広島県立総合技術研究所東部工業技術センター研究報告　No. 38（2025）
技術ノート

事前深層学習モデルの転移学習による能力比較（第 2 報）

DINO

廣川　勝久

Transfer Learning and Fine-tuning for Pre-trained Deep-learning Models (II)

DINO: Self-distillation with No Labels

HIROKAWA Katsuhisa

本報告では、自然言語処理用の Transformer 技術を画像認識に応用した Vision Transformer の自己教師あり事前学習手法である DINO: Self-distillation with No Labels [arXiv:2104.14294 (2021)]を実装する。Vision Transformer は高い性能を示す一方で、従来は大量のラベル付きデータを必要とするという課題があった。DINO はラベルを用いずに学習が可能な手法であり、本報告では、学習後の Vision Transformer が生成する Attention Map が、ラベルなしにもかかわらず画像中の物体領域を識別可能であることに着目し、背景など画像の特徴の違いが物体領域検出能力に与える影響について評価する。

キーワード：DINO、 self-distillation、 ViT、 Transformer、 Attention Map

1. 緒　　言

近年、Transformer は多層パーセプトロン（**M**ulti-**L**ayer **P**erceptron: **MLP**）や畳み込みニューラルネットワーク（**C**onvolutional **N**eural **N**etwork: **CNN**）に続く新たな深層学習技術として注目を集めている [1]。Transformer は、もともと自然言語処理を目的として開発されたモデルであり、CNN において ReLU 関数やスキップ接続の導入によって多層化が可能となったのに対し、Transformer は Scaling Law の発見により、ChatGPT に代表されるような大規模言語モデルの実現を可能とした [2]。

こうしたモデルの大規模化に伴い、高性能な深層学習モデルの知識を、より軽量なモデルへ効率的に転移させるための技術も求められており、知識蒸留はその代表的手法の一つである。従来の知識蒸留では、大規模な学習済みモデル（教師モデル）を用いて、小規模なモデル（生徒モデル）を学習させることで、学習済み知識を圧縮・転送することが可能となる。

本稿で検討する Self-**di**stillation with **No** Labels: **DINO** は、蒸留を応用した手法であるが、モデルの圧縮ではなく、ラベル付けを必要としない自己教師あり事前学習を目的としている [3]。DINO は、ラベルのない大規模な画像データに対して学習を行うことが可能である。特に、基盤モデルとして **V**ision **T**ransformer: **ViT** [4]を用いた場合には、得られた Attention Map が画像中の物体領域を示すことが報告されており、本研究では、DINO の物体領域検出能力に着目し、背景の有無や画像解像度の違いがこの能力に与える影響について評価を行う。

2. DINO

2.1 DINO の概要

ViT は Transformer の Encoder モジュールから構成されており、その性能を十分に引き出すためには、大量の学習用画像データが必要とされる [5,6]。しかし、これらのデータに対するラベリング作業は極めて高コストであり、大きな障壁となっている。本稿で報告する DINO は、ViT のような画像モデルに対してラベル不要の自己教師あり事前学習手法であり、画像を入力するのみで学習を進めることが可能である。これにより、ラベリングに要するコストを大幅に削減できる。一度 DINO により事前学習を完了させれば、次に少量のラベル付き画像を用いることで、転移学習やファインチューニングによる画像モデルへの適用が可能となる。さらに注目すべきは、ViT を用いた DINO 学習において、得られた Attention Map が画像中の物体の領域を表している点である。各物体はクラスタリングされるとともに、Attention モジュールが物体の領域を捉える能力を示すことが確認されている。この性質は、学習後の ViT を基盤モデルとしてセマンティックセグメンテーションモデルなどへ実装することも期待される。

2.2 DINO の基本構造

DINO は、教師モデルと生徒モデルの 2 つのニューラルネットワークから構成される自己蒸留型の学習手法であり、ラベルなしでの自己教師あり学習を可能とする。本モデルの全体構造を**図 1** に示す。従来の知識蒸

