JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 18, No. 1, March 2024, pp. 29-35

Microplastic Binary Segmentation with Resolution Fusion and Large Convolution Kernels

Jaeheon Jeong and Gwanghee Lee
Department of Computer Convergence, Chungnam National University, Daejeon, Korea
tizm423@o.cnu.ac.kr, manggu251@o.cnu.ac.kr

Jihyun Jeong, Junyoung Kim, and Jinsol Kim
Edam Environmental Technology, Daejeon, Korea
jjh2980@gmail.com, qlr9657@gmail.com, solyi0827@gmail.com

Kyoungson Jhang*
Department of Artificial Intelligence, Chungnam National University, Daejeon, Korea
sun@cnu.ac.kr

Abstract
The term "microplastic" refers to plastic particles with a length or diameter of less than 5 mm that do not easily decompose in the natural environment and persist for a long time. These microplastics have adverse effects on the marine ecosystem when they enter the ocean. Therefore, it is necessary to estimate the amount of microplastics in rivers and sewers and to block the outflow of microplastics in areas where they are found to be present at high levels. However, estimating the amount of microplastics first requires detecting these particles, which is not an easy task to complete efficiently and accurately due to their small size and the difficulty involved in distinguishing them from organic materials. The current study therefore proposes a new model structure for microplastic segmentation. This model uses the multi-resolution fusion module (MRFM), which is known to significantly contribute to the segmentation performance in HRNet, and this model employs the EfficientNetV2B3 model as a backbone. We also utilize large convolution kernels to achieve better feature extraction from the inputs of three resolution stages that are closer to the input image resolution. The experimental results showed that the model using two MRFMs outperformed the model using feature pyramid network in the head network, with improvements of 3.28% in IoU and 2.67% in F1-score.

Category: Computer Graphics / Image Processing

Keywords: Deep learning; Computer vision; Binary segmentation; Microplastic

I. INTRODUCTION

Plastic waste, which is found both on land and at sea, has long been a focus of environmental concerns. However, tiny plastic particles known as "microplastic" have more recently been highlighted as particularly harmful pollutants due to their effects on marine biota [1-3].

Not only are microplastics small in size, but they also

Open Access http://dx.doi.org/10.5626/JCSE.2024.18.1.29 http://jcse.kiise.org

Received 02 January 2024; **Accepted** 12 March 2024
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

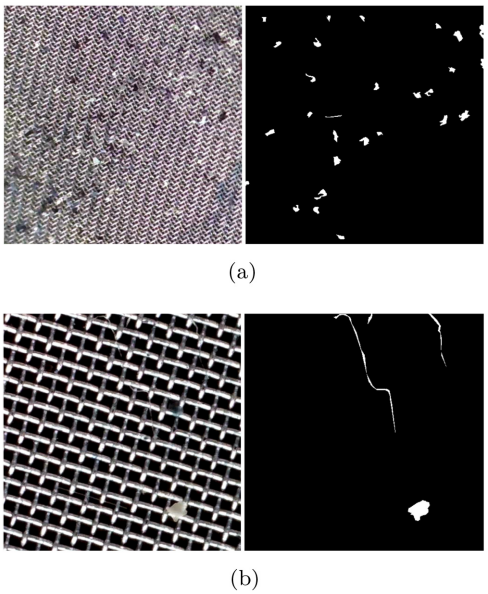

Fig. 1. Input images on the left side along with ground truths on the right side for the corresponding input image: (a) an input example primarily containing microplastic fragments and (b) an input example primarily containing microplastic fibers.

exist in both pelagic and benthic ecosystems and are ingested by a variety of marine organisms [4]. The issues associated with microplastics are not only in the transfer of pollutants through the food chain but also in the fact that microplastics can absorb contaminants from water and pass them on to other trophic levels through bioaccumulation. Microplastics consist of toxic additives and monomers and have a relatively large surface area to volume ratio, making them particularly effective at absorbing hydrophobic pollutants in water. They can absorb toxins when ingested by phytoplankton and corals, and they can produce toxic substances like phycotoxins, thus posing direct and indirect threats to the health of a wide variety of marine animals as well as humans [5, 6]. It is therefore necessary to manage the detection and filtration of microplastic particles in rivers, sewers, etc. It is expected that deep learning can be used for the automation of microplastic detection with sufficient performance. This involves using semantic segmentation technology and training a segmentation deep learning model using images of microplastics filtered through mesh structures.

