Application of the Prompt Engineering-assisted Generative AI for the Drone-based Riparian Waste Detection

Shijun PAN[1], Keisuke YOSHIDA[2], and Takashi KOJIMA[3]

[1]Member of JSCE, Doctoral Student, Graduate School of Environmental and Life Science, Okayama University
(Tsushima-naka 3-1-1, Kita-ku, Okayama 700-8530, Japan)
E-mail: p4b36znn@s.okayama-u.ac.jp (Corresponding Author)
[2] Member of JSCE, Associate Professor, Graduate School of Environmental and Life Science, Okayama University
E-mail: yoshida.k@okayama-u.ac.jp
[3] Senior Researcher, TOKEN C.E.E. Consultants Co.,Ltd.

Waste pollution detection has emerged as one of the crucial environmental concerns in recent years, and the accuracy of this practical application has been significantly improving with advancements in deep learning (DL) algorithms. To efficiently detect and quantify waste over large areas, the use of unmanned aerial vehicles (UAVs) has become essential. However, UAV flights and real-world image collection pose challenges that demand expertise, significant time, and financial investments. These challenges are particularly prominent in specialized applications such as waste detection, which rely on large amounts of data. Notably, the availability of adequate and accurately labeled data is vital for the performance of object detection models. Therefore, the identification and acquisition of suitable training data are critical objectives of this study. While ensuring data quality, AI-Generated Content (AIGC), specifically derived from Stable Diffusion, is emerging as a promising data source for DL-based object detection models. This research employed the Stable Diffusion to generate images by utilizing the prompts generated from specified images. Subsequently, the public dataset-based existing trained model automatically labeled the AIGC, which were then assigned corresponding labels in a uniform ratio for training, validation, and testing purposes. To assess the performance differences between the generated dataset and the dataset collected from real-world scenarios, several benchmark datasets were used for accuracy evaluation in this work. The results revealed that the AIGC exhibited superior accuracy in identifying high Ground Sample Distance (GSD) targets in simple backgrounds compared to the realistic collected dataset (F1 score-based). The results demonstrate the potential of AIGC in providing data for object detection models.

Key Words : *Prompt Engineering, Object Detection, Riparian Waste Detection, Stable Diffusion, UAV*

1. INTRODUCTION

Recently, waste pollution in water ecosystems has emerged as a global environmental problem. One of the primary factors contributing to its occurrence is the phenomenon of indiscriminate waste dumping. And monitoring waste pollution along riverbanks with a better cost-performance way is an emergency need for the riparian management. In response to this issue, drones and artificial intelligence (AI) technologies, including You Only Look Once version 5 (YOLOv5), have been employed for the study of waste pollution monitoring at riverbanks[1), 2)].

These technologies have provided valuable insights into the extent of waste pollution. Nonetheless, certain challenges remain, such as the scarcity of data required for training the YOLOv5 due to difficulties

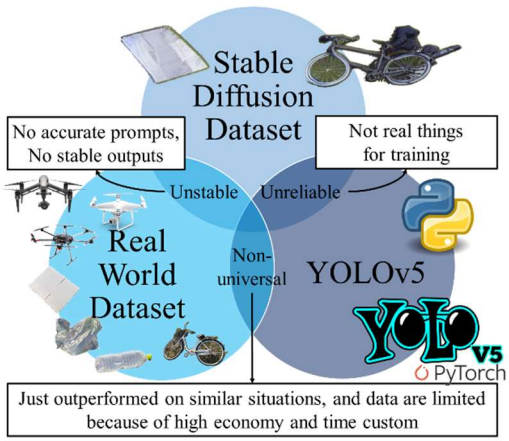

Fig.1 Existing problems among the current datasets and YOLOv5.