2025. 7. 18　加工技術研究部　受理

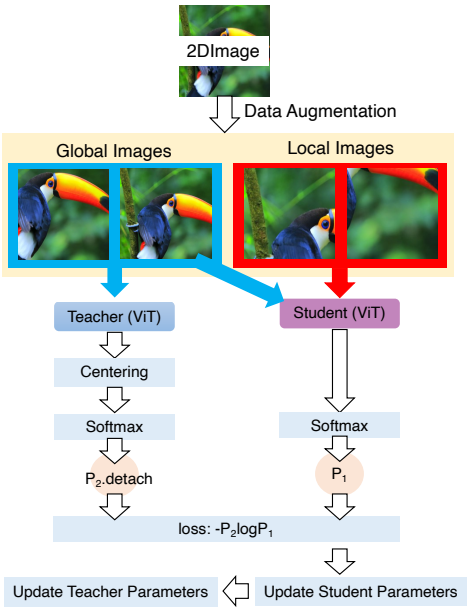


図 1　DINO の構造

留とは異なり、DINO では生徒モデルの学習後に、得られた重みパラメータを用いて教師モデルのパラメータが更新される。

学習時に入力された画像は、データ拡張が行われ、それぞれの拡張画像が教師モデルおよび生徒モデルの双方に入力される。データ拡張の際には、1 枚の画像から、元画像のほぼ全体を含むようなグローバル画像（スケール 50%以上）と、局所的な領域を切り出したローカル画像（スケール 50%未満）が複数生成される。これらの拡張画像のうち、すべての画像は生徒モデルに入力される一方、教師モデルはグローバル画像のみを受け取る。

ネットワークからの出力は、分類カテゴリ数Kに対応する確率分布として得られ、出力ベクトルは softmax 関数により次式のように正規化される

$$P(x)^{(i)} = \frac{\exp\left(\frac{g_\theta(x)^{(i)}}{\tau}\right)}{\sum_{k=1}^{K}\exp\left(\frac{g_\theta(x)^{(k)}}{\tau}\right)} \tag{1}$$

ここで、$g_\theta(x)^{(i)}$は入力xに対するネットワーク出力のi番目の要素、θは学習パラメータ、τは softmax 関数分布の先鋭化係数(sharpening)を示す。教師モデルと生徒モデルの確率分布をそれぞれ$P_t(x)$ $P_s(x)$とすると、DINO における学習は、両分布間の cross-entropy loss を最小化するように生徒モデルのパラメータθsを最適化することで実行される。

$$\min_{\theta_s} H\left(P_t(x), P_s(x)\right) \tag{2}$$

$$\because H(a,b) = -a\log b$$

この結果、生徒モデルは教師モデルの出力分布を模倣するように学習が行われる。一方、教師モデルのパラメータθtは生徒モデルのパラメータθsを用いて以下のように更新される。

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \tag{3}$$

λにはコサインスケジュールが適用され、学習の過程で 0.996 から 1 へと徐々に増加する。さらに、学習の安定性を確保するため、教師モデルに限って出力に対して centering 処理が適用される。この処理では、出力の平均値をcし、以下のように更新する。

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i) \tag{4}$$

ここでmは減衰率、Bはバッチサイズを示す。

(1)式で示した教師モデルにおける出力$g_{\theta_t}(x)$は実際には以下のようにバイアスcによって補正され、softmax 関数に入力される。

$$g_{\theta_t}(x) \leftarrow g_{\theta_t}(x) - c \tag{5}$$

3. DINO の実装

3.1 学習データ

DINO の性能評価に際し、**MNIST**: **M**odified **N**ational **I**nstitute of **S**tandards and **T**echnology[7]、CIFAR-10[8]、および STL-10[9]の 3 種類の画像データセットを比較対象として用いた。CIFAR-10 は、動物や車両などを含む 10 種類のクラスから構成され、各クラスについて 5000 枚の学習用画像および 1000 枚の評価用画像が提供されている。画像は全てカラーであり、解像度は 32×32 画素である。一方、STL-10 も同様に 10 クラスから構成されるカラー画像データセットであるが、画像の解像度は 96×96 画素と、CIFAR-10 の 9 倍に相当する。STL-10 には各クラスあたり 500 枚の学習用画像、800 枚の評価用画像に加えて、ラベルなしの画像が 10 万枚含まれている。

