H-023

Vision Transformerを用いた鍛造部品の不良品検出
Detection of Defective Forged Parts Using Vision Transformer

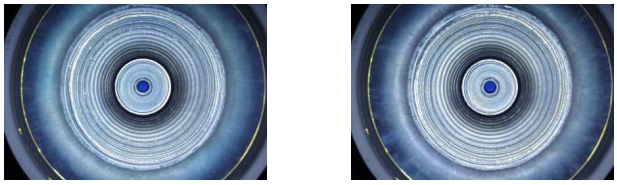
高木 裕也* 藤田 和弘* 中川 真言† 世継 武志†
Yuya TAKAGI Kazuhiro FUJITA Makoto NAKAGAWA Takeshi YOTSUGI

1 はじめに

工業製品において外観検査は，安全性や信頼性などの観点から必要な工程である．しかし，現在，外観検査の多くは人間による目視で行われている．そのため，検査員の習熟度や疲労などによる検査精度のバラツキから，品質を一定に保てないことや，人手不足といった課題が挙げられる．そこで，画像検査による外観検査の自動化によって，これらの課題を解決することを目的としてVision Transformer(ViT)を用いた鍛造部品の不良品検出について研究を行った．また，鍛造部品の製造で用いられるプレスの金型が摩耗することで，製造時期によって表面テクスチャが変化することが考えられる．そのため，画像検査機の再学習が求められるが，再学習のスパンが短いとコストがかかるためロバストなモデルが求められる．そこで，近年，画像タスクにおいて優秀な成績をあげているViTと，従来から用いられてきた畳み込みニューラルネットワーク(CNN)について学習外データについてのロバスト性の比較を行う．

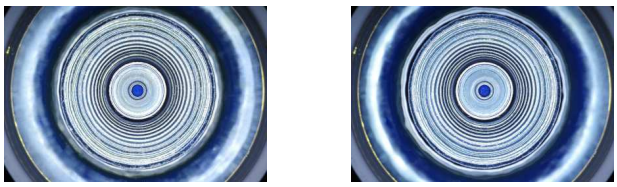
2 データセット

本研究では，高橋金属(株)提供の鍛造部品の底面部と側面部が写っている画像をデータセットとする．側面部について傷や欠け，打痕が見られる不良品画像と，それらが見られない，または無視できる程度の大きさである良品画像でラベル付けされている．用いるデータセットは大きく分けて3つあり，それぞれ2022年に撮影されたデータ，2021年に撮影されたデータ，2020年に撮影されたデータがある．2022年のデータは，2,304画素×1,536画素の，2,959枚(不良品画像:984枚，良品画像:1,975枚)である．この内，学習用データ2,048枚，検証用データ256枚，評価用データ655枚に分割されている．不良品画像の例を図1(a)，良品画像の例を図1(b)に示す．



(a)不良品画像 (b)良品画像
図1. サンプルデータ

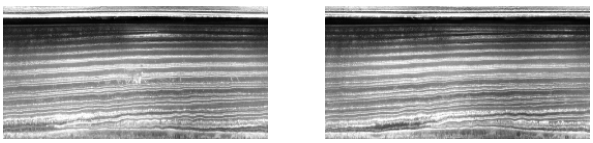
2021年のデータは，4,608×3,072画素の画像150枚(不良品画像:50枚，良品画像100枚)，2020年のデータは，4,608画素×3,072画素の画像294枚(不良品画像:152枚，良品画像:142枚)，2021年に撮影した良品画像を図2(a)，2020年に撮影した良品画像を図2(b)に示す．



(a)2021年，良品画像 (b)2020年，良品画像
図2. サンプルデータ

3 手法

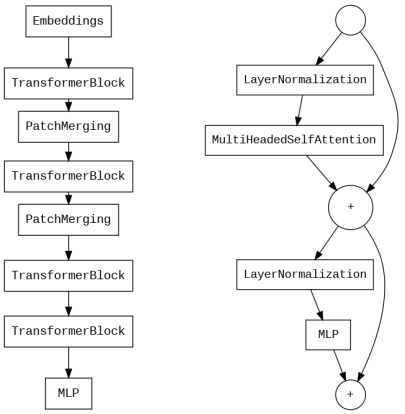
データセットの不良品画像は側面部についての欠陥だけなので，原画像から側面部のみが写った画像に変換する．原画像を鍛造部品の中心を原点として極座標変換を行い，横軸を角度，縦軸を動径とした178×256画素にリサイズし，上部50画素を除去することで，128×256画素の画像にする．その後，グレースケール化を行う．変換後の不良品画像の例を図3(a)，変換後の良品画像の例を図3(b)に示す．



(a)変換後の不良品画像 (b)変換後の良品画像
図3. 前処理後の画像

*龍谷大学, Ryukoku University
†高橋金属株式会社, Takahashi Metal Industries Co., Ltd.

