Unsupervised Crack Detection in Large Stamped Metal Products Via Spatial Transformer and U2-Net with Masked Attention and Patch Positional Encoding

PENGHUA ZHANG, YESEUL KONG and GYUHAE PARK

ABSTRACT

Crack detection is crucial for the quality assessment of metal products. While supervised learning methods are commonly employed to automate this process, their performance is constrained by the limited availability of large and diverse crack datasets. Unsupervised learning has shown exceptional performance in anomaly detection by relying solely on healthy data during training. For large products, cracks are small in size compared to the entire object, making image patches a more suitable representation. However, the large diversity of image patches introduces significant challenges for unsupervised learning methods. This study proposes a spatial transformer network utilizing binary masks from Segment Anything Model to reduce the complexity of image patches. To perform crack detection, we propose a U2Net- based model with neighbor masked attention at different scales to learn the distribution of healthy data from discrete representations extracted by Vector Quantized Variational Autoencoder (VQ-VAE). Additionally, patch positional encoding is incorporated to enhance the model's ability to match patches and the distribution. Features that deviate from the learned distribution are identified as cracks. We resemble a real manufacturing setting and capture images from large stamped metal panels. Comprehensive experiments are conducted, and results indicate the effectiveness of the proposed method and individual modules.

INTRODUCTION

Stamping is a widely used manufacturing process that transforms metal sheets into various products. When producing products with complex geometries, cracks occasionally occur due to the characteristics of stamping machines[1]. The presence of cracks can result in significant issues, such as safety and structural integrity[2]. However, inspections are currently carried out by human inspectors, which cannot ensure consistent reliability and efficiency over time. Therefore, it is necessary to develop an automated crack detection system.

Supervised deep learning approaches[3–7] have been extensively employed to inspect metal surfaces. However, these approaches require a large number of labeled

Penghua Zhang, Yeseul Kong, Gyuhae Park, School of Mechanical Engineering, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju, South Korea

and balanced data to achieve reliable performance. Moreover, these models struggle to detect unseen types of cracks, posing a significant challenge for generalization.

Unsupervised deep learning methods, which rely solely on healthy data during training, are commonly used in anomaly detection scenarios where the number of anomalous samples is significantly smaller than that of healthy ones. These methods can be divided into two categories: feature-based and reconstruction-based approaches. The feature-based methods use pretrained encoders to extract features from healthy data and detect anomalies by identifying deviations, for example, a teacher-student framework[8] compares output from a generalized teacher network and a distilled student network, while another approach[9] measures the similarity between test feature vectors and those vectors stored in a memory bank during training. However, these approaches are computationally expensive, particularly when applied to large datasets such as our case. When inspecting large stamped products, cracks are relatively small, so high-resolution images are captured and cropped into smaller patches to facilitate the detection. The resulting large number of patches imposes a huge demand for computational resources. Moreover, crack-like features such as holes and shadows further complicate detection using these methods. Reconstruction-based approaches[10–12] assume that models trained only on healthy data will fail to reconstruct anomalous patterns. These methods typically adopt an encoder-decoder architecture to reconstruct the original input and identify anomalies based on reconstruction errors. Despite their effectiveness in some situations, these methods face limitations in crack detection on large stamped panels. Cracks are sometimes finely reconstructed, leading to missing detections and all pixels are reconstructed based on an input image, resulting in crack indications outside the actual crack area.

To address the above issues, we proposed an unsupervised deep learning framework tailored to detect cracks in large stamped products. First, to reduce the complexity of image patches, we employed a Spatial Transformer Network (STN)[13] to align the product positions in images. We used masks of the target object as input, which can avoid the influence of surface textures and varying light conditions. Second, we optimized a Vector Quantized Variational Autoencoder (VQVAE)[14] by incorporating an additional perceptual loss to obtain reconstructions with high fidelity. Cracks were allowed to be reconstructed. Third, we proposed a segmentation model with neighbor-masked attention to estimate the distribution of the extracted features from healthy data and resampled those that deviated from the distribution during test. Instead of comparing the reconstruction with the original input, we compared it with the decoded image obtained after resampling to generate a cleaner crack indication.

