handwritten marks with printed text differ from those in actual exam papers. As a result, although synthetic datasets are useful for model pretraining, their generalization to real-world applications is constrained.

The automated cleaning plus manual verification approach proposed herein substantially reduces human labor while ensuring consistent GT image quality. Additionally, preprocessing steps such as paper bleaching, watermark removal, and moiré pattern elimination further enhance the suitability of the data for end-to-end model training.

4. Model

4.1 ResNetV2-UNet

Under the constraint of no handwritten mask supervision, we formulate the handwritten mark removal task as an end-to-end image-to-image mapping problem, i.e., directly learning the mapping from the observed image $x$to the clean image y:

$$\hat{y} = G_\theta(x), G_\theta: x \mapsto \mathcal{Y} \tag{9}$$

where $x \in X \subset \mathbb{R}^{H\times W\times C}$ denotes the observed image tensor with height H, width W, and Cchannels. In the observed image, handwritten marks act as local perturbations superimposed on the original content. The clean image $y \in \mathcal{Y} \subset \mathbb{R}^{H\times W\times C}$has the same spatial dimensions and number of channels as the input. The objective is to recover the original structure and semantic information while removing the handwritten interference. $\hat{y} \in \mathcal{Y}$denotes the predicted output of the generator network G_θ, which can adopt architectures such as U-Net or the generator of a conditional generative adversarial network (cGAN). The network produces the restored image on a pixel-wise basis using convolution, downsampling, upsampling, and skip connections.

In this study, we adopt a U-Net with integrated ResNetV2[7] modules as the baseline model. The U-Net[8] is a classical image reconstruction architecture that extracts multi-scale features and restores spatial resolution through symmetric downsampling and upsampling paths, The network structure is shown in Figure 1. To enhance feature representation and improve the training stability of deep networks, we incorporate ResNetV2 modules within each convolutional block. ResNetV2 employs a pre-activation residual structure, where batch normalization (BN) and nonlinear activation precede the convolution operation. Residual learning is performed via an identity shortcut as follows:

$$\mathrm{y}_l = \mathcal{F}(\mathrm{x}_l, \mathcal{W}_l) + \mathrm{x}_l \tag{10}$$

where $\mathrm{x}_l$and $\mathrm{y}_l$are the input and output of the l-th residual block, $\mathcal{F}(\cdot)$represents the residual mapping implemented by the sequence of pre-activation BN, activation, and convolution layers, and $\mathcal{W}_l$denotes the weights of the block. By integrating ResNetV2 into U-Net, the model can stably propagate gradients, extract richer hierarchical features, and generate pixel-wise restored images with improved fidelity.

where $f_l \in \mathbb{R}^{H_l\times W_l\times C_l}$denotes the feature map at the l-th layer, and $\mathcal{F}_{\mathrm{ResV2},l}(\cdot)$represents the residual mapping function of the l-th ResNetV2 block. The pre-activation design improves gradient flow, stabilizing the training of deep networks while preserving fine-grained features from shallow layers, which is beneficial for capturing local textures and edge information of handwritten marks.

The encoder extracts multi-scale semantic features while reducing spatial resolution by sequentially applying ResNetV2 convolutional blocks $\mathcal{R}_l(\cdot)$ followed by strided convolution downsampling operations $\mathcal{P}_l(\cdot)$:

$$f_l = \mathcal{P}_l\big(\mathcal{R}_l(f_{l-1})\big), l = 1, \dots, L \tag{11}$$

In the decoder, transposed convolution operations $U_l(\cdot)$are used to upsample the features and restore spatial resolution. Features from the corresponding encoder layers are fused via skip connections $\oplus$, enabling joint reconstruction of local details and global semantic information:

$$g_l = \mathcal{R}_l(U_l(g_{l+1}) \oplus f_l), l = L, \dots, 1 \tag{12}$$

Finally, the restored prediction is obtained by a convolutional mapping $\mathcal{C}_{\mathrm{out}}$:

$$\hat{y} = \mathcal{C}_{\mathrm{out}}(g_1)$$

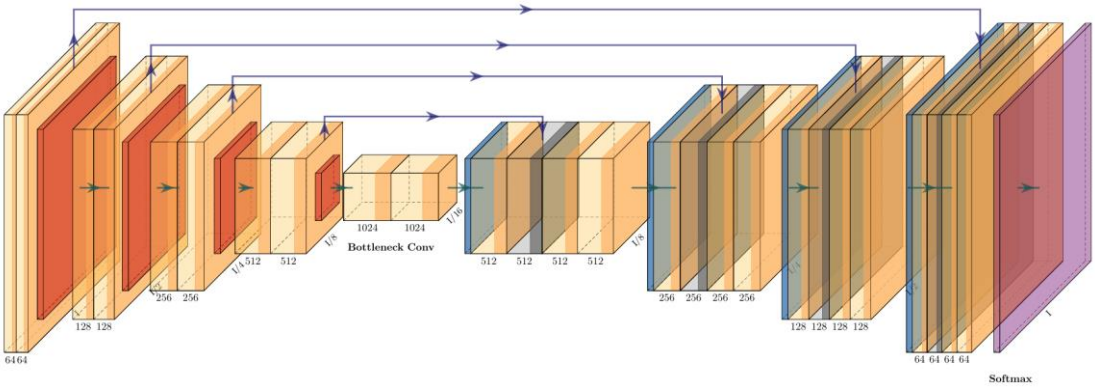

Figure 1. UNet model

To guide the model training, a composite loss function is employed to optimize the network. This loss function combines the L1 loss (mean absolute error) and MSE loss (mean squared error) with corresponding weights. Its mathematical formulation is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{L1}} + \lambda_2 \cdot \mathcal{L}_{\text{MSE}} \tag{13}$$

where

$$\mathcal{L}_{\text{L1}} = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |, \mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{14}$$

The L1 loss measures the absolute difference between predictions and ground truth, offering strong robustness to outliers and helping to preserve fine image details. The MSE loss emphasizes structural consistency, promoting improvement in the overall image quality.