Microplastics can be categorized into two forms: fragments and fibers. As depicted in Fig. 1(a), fragments are generally small in size, while fibers are generally thin and long, as shown in Fig. 1(b). Fragment-type microplastics, which are radially shaped, can be easily distinguished by a model that has been trained using a convolutional neural network with an effective receptive field that resembles a Gaussian distribution. However, fiber-type plastics, which are narrow and long and thus exist over a relatively wide area, pose difficulties for training and accurate recognition. Plastics with high reflectivity and metal mesh structures can also reflect light, further hindering precise detection. To overcome these challenges, we propose a new segmentation deep learning model that effectively combines the feature extraction capabilities of EfficientNet with the characteristics of the multi-resolution fusion module (MRFM) from HRNet [7], which is capable of classifying subjects at various scales.

In the next section, we introduce the U-Net, a neural network model used for microplastic segmentation, and we explain the effective receptive field and MRFM. Section III presents the experimental results of five models, including the proposed segmentation model, along with an introduction to the microplastic dataset. The final section provides a summary of the proposed method and outlines future research directions.

II. MICROPLASTIC SEGMENTATION

Binary semantic segmentation is the process of separating the areas of the background and the target object. Through this process, the amount of microplastics in the sample can be estimated based on the ratio of the microplastic area to the total image area.

A representative neural network architecture that is often used in binary semantic segmentation is the encoder-decoder structure. In this structure, the encoder embeds information that is useful for object recognition by reducing the size and increasing the channels of the input image, while the decoder outputs the prediction results by expanding the embedded features back to the size of the input image and reducing the channels.

In the encoder, a maxpooling layer is used to reduce the size of the input image. The maxpooling layer outputs the largest value within the kernel area, with the kernel typically set to a height of 2, a width of 2, and a stride of 2. With repeated applications of the maxpooling layer, the size of the feature map decreases significantly. As the size of the feature map decreases, it becomes more difficult to retain detailed information, such as very small particles or thin shapes. This problem can be solved by using a skip connection, which involves concatenating features from the front layers of the encoder, which have undergone fewer maxpooling layers, with features of the same resolution in the decoder. U-Net [8] is composed of an encoder and a decoder, as shown in Fig. 2, and it includes a skip connection that copies and joins features of the same resolution from the encoder to the corresponding resolution in the decoder. The horizontal arrows in Fig. 2 represent the role played by the skip connections. The feature pyramid network (FPN) as shown in Fig. 3 [9] makes use of the idea of employing intermediate features from the encoder in the decoder in a manner similar to U-Net. However, it combines results embedded at various

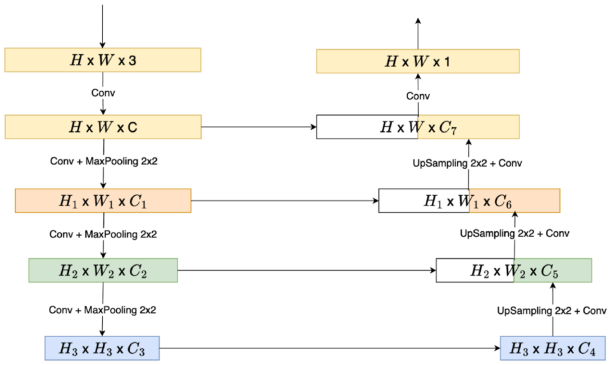

Fig. 2. U-Net architecture.

