2024 Visual Media Final Report
Choice 2

37-235079 CAO Wei

Paper 1: X. Wu, D. Hong and J. Chanussot, "UIU-Net: U-Net in U-Net for Infrared Small Object Detection," in IEEE Transactions on Image Processing, vol. 32, pp. 364-376, 2023, doi: 10.1109/TIP.2022.3228497.

Summary and Reflections:

Object recognition and classification are important topics in the field of image processing. Various methods can be used to acquire images, such as active imaging methods like Synthetic Aperture Radar, or passive methods like infrared imaging and various optical sensors. A current popular trend is using machine learning and neural networks for object recognition in images. Among these, the U-Net architecture, a type of Convolutional Neural Network, has become a classic in the image processing field due to its efficient use of multi-scale features and excellent performance. It is widely used in medical image analysis and other fields that require precise segmentation. Additionally, I have noticed that the U-Net structure is flexible and diverse, allowing for the addition or modification of modules to create new, high-performance architectures that better suit different image recognition and segmentation tasks.

In various image detection applications, infrared image target detection is widely used in daily security and military fields. This paper aims to detect small targets in infrared images more accurately. It points out that existing learning-based infrared small target detection methods mainly rely on classification backbone networks, which may lose small targets as network depth increases and limit feature distinguishability. Moreover, small targets in infrared images often exhibit varying brightness, posing a challenge to obtaining accurate target contrast information. Therefore, this paper proposes a novel framework to achieve better performance.

Firstly, it innovatively proposes a "U-Net in U-Net" framework, abbreviated as UIU-Net, by embedding a smaller U-Net within a larger U-Net framework to achieve multi-level and multi-scale representation learning. The UIU-Net model consists of two modules: the Resolution Maintenance Deep Supervision (RM-DS) module and the Interaction-Cross Attention (IC-A) module. RM-DS integrates residual U-blocks into the deep supervision network to generate deep multi-scale resolution-maintaining features while learning global contextual information. IC-A encodes local contextual information between low-level details and high-level semantic features. Its structure is shown in Figure 1.

Secondly, experimental results show that UIU-Net outperforms several state-of-the-art infrared small target detection methods on two infrared single-frame image datasets (SIRST and synthetic datasets). Furthermore, UIU-Net demonstrates strong generalization performance on video sequence infrared small target datasets (e.g., ATR ground/air video sequence datasets).

Therefore, considering the practical significance of infrared small target detection, I believe the innovative framework and excellent performance are the reasons this paper was accepted.

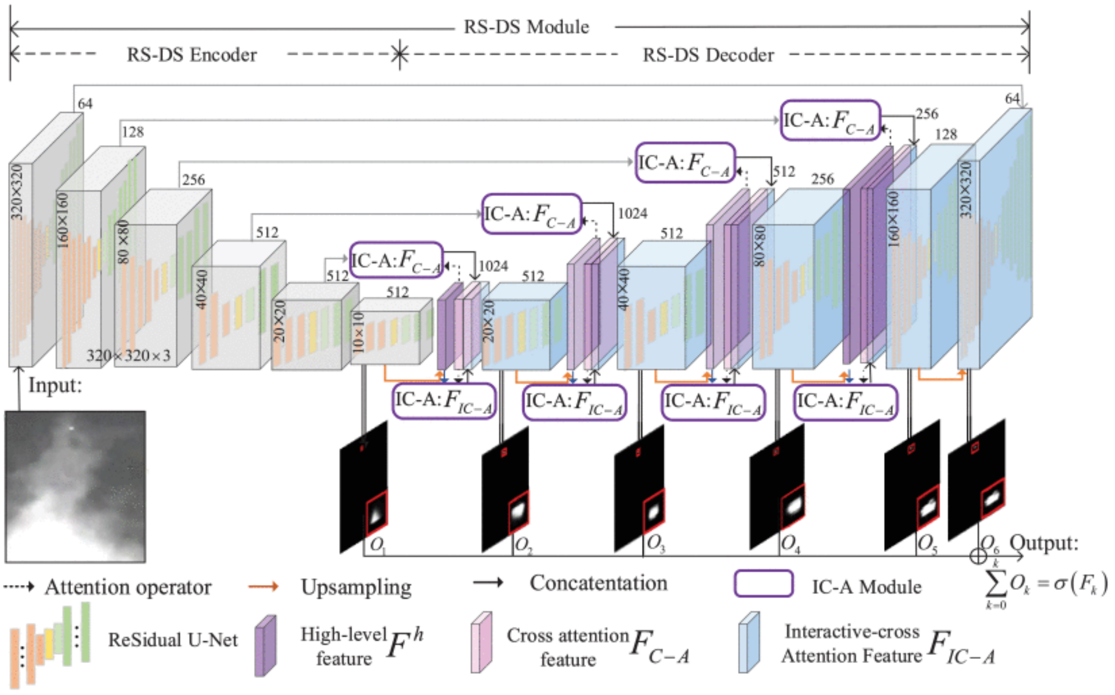

Fig.1 Overview of the proposed UIU-Net for infrared small object detection.