4.2 Experiments

Based on the 12,377 training samples obtained through data augmentation in Section 3.2, the model was first pretrained to sufficiently initialize network parameters. During the pretraining phase, the network was trained for 10 batches to allow it to preliminarily learn local textures and structural features of the images.

Subsequently, formal training was conducted on a dataset containing 4,165 samples for 20 batches, using the U-Net model for comparative experiments. To ensure the scientific validity of the training and evaluation process, the dataset was split into training and test sets at a 9:1 ratio.

A staged training strategy was employed to improve the model's adaptability to structural features at different scales. In the first 5 batches, images of size 640 × 640 were used to stabilize the learning of local features. For the subsequent 15 batches, images of size 1024 × 1024 were used to enhance the model's ability to capture large-scale structures and global textures. All experiments were conducted on an NVIDIA V100 16GB GPU, with the software environment comprising PyTorch 2.3.1 and CUDA 11.8. Training parameters were set as follows: batch size of 8, learning rate of 1×10^{-4}, and AdamW optimizer, with an initial image size of 640 × 640, The loss behavior is shown in Figure 2.

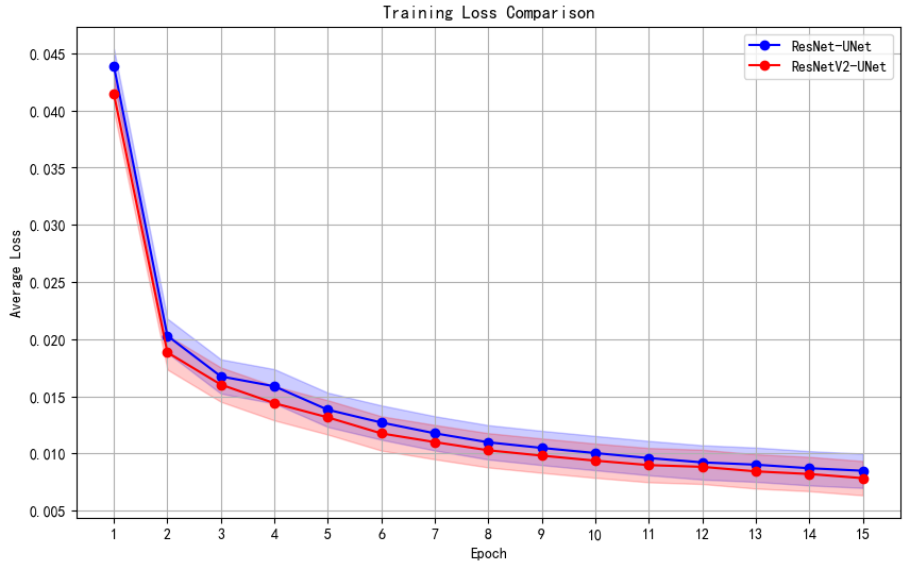

Figure 2. Training loss of models

5. Conclusion

This paper addresses the problem of handwritten marks interfering with the understanding of printed content in digital documents and proposes an end-to-end handwritten text removal method based on a U-Net enhanced with ResNetV2 modules. By leveraging multi-scale feature extraction, residual learning, and skip connections, the model effectively removes handwritten annotations while preserving printed text and document layout. In addition, a high-quality, large-scale handwritten text removal dataset is constructed and publicly released, providing a standardized benchmark for evaluation and reproducibility. Experimental results demonstrate that the proposed method efficiently removes handwritten traces while maintaining document structure and visual consistency, thereby improving the usability of digital documents. This study offers new insights for the development of handwritten text removal techniques and provides technical support for educational resource digitization, smart learning, and document management applications.

References

[1] Nakamura T, Zhu A, Yanai K, Uchida S. Scene Text Eraser[C]. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017: 832-837. doi: 10.1109/ICDAR.2017.141.

[2] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[3] Zhang S, Liu Y, Jin L, Huang Y, Lai S. EnsNet: Ensconce Text in the Wild[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019; 33(01): 801-808. doi: 10.1609/aaai.v33i01.3301801.

[4] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016: 770-778. doi: 10.1109/CVPR.2016.90.

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[Preprint]. arXiv:1409.1556, 2014.

[6] Huang L, Fink GA, Jain R, Kise K, Zanibbi R. EnsExam: A Dataset for Handwritten Text Erasure on Examination Papers[C]. In: Fink GA, Jain R, Kise K, Zanibbi R, eds. Document Analysis and Recognition - ICDAR 2023. Lecture Notes in Computer Science, vol 14189. Cham: Springer; 2023. doi: 10.1007/978-3-031-41682-8_29.

[7] He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks[C]. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol 9908. Cham: Springer; 2016. doi: 10.1007/978-3-319-46493-0_38.

[8] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Cham: Springer; 2015. doi: 10.1007/978-3-662-54345-0_3.