Stable Diffusion を用いたドローン河川巡視用 AI のデータ増強と特徴量補完の可能性検証

高橋　悠太[1*]・藤井　純一郎[1]・天方　匡純[1]

[1]正会員 八千代エンジニヤリング株式会社（〒111-8648 東京都台東区浅草橋 5-20-8 CS タワー）
*E-mail: yt-takahashi@yachiyo-eng.co.jp (Corresponding Author)

土木分野における特異な物体検知はその多様さに比べてデータが少ない．ドローン河川巡視はドローンにより広大な河川領域を撮影し，一般ごみを含む不法投棄などを AI により検知することになる．常時飛行でない場合，空撮で捉えられることは少なく，一時的な不法占用等はさらに困難となる．既往の研究において，画角は異なるが地上で撮影された画像を学習データに加えることで学習を改善する効果が確認されているが，地上撮影であっても画像数が要求される．本研究では，学習を改善するデータ増強用の画像を Stable Diffusion により生成し，データ増強および空撮画像データセットにない特徴量が補完可能か検証を行った．

Key Words: *ドローン河川巡視, 対象物検知（Object Detection）, Stable Diffusion, データ増強*

1. はじめに

我が国は多くの多様性ある河川を有し，効率的な巡視のため，ドローン(UAV)と AI を用いた河川巡視の高度化が検討されている [1]．AI にとって，多様な背景の中から，多様な対象を検出する必要があるため，多様で大量の学習データを必要とする．ドローンによる空撮に関する制限は未だ多く，データを増やすことは容易でない．空撮機会が限られる場合，単純に検知対象の多様性が失われるだけでなく，一時的な不法占用などを画像に捉えることは殊更に困難となる．

ドローン河川巡視 AI の学習改善を目的とするデータ増強手法が既往研究によって提案されている．例えば，河川維持管理データベース RiMaDIS[2]に登録されている既に地上で実施されている河川巡視において撮影された画像を空撮画像に追加して学習することで学習が改善されることが確認されている [1]．ただし，この場合においても画像数や学習対象には限りがあった．そのため，本研究では Stable Diffusion[3]に注目し，生成した画像を学習に加え，学習が改善可能か二つの検証を行う．

第一に，学習改善には単純に空撮に類似した画像を生成・追加することで量的なデータ増強することが可能か検証する．既往研究 [1,4]で得られた学習セットとモデルに対し，生成画像を追加して学習し，推論結果の変化を評価する．第二に，生成できる対象であれば，学習の対象データを生成して学習することで，空撮画像データセットに不足する特徴量を補完することが可能か検証を行う．既往モデルの内，単純なデータ増強によって学習したモデルと，第一の検証で学習した提案モデルを，未学習あるいは学習ケースの少ないデータに対して適用し，それぞれ検知結果を比較する．すなわち，生成 AI による生成画像をドローン河川巡視 AI のデータ増強および特徴量の補完が可能か検証を行う．

Stable Diffusion は Latent Diffusion モデルを利用した生成モデルで，Transformer [5]を用いてテキストデータをベクター化して画像データを学習した空間に埋め込む．プロンプトと呼ばれる入力テキストにより，求める画像をモード [6]や手などの特徴量が崩壊することなく得られるように開発された．生成に用いた Stable Diffusion モデルはローカル環境で使用可能な日本語化ツールが存在し [7]，本論文では 2023 年 1 月段階での Git の訳を基準とする．使用する日本語ツールは Stable Diffusion 以外のモデルも使用可能であるが，今回は Stable Diffusion v-21-768-pruned モデル [8]を safetensor 形式で使用した．

本研究では Stable Diffusion で生成した画像を空撮画像に追加して学習して得られたモデルを，既往研究のモデルと比較する．少数の空撮データで学習した既往研究モデルでは未学習の空撮画像に対する Average Precision (AP) を，データ増強によって大量の空撮画像を学習した既往

研究モデルと未学習の空撮データに対する推論結果を定性的に比較することで，生成 AI で生成した画像をドローン河川巡視 AI の学習に用いることでデータ増強および特徴量補完が可能か検証を行った．