It consists of two main modules: the resolution-maintenance deep supervision (RM-DS) module and the interactive-cross attention (IC-A) module. RM-DS improves global context representations by learning deep multi-scale features. IC-A encodes RM-DS features to further enhance local context representations.

Paper 2: C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet and M. Norouzi, "Image Super-Resolution via Iterative Refinement," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 4713-4726, 1 April 2023, doi: 10.1109/TPAMI.2022.3204461.

Summary and Reflections:

Like the previous paper, this article also utilizes the U-Net framework, with the goal of improving the accuracy and effectiveness of image processing. Specifically, this paper aims to achieve image super-resolution. Image super-resolution refers to the technique of generating high-resolution images from low-resolution ones, with the goal of enhancing image clarity and detail, resulting in high-resolution images with superior visual quality. Image super-resolution has important applications in various fields, including medical imaging, satellite image processing, video surveillance, and image enhancement.

In previous research, deep generative models have succeeded in learning complex empirical distributions of images. However, under these methods, autoregressive models are too costly for high-resolution image generation, normalizing flows (NF) and variational autoencoders (VAE) often produce suboptimal sample quality, and generative adversarial networks (GAN) require carefully designed regularization and optimization techniques to control instability and mode collapse. Therefore, this paper proposes a method for image super-resolution through iterative refinement, called SR3. The working principle of SR3 is to learn to convert a standard normal distribution into an empirical data distribution through a series of refinement steps.

SR3 employs denoising diffusion probabilistic models (DDPM) and applies them to image-to-image translation, achieving super-resolution through a stochastic iterative denoising process. The output images are initialized with random Gaussian noise and iteratively refined using a U-Net architecture trained on denoising at various noise levels. In the Figure 2 SR3 architecture, this paper uses residual blocks from BigGAN to replace the original DDPM residual blocks, rescales skip connections by $\frac{1}{\sqrt{2}}$, and increases the number of residual blocks and channel multipliers at different resolutions. The model inputs noisy high-resolution images and low-resolution conditional images that have been interpolated and up sampled to the target resolution.

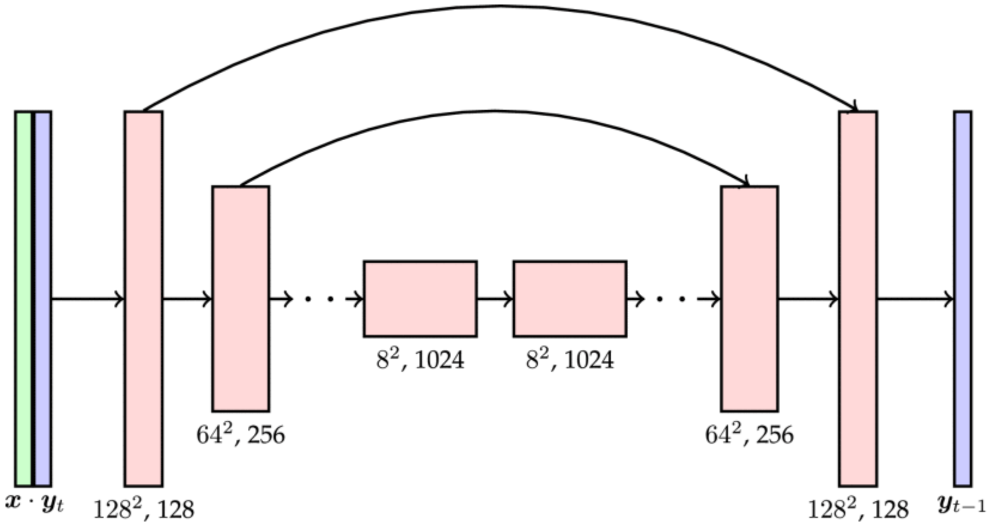

Fig.2 Depiction of U-Net architecture of SR3.

The low-resolution input image x is up-sampled to the target resolution using bicubic interpolation, and concatenated with the noisy high resolution output image y_t. We show the activation dimensions for a 16×16 → 128×128 super resolution model. We perform self-attention on 16×16 feature maps.

The results show that SR3 performs excellently on super-resolution tasks for faces and natural images at different magnification factors. In human evaluation on the standard 8× face super-resolution task (CelebA-HQ dataset), SR3 achieves a "fool rate" close to 50%, indicating photo-realistic outputs, while GAN-based baselines do not exceed a "fool rate" of 34%. In the 4× super-resolution task on the ImageNet dataset, SR3 outperforms baseline methods in both human evaluation and classification accuracy of high-resolution images by a ResNet-50 classifier trained on these images. Additionally, SR3 shows excellent performance in cascaded image generation, combining a generative model with a super-resolution model to achieve competitive FID scores in the class-conditional 256×256 ImageNet generation challenge.

In conclusion, I believe the value of this paper lies in its outstanding innovation, practical application value, performance advantages, and technical details and theoretical contributions. I think these are the reasons for its acceptance.