学習用データ2,048枚を用いてViTと，CNNを用いて学習済みモデルを作成する．CNNモデルにはSEResnetを採用した．本研究で用いたViTの構造を図4に示す．



(a)ViT　　(b)TransfomerBlock

図4. **本研究におけるViTの構造**

PatchMerging層によって，2×2の隣接するパッチを1つのパッチにまとめることで，様々な画像サイズから特徴量を学習するようにしている．

4　識別実験結果

作成した学習済みモデルを用いて，2022年の評価用データ，2021年のデータ，2020年のデータについて，良品画像および不良品画像に対して識別を行った．それぞれのデータに対するAUCの値を表1に示す．

表1. **ViTとCNNのAUC**

	2022年	2021年	2020年
ViT	1.00	0.98	0.86
CNN	1.00	0.64	0.66

表1より，ViTとCNNともに学習データと同じ時期に撮影された2022年評価用データはAUCが1.0であり，高い精度に識別ができている．2021年のデータの関しては，ViTのAUCは0.98，CNNのAUCは0.64であり，ViTは高い値となっている一方で，CNNは大きく値が低下している．2020年のデータについては，ViTのAUCが0.86である．また，CNNのAUCは0.66であり，ViTとCNNどちらもAUCの値が低下しているが，ViTの方がCNNよりもAUCが高い．

ViTとCNNについて，それぞれのデータに対する識別のROC曲線を図5に示す．

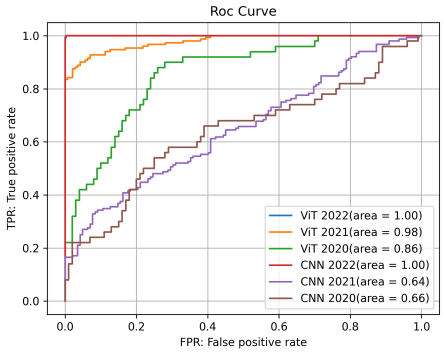
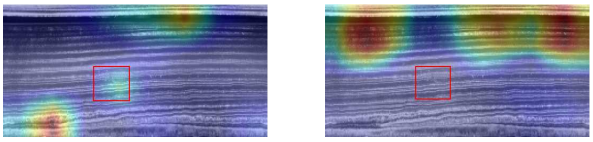


図5. **ROC CURVE**

図5より，2022年の評価用学習データと比べて，不良品画像の誤識別率を0に抑えた上での良品画像の誤識別率が，学習データとは異なる撮影時期のデータだとViTとCNNどちらも増加している．しかし，ViTはCNNよりも良品画像の誤識別率は低い．

次にViTおよびCNNの識別における注目箇所の可視化を試みた．ViTの4層あるTransformer Blockの内，3層目のTransformer BlockのAttention Mapの平均を図6(a)，CNNのGrad-CAMの結果を図6(b)に示す．



(a)ViT　　(b)CNN

図6. **注目箇所の可視化**

図6より，CNNは傷とは関係のない部分に注目しているように見られる．ViTも傷とは関係のない部分について注目しているが，傷の付近ついても注目している．

5　まとめ

本研究では鍛造部品の側面部についてViTとCNNを用いて，不良品画像および良品画像について識別を行った．学習データと同時期に撮影された評価用データに対する識別結果としてはViT，CNNともに正確な識別が行えた．学習データと異なる撮影時期の鍛造部品のデータについては識別精度が低下してしまうが，ViTはCNNに比べて精度の低下が抑えられており，よりロバストなモデルが作成できた．今後は，学習データと異なる撮影時期のデータに対して，不良品画像の誤識別率を0に抑えた上での良品画像の誤識別率を低くできるような学習方法や，アーキテクチャについて検討していきたい．