2. 生成・学習条件

(1) 生成条件

生成画像は Stable Diffusion の txt2img 機能を用い，プロンプトに入力して生成した．GUI として，ローカル環境で使用可能な日本語化ツール [7]を使用した．生成する画像は RiMaDIS に不法投棄として登録されることの多い対象のうち，ペットボトル(plastic bottle)，ゴミ袋(plastic bag)，三角コーン(safe cone)，段ボール箱(cardboard box)の四種類について，既往研究 [9]を参考にしてプロンプトに入力した．空撮風画像を得たいため，空撮画像(aerial image)を入力した．空撮画像の場合，対象は小さく写る必要があることから，Small と入力した．また定冠詞 a を入力し，対象が大量に生成されることを防止した．ただし 2-3 個程度の複数出現は許容した．Negative ワードに many, garbage を追加することで，同じく大量のゴミが生成され，実際の環境とも異なる上，検知結果の検証が困難な画像が生成されることを防止した．背景について，通常河川敷の状況は各河川で異なるため，今回草原(grass)，岩場(rock)，粘土(clay)，土(dirt)の四種類を入れ替えて生成した．これら背景は任意に設定可能である．また河川敷であることが強調されるよう，river side を Positive に入力した．すなわちペットボトルが岩場にある画像を生成したい場合，例えばプロンプトは a small plastic bottle, on rock field, in river side, aerial image.となる．画像サイズは 768*768 pixel，Diffusion モデルにおける拡散過程数を示すサンプリング回数は 30，生成画像のプロンプトへの依存度である CFG スケールは 14 とした．その他，画像生成に関する条件は GUI のデフォルトを使用した．本研究において，特徴量が崩壊した画像は学習の妨げとなるため，明らかな崩壊の有無を基準として，生成後にクレンジング(データ洗浄)を行った．一部三角コーンの生成画像に崩壊が見られたため追加生成し，各背景・対象毎 25 枚を最低限学習できるよう生成した．これらの生成画像は入力条件とモデル・バージョンについて，同一であれば本研究で用いた画像を再生成できる．

(2) 検知モデルの学習条件

学習データセットの配分は**表-1** に示す．以下特に断りのない限り，数値や条件については，既往研究 [1,4,9]を基に決定した．既往研究と比較することから，検知モデルは Faster R-CNN[10]を使用した．現在 Transformer を用いるなどして高精度のモデルが存在するが，発展途上であり，学習にさらなるデータが必要であるため，本研究では CNN を用いた基本的な Region Proposal Network (RPN)を持ち，精度が期待できる Faster R-CNN を用いた．生成画像のアノテーションは人間が四種類の対象物と考えられると認識した対象に行った．データ増強は空撮画像から対象物の周辺を含めて切り出す範囲をランダムに変更する，切り出した画像を反転するなどで得た．生成で得られた少数の空撮画像を学習したモデルを Benchmark，既往研究 [3]で得られたデータ増強済み学習モデルを Augmented, Benchmark の学習データセットと同程度の空撮画像データセットに生成画像を加えて学習したモデルを Proposal とする．Benchmark は空撮画像が 280 枚，生成画像は 0 枚，計 280 枚で学習されている．Augmented では空撮画像が 10209 枚，生成画像は 0 枚，計 10209 枚で学習されている．これに対し，Proposal は空撮画像 250 枚，生成画像 443 枚，計 693 枚で学習した．Benchmark と比較し，空撮画像が少なくなっているが，これは Proposal にとって不利な条件である点に留意する．学習条件は既往研究 [1,4,9]と同様の条件を用いた(**表-2**)．推論画像は未学習の空撮画像を Benchmark および Proposal でそれぞれ 50 枚，Augmented では 6807 枚を使用した．推論画像枚数はデータセット内の空撮画像枚数に合わせて決定した．学習す