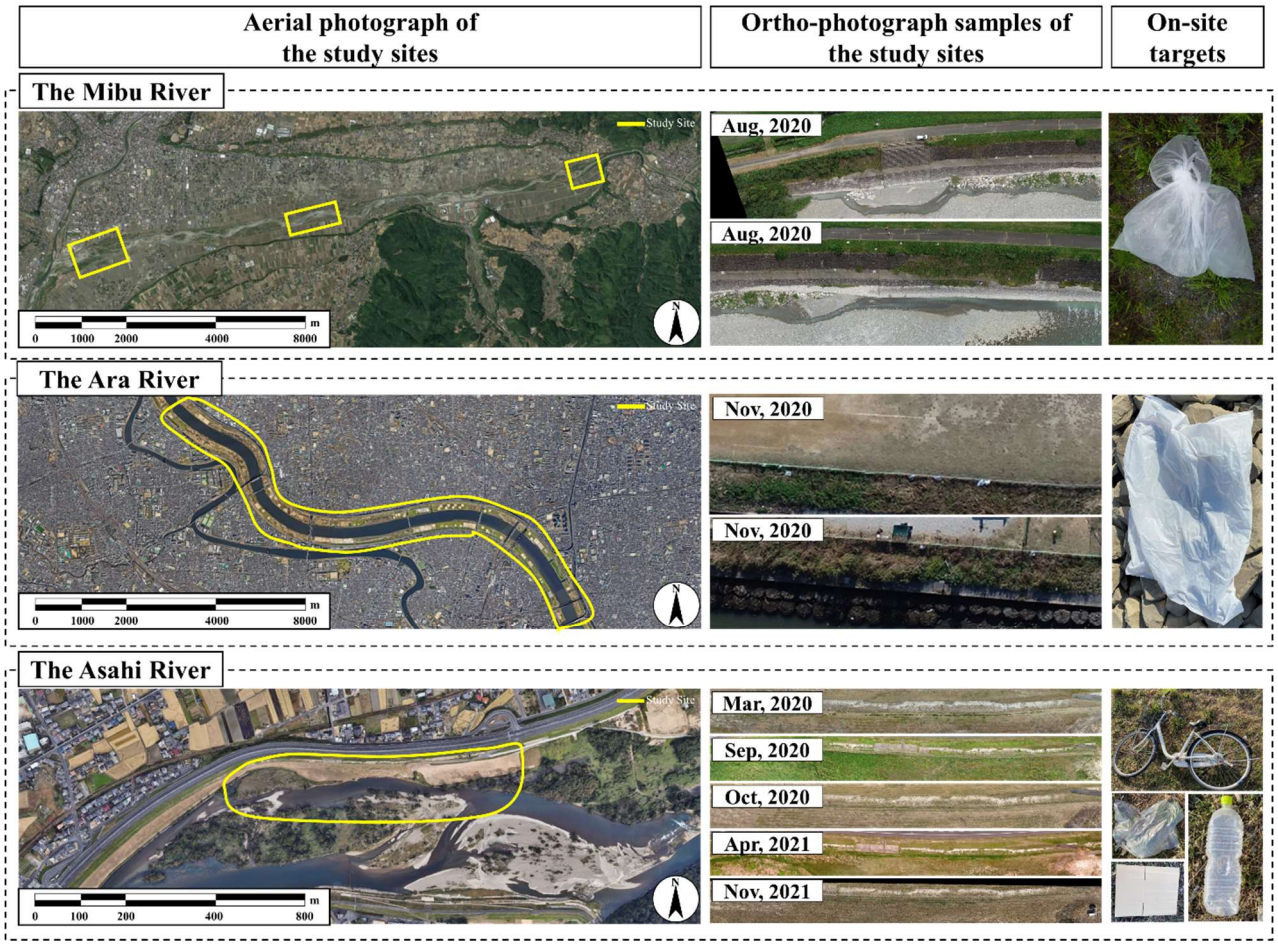

Fig.2 Aerial-, ortho-photograph and on-site targets of the study sites from up to down side (i.e., the Mibu River, the Ara River and the Asahi River). Noteworthy, among the Ortho-photograph in the Asahi River, only the Nov, 2021 consists the On-site targets. Except of the Nov, 2021, the other data in the Asahi River are prepared for the background change operation (i.e., Background Images). Aerial photographs are from Google Map; Ortho-photographs are from original.

in collecting high-quality drone images featuring specific waste targets.

In particular, as depicted in **Fig.1 (left)**, the collection of Real World Dataset necessitates a significant amount of equipment, such as UAVs, and the placement of specific targets on the site, such as Bikes, Cardboards, PET Bottles, and Plastic Bags. Due to the limitations of only using the Real World Dataset for model training, the YOLOv5 in **Fig.1 (right)** can just focus on the features in the limited dataset, which may lead to misclassification of other targets that have been not included in the training (i.e., non-universal training).

As shown in **Fig.1 (top)**, one of the open-source image-based generative AI models, Stable Diffusion[3] model, that uses deep learning (DL) text-to-image technology. It is designed to generate detailed images based on text descriptions (i.e., prompts) and can also be utilized for tasks like image to image (i.e., img2img) translation guided by prompts. In this study, the Stable Diffusion Dataset was generated based on the features of the targets in the Real World Dataset.

It is important to note that the quality of the Stable Diffusion Dataset primarily depends on the well-performed and accurate prompts. And it cannot be stable-generated just based on the Real World Dataset directly using img2img function. These unstable outputs may have contributed to lead to the unreliable trained model.

Although the aforementioned issues have existed in the practical application, there has been no comparison conducted to assess the trained YOLOv5 derived from the Real World Dataset and Stable Diffusion Dataset using a benchmark-based evaluation approach. In this research, the authors focus on the possibility of replacing or enhancing the Real World Dataset with the Stable Diffusion Dataset during the training of the YOLOv5 for the detection of real targets in practical waste pollution detection.