METHODOLOGY

Image Alignment

Product positions may vary slightly in the production line. Cropping the original high-resolution images into smaller patches can result in an excessive number of patch variations. To reduce the complexity of these patches, we employed a Spatial Transformer Network (STN) to align product positions in images as shown in Figure 1. Surface textures and varying light conditions present challenges for reliable image alignment. To avoid these influences, we employed Segment Anything Model (SAM)[15] to extract masks of products and use these masks as input to STN. However, the generalized SAM often fails to segment the products precisely, particularly in

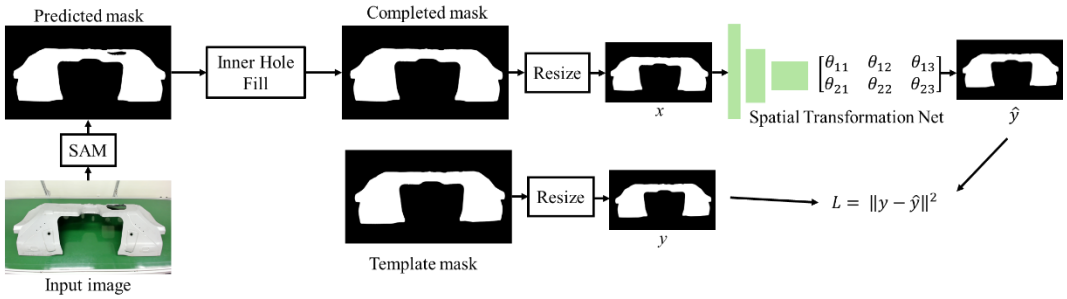

Figure 1 Image alignment using a spatial transformer network

handling inner features. Therefore, we refined the predicted masks by filling in inner features before inputting them to the STN. To further reduce segmentation bias across different samples, we downsampled the masks. The STN was then trained to learn an affine transformation matrix that aligns each mask to a selected reference template. The network was optimized by minimizing the difference between the downsampled template and the aligned mask produced after applying the transformation matrix. The loss function is defined as:

$$L = \|y - \hat{y}\|^2 \tag{1}$$

where y is the downsampled template mask and $\hat{y}$ is the aligned mask

The small mask shares the same spatial transformation as the original input image. Therefore, the original input image can be aligned by the same transformation matrix of the mask.

Image Reconstruction and Distribution Estimation

The mainstream reconstruction-based methods aim to suppress crack patterns and identify cracks by the reconstruction errors. In contrast, our approach allowed the model to reconstruct cracks explicitly. We utilized a segmentation network to model the distribution of healthy data and resample features that deviate from this distribution in the latent feature space. Cracks were then detected by comparing the original reconstruction and the decoded image after resampling. Compared to most of the reconstruction-based methods, our approach produced a cleaner and more precise crack indication map.

We employed the VQ-VAE architecture for image reconstruction, as illustrated in Figure 2. Image patches were first passed through an encoder to obtain latent features, denote as z_e.Then a codebook was used to quantize these latent features into discrete representations by selecting vectors in the codebook with the smallest Euclidean distance as new representations. However, the original model often generates blurred reconstructions with small cracks missing, especially when handling high-resolution inputs. To improve the reconstruction quality and reduce the redundancy in codebook vectors, VQGAN[16] incorporated a discriminator and perceptual loss. The discriminator was employed to distinguish between real and reconstructed images, encouraging the model to generate outputs that appear more realistic. The perceptual loss was leveraged to optimize the distance between extracted features from the input and those from the reconstruction. However, in our experiments, we observed that the