3.2 DINO の実装

学習用入力画像に対してはまずデータ拡張処理を施し、得られた拡張画像を用いて学習を行う。すべての拡張画像は生徒モデルに入力され、それぞれの画像に対応したクラスごとの確率分布が出力される。一方で、拡張画像のうちグローバル画像のみが教師モデルに入力され、クラス分類が実行される。

```
x = Aug(images)
model_student.zero_grad()
```

```
s = model_student(x)
with torch.no_grad():
    t = model_teacher(x[: batch_size *
                                global_view])
loss = H(t, s, C)
loss.backward()
optimizer.step()
```

以下に示すデータ拡張では、まず 2 枚のグローバル画像を生成し、それぞれからローカル画像を生成する。生成されたローカル画像は、グローバル画像の後に連結され、最終的な入力画像が構成される。

```
img1 = self.large_image(x)
img2 = self.large_image(x)
x = torch.cat([img1, img2], dim=0 )
for _ in range(local_view // 2):
    simg1 = self.small_image(img1)
    simg2 = self.small_image(img2)
    x = torch.cat([x, simg1, simg2], dim=0)
```

(1)式の生徒モデルおよび教師モデルの出力確率分布は、それぞれ以下のように計算される。

```
s = F.softmax(s / TPS, dim=1)
t = F.softmax((t - center) / TPT, dim=1)
```

ここで、TPS および TPT は先鋭化係数、教師モデルには(4)、(5)式の centering 係数を組み込む。(2)式の cross-entropy loss は生徒モデルと教師モデルの確率分布の誤差計算ため、損失関数には Kullback–Leibler divergence を用いる。

```
tmp=F.kl_div(torch.log(s[j]), q, None, None,
'batchmean')
```

なお、教師モデルと生徒モデルは同一の初期パラメータで学習を開始し、(3)式の学習に伴う教師モデルの更新は以下により行われる。

```
teacher_state_dict[name] = l *
teacher_state_dict[name] + (1 - l) * param
```

また、教師モデルにのみ適用される centering 処理は、グローバル画像に対する出力確率分布を用いて以下のように更新される。

```
C = M * C + (1 - M) * t.mean(dim=0)
```

4. DINO の性能評価

4.1 MNIST データセットによる性能評価

ViT を基盤とする DINO による自己教師あり学習では、得られた Attention Map が画像中の物体領域を捉える能力を有していることが報告されている。本節では、まず MNIST データセットを用いて、DINO が物体領域をどのように認識するかを検証する。

本実験では、先行研究と同様に、コンパクトな ViT 基盤モデルを DINO に採用し、学習を行った。使用した主な ViT パラメータは**表 1** に示す。最初に MNIST データセットに含まれる 5 万枚の手書き数字画像を用い、200 エポックにわたって学習を実施した。データ拡張においては、1 枚の入力画像からグローバル画像 2 枚およびローカル画像 8 枚の計 10 枚の拡張画像を生成した。モデルのパラメータとしては、先鋭化係数 TPS および TPT にそれぞれ 0.9 および 0.05、センタリング減衰率 M に 0.9、教師モデルのパラメータ更新係数 *l* には 0.996 を設定した。

学習終了後、未学習画像を教師モデルに入力し、Attention スコアを可視化した結果を**図 2** に示す。**図 2 (a)** は DINO による学習後に入力した未学習画像を、**図 2 (b)** はその画像に対応する Attention Map を、**図 2 (c)** は Attention Map を入力画像に重ね合わせた注視領域をそれぞれ示している。

これら図の比較から、DINO が画像中の数字の構造を注視していることが確認された。このことは、ViT ベースの DINO が自己教師あり学習において、ラベルなしに物体領域を効果的に捉える能力を有することを示している。