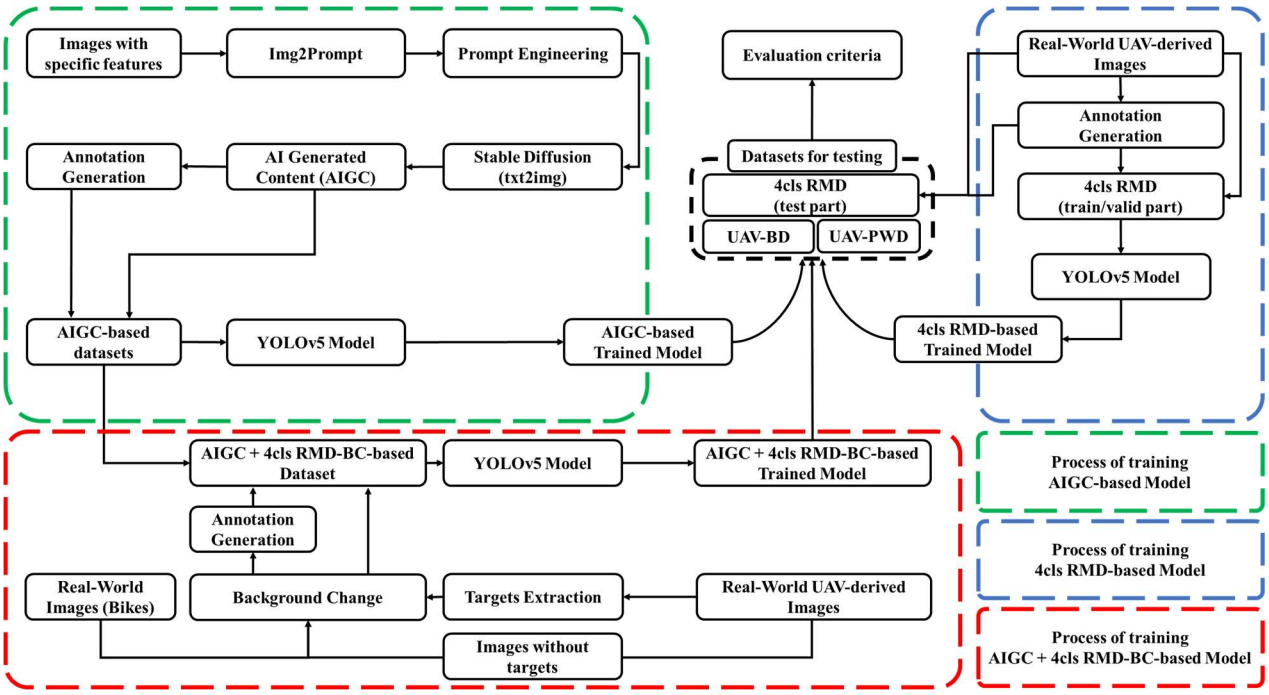

Fig.3 Process of assessing the AIGC and Real World Dataset-based models with benchmark datasets (i.e., AIGC, AI Generated Content or Stable Diffusion Dataset; 4cls RMD, River Monitoring Dataset with 4 classes waste pollution; BC, Background Change).

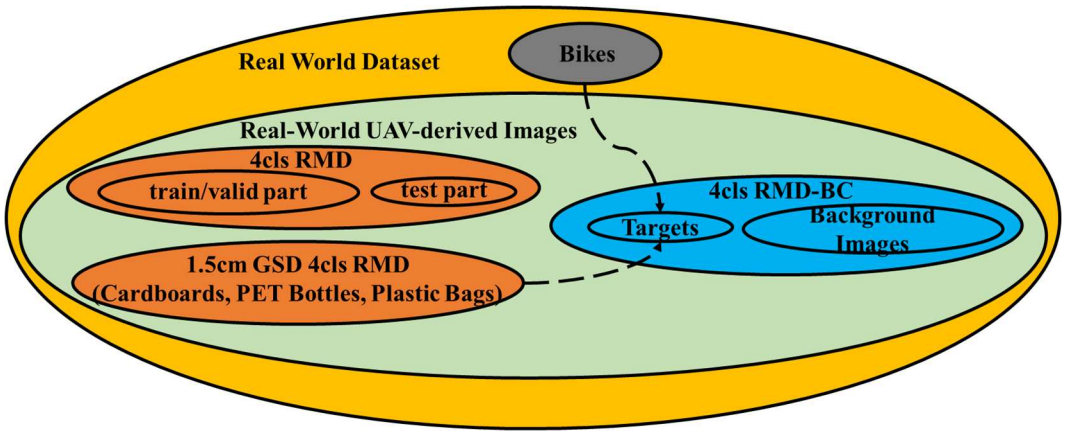

Fig.4 Composition of the Real World Dataset.

2. STUDY SITE AND METHODS

(1) Study Sites

Fig.2 (left) displayed the aerial photographs of the study sites, which are located in the Mibu River, the Ara River and the Asahi River, from up to down sides, individually. And these three state-controlled first-class rivers in Japan that flows through Nagota, Tokyo and Okayama Prefecture. To understand the detailed situations of these sites, thus in the **Fig.2 (middle),** ortho-photograph samples are also performed in this work. The **Fig.2 (right)** showed the on-site targets in this research of individual location. And the targets are mainly around Bikes, Cardboards, PET Bottles, and Plastic Bags.

(2) Flow Chart of Research Process

Generally, this research is separated into four main sections in the **Fig.3**. In the **Fig.3 (left)**, AIGC-based Model is mainly derived from the Stable Diffusion Dataset that is generated by the txt-based prompts (i.e., txt2img). The first step of generating the images with features in need is to capture the images that are matching the requirements. Then applying the website with img2prompt function to extract the information of the images (i.e., CLIP Interrogator online version in this research). CLIP (Contrastive Language-Image Pre-training) is a neural network trained on a variety of (image, text) pairs, that can predict the

Fig.5 Samples of the AI Generative Content (AIGC).

Table 1 Components of the prompts in this work.

Components	Samples
Subjects	Bikes, Cardboards, PET Bottles, Plastic Bags
Resolution	High resolution, 8k camera
View angle	Bird view, UAV view
Area	Riparian area

most relevant text snippet given an image[4]. CLIP can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3. Worth mentioning, these prompts derived from the CLIP Interrogator can just provide approximate information. For the AIGC with more detailed information, Prompt Engineering is necessary. In this research, after the comparison of several AIGC samples derived from the Prompt Engineering, the key words that can indicate the reasonable results have been confirmed (e.g. UAV, 8k, super detailed and high resolution).

Based on the AIGC derived from the Stable Diffusion (i.e., txt2img function), the annotation Generations are also important for the model training. All the annotation generations for the AIGC are based on the public dataset-based garbage reorganization standard[5] with similar feature. After collecting the AIGC and corresponding annotation generations, the authors used the Roboflow (i.e., an online platform to pre-process the dataset) to preprocess the AIGC-based Dataset.