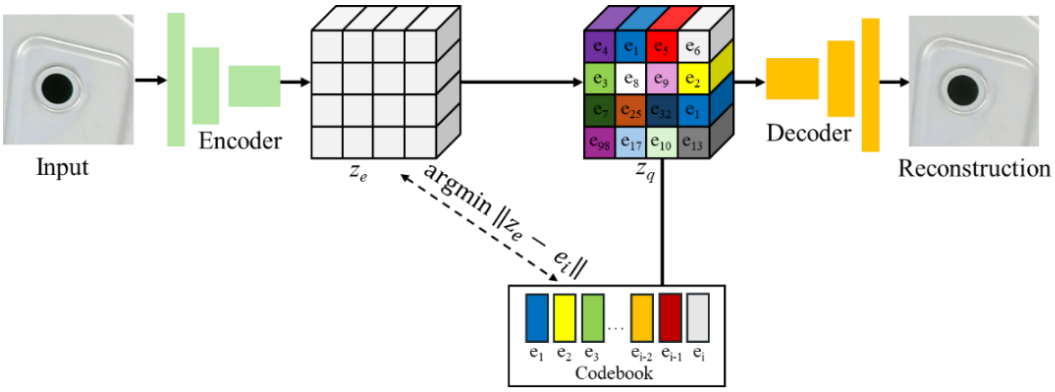

Figure 2 Image reconstruction

discriminator introduced undesirable effects such as spatial shift in patterns, pattern diversity, and failures in reconstructing cracks. Therefore, we trained the VQ-VAE using the following loss function:

$$L = \|x - \hat{x}\|_2^2 + \alpha\|sg[e] - z_e\|_2^2 + \beta L_{percptual}(f(x), f(\hat{x})) \quad (2)$$

where x is the input, $\hat{x}$ is the reconstruction, sg[] denotes the stop-gradient operator that passes zero gradient during the backpropagation, e represents vectors in the codebook, z_e denotes encoder-extracted features. The function f represents layers of the model used for the perceptual loss. Coefficients α and β were employed to balance the contributions of each term. The codebook was updated by the exponential moving average[17].

After obtaining discrete representations from the encoder, we modeled their distribution using maximum likelihood estimation, depicted as follows:

$$p(c|z_q) \quad (3)$$

where the $c \in \{1, 2, \ldots, n\}$ is the index of vectors in the codebook.

The model was designed to output, at each spatial position j, a probability distribution over all codebook entries. Rather than using autoregressive models that use preceding information to determine the current feature, we proposed an assumption that all pixels can contribute to the probability of the current pixel:

$$p(c_j) = p(c_j|z_{q1}, z_{q2}, \ldots, z_{qj}, \ldots, z_{qk}) \quad (4)$$

where $k = H \times W$ is the total number of spatial locations in the quantized latent map.