4.2 解像度の異なるデータセットに対する性能評価

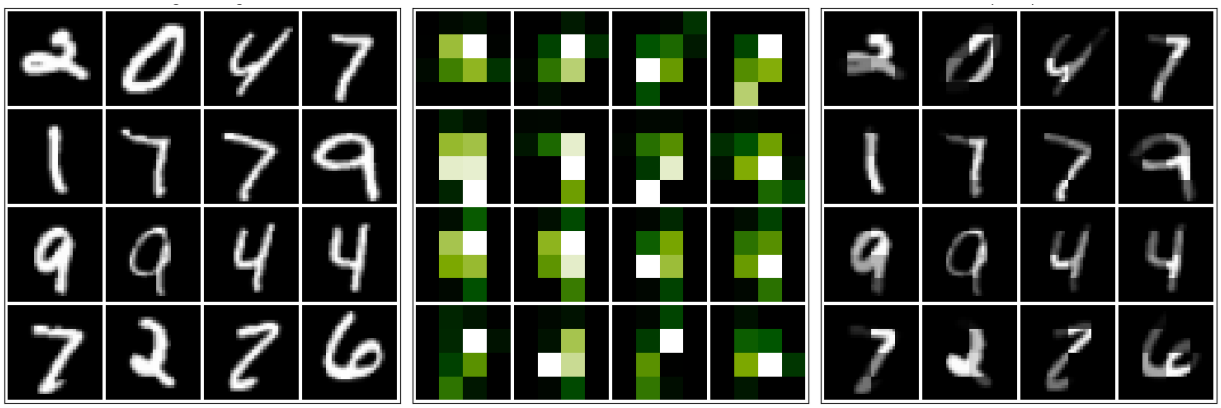
前節において、MNIST データセットのような背景を含まない画像に対し、DINO がラベルなしで物体領域を捉える能力を有することを確認した。本節では、背景を含む一般的なカラー画像に対しても同様の能力が発揮されるかを検証するとともに、入力画像の解像度の違いが領域検出性能に与える影響についても評価を行う。

ViT の基盤モデルには、MNIST データセットに用いたものと同一の構造を採用した。ただし、入力画像が RGB カラーであるため、最初の畳み込み層の入力チャンネル数を 3 に変更し、注視領域の細分化のため画像分割数は 8×8 とした。画像解像度については各データセットの仕様に合わせて設定した。

まず、解像度が 32×32 画素で構成される CIFAR-10 デ

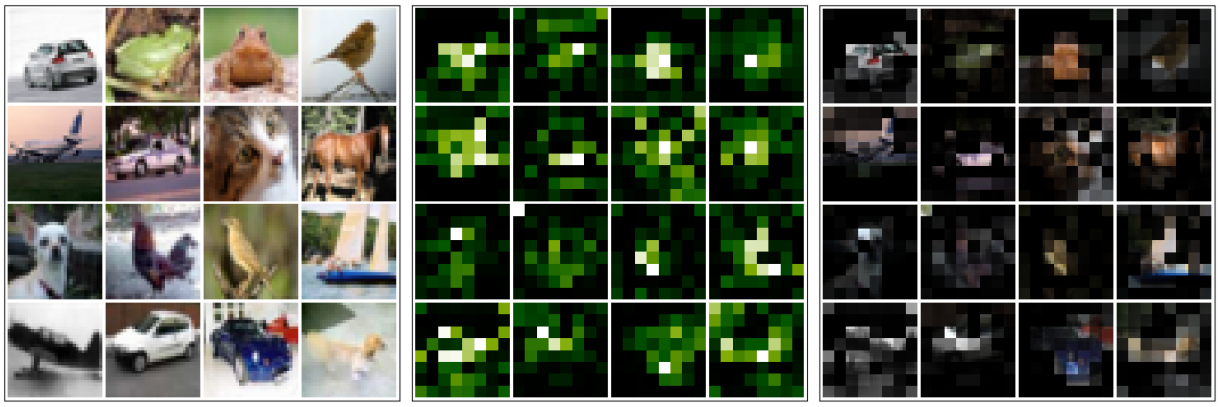
表 1　ViT 基盤モデルの主なパラメータ

	パラメータ
Datasets	MNIST
画素	32画素×32画素
ラベル数	10
学習データ数	5万画像
画像分割数	4×4
ヘッド数	4
Hidden size D	192
MLP size	768
レイヤー数	12
パラメータ数	5.3M