Continually in the **Fig.3 (right)**, Real-World UAV-derived Images are separated into three parts (i.e., train/valid and test part). And 4cls RMD-based Trained Model is derived from the train/valid part in this dataset. Remarkably, annotation generations were mainly based on the practical situation of the on-site targets.

Shown in the **Fig.3 (down)**, several targets were extracted from the Real-World UAV-derived Images, and combined the images without targets to generate the images with the Background Change. And the Bikes targets are not enough, the supplement of the Real World Dataset (i.e., Bikes) are necessary. After the generation of the annotations, the AIGC + 4cls RMD-BC-based Model can be trained based on the dataset combination of the AIGC and 4cls RMD-BC-based Dataset. In general, as performed in the **Fig.3 (middle)**, the mentioned trained models need to be evaluated by the following three datasets: UAV-BD[6], UAV-PWD[7] and 4cls RMD (test part) for the evaluation criteria, individually.

For understanding the sections in the Real World Dataset, the **Fig.4** explained the relationship among each section in the dataset. Firstly, Real-World UAV-derived Images have two sections (i.e., 4cls RMD with train/valid/test parts, 4cls RMD-BC with background images and targets). Except of the mentioned images, there are also Targets (i.e., Bikes) existing for the supplements.

(3) Models

Mainly the Stable Diffusion consists of three main components: the variational autoencoder (VAE), U-Net, and an optional text encoder. The Stable-Diffusion-v1-5 checkpoint used in this research was initialized with the weights of the Stable-Diffusion-v1-2 checkpoint[8] and subsequently fine-tuned on 595k steps at resolution 512px ×512px on "laion-aesthetics v2 5+" (i.e., 600M image-text pairs with predicted aesthetics scores of 5 or higher in the LAION 5B dataset) and 10% dropping of the text-conditioning to improve classifier-free guidance sampling.

The You Only Look Once (YOLO) version 5 model (i.e., YOLOv5), which is an open-source software based on convolutional neural networks (CNNs) with optimal detection accuracy and reasonable computational complexity. Based on the mentioned issues, YOLOv5 was chosen as the model for object detection training model in this work.

(4) Datasets for training/validation

Quality and quantity of the AIGC were mainly controlled by the model-related parameter setting in the Stable Diffusion web UI. The model-related parameters setting were mainly adjusted derived from the total computational time-consuming and VRAM (i.e., GPU memory). The generated samples are performed in the **Fig.5** derived from the specified prompts. As performed in **Table 1**, the prompts used in this research include three main components: subject, resolution, view angle, and area.

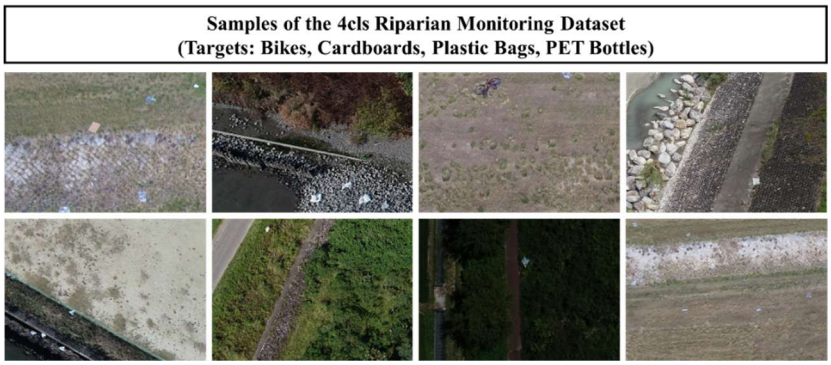

Fig.6 Samples of 4cls RMD (i.e., River Monitoring Dataset).

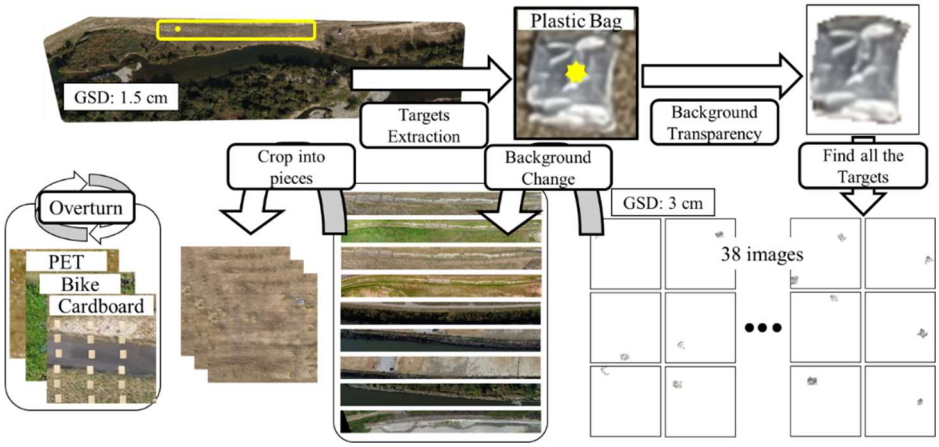

Fig.7 Process of generating 4cls RMD-BC (i.e., River Monitoring Dataset-Background Change).