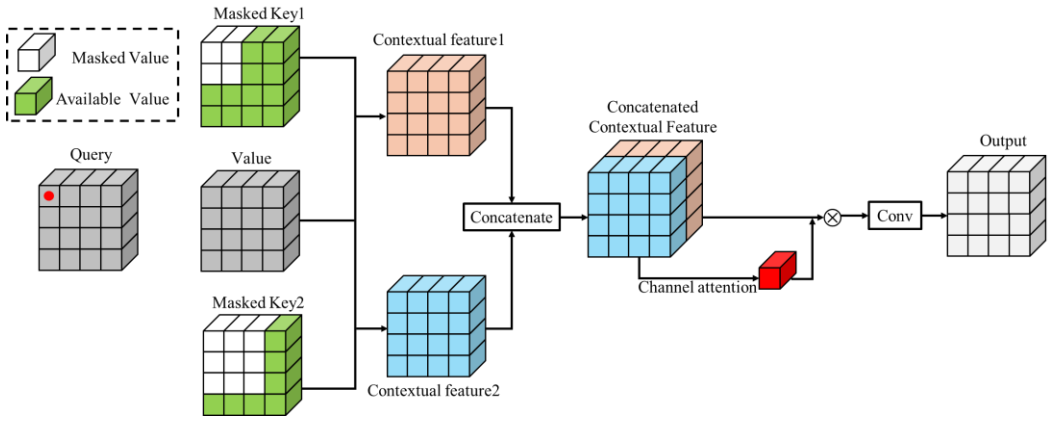

Figure 3 Multi-scale neighbor-masked attention

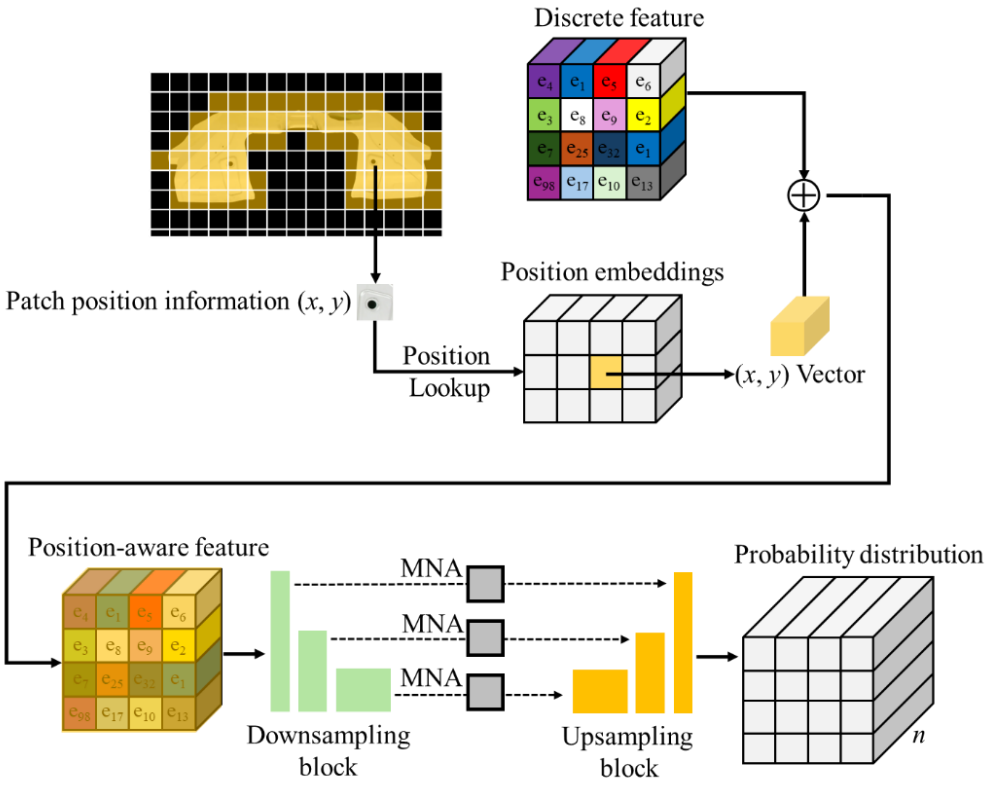

Figure 4 Distribution estimation model

However, since the input z_q is discrete, the model can easily match the input and its corresponding index. If the input and index are both accessible, the model may trivially learn a mapping function that assigns a high probability to the matching entry even if it represents a crack:

$$p\left(c_j = k \middle| z_{qj} = e_k\right) \approx 1 \tag{5}$$

To address this issue, we proposed a multi-scale neighbor-masked attention (MNA) to prevent the model from accessing the content of the current pixel and its neighboring regions are also masked because they share big similarities. As illustrated in Figure 3, Contextual features after applying the mask mechanism in different scales were concatenated and passed through a channel attention module to obtain different weights for each channel based on their importance.

To further enhance the distribution estimation, patch positional information was added to the extracted discrete representation as shown in Figure 4. U2net[18], a powerful segmentation model, was employed as the backbone of our framework. The proposed MNA module was inserted into skip connection layers between the downsampling and upsampling process.

Crack detection

During testing, crack candidates in the latent space were identified by analyzing whether features deviate from the learned distribution of healthy data. These candidates were resampled by the most dissimilar vectors in the codebook. By passing the original features and the resampled features to the same decoder, we obtained the original reconstruction and the resampled reconstruction. Since only several specific vectors