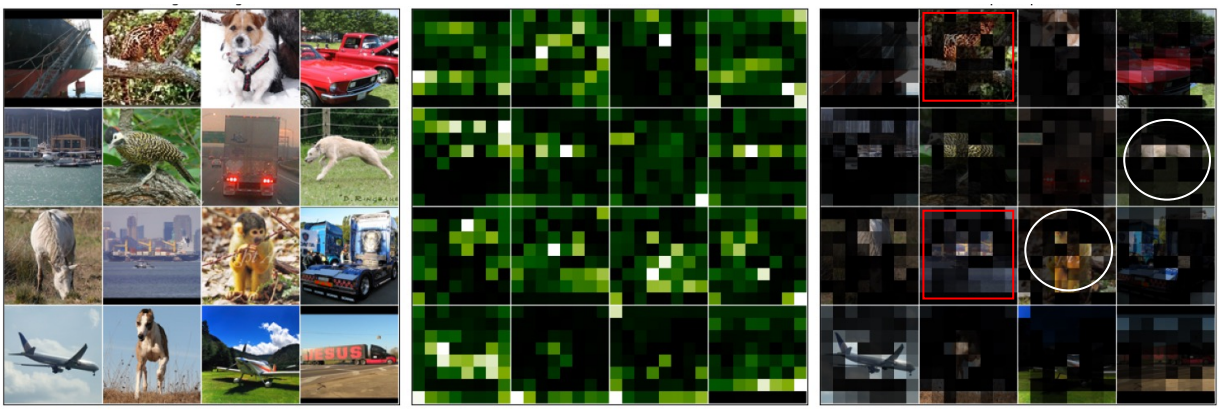
(a) 入力画像　(b) Attension Map　(c) 注視画像

図 2 MNIST の Attention 領域



(a) 入力画像　(b) Attension Map　(c) 注視画像

図 3 CIFAR-10 の Attention 領域



(a) 入力画像　(b) Attension Map　(c) 注視画像

図 4 STL-10 の Attention 領域

ータセットに対し、学習用画像 5 万枚を用いて 1000 エポックにわたり学習を実施した。このときの教師モデルの学習パラメータ数は 5.37MB であった。学習後、ランダムに選択した 16 枚の未学習画像を教師モデルに入力し、得られた Attention スコアを可視化した結果を**図 3** に示す。**図 3 (a)** は入力画像、**図 3 (b)** はそれに対応する Attention Map、**図 3 (c)** は Attention Map を重ね合わせた注視領域をそれぞれ示している。

この結果から、CIFAR-10 のように背景情報が含まれる画像においても、DINO はラベルなしで対象物体の領域検出が可能であることが確認された。

STL-10 データセットは、解像度 96×96 画素の高解像度画像から構成されており、本実験ではランダムに選んだラベル付けのない 5 万枚の画像を学習に用いた。STL-10 は CIFAR-10 と同様に 10 クラスを含むが、画像解像度は CIFAR-10 に比べて 9 倍の画素数からなる。それにもかかわらず、使用した ViT モデルの学習パラメータ数は 5.43MB と、CIFAR-10 使用時と比較してわずかな

増加にとどまった。

ラベル付けのない STL-10 に対して DINO による事前学習を行い、得られた Attention スコアを可視化した結果を**図 4** に示す。CIFAR-10 では、物体が画像中央に大きく配置されているため、多くの場合、Attention は物体領域に集中していた。一方、STL-10 においては、背景の占める割合が大きい画像も多く見られ、Attention が背景領域にも向けられている例が確認された。

図中の白丸印で示した画像は、物体の占有率が高いため、Attention は主に物体領域に集中的に向けられている。対照的に、赤四角印の画像では、物体と背景の両方に Attention が分散しているのが分かる。他の一部の画像では背景のみに着目している例も見られた。

STL-10 は CIFAR-10 と同様に 10 クラスから構成されており、学習モデルのパラメータ数はわずかに STL-10 の方が多い。一方、画像解像度は CIFAR-10 に比べておよそ 9 倍の画素数を有する。**図 5** に、両データセットに対する DINO の学習曲線を示す。ラベルなしの 5 万枚の学習データに基づく結果からは、画素数が 9 倍であるにもかかわらず、学習性能には大きな差は見られず、STL-10 の方がわずかに高い精度を示した。