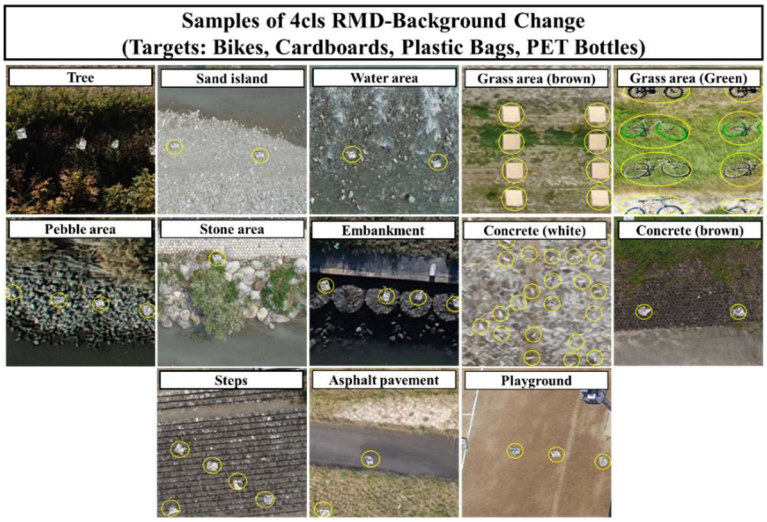

Fig.8 Samples of 4cls RMD-BC.

Table 2 Dataset-based composition of each case.

	AIGC	4cls RMD-BC	4cls RMD
Case-1	○	×	×
Case-2	○	○	×
Case-3	×	×	○

(1)

Image Numbers	Case-1	Case-2	Case-3
Bikes	500	1042	155
Cardboards	500	986	452
PET Bottles	500	997	309
Plastic Bags	500	928	2605

(2)

Table 3 Model-related parameter setting.

Parameters	Configuration (Stable Diffusion)	Configuration (YOLOv5)
Operating system (Version)	Windows (Win 11)	Linux (Ubuntu 20.04.4 LTS)
Model version	v1-5-pruned-emaonly.safetensors	v 6.0
Image-size (pixel)	608 × 608	1024 × 1024
Initial learning rate	-	0.01
Final learning rate	-	0.1
Optimizer	-	SDG
Momentum	-	0.937
Batch size	1	12
Batch count	100	-
Epochs	-	500
Patience	-	100

Table 4 Performance measurement TP, TN, FP, FN are the parameters used in the evaluation of Recall (R), Precision (P), F1.

True \ Predicted	Positive	Negative
Positive	True Positive (*TP*)	True Negative (*FN*)
Negative	False Positive (*FP*)	True Negative (*TN*)

$$Recall\ (R) = \frac{TP}{TP + FN} \quad (1)$$

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (2)$$

$$F1 = \frac{2 \times R \times P}{R + P} \quad (3)$$

$$mAP_{\ IoU} = \frac{1}{N} \sum_{\substack{k=1 \\ k \in (1,\,2 \ldots, N)}}^{N} AP_{k_{IoU}} \quad (4)$$

$AP_{k_{IoU}}$: AP of class k under the IoU threshold.
N: Number of all the classes (class is 1 in this study).

The images of the 4cls RMD were taken by multiple drones (i.e., Inspire2, Phantom4 Pro, Zenmuse X4s) with different sensors (i.e., Zenmuse X4s and Z3) on three riparian areas using multiple camera angles (i.e., 45°, 60°, 75°) and GSDs (i.e., 2-, 3-, 4- cm). As performed in the **Fig.6,** the before-mentioned four garbage are all concluded in the sample images.

As the supplement of the AIGC, the 4cls RMD-BC followed the steps in **Fig.7**. Extracting all the Plastic Bags and replacing the background using anther UAV-derived image without Plastic Bags. As a final

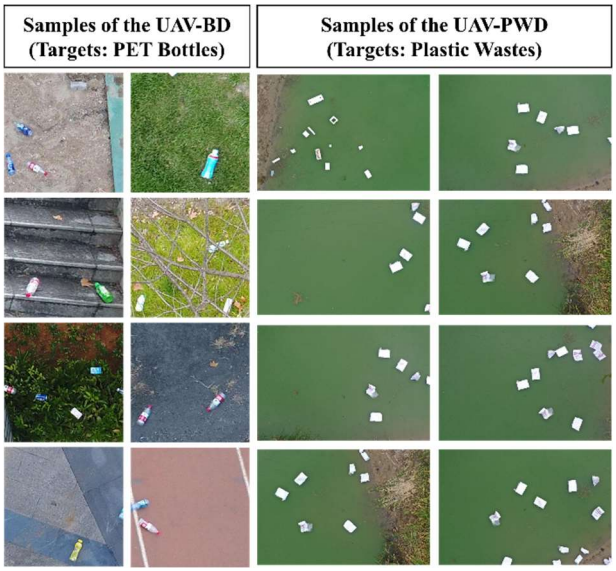

Fig.9 Samples of the images derived UAV-BD and UAV-PWD, mainly bottles and plastic waste pollution.

point, cropping the background-changed images into pieces, and overturning the same operation on the other targets. Shown in the **Fig.8**, there are thirteen kinds of backgrounds have been collected for supplement. Worth mentioning, not only natural also artificial environment has been collected in the dataset.

As displayed in the **Table 2 (1) & (2)**, three cases with specificed image numbers have been considered in this research for confirming the effect of the AIGC in detecting the Real World Dataset. Case 1 and Case 2 consist the Stable Diffusion Dataset, and Case 3 is totally derived from Real World Dataset.

(5) Model-related parameter setting

The details of the parameters setting derived from the Stable Diffusion and YOLOv5 have been performed in the **Table 3**. The Stable Diffusion is using the pre-trained model that was downloaded from the Hugging face (i.e., v1-5-pruned-emaonly.safetensors), is an American company that develops tools for building applications using machine learning.