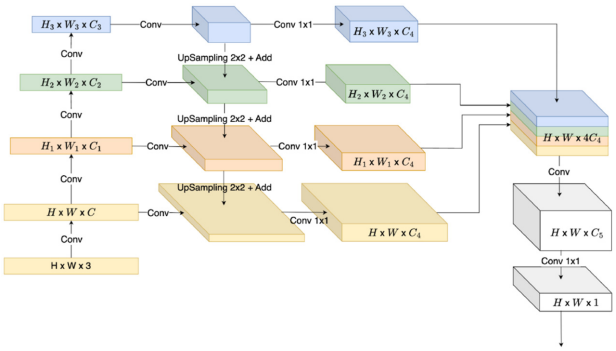

Fig. 3. FPN architecture.

sizes through concatenation, so it could potentially achieve better performance than U-Net. The backbone network used for feature extraction can also be replaced with different models, thus allowing for performance or computational load to be adjusted according to an application's particular needs.

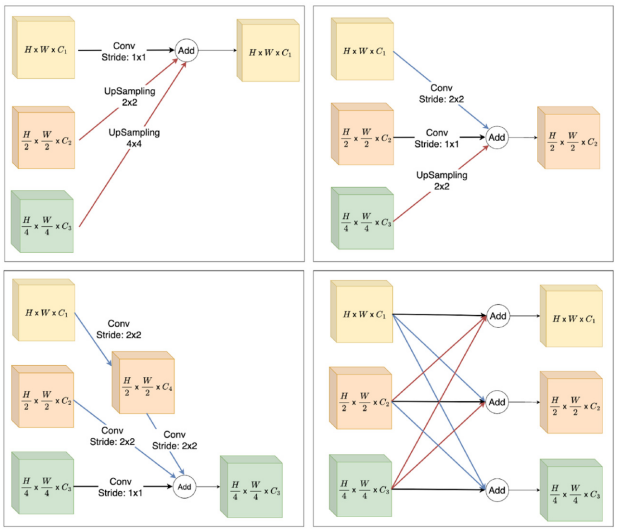

Fig. 4. The architecture of the multi-resolution fusion module.

The MRFM used in HRNet was developed to extract features from various resolutions. As shown in Fig. 4, it fuses high and low resolutions through upsampling and convolution with a stride of 2. Since MRFM maintains the resolution even after fusion, it is possible to design a structure that repeats MRFM two or more times to extract more features.

A. Effective Receptive Field

A receptive field is the input area that influences a single value of the tensor resulting from a convolution operation. Fig. 5 shows the receptive field after performing two convolution operations consecutively with a 3×3 kernel size. Ultimately, a specific pixel in the third layer is influenced by an area 5 pixels in height and 5 pixels in width from the first layer. This can be visualized as shown with the numbers in the first layer of Fig. 3. The visualized values follow a Gaussian distribution, and the more times they are applied, the greater the influence they have on the final output value. The part of the receptive field that actually influences the output is referred to as the effective receptive field [10].

The size of the effective receptive field is directly proportional to both the depth of the model and the size of the kernels used. Among these two factors, increasing the depth of the model leads to several issues such as overfitting, increased computational costs, and problems with gradient vanishing and exploding. Meanwhile, although increasing the size of the kernel has the disadvantage of increased computational costs, it can successfully increase the effective receptive field without significantly increasing the depth of the model. Therefore, the model proposed in this paper applies large kernels of sizes 21×21, 15×15, and 13×13 to the intermediate layer outputs of 512, 256, and 128 pixels in the backbone network to effectively increase the effective receptive field.