また、学習曲線からは 1000 エポック時点でも学習が完全には収束していないことが確認される。一方、STL-10 のラベル付き学習用画像を用いた場合には、およそ 200 エポック程度でも学習が収束しており、ラベルなし学習によるモデルと同等の Attention スコアが得られることを Attention Map より確認した。この学習では学習条件を統一するため 1 エポックに 5000 枚の画像からランダムに 5 万枚を選び用いた。

さらに、DINO に用いた ViT の基盤モデルを STL-10 のラベル付き教師画像（5000 枚）を用いて単体で学習させた場合の学習曲線を**図 6** に示す。この実験ではデータ拡張も適用したが、学習初期段階から過学習が発生しており、十分な学習が行えなかった。なお、DINO と ViT は異なる損失関数を用いるため、厳密な比較は困難であるが、学習データが少ない場合には、DINO による事前学習を導入した方が、より安定した学習が期待される。

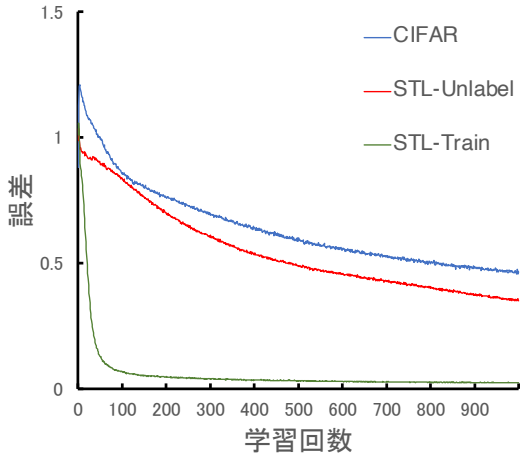


図 5　各データセットに対する学習曲線

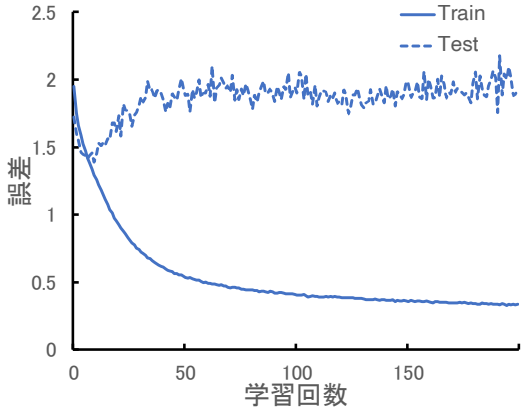


図 6　STL-10 に対する ViT の学習曲線

5. 結　　言

本報告では、自己教師あり学習手法である DINO に ViT 基盤モデルを用い、複数の異なる解像度および背景構造を持つ画像データセットに対してその学習を検証し、Attention Map を可視化・比較を行った。本報告の主な結果は以下のとおりである。

1. CIFAR-10 および STL-10 データセットに対して、同一構造の ViT モデルを用いて DINO による学習を行い、得られた Attention スコアを可視化した。特に、背景が画像全体の大部分を占める STL-10 において、Attention が背景領域にも分散する傾向があることを確認した。
2. STL-10 に対し、ラベル付き画像（5000 枚）を用いた DINO による自己教師あり事前学習と ViT の単体学習を比較した。学習曲線の比較から、ラベル付きデータによる ViT の単体学習では初期段階から過学習が生じた一方、DINO によるラベルなし事前学習ではより安定した学習が可能であることを示した。少量データ環境における DINO の有効性が示された。

文　　献

1) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin: Attention is all you need, arXiv:**1706.03762** (2017).

2) Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei: Scaling Laws for Neural Language Models, arXiv:**2001.08361** (2020).

3） Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, Armand Joulin: Emerging Properties in Self-Supervised Vision Transformers, arXiv:**2104.14294** (2021).

4） Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:**2010.11929** (2020).

5） 廣川 勝久：広島県立総合技術研究所東部工業技術センター研究報告, 事前深層学習モデルの転移学習による能力比較（第１報）, **37**（2024）.

6） 廣川 勝久：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による画像認識（第１報）, **38**（2025）.

7） THE MNIST DATABASE of handwritten digits : http://yann.lecun.com/exdb/mnist/

8） The CIFAR-10 dataset: https://www.cs.toronto.edu/~kriz/cifar.html

9） STL-10 dataset: https://cs.stanford.edu/~acoates/stl10/