〔様式第４号の１〕

令和6年　3月２１日

令和５年度　学生自主研究成果報告書

教　育　本　部　長　様

学生自主研究グループ名	Fake Photo	
研究課題名	画像セグメンテーションにおける，架空画像のみを使った学習および性能向上の検討	
研究代表者（学生）	学籍番号	B25P019
	氏　名	小西　遼翔
指導教員	学　科	知能メカトロニクス学科
	氏　名	伊　藤　　亮　准教授

学生自主研究の報告書を別紙のとおり提出します。

〔様式第４号の２〕

画像セグメンテーションにおける，架空画像のみを使った学習および性能向上の検討

学部名　学科名
システム科学技術学部情報工学科１年
小西　遼翔

指導教員　学部名　学科名
システム科学技術学部　知能メカトロニクス学科
伊藤　亮　准教授

1　はじめに

セグメンテーションモデルの深層学習には、画像と対応した教師画像が必要である。バーチャル環境を利用しこれらを同時に作成することが可能である。通常セマンティックセグメンテーションモデルの学習には人が対象物の範囲を示した教師画像を用いる。そのため、大量の画像を必要とする深層学習を行う際にバーチャル画像を用いるとアノテーションの作業に必要だった工数が削減され、より短期間で学習用の画像を用意することができる。よって、バーチャル画像の利用はセグメンテーションモデルの準備期間の短縮につながる。しかし、バーチャル画像をそのまま使用すると実際の使用場面での精度が低い。そこで、バーチャル画像を加工することで精度の向上があるのか検証する。加工手段として Stable Diffusion の img to img を用いる。

2　研究方法の説明

2.1　画像の準備

深層学習を行うための教師画像としてグレースケールの教師画像を作成した。トマトの果実付近の幹、トマト果実、トマトのへたの 3 つを明度の違いとして示した教師画像を作成した。未加工の場合と SDXL を使用した場合のどちらもこの教師画像を用いる。

評価用の画像として実際のトマトの画像を用意し、8 枚の画像に対し LabelMe を使用してアノテーションを行った。LabelMe とは画像アノテーションを行うオープンソースのソフトウェアである．画像上の物体を囲み，クラス名を指定すると json ファイルが出力される．この Json ファイルを Png ファイルに変換して評価用のデータとして用いる．

2.2　画像の加工

使用するバーチャル画像は 1024 ピクセル四方、枚数は 3000 枚である。これらの画像に対して Stable Diffusion XL（以後 SDXL と記述）による処理を行った。表 1 に指定内容を示す。

2.3　　モデルの学習

2 つの学習モデルでセマンティックセグメンテーションモデルの追加学習を行った。ファインチューニングという手法を用いて学習済みのモデルに新たに層を追加し、モデル全体の重みの更新を行った。用いる Unet モデルは画像データベースである Imagenet でエンコーダを事前学習している。

モデルによる学習結果の差があるかを確認するために mobilenetv3_Large と efficientnet-b3 のモデルを用いてそれぞれ学習を行った。

表 1　画像生成時の指定内容

サンプリングメソッド	DPM++ 2M Karras
サンプリングステップ数	30
生成画像サイズ	1024
CFGスケール	7
Denoising strength	0.75
ControlNet	cany（入力した画像の線画を出力）
Cany Low Threshold	100
Cany How Thresholc	60

3　　結果

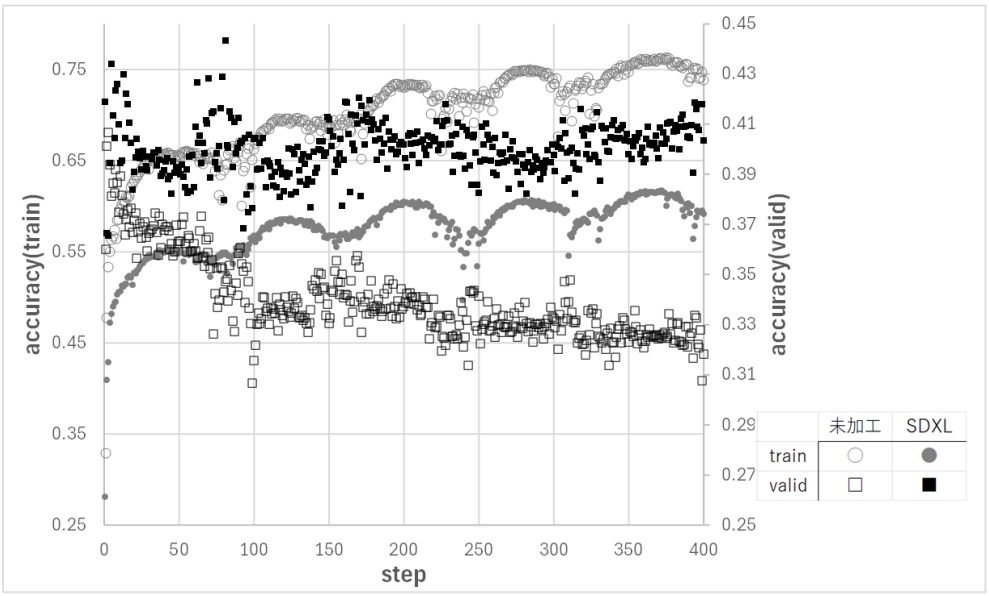

図 1　SD の処理の有無による学習の結果(mobilenetv3_Large)