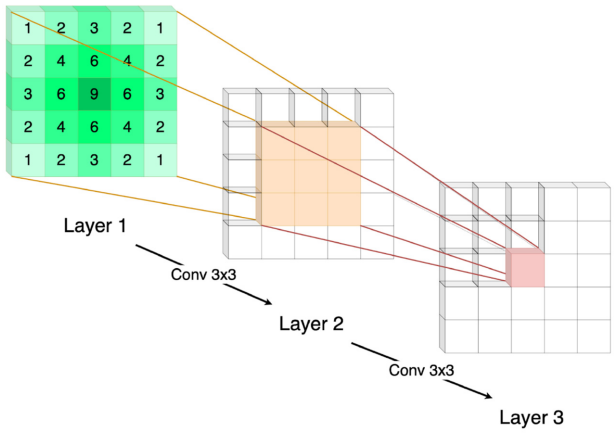

Fig. 5. Illustration of receptive field and effective receptive field.

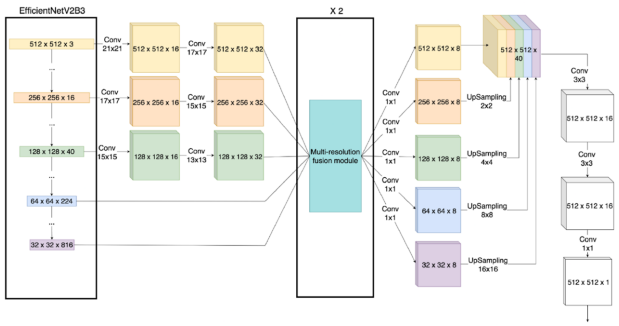

Fig. 6. Overall structure of the proposed segmentation model.

B. Model Architecture

The model structure proposed in this paper, as shown in Fig. 6, includes the EfficientNetV2B3 [11] as the backbone network, along with the MRFM and other convolution layers. This specific backbone network is selected for its proven performance in a variety of problems, including ImageNet classification. The head network utilizes HRNet's MRFM, which fuses feature maps of five different resolutions. These fused feature maps from five stages are concatenated into one tensor, which then goes through three convolution layers. The input image resolution is 512×512, and the outputs at each downsample point of EfficientNetV2B3 are used as inputs to the head network. As shown in the middle left in Fig. 6, for the outputs of the backbone with heights and widths close to the input resolution, i.e., 512, 256, 128 pixels, larger convolution kernels are used to better extract features of thin and long objects like fibers. The convolutions used here have large kernel sizes that have been set to 21, 17, 15, and 13 pixels with the aim of increasing the effective receptive field. Subsequently, using the outputs of the backbone with heights and widths of 64 and 32 pixels, the MRFM is used to fuse outputs across all resolutions, ultimately extracting a wealth of features. Finally, through concatenation, the features of all resolutions are combined, and the result is produced by final three convolutions.

III. EXPERIMENTS

A. Dataset

To achieve accurate identification and detection of microplastics, it is important to remove organic materials attached to the sample or on the surface of microplastics. Therefore, it is essential to include a preprocessing step wherein chemicals are used to remove organic matter before detecting microplastics. After this treatment, microplastic samples were obtained using a self-made automated microplastic detection device (with five stages of filter sizes: 500, 250, 134, 63, and 25 μm). Microplastic sample image data were measured using a stereo microscope, and images were repeatedly taken at the same position after dividing the filter mesh into nine to 36 sections and then fixing it at a constant position.

The obtained image data, comprising images of microplastics trapped in the mesh at various resolutions, were used for training and evaluation. The dataset contains 13,711 images of size 512×512×3, of which 10,959 images were used as training data and 2,752 were used as test data.

B. Experimental Results

Since there was only a small amount of image data, data augmentation techniques were applied. The data augmentation techniques applied included a 50% probability of a random horizontal flip, 80% probability of Cutout [12], random rotation between 0° and 360°, 50% probability of Gaussian random blurring with a 9×9 size, random contrast adjustment between 80% and 120%, random saturation adjustment between 50% and 150%, and random brightness adjustment between -20% and 20%.