表-1 学習データセットの配分

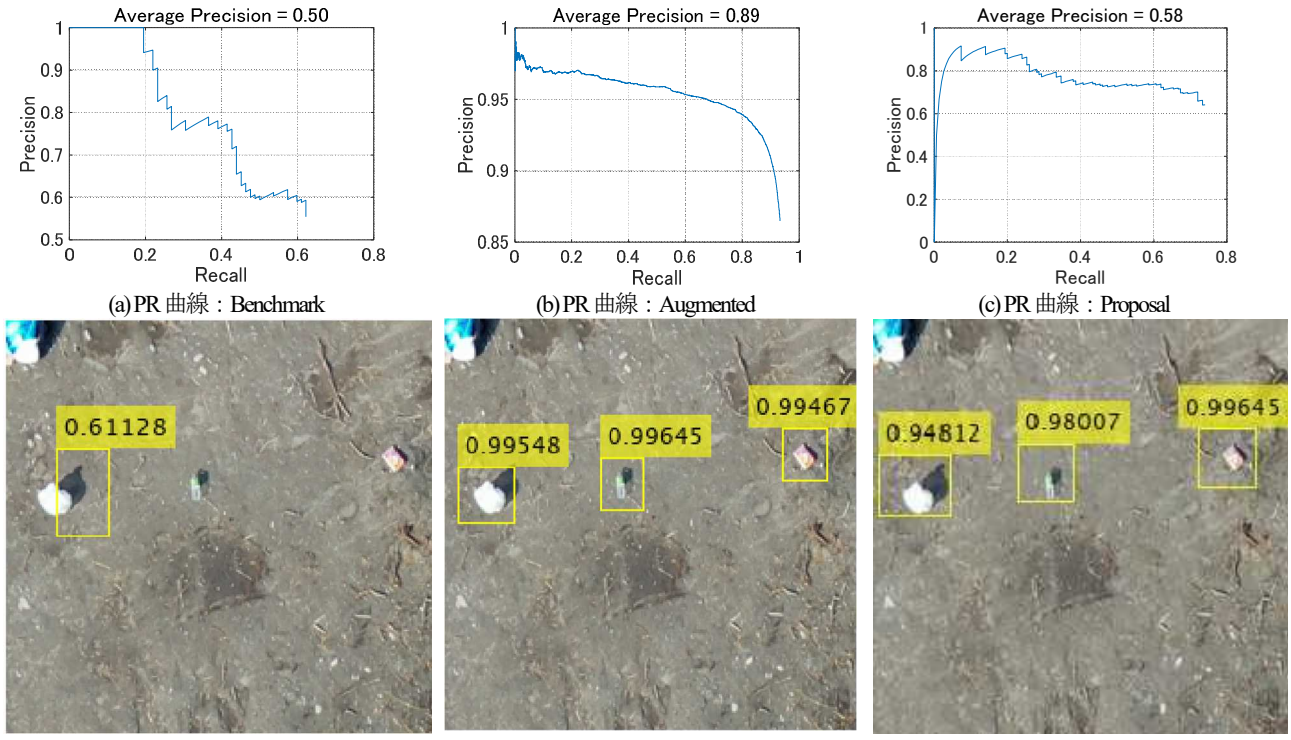
	空撮	生成	計
Benchmark	280	0	280
Augmented	10209	0	10209
Proposal	250	443	690

単位：枚

表-2 学習条件

Optimizer	SGDM
Epoch	10
Learning Rate	0.0001
Backbone	ResNet50
Extraction Layer	40 relu

図-1 生成画像：岩場の上のペットボトル



(a) PR 曲線：Benchmark　(b) PR 曲線：Augmented　(c) PR 曲線：Proposal

(d) 推論結果の一例：Benchmark　(e) 推論結果の一例：Augmented　(f) 推論結果の一例：Proposal

図-2 Benchmark, Augmented, Proposal モデルの推論結果：PR 曲線(a),(b),(c)と推論結果の一例(d),(e),(f)

る空撮画像枚数が少ない Benchmark および Proposal は 20%程度，大量に学習している Augmented は既往研究[1,4,9]を参考に 40%程度とした．Backbone に用いた ResNet50 は ImageNet で事前学習している．学習クラスは不法投棄として考えられる対象を 1 つのクラスとして学習している．空撮画像には三角コーンが含まれていないため，データセットに不足する特徴量補完の検証に用いる．

3. 検証

(1) 検証 1：Benchmark と Proposal の比較

Benchmark(既往研究の結果)と，Proposal の PR 曲線と Average precision(AP)および Benchmark と同一画像に対する推論結果を図-2 に示す．図-2(a)において Benchmark の AP が 0.50 に対し，図-2(c)が示す通り Proposal では 0.08 上昇していることが確認できる．次に同じ画像に推論を行った場合の結果を比較する図-2(d)に示す通り Benchmark ではほとんど検知できていないことが分かる．推論画像には左からゴミ袋，ペットボトル，雑誌が置いてある．この時，左上の物体は教師として設定していないため，今回検知できていなくとも問題ない．一方，図-2(f)が示す Proposal の推論結果ではビニール袋やペットボトル等が検知されている．したがって，生成画像を加えることでデータセットを増強し，学習を改善可能なことが分かる．

(2) 検証 2：Augmented と Proposal の比較

Augmented と，Proposal の PR 曲線と Average precision (AP) および Benchmark と同一画像に対する推論結果を図-2(b),(e)に示す．Augmented モデルは既往研究結果[3]を示している．AP は 0.89 となっており，Proposal と比較しても高い．推論結果の一例を比較すると，ほとんど変化はないが，検知対象を検知範囲のより中央で捉えていることが分かる．このことから，データ増強によって大量に学習データを用意することができれば，大きく AP を改善することができる．ただし，このようなデータ増強で得られたモデルは特徴量に偏りが生じ，同じ対象で異なる背景や，照度の違いによって写り方が異なる場合に検知できない過学習に陥る場合もあることに留意する．