(6) Evaluation method

As shown in the **Table 4,** the binary confusion matrix has four entries: the number of true positive (TP) and true negative (TN) samples, which are respectively those that are correctly detected as positive and negative, and the two error categories of false positive (FP) and false negative (FN) samples, which represent the number of negatives incorrectly detected as positives.

When using the YOLOv5 to detect the garbage, it is important to choose evaluation measures for this object detection task. Here, as shown in the **Equation (1) & (2)**, both Precision and Recall should be considered as the measure that the model can accurately

Table 5 Dataset-based composition of each case.

Test Datasets	Cases	Image size	P	R	F1	mAP50	mAP50-95
4cls RMD (test part)	1	1024	0.727	0.721	0.724	0.744	0.377
	2		0.823	0.716	0.766	0.8	0.402
	3		0.952	0.893	0.922	0.966	0.783
UAV-PWD	1	1024	0.807	0.864	0.835	0.87	0.508
	2		0.834	0.861	0.847	0.891	0.535
	3		0.554	0.657	0.601	0.421	0.176
UAV-BD	1	342	0.771	0.746	0.758	0.764	0.328
	2		0.744	0.689	0.715	0.727	0.31
	3		0.575	0.533	0.553	0.46	0.179

Table 6 4cls RMD (test part)-derived class-based results using Case 2.

Class	Image	Instances	P	R	F1	mAP50	mAP50-95
Bikes	38	38	0.317	0.395	0.352	0.213	0.060
Cardboards	114	138	0.647	0.551	0.595	0.656	0.342
PET Bottles	68	73	0.518	0.691	0.592	0.589	0.241
Plastic Bags	713	1615	0.775	0.736	0.755	0.767	0.401

detect the garbage or not, Precision and Recall value depend on the factors from the **Table 4** basically. And **Equation (3)** performed the harmonic mean of Precision and Recall, that is main evaluation criteria in this research.

The mean Average Precision (mAP) in **Equation (4)** provides an overall assessment of the YOLOv5's performance in detecting the garbage accurately and consistently derived from Precision and Recall. mAP50 and mAP50-95 are two variants of the mAP metric, where the numbers indicate the IoU threshold used for evaluating the model. The mAP50 uses an IoU threshold of 0.5, while mAP50-95 uses a range of IoU thresholds from 0.5 to 0.95.

(7) Datasets for testing

Except for the 4cls RMD (test part), two public datasets have been prepared for testing. **Fig.9** performed the samples of the images derived from UAV-BD and UAV-PWD. UAV-BD has eight types of backgrounds to be selected to collect the images (i.e., Ground, Step, Bush, Land, Lawn, Mixture, Sand, and Playground). And UAV-PWD has just one type of background (i.e., water area) without the complex feature. Compared with the complicated color and textures of the backgrounds in UAV-BD, UAV-PWD is comparably much simpler than UAV-BD. In other words, UAV-PWD has a simple background than UAV-PWD. Based on the results derived from these two test datasets, this work can measure the ability of the AIGC-based models to detect the targets both in simple and complex backgrounds.

3. RESULTS AND DISCUSSION

This study is mainly discussing waste pollution detection using UAVs aided with deep learning algorithms. And the authors also explored the challenges of collecting and labeling training data for waste pollution detection models and introduce AIGC as a potential data source. The Stable Diffusion, a text-to-image model, is used to generate images based on specified prompts.

The prompts are derived from the existing images, and the AIGC is automatically labeled using a pre-trained object detection model. The generated dataset is then utilized to train object detection models for the detection of the waste pollution. In summary, this study compares the performance of the AIGC-based Dataset with Real World Datasets using benchmark datasets for evaluation.

Performed the results of using 4cls RMD (test part) for testing in the **Table 5**, Case 3 showed the dominant high accuracy (i.e., F1 value) than Case 1 and 2 derived from AIGC. And Case 2 has improved from Case 1 because of using the Real World Dataset with background change. As shown in **Table 6**, because of the limited additional targets-based colors/shapes (i.e., Bikes), Bikes have not been detected with comparably low F1 value using Case 2.

On the other hand, the results in the **Table 5** derived from UAV-PWD and UAV-BD indicate that the AIGC-based Dataset (i.e., Case 1, 2) showed superior accuracy in detecting waste pollution on the simple backgrounds (i.e., water area) compared to the

Fig.10 Samples of the results derived 4cls RMD using Case 2.

Table 7 1.5 cm GSD 4cls RMD-derived class-based results using Case 1.

Class	Image	Instances	P	R	F1	mAP50	mAP50-95
Bikes	2	2	0	0	0	0	0
Cardboards	2	2	0.978	1	0.989	0.995	0.566
PET Bottles	2	2	1	1	1	0.995	0.224
Plastic Bags	3	3	0.972	1	0.986	0.995	0.714

Fig.11 Samples of the results derived 1.5 cm GSD 4cls RMD using Case 1.