Alongside the U-Net model, four additional models were used in the experiment to sum to a total of five models: one using EfficientNetV2B3 pretrained on ImageNet [13] as the backbone with FPN in the head network, and three proposed models where MRFM was applied once, twice, or three times, respectively, in the head network. The training was conducted for 50 epochs, and the performance measures compared were intersection over union (IoU), recall, precision, and F1-score. The experimental results are summarized in Table 1. In the table, the number of parameters for each model is shown as number of parameters, while the computational cost for a single prediction is presented in FLOPs. The proposed models have a similar number of parameters to U-Net but a lower computational cost compared to the Backbone+FPN model. In terms of performance, the proposed models consistently show better results in IoU, precision, recall, and F1-score compared to U-Net and the Backbone+FPN model. Among the proposed models, the Backbone+MRFM×2 model, which repeats MRFM twice in the head network, showed the best performance in terms of all aspects. The Backbone+MRFM×3 model, which repeats MRFM three times, showed lower performance in IoU, precision, and F1-score compared to the Backbone+MRFM×1 model, which only uses MRFM once.

Fig. 7 shows the segmentation prediction results for each model on actual input examples containing Fiber and Fragment. Fig. 7(a) shows the input example whereas Fig. 7(b) shows the Ground Truth image, where the white areas on a black background represent microplastics. Fig. 7(c) presents the prediction result using the U-Net