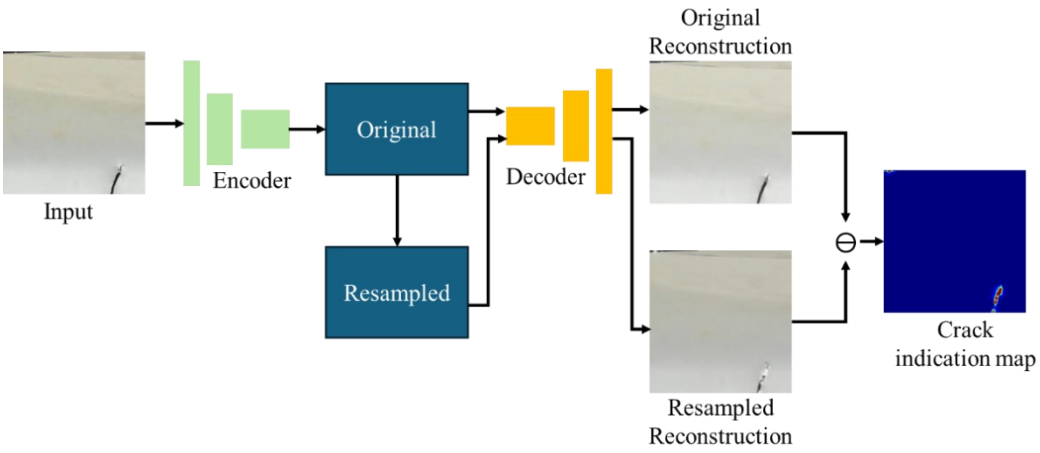

Figure 5 Crack detection

were altered during resampling, a clean crack indication map could be obtained by comparing the two reconstructions as illustrated in Figure 5.

EXPERIMENT

Real stamped products were used to evaluate the effectiveness of the proposed method. Images were captured at a resolution of 3×3840×2160. These high-resolution images were then cropped into 256×256 patches. Crack detection performance was compared with our previous work[2]. We used Area under the Receiver Operating Characteristic Curve (AUROC) and Precision-Recall Area under the Curve (PR-AUC) to evaluate classification performance and Intersection over Union (IoU) to assess localization accuracy. Table 1 presents the crack detection performance comparison between our proposed method and the previous approach. The proposed method achieved substantial improvements across all evaluation metrics. Specifically, the proposed framework achieved an AUROC of 95.25%, significantly higher than the 87.99%. In terms of PR-AUC, which is more appropriate under class imbalance, our method reached 75.30%, showing a notable improvement over the previous result of 47.44%. For the localization performance, our approach yielded 28.28%, which is much higher than that of the previous work. As illustrated in Figure 6, our proposed methods achieved more accurate crack detection while suppressing false positives.

Table 1 Crack detection performance comparison

Model name	Image AUROC	Image PR-AUC	IoU
VQ-VAE2-PixelSNAIL[2]	87.99%	47.44%	9.42%
Ous	**93.67%%**	**78.06%**	**26.64%**