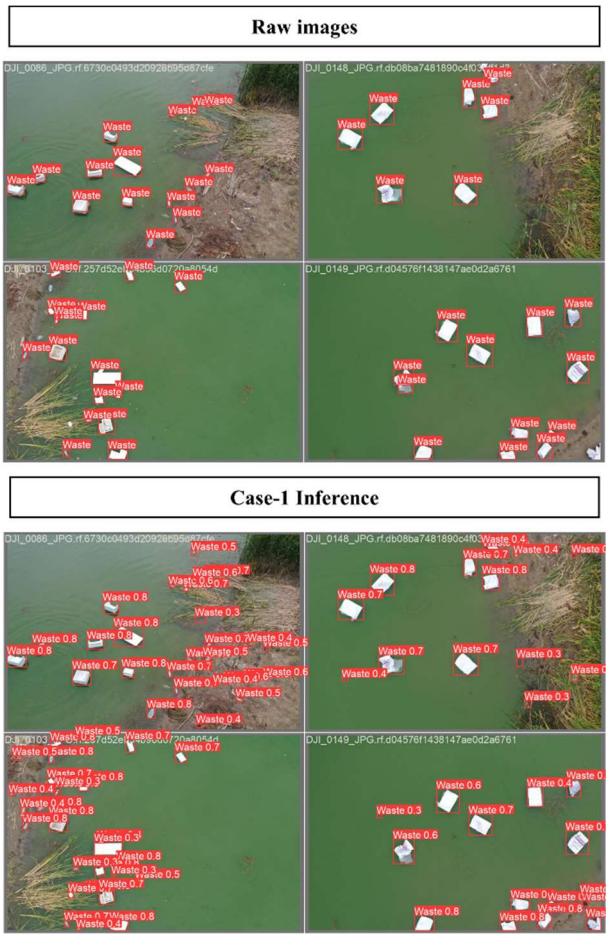

Fig.12 Samples of the results derived UAV-PWD using Case 1.

Case 3. In the case of UAV-BD, even Case 2 has increased the data amount, Case 1 also outperformed both in Precision and Recall value. The increased background-change images in Case 2 have almost the same targets (i.e., cropped images including Bikes, Cardboards, Plastic Bags, PET Bottles), which reduced the F1 score of the trained model in detecting the targets with complex features (i.e., different colors, complicated shapes). Generally speaking, if the background of the test dataset is simple, more targets for training even similar could improve the F1 score. On the contrast side, the more complex features the targets of test datasets have, the more data with complex features need to be added to the training dataset.

4. CONCLUSION

In this study, to some content, using the AIGC can support (i.e., replacing or enhancing) the UAV-based Real World waste pollution detection tasks. Especially with the assistance of the Prompt Engineering, the images with specified targets can be generated with purposes. But there are also some limitations that cannot be solved yet. The pre-trained model for generating annotations for the AIGC is just one dataset with specified features, and the generated annotations are totally derived from the features of this dataset. Alternatively, if the pre-trained model changed, the generated annotations can also be an unstable factor for training a model derived from the AIGC.

In conclusion, without UAV flight and manual annotation for the train/valid dataset, the AIGC in this work can also apply in the riparian monitoring tasks of detecting the waste pollution in the simple backgrounds with a comparable high F1 value. The AIGC showed efficiency rather than UAV- or vehicle-based data collection process and also can reduce the burden of the professional civil engineering staff.