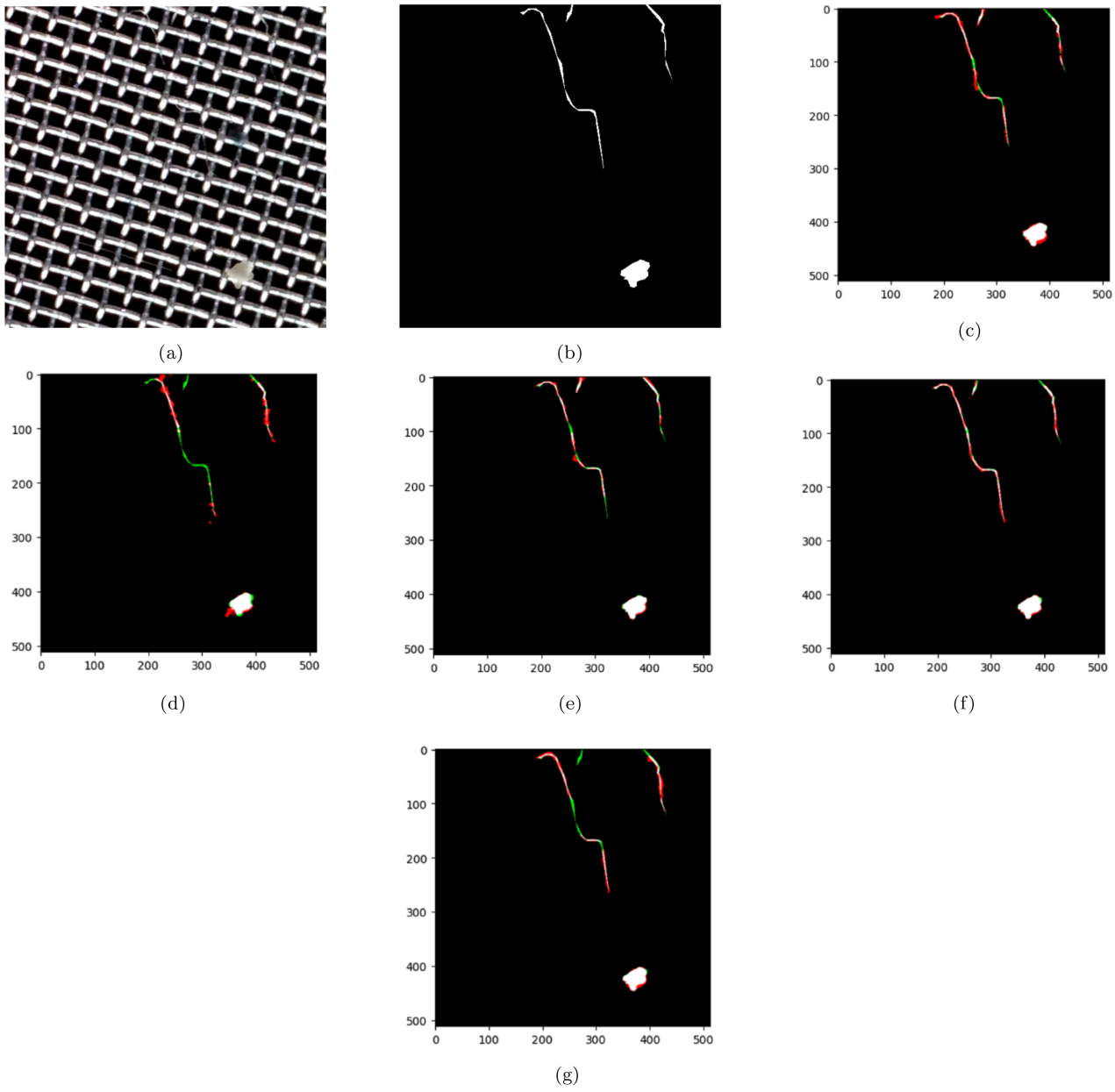

Fig. 7. Comparison of input image and ground truth with the outputs of the five models: (a) input image, (b) ground truth, (c) prediction result of the U-Net model, (d) prediction result of the Backbone+FPN model, (e) prediction result of the Backbone+MRFM×1 model, (f) prediction result of the Backbone+MRFM×2 model, and (g) prediction result of the Backbone+MRFM×3 model.

Table 1. Experimental results comparing five models in terms of IoU, recall, precision, and F1-score shown together with FLOPs and the number of parameters of each model

Model	FLOPs ($\times 10^9$)	# Parameters ($\times 10^6$)	IoU (%)	Recall (%)	Precision (%)	F1-score (%)
U-Net	55.841	7.85	58.73	77.36	80.88	79.08
Backbone+FPN	220.119	4.51	59.86	80.87	81.62	81.25
Backbone+MRFM ×1	104.513	7.45	62.15	81.03	83.82	82.40
Backbone+MRFM ×2	112.029	8.86	63.14	82.19	85.71	83.92
Backbone+MRFM ×3	119.546	10.27	60.87	82.19	81.43	81.81

model, while Fig. 7(d) shows the prediction result of the model with EfficientNetV2B3 as the Backbone and FPN in the head network. Fig. 7(e)–7(g) each display the prediction results of models applying MRFM once, twice, and three times, respectively, in the head network of the Backbone. As can be seen in the results presented in Fig. 7, models using MRFM in the head network can more accurately separate fibers and fragments compared to the U-Net or Backbone+FPN models. Moreover, among the models with MRFM applied in the head network, the one repeating this module twice was observed to separate fibers more accurately than the models repeating it either once or three times.

IV. CONCLUSION

Herein, we have proposed microplastic segmentation models that employ EfficientNetv2B3—which is known to be effective for problems like ImageNet classification—as the backbone and utilize one, two, or three MRFMs from HRNet in the head network. Compared to the U-Net model or models using FPN in the head network, our models showed superior performance in terms of IoU, precision, recall, and F1-score. The model applying two MRFMs in the head network exhibited the highest performance. The model with three MRFMs showed lower performance than the model with just one MRFM, aside from in recall.

The proposed segmentation model in this study can be extended to instance segmentation for tasks involving the distinction between fragments and fibers and the counting of objects.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ACKNOWLEDGEMENTS

This work was supported by Chungnam National University.