4. 考察

(1) 検証 1：Benchmark と Proposal の比較

AP を比較すると，生成された空撮風の画像を加えただけの Proposal において改善がみられた．このことから生成画像であっても学習データ量を補完することが可能であることが確認できる．ここで，簡単に増やすことのできない空撮画像の枚数に対し，生成画像は倍近い枚数を学習に使用している点に注目する．学習に対する悪影響について，本研究の結果からは明らかでないが，今後の検討で画像枚数を増やしていく実験により検証することができると考えられる．

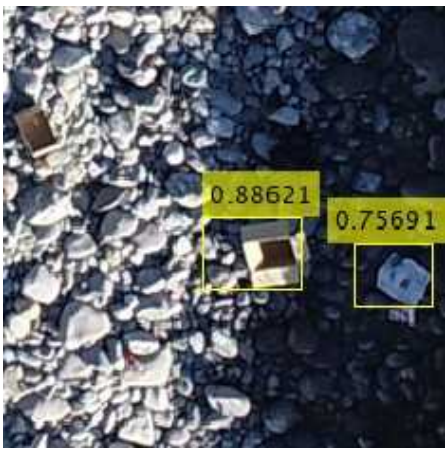
AP の改善効果について，既往研究[4]と比較しても高くはないが，既往研究において用いられた手法は本研究で検証した手法と排他的ではないため，組み合わせること

でさらなる改善効果が期待できると考えられる．

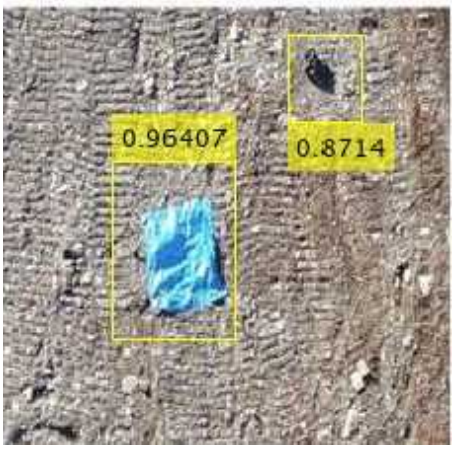
(2) 検証 2：Augmented と Proposal の比較

図-3 および図-4 にそれぞれ Augmented，Proposal モデルで推論した結果を並べて示す．推論画像は背景について，Augmented では全く学習していない．黄色枠内は Confidence を示す．閾値は図-4(c)，(d)が 0.16，それ以外は既往研究[4]と同様の 0.6 以上とした．図-4(c)，(d)において，対象となる物体が検知される最低の閾値を基準として 0.16 に設定した．図-3 について，Augmented で推論した(a)に対し，極めて少数の空撮画像と生成画像で学習した Proposal では右側に写る新聞紙を検知できている．これは Augmented の学習データセットで相対的に少なかった特徴量について，生成画像を学習することによって補完できていることが示唆される．一方，図-3(c)，(d)においては Proposal では右上に写る黒いゴミを検知できていない．そのため，空撮画像が少なすぎる場合については，生成画像のみによる補完効果は期待できないことを示唆している．図-4(a)，(b)について，Augmented では学習データセットに含まれていなかった三角コーンが検知できているかを検証する．Augmented では三角コーンが検知できていないが，生成画像で三角コーンを学習した Proposal では検知できていることから，生成画像学習によるデータセットに不足する特徴量補完の可能性，すなわち，空撮画像に捉えることができなかった対象であっても，生成画像を学習することで補完できる可能性が確認できる．最後に図-4(c)，(d)では Augmented，Proposal 共に検知対象を検知できていない場合を示す．これらの画像について，類似する画像を Augmented では学習されておらず，Proposal では生成画像によって学習している．ただし，対象物が背景の岩場に紛れ，また日照によって判別が難しくなっているため，生成画像による補完が十分機能していないことを確認できる．これらの結果から，AP を比較すると Proposal は汎化性が十分とは言えないが，生成画像による特徴量の補完可能性が示されたと考えられる．加えて，全く学習していない背景画像に対し，Augmented は Proposal と同程度の検知結果となっている点について，空撮画像のデータ増強は効果的であるが，汎化性を高めることを考慮すると，今回学習した枚数(10209 枚)も必要ではなく，本実験における三角コーンのように，データセットに不足する特徴量を生成画像によって補い，学習データを圧縮しながら汎化性を高められ