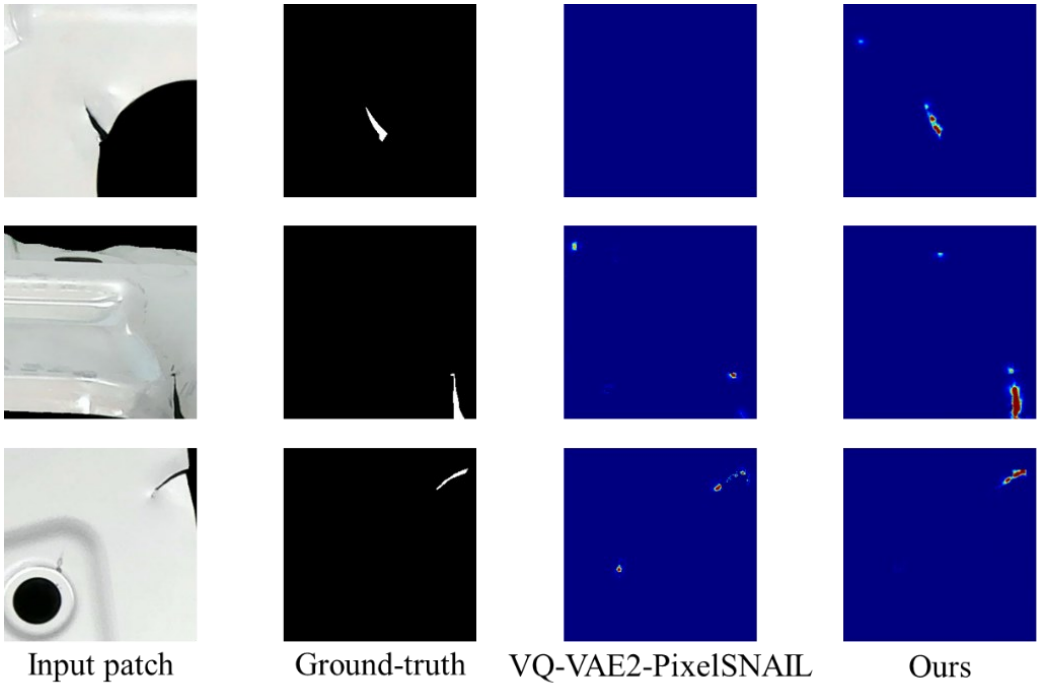

Figure 6 Detection result comparison

CONCLUSION

In this study, we proposed an unsupervised deep learning framework to detect cracks in large stamped products. We employed a spatial transformer network to reduce the complexity of cropped image patches. To improve reconstruction quality, we incorporated the perceptual loss to VQ-VAE. In addition, we proposed the multi-scale neighbor-masked attention module to prevent the model collapse and integrate it with U2Net as the backbone to explicitly estimate the distribution of healthy data. Cracks were identified by analyzing whether features deviate from the learned distribution. Through experiments and comparison with the other model, it was shown that the proposed framework achieved high crack detection performance.

ACKNOWLEDGEMENT

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through the Agriculture and Food Convergence Technologies Program for Research Manpower development, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(project no. RS-2024-00397026). This work was also supported by the Korea Planning & Evaluation Institute of Industrial Technology(KEIT) grant funded by the Korea government (RS-2024-0050-7451)

REFERENCES

1. Dong X, Zhang C, Wang J, et al (2024) Real-time detection of surface cracking defects for large-sized stamped parts. Comput Ind 159–160:104105. https://doi.org/10.1016/j.compind.2024.104105

2. Zhang P, Ryu H, Miao Y, et al (2024) Robust unsupervised-learning based crack detection for stamped metal products. J Manuf Syst 73:65–74. https://doi.org/10.1016/j.jmsy.2024.01.003
3. LSKA-YOLOv8——a lightweight steel surface defect detection algorithm based on YOLOv8 improvement(2024).pdf
4. Yun JP, Shin WC, Koo G, et al (2020) Automated defect inspection system for metal surfaces based on deep learning and data augmentation. J Manuf Syst 55:317–324. https://doi.org/10.1016/j.jmsy.2020.03.009
5. Yun JP, Shin WC, Koo G, et al (2020) Automated defect inspection system for metal surfaces based on deep learning and data augmentation. J Manuf Syst 55:317–324. https://doi.org/10.1016/j.jmsy.2020.03.009
6. Liu T, He Z, Lin Z, et al (2024) An Adaptive Image Segmentation Network for Surface Defect Detection. IEEE Trans Neural Netw Learn Syst 35:8510–8523. https://doi.org/10.1109/TNNLS.2022.3230426
7. Li Z, Wei X, Hassaballah M, et al (2024) A deep learning model for steel surface defect detection. Complex and Intelligent Systems 10:885–897. https://doi.org/10.1007/s40747-023-01180-7
8. Bergmann P, Fauser M, Sattlegger D, Steger C (2020) Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 4182–4191. https://doi.org/10.1109/CVPR42600.2020.00424
9. Roth K, Pemula L, Zepeda J, et al (2022) Towards Total Recall in Industrial Anomaly Detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2022-June:14298–14308. https://doi.org/10.1109/CVPR52688.2022.01392
10. Bergmann P, Löwe S, Fauser M, et al (2019) Improving unsupervised defect segmentation by applying structural similarity to autoencoders. VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 5:372–380. https://doi.org/10.5220/0007364503720380
11. You Z, Cui L, Shen Y, et al (2022) A Unified Model for Multi-class Anomaly Detection. Adv Neural Inf Process Syst 35:1–19
12. Liu Y, Zhuang C, Lu F (2021) Unsupervised Two-Stage Anomaly Detection
13. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. Adv Neural Inf Process Syst 2015-Janua:2017–2025
14. Vinyals O, Kavukcuoglu K (2017) Neural Discrete Representation Learning
15. Kirillov A, Mintun E, Ravi N, et al (2023) Segment Anything
16. Esser P, Rombach R Taming Transformers for High-Resolution Image Synthesis
17. Razavi A, van den Oord A, Vinyals O (2019) Generating diverse high-fidelity images with VQ-VAE-2. Adv Neural Inf Process Syst 32:
18. Qin X, Zhang Z, Huang C, et al U 2 -Net : Going Deeper with Nested U-Structure for Salient Object Detection