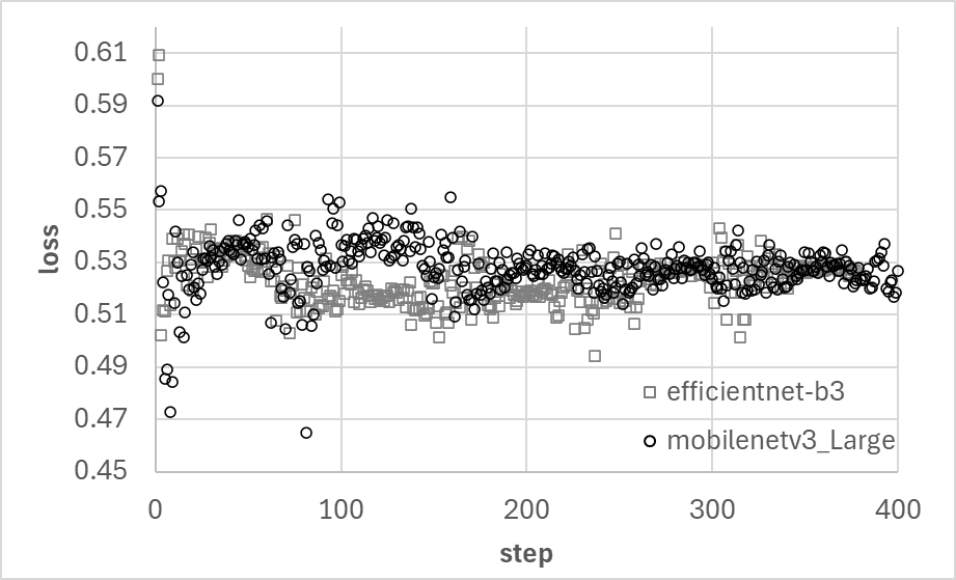

図 2　SDXL を使用した際の U-net のモデルによる違い

図 1 では学習が進むごとのモデルの推論結果の正確性について 4 項目を表示している。セグメンテーションモデルの学習では正確性の値が大きいほど良い。

valid_未加工と valid_SDXL を比べると全体的に SDXL による加工をしたほうが accuracy の値が大きい。step が進むごとに accuracy の値が減少している。一方、SDXL による処理をすると過学習が起きずに、0.65 付近の値を取り続けている。

train_未加工と train_SDXL を比べると未加工のほうが同じ step 数のときにより高い精度を持っている。step 数が増えるにつれて差がより大きくなっている。

図 2 では加工した画像を用いた学習を行った際の valid の値である。どちらも 260step からは値の変化は少ないが、efficientnet-b3 は 100step の時点で loss の値が安定している。

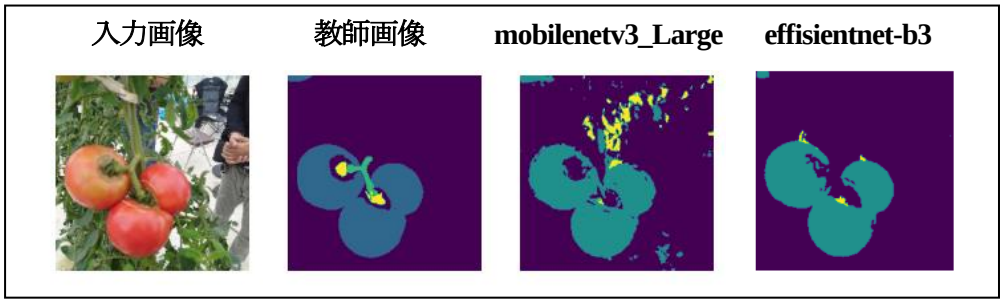

図 3 から effisientnet-b3 を用いると誤認識が少ないことが読み取れる。トマトの果実は認識できているが幹やトマトのへたの部分はほとんど認識されていない。

4　考察

4.1　SDXL の使用による変化

SDXL で加工をしたことによって学習画像への最適化がされにくくなった。それによって実際の写真に対する正確性である valid の値が変化しにくくなったと考えられる。未加工の画像を用いた場合に学習が進むごとに valid の正確性が減少した原因はバーチャル画像への最

適化が進み現実のトマトの画像と共通してみられる特徴が減少したことだと考えられる。SDXL の加工によって学習画像と現実のトマトとの差異が減少したにも関わらず valid の値が train の値ほど増加しなかった原因として SDXL での加工時に画像内でのトマトの位置がずれたことが考えられる。Control net の cany を用いて入力画像の線画をもとに画像を生成させた。この作業は位置を変えずに本来のトマトの色に近づけることを目的としている。しかし、まれにトマトが消えたりなかった場所に生成されたりすることがある。学習画像にずれが生じるとトマト以外の部分をトマトと認識し、反対にトマトをトマトと認識できないようなモデルとなる。とくにへたや周辺の幹は画像に占める割合が小さく、SDXL の加工により大きくずれたと考えられる。

モデルの差は 400step の学習を終えた時点では見られないが、efficientnet-b3 はより早く loss の値が安定する傾向がある。同じ画像による学習でも学習時の画像サイズの変化によって過程が変化していることがわかる。400step の学習を終えた時点では loss の値に差はないが、図 3 に示すように誤検知が少ない。位置の特定を目的に使用する場合、クラスの誤検知は精度に大きく影響するため efficientnet-b3 が適しているといえる。

4.2 課題に対する提案

上記の課題はトマト、へた、周辺の幹のそれぞれの正解画像を ControlNet に読み込ませることで解決できる。特定クラスの明度を上げて正解画像の示す範囲を囲うように線画を出力させることで出力時に画像のずれが抑えられる。

5 まとめ

研究の目的であるセグメンテーションモデルの精度向上は達成できなかった。SDXL の加工による変化は過学習が起きないことだった。また、学習に使用するモデルによっては数値上では精度が変わらなくても、クラスの誤検知が発生しづらいという差がみられた。