1T-03

合成画像を入力としたPix2Pixモデルによるセグメンテーション画像拡張手法の提案

中根睦仁[†1] 平原健太郎[†2] 黒田剛士[†3] 岩城洋平[†3] 内海智仁[†3]
野村祐一郎[†4] 峰野博史[†4, †5]

静岡大学情報学部[†1] 静岡大学院総合科学技術研究科情報学専攻[†2] ヤマハ発動機株式会社[†3]
静岡大学学術院情報領域[†4] グリーン科学技術研究所[†4, †5]

1. はじめに

近年，農作物のセグメンテーション画像を用いた研究が盛んにおこなわれている．セグメンテーション画像を用いた研究を行うためには，多様なデータの大量収集や対象物1つ1つにピクセル単位でのセグメンテーションを行うため，すべての収集データに対して作業を行うには大量の時間や労力がかかる．

そこで，本研究では合成画像を入力としたPix2Pixモデルによるセグメンテーション画像拡張手法を提案する．高解像度の実圃場画像に対して512*512のパッチ化を施して学習したControlNet[1]に対して，エッジ化したワインブドウの食品サンプル画像とエッジ化した実圃場画像を合成する．これをControlNetの入力とし生成した画像を実圃場画像の一部に合成することで，セグメンテーション画像を拡張する．この手法を用いた画像拡張によって必要なセグメンテーション画像は少なく済み，研究者のセグメンテーションにかかる時間や労力の低減を目指す．

2. 関連研究

2.1 ControlNet

ControlNetは，拡散モデルであるStable Diffusion[2]をベースに作られており，画像の構図を保ったまま画像生成が可能になったモデルのことである．しかし，高解像度画像を学習する際，画像サイズを512*512にリサイズするため，高解像度画像が縮小し画質の低下や，意図した生成が行われない課題が存在する．

2.2 動物と背景との合成画像を入力とした画像生成

動物と背景との合成画像を入力とした画像生成にはBi-ControlNet[3]がある．動物のポーズ推定を行うモデルPASy-nにControlNetを組み込んだSPAC-Netの中に存在するBi-ControlNetは，動物と背景のHED境界を別々に検出することで，生成データの精度と安定性を向上させることができる．シマウマのように明確な縞模様を示す動物に対して，実画像のみを使用した場合とSPAC-Netを適用したデータセットの平均精度を比較した結果，実画像のみは78.7%に対して，SPAC-Netを適用したデータセットは96.3%を示した．動物と背景を対象とした合成画像を入力として用いる画像生成を行っている研究は存在するが，植物と背景に対して上記内容が適用された例はない．†

3. 提案手法

3.1 概要

本研究では，植物と背景に対して合成画像を入力としたControlNetによるセグメンテーション画像拡張手法を提案する．提案手法は大きく4つのステップから成る．

3.2 画像のエッジ化

第1ステップでは，実圃場を撮影した画像を512*512でパッチ化を行った後にワイン圃場のエッジ画像を生成する．これによってブドウの房や葉などの特徴をより多く抽出することができる．撮影したデータセットはfront，up，sideの3種類の画角を用いる．frontは圃場に対して斜め前から撮った画像，sideは圃場に対して垂直にとった画像，upは圃場を下から見上げるようにとった画像である．エッジ化にはCanny[4]を使用する．

3.4 対象画像学習

第2ステップでは，第1ステップで得たエッジ画像を入力画像，元画像を正解画像として，その入出力関係をControlNetで学習する．拡張という観点からより実圃場に近い画像を生成するため今回は学習済みのモデルに対して追学習を行う．

3.5 画像拡張

第3ステップでは追学習したControlNetに対してエッジ化されたセグメンテーションブドウの食品サンプル画像とエッジ化された実圃場画像の一部を合成した画像を入力とすることでセグメンテーション用画像の生成を行う．なお食品サンプルの房画像に対してはセグメンテーション情報を基にセグメンテーション部分以外を透明化した．合成に関しては実圃場画像のセグメンテーション部分にエッジ画像同士の合成を行う．これにより房のセグメンテーション情報を保持したままブドウの房を背景に馴染むように画像を生成できる．

Proposal of Segmentation Image Augmentation Method using Pix2Pix Model with Composite Images
†1 Chikahito Nakane, Faculty of Informatics, Shizuoka University
†2 Kentaro Hirahara, Graduate School of Science and Technology,Shizuoka University
†3 Tsuyoshi Kuroda , YAMAHA Motor Co., Ltd
†3 Yohei Iwaki, YAMAHA Motor Co., Ltd
†3 Tomoyoshi Utsumi, YAMAHA Motor Co., Ltd
†4 Yuhichiro Nomura, College of Informatics, Academic Institute, Shizuoka University
†4, †5 Hiroshi Mineno, College of Informatics, Academic Institute, Shizuoka University, Research Institute of Green Science and Technology, Shizuoka University

3.6 生成画像を既存のセグメンテーションモデルに適用

第 4 ステップでは，生成した圃場画像を実圃場画像と組み合わせたデータセットを作成し，既存のセグメンテーションモデルに適用することで，本手法の有効性を示す．

4. 実験

4.1 Canny を用いたエッジ画像の生成

画像サイズ 5120*2880 で撮影した実圃場画像に対して 1 枚あたりの画像サイズが 512*512 になるように画像を切り出した．このとき高さに余りが出てしまうため元画像 1 枚に対して幅 10 枚，高さ 5 枚の画像に切り出した．切り出した画像 1440 枚に対して Canny を用いてエッジ化した．エッジ化前とエッジ化後の画像をそれぞれ対応付け，1420 枚を学習画像，20 枚を検証用の画像として ControlNet での学習用データセットとした．