(a) Augmented で推論



(b) Proposal で推論



(c) Augmented で推論

(d) Proposal で推論

図-3 Augmented, Proposal モデルの推論結果比較：Proposal で改善(a),(b)，Proposal で一部失敗(c),(d)

(a) Augmented で推論

(b) Proposal で推論

(c) Augmented で推論

(d) Proposal で推論

図-4 Augmented, Proposal モデルの推論結果比較：Proposal で検知可能(a),(b)，両モデル共に検知失敗(c),(d)

る可能性が示唆される．

5. まとめ

ドローン河川巡視 AI に必要となる学習データを増強するため，Stable Diffusion で生成した画像を学習データセットに追加し，学習が改善可能か，2 つの検証を行った．第一検証として，空撮画像のみで学習した Benchmark モデルと生成画像を追加した Proposal モデルで未学習画像に対し推論を行った．第二検証として，データ増強した空撮画像を大量に学習した Augmented モデルと Proposal モデルを用いて推論結果を比較した．この時，推論画像は Augmented モデルでは全く未学習の背景・対象の画像だが Proposal モデルでは生成画像で対象の特徴量を学習している．上記の検証から得られた知見を以下に示す．

1) 第一検証において，Benchmark および Proposal モデルの PR 曲線から得られた Average Precision と同一画像への推論結果の定性的比較から，生成画像追加による増強効果を確認した．

2) 第二検証において，Augmented モデルと Proposal モデルの推論結果を定性的に比較し，Augmented モデルでは検知できなかった対象に対して Proposal モデルが検知できる場合があることを確認し，単なるデータ増強だけでは不足する特徴量の多様性を生成画像で補完できる可能性が示唆された．

今後の課題として，他の生成モデルで生成した画像を用いた場合との比較や，実写画像と生成画像の割合の影響を考慮した検証が考えられる．また，今回の検討では生成画像は画像サイズに対して大きく写ってしまい，実際の空撮画像サイズとは異なる．サイズや位置を調整するにはプロンプトエンジニアリング [11]する，あるいは生成 AI によってさらに背景を延伸させるといった手法が考えられ，これら手法を活用した性能向上手法についても検討する．

参考文献

1) 高橋悠太，藤井純一郎，天方匡純，山下隆義：UAV と画像認識 AI による河川巡視を補う地上画像の特徴量とその利用法検討，AI・データサイエンス論文集，第 1 巻，J1 号，pp.580-587，2020.
2) 森永泰司，鈴木克尚，沼田太郎，山口修平，長坂健，星尾日明：RiMaDIS の構築と運用および今後の展開，河川技術論文集，第 26 巻，2020.
3) Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, arXiv:2112.10752, 2021.
4) Takahashi, Y., Fujii, J., Amakata M., and Yamashita, T.: An application of AI technology to UAV's river patrol and the features value of datasets, SHMII-10, 2021.
5) Vaswani, A., Shazeer N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, arXiv:1706.03762, 2017.
6) Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial nets, In Advances in Neural Information Processing Systems 27, pp 2672–2680. Curran Associates, Inc., 2014.
7) https://github.com/AUTOMATIC1111/stable-diffusion-webui
8) https://huggingface.co/stabilityai/stable-diffusion-2-1
9) 高橋悠太，藤井純一郎，天方匡純：StableDiffusion などの画像生成 AI を用いたドローン河川巡視用 AI のデータ増強，人工知能学会全国大会，2023.
10) Ren, S., He, K., Girshick, R., and Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint, arXiv:1506.01497, 2015.
11) Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, and GLIGEN, Y. J.: Open-Set Grounded Text-to-Image Generation, arXiv:2301.07093, 2023.

(Received June 30, 2023)
(Accepted September 15, 2023)

The data augmentation and supplement of feature for UAV river patrol AI using generated image by Stable Diffusion

Yuta Takahashi, Junichiro Fujii, and Masazumi Amakata

The data in the civil engineering field is less data with much variety. Drone river patrols should fly the vast river areas and AI must detect illegal dumping, including general garbage. The patrol drones are not constantly flying and they are rarely captured by aerial images. Thus, it is even more difficult to detect temporary target such as illegal occupation. Previous study has confirmed that the addition of images taken on the ground with different angles of view to the learning data improves learning. However, the number of images is required for training even if the images are taken on the ground. In order to improve the learning of the detection model, this study verified whether the image for data augmentation and the supplement of feature which is less in dataset by generated images by Stable Diffusion.