目視検査を模擬した単純タスクにおけるラベル欠損の検出精度の比較 ～深層学習が人間の目視と同等となるための訓練サンプル数の検証～

加 藤 直 人**, 井 上 路 子**, 西 山 正 志**, 岩 井 儀 雄**

Comparison of the Accuracy of Defect Detection in a Simple Task of Label Inspection
–Investigation of the Number of Training Samples for Deep Learning to be equivalent to Human Visual Capability–

Naoto KATO, Michiko INOUE, Masashi NISHIYAMA and Yoshio IWAI

In this paper, simple tasks mimicking visual label inspection are described to compare the accuracy of humans with that of deep learning techniques. The number of training samples that are required to obtain equal or higher accuracy as the inspection by humans is investigated using the simple task. In our method, letters printed on test labels are represented as symbols. The variations in the symbols are controlled by changing the angle of rotation, the defective position, and the defect rate. Training samples consisting of images and defect bounding boxes are automatically generated. The experimental results have shown that the number of training samples was needed to be in the order of several thousand to obtain equal or higher accuracy of humans in the simple task. They have been also demonstrated that the number of training samples was needed to be in the order of tens of thousands when the defect rate of the symbols was low.

Key words: Visual label inspection, defect rate, deep learning, number of training samples.

1. はじめに

近年，工場の労働力不足を解消するため，人手で行われる作業の自動化[1,2,3,4]が進められている．これらの作業の中でも，本論文では目視検査に着目する．目視検査とは，確認対象となる部品やラベルの表面に存在する欠損を見つける作業のことである．この目視検査は様々な工程で行われている．具体的には，基板や外枠などの部品を製作する工程[5,6,7]，部品を組み合わせて製品を仕上げる工程[8,9]，製品にラベルを貼る工程[10,11]で目視検査が必要とされている．

本論文では，多様な目視検査の中でも，ラベル欠損の検出に焦点を当て議論を進めていく．ラベルは，例えばノートパソコンや薬品ケースなど，その製品自体に貼り付けられている．そのラベルには，製品情報をユーザへ正確に伝達するため，文字が記載されている．もしラベル欠損が生じると，ユーザはラベル内の文字を読み取ることができなくなる．このため，製品の出荷前にラベル欠損の検出を高精度で行うことが要求されている．

ラベル欠損の検出の自動化に必要な要素技術の一つとして，深層学習[12,13,14]が適用されることが多い．欠損検出のネットワークモデルを学習するために，画像と教師信号のペアからなる訓練サンプルを準備する．製品の表面に貼られたラベルを撮影することで画像を獲得し，経験を積んだ専門家が，画像中の欠損の有無を教師信号として目視で付与する．深層学習で高い精度を得るためには，この訓練サンプルが大量に必要となる．ただし，訓練サンプルを収集するためにはコストが発生する．深層学習の精度を維持したまま収集コストを抑えるために，必要最小限となる訓練サンプル数の当たりを付けることが望まれている．これまでに，いくつかのガイドライン[15,16,17]が提言されており，その中では訓練サンプルを適切に準備することが重要と言及されている．ただしガイドラインの中では，訓練サンプルについて，精度が得られるまでの個数の当たりの付け方を言及していなかった．ラベル欠損の検出に必要な訓練サンプルの個数について，当たりの付け方を論じた文献は，我々が調べた限りではあるが存在しなかった．別目的ではあるが物体認識のために文献[18,19]では，フラクタル幾何学を用いて数千万個の訓練サンプルを，自動で生成する試みがなされている．物体認識の精度は，実際の自然画像の訓練サンプルを用いた場合に差し迫ることが報告されている．人間との精度比較は考察されていないものの，物体認識における訓練サンプル収集の見通しを立てるために，有益な方向性が示唆されていると捉えることができる．ただしフラクタル幾何学は，ラベル内の文字を表現することに適しているとは言えない．以下では，ラベル欠損の検出における訓練サンプルの個数の当たりを付けることで，収集の見通しを立てる方法を考える．