4.2 ControlNet を用いたセグメンテーション画像の生成

作成した学習用データセットを ControlNet に入力して学習を行った．学習率を 2e-4，バッチサイズを 8，ステップ数を 10000 として学習した．最も高精細な画像を出力するモデルを使用するために生成された画像を LPIPS[5]，DreamSim[6]を用いて評価しそのスコアが良いステップ数のモデルを使用した．LPIPS，DreamSim は従来の画像評価指標に比べ，色味や構図を重視し，より人間の知覚や整合性に似た評価指標である．圃場を生成するにあたり，様々な画角や形状のブドウを生成するため上記の評価指標を用いた．LPIPS ではステップ数 4700 時に最高スコア 0.24 を示し，DreamSim ではステップ数 5200 時に最高スコア 0.19 を示した．図 1 に各評価指標最高スコア時の生成結果を示す．生成された画像に対して，房や葉の大きさが元画像から乖離していない点と，房や葉の生成箇所が元画像と同じ箇所にできている点から，エッジ画像に沿った自然なブドウ圃場を生成可能であることが示された．図 2 に画像サイズ 5120*2880 と 512*512 にパッチ化した画像で生成品質を比較した．画像サイズ 5120*2880 の実圃場画像とパッチ化した画像実圃場画像のエッジ化，生成画像の比較を示す．画像サイズ 5120*2880 の方は 512*512 にリサイズして学習しているため，エッジが潰れブドウと葉部分の境界がとても難しい．しかし，パッチ化した画像の方では，葉部分とブドウ部分のエッジを十分に捉えることができており，より正確に画像を生成できていた．これにより房や葉の細かい部分まで生成が可能であることを確認できた．

5. おわりに

本研究では合成画像を入力とした Pix2Pix モデルによるセグメンテーション画像拡張手法の提案を行った．課題であった高解像度画像の学習，生成をパッチ化することにより定性的にも，定量的にも高精度なブドウ画像が生成できることが示せた．

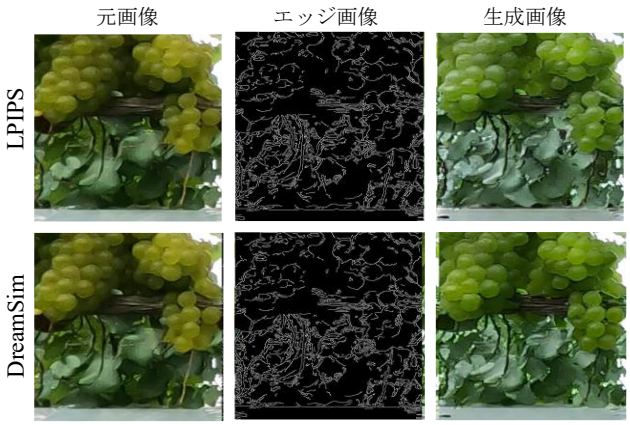

図 1 各評価指標最高スコア時の生成結果

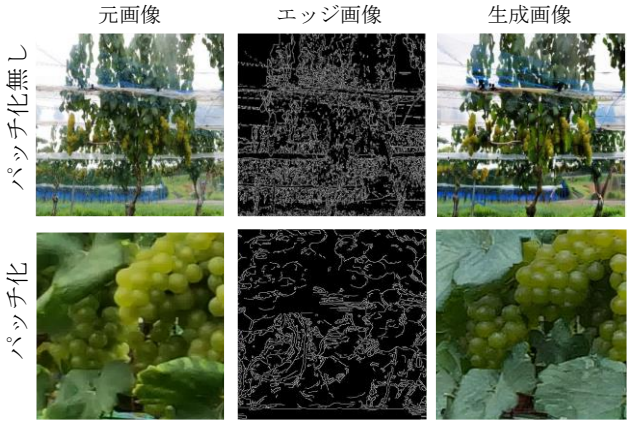

図 2 パッチ化の有無による画像の生成結果

今後の課題としては，食品サンプルを用いた画像を使用した画像生成や，実圃場画像と生成画像を組み合わせたデータセットを既存のセグメンテーションモデルに適用した際の評価，学習モデルの改善，より自然な背景を持つ圃場画像生成手法を検討していく．

謝辞

本研究の一部は，JST 創発的研究支援事業ならびに静岡大学グリーン科学技術研究所プロジェクト研究支援を受けたものである．

参考文献

[1] Zhang, L., Rao, A. and Agrawala, M.: Adding conditional control to text-to-image diffusion models, Proc. IEEE/CVF ICCV, pp. 3836- 3847 (2023).
[2] Stability AI: Stable diffusion v2 model card, stable-diffusion2-depth, Hugging Face, Hugging Face(online),available from <https://huggingface.co/stabilityai/stable-diffusion-2-depth> (accessed 2023-11-02).
[3] Jiang, L., Ostadabbas. S.: Synthetic Pose-aware Animal ControlNet for Enhanced Pose Estimation, arXiv:2305.17845 (2023).
[4] OpenCV: Canny Edge Detection, OpenCV(online), available from <https://docs.opencv.org/4.x/da/d22/tutorial_py_canny.html> (accessed 2023-11-02).
[5] Zhang, R., Isola, P., Efros, A.A., et al.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, Proc. *IEEE CVPR*, pp.586-595 (2018).
[6] Fu, S., Tamir,N., Sundaram, S., et al.: DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data, arXiv:2306.09344 (2023).