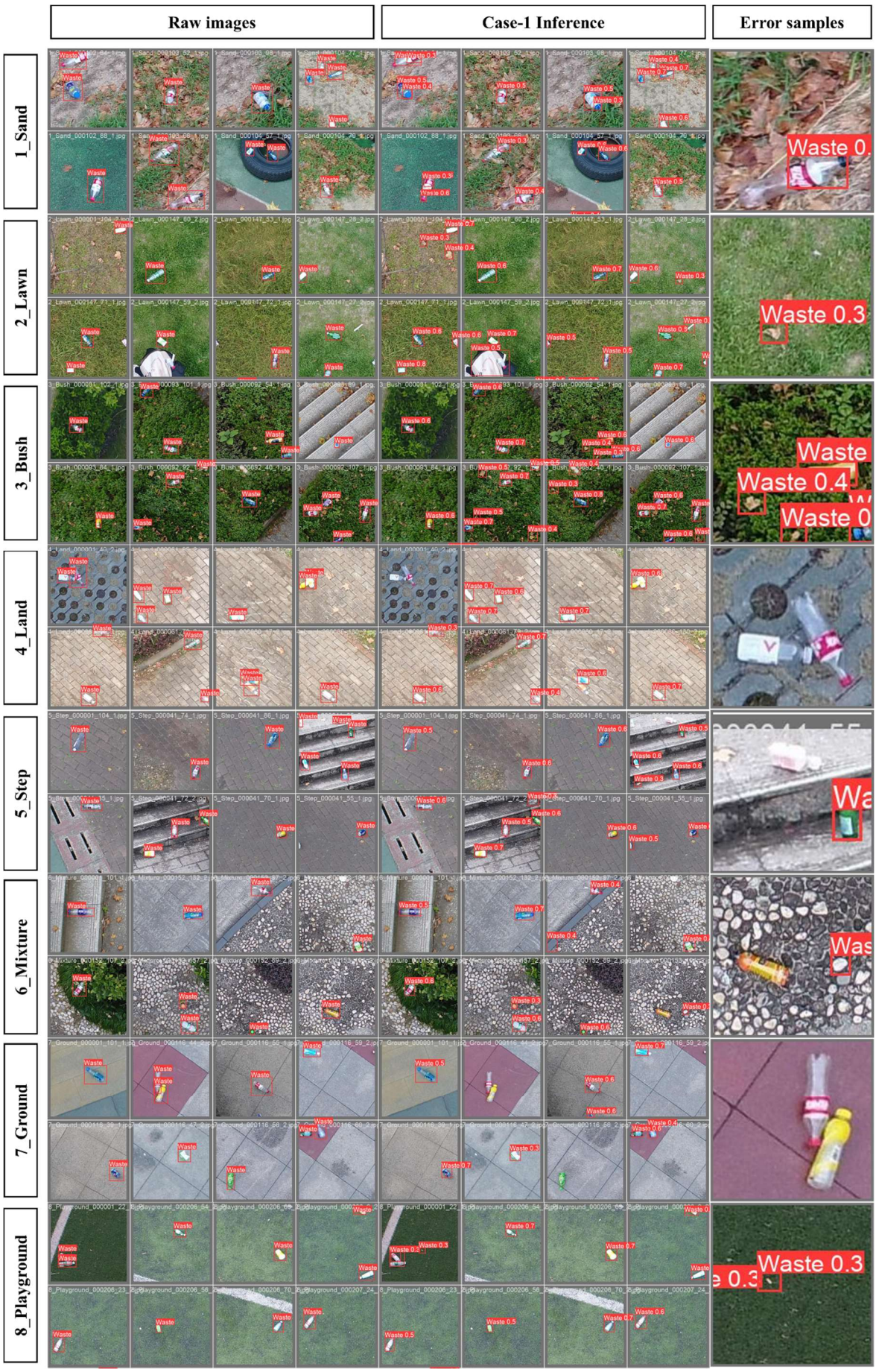

Fig.13 Samples of the results derived UAV-BD using Case 1.

Table 8 UAV-BD-derived background-based results using Case 1.

Background	Images	Instances	P	R	F1	mAP50	mAP50-95
1_Sand	2704	4630	0.758	0.800	0.778	0.806	0.385
2_Lawn	5778	8424	0.860	0.904	0.881	0.911	0.480
3_Bush	1812	3254	0.721	0.774	0.747	0.724	0.326
4_Land	1538	2365	0.709	0.587	0.642	0.636	0.269
5_Step	1325	2198	0.737	0.648	0.690	0.716	0.315
6_Mixture	3702	5205	0.698	0.625	0.659	0.658	0.272
7_Ground	4355	6246	0.692	0.644	0.667	0.664	0.262
8_Playground	4180	5178	0.889	0.862	0.875	0.907	0.442

In the near future, the more detailed and accurate prompts that can increase the accuracy of detecting the targets in complex backgrounds are looking forward to being applied in practical riparian monitoring tasks.

5. FUTURE WORK

Fig.10 has misclassified the rocks, concrete blocks, and electric wires protectors as wastes. This phenomenon has indicated the limitations of the AIGC-based Dataset, that if the non-waste targets with waste-similar-outlines in the test datasets have not been trained in the model, it is difficult for the trained model to separate the wastes and non-waste-targets. Based on the mentioned issues, in the future works, dataset supplements of the images with waste-similar-outlines are necessary.

Considering the possibility of improving the F1 value derived from the AIGC using a lower GSD value, 1.5 cm GSD 4cls RMD with detailed information has been utilized for confirmation. As shown in **Table 7**, waste pollution samples in 1.5 cm GSD 4cls RMD with 90° camera angle have been inferred by Case 1. Except for the Bikes class, all the other targets were detected with almost 1.0 F1 value using 0.45 IoU and 0.1 Confidence threshold. The reason of mis-detecting the Bikes is mainly based on the prompts. The results can be improved if prompts with more details are used.

As performed in **Fig.11**, although the Bike as a whole target has not been detected using the mentioned IoU and confidence threshold, the tire part has been seen with 0.3 Confidence. Based on this information, the Bike class can be considered to be annotated part by part to increase the accuracy, and if the IoU value can be changed from the default value used in this study (i.e., 0.45) to a lower value, the accuracy can also be improved. As shown in **Fig. 12**, all the plastic wastes have been detected, on the other hand, **Fig. 13** performed several left-unnoticed wastes in the groups of 4_Land, 5_Step, 6_Mixture, 7_Ground, individually. Respondly, as displayed in the **Table 8**, F1 value of all the groups with left-unnoticed bottles are lower than 0.7. The wastes in all the natural or similar-natural background can be detected with comparatively high F1 value derived from the prompt in this study (i.e., riparian area). In the future, it is necessary to expand the scope of the prompts in the AIGC systematically for expanding the application.

ACKNOWLEDGMENT:
This research was supported in part by the Electric Technology Research Foundation of Chugoku. 4cls RMD (train, vald and test part) were provided by Prof. Nishiyama from Okayama University.

REFERENCES

1) Pan, S., Yoshida, K., Boney, A. S., & Nishiyama, S. The Application of Drone-Assisted Deep Learning Technology in Riverbank Garbage Detection. *Journal of Japan Society of Civil Engineers, Ser. B1 (Hydraulic Engineering)*, *78*(2), I_133-I_138. 2022.
2) https://github.com/ultralytics/yolov5.
3) https://github.com/AUTOMATIC1111/stable-diffusion-webui.
4) https://github.com/OpenAI/CLIP.
5) Maharjan, N., Miyazaki, H., Pati, BM., Dailey, MN., Shrestha, S., Nakamura, T. Detection of River Plastic Using UAV Sensor Data and Deep Learning. *Remote Sensing*. 14(13):3049. 2022.
6) Wang, J., Guo, W., Pan, T., Yu, H., Duan, L., & Yang, W. Bottle Detection in the Wild Using Low-Altitude Unmanned Aerial Vehicles. *2018 21st International Conference on Information Fusion (FUSION)*, 439-444. 2018.
7) Han, W.L. UAV data monitoring plastic waste dataset. V1. *Science Data Bank*. 2021
8) Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684-10695). 2022.

(Received June 30, 2023)
(Accepted August 31, 2023)