深層学習を適用するために必要な訓練サンプルの個数は，ラベル欠損の検出の難しさによって変化する．検出の難しさは，対象となる欠損が取り得る大きさや向きなどの変動要因で決まる．深層学習を用いてラベル欠損を検出する場合，その欠損が取り得る変動要因を訓練サンプルへ十分に含めることが出来れば，安定した精度を得ることができる．例えば，ラベル欠損の検出が容易な場合，必要な訓練サンプルの個数は少なくなる．それに対して，ラベル欠損の検出が困難な場合，必要な訓練サンプルの個数は多くなる．様々な変動を含むラベル欠損の事例を多く集め，自動化に必要な訓練サンプルの個数を，それぞれの事例で実験的に求めるケーススタディが考えられる．しかしこの方法では，教師信号の付与作業を伴う訓練サンプルの収集を大量に行う必要がある．

そこで本論文では，ラベル欠損の検出を簡単化した単純タス

* 原稿受付 令和 2 年 5 月 8 日
* 掲載決定 令和 2 年 8 月 31 日
** 鳥取大学大学院工学研究科 (鳥取市湖山町南 4 丁目 101)

クを設計し，訓練サンプルを自動生成することで，人間の目視と同等以上の精度となるために必要な訓練サンプルの個数を検証する．このために，単純タスクでの欠損検出の精度を求め，人間と深層学習との間で精度を比較する．単純タスクでは，ラベルに記載されている文字をシンボルで表現する．このシンボルに含まれる欠損の向きや大きさを可変とすることで，ラベル欠損が取り得る変動を制御する．画像と教師信号のペアからなる訓練サンプルを自動で生成し，欠損の変動要因について検出精度を評価する．本論文の単純タスクでは，教師信号の付与作業を伴う訓練サンプル収集を，人手で行うことなく自動で行える利点がある．ラベル検査で起こり得る変動要因を完全に網羅できる訳ではないが，必要な訓練サンプルの個数を，少ないコストで実験的に見積もれる可能性を，単純タスクは持つと考える．以下，2. では単純タスクの詳細を説明する．3. では人手による単純タスクの精度を測定し，4. では深層学習による単純タスクの精度を測定する．次に，5. では人手と深層学習による単純タスクの精度を比較し，6. では単純タスクの設計について考察し，7. でまとめる．

2. 単純タスク

2.1 ラベル欠損の検出タスクの簡単化

製品情報をユーザへ適切に伝えるため，ラベルには様々なフォントで文字が記載されている．本論文で考える単純タスクでは，Fig. **1** (a) に示す 4 つの記号 (クラブ，スペード，ダイヤ，ハート) からなるシンボルで文字を表現する．実際の文字は種類が非常に多いため，ここではタスクの簡単化を狙い上記の 4 クラスの記号を用いることにする．シンボルを 1 度刻みで 360 度回転させることで，Fig. **1** (b) に示すように，シンボルの様々な見え方の変化を表すことを狙う．実際のラベルには文字が並んで記載されている．本論文で扱う単純タスクでは，シンボルが等間隔に並んで記載されており，それぞれの位置とサイズは固定されているものとする．

単純タスクで取り扱う欠損について考えるために，実際のラベル欠損について述べる．ラベル欠損の具体例として，破れ，穴あき，めくれ，印字スレ，文字欠けが挙げられる．これらの欠損が存在する場合，文字の判読性が低下し誤読が生じることや，可読性が低下し文字と認識できなくなることがある．ラベル欠損の変動を表す主な要因として，1. で述べたように，向きと大きさが挙げられる．以下では，それぞれの変動要因について，何が生じるかの例を用いて説明した後に，単純タスクにおける欠損の表現方法について述べる．

まず，ラベル欠損の変動要因の 1 つである欠損の向きについて “あ” という文字を例に説明する．“あ” の上部に欠損が存在すると判読性が低下し，“め” と誤読する場合が生じる．また，“あ” の下部に欠損が存在すると可読性が低下し，文字と認識できなくなる．欠損の向きを簡単に表現するために，本論文の単純タスクでは，Fig. **1** (c) に示すように，シンボルの上下左右の 4 方向から，いずれか 1 つを選択する．

次に，ラベル欠損の変動要因である大きさについて説明する．“あ” で小さな欠損が存在したとしても判読性は維持され，“あ” と読むことはできる．ただし，文字フォントの美しさが損なわれる課題が生じる．さらにその欠損が大きくなると可読性が低下し，“あ” と認識できなくなる．欠損の大きさを簡単に表現するために，本論文の単純タスクでは，Fig. **1** (d) に示すように，指定した割合でシンボルに欠けを生じさせる．

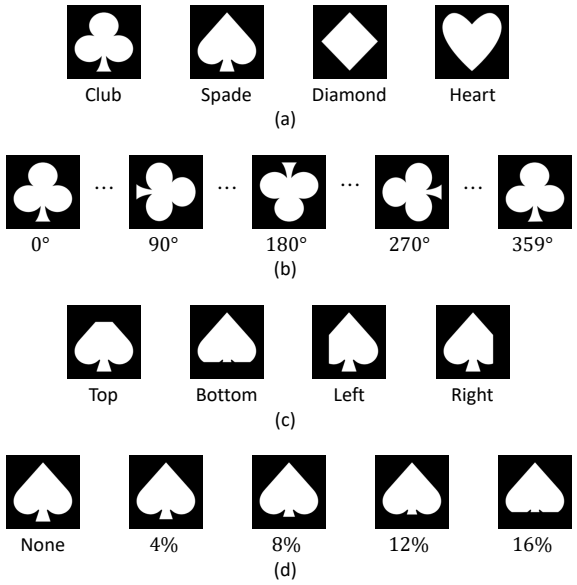


Fig. **1** Examples of the parameters used to generate anomalous symbols in the simple task of the four marks (club, spade, diamond, and heart). We show the suits in (a), the angle of rotation in (b), the defective positions in (c), and the rate of defect in (d).

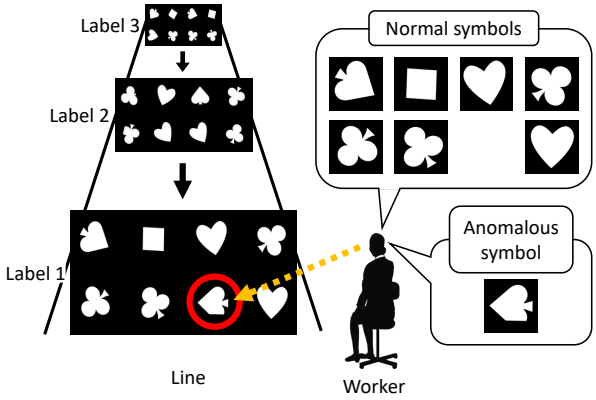


Fig. **2** Overview of the simple task of label inspection by participants.

複数の文字を並べて記載しているラベルでは，最初の文字が欠損する場合があれば，最後の文字が欠損する場合もある．本論文の単純タスクでは，欠損が存在するシンボルを，ラベル内でランダムに配置することにより，欠損の位置を制御する．

実験協力者に与えるタスクの内容について説明する．実験協力者は，ラベルに記載されているシンボルの中から，欠損のある異常シンボルの位置を検出する単純タスクに取り組む．その例を Fig. **2** に示す．ここでは，ラインに流れている製品へ貼られたラベルに対して目視検査を行う状況を想定している．実際のラベルは主に製品の平面に貼られていることが多いため，本論文では平面ラベルを検査対象とする．以下では，単純タスクで用いるラベルを刺激画像と呼ぶ．次節にて，刺激画像を生成する方法について説明する．

2.2 刺激画像の生成

刺激画像を生成するために，シンボルを横 8 個，縦 4 個の格子状に並べて配置した．画像生成の流れを以下で述べる．

S1. ある 1 枚の刺激画像中に含まれる異常シンボルの個数を決定した．異常シンボルの個数として，0，2，4，6 個の 4 通りの中からランダムに 1 つを選択した．