REFERENCES

1. E. J. Carpenter and K. L. Smith, "Plastics on the Sargasso Sea surface," *Science*, vol. 175, no. 4027, pp. 1240-1241, 1972. https://doi.org/10.1126/science.175.4027.1240
2. M. R. Rands, W. M. Adams, L. Bennun, S. H. Butchart, A. Clements, D. Coomes, et al., "Biodiversity conservation: challenges beyond 2010," *Science*, vol. 329, no. 5997, pp. 1298-1303, 2010. https://doi.org/10.1126/science.1189138
3. W. J. Sutherland, M. Clout, I. M. Cote, P. Daszak, M. H. Depledge, L. Fellman, et al., "A horizon scan of global conservation issues for 2010," *Trends in Ecology & Evolution*, vol. 25, no. 1, pp. 1-7, 2010. https://doi.org/10.1016/j.tree.2009.10.003
4. K. Betts, "Why small plastic particles may pose a big problem in the oceans," *Environmental Science & Technology*, vol. 42, no. 24, pp. 8995-8995, 2008. https://doi.org/10.1021/es802970v
5. Y. Mato, T. Isobe, H. Takada, H. Kanehiro, C. Ohtake, and T. Kaminuma, "Plastic resin pellets as a transport medium for toxic chemicals in the marine environment," *Environmental Science & Technology*, vol. 35, no. 2, pp. 318-324, 2001. https://doi.org/10.1021/es0010498
6. E. V. Morse, "Paralytic shellfish poisoning: a review," *Journal of the American Veterinary Medical Association*, vol. 171, no. 11, pp. 1178-1180, 1977.
7. J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349-3364, 2021. https://doi.org/10.1109/TPAMI.2020.2983686
8. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
9. T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2117-2125. https://doi.org/10.1109/CVPR.2017.106
10. W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 29, pp. 4898-4906, 2016.
11. M. Tan and Q. Le, "EfficientNetV2: smaller models and faster training," *Proceedings of Machine Learning Research*, vol. 139, pp. 10096-10106, 2021.
12. T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017 [Online]. Available: https://arxiv.org/abs/1708.04552.
13. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015. https://doi.org/10.1007/s11263-015-0816-y

Jaeheon Jeong https://orcid.org/0009-0003-9918-3192

Jaeheon Jeong is a senior student in the Division of Computer convergence at Chungnam National University. He is interested in computer vision and deep learning, and his research interests include semantic segmentation and object detection.

Gwanghee Lee https://orcid.org/0000-0003-3023-4230

Gwanghee Lee has been in progress of Ph.D. course in Computer Science at Chungnam National University since 2021. His interested research is Computer Vision and Deep Learning. His research interests include computer vision deep learning, face alignment, pose estimation, object detection, keypoint detection, semantic segmentation, instance segmentation, and face recognition.

Jihyun Jeong https://orcid.org/0000-0001-6168-3800

Jihyun Jeong received Ph.D. degree in Department of Environmental Engineering from Chungnam National University in 2018. Since 2021, she has been working as a CEO at EDAM Environmental Technology Co. Ltd., Deajeon, South Korea. Her research focused on analysis of microplastic in water and soil.

Junyoung Kim https://orcid.org/0009-0007-8765-2448

Junyoung Kim is an assistant researcher of EDAM Environmental Technology. He received an M.S. degree in Department of Environmental Engineering from Chungnam National University in 2020. He is interested in analysis of microplastic in water and soil.

Jinsol Kim https://orcid.org/0009-0009-3876-8115

Jinsol Kim is an assistant researcher of EDAM Environmental Technology. He received an M.S. degree in Department of Environmental Engineering from Gyungsang National University. He is interested in analysis of microplastic in water and soil.

Kyoungson Jhang https://orcid.org/0000-0001-5659-0503

Kyoungson Jhang received B.S., M.S., and Ph.D. degrees in Department of Computer Engineering from Seoul National University in 1986, 1988, and 1995, respectively. Since September 2001, he has been working as a professor for the Department of Computer Science and Engineering at Chungnam National University, Daejeon, Korea. His research focuses on computer vision and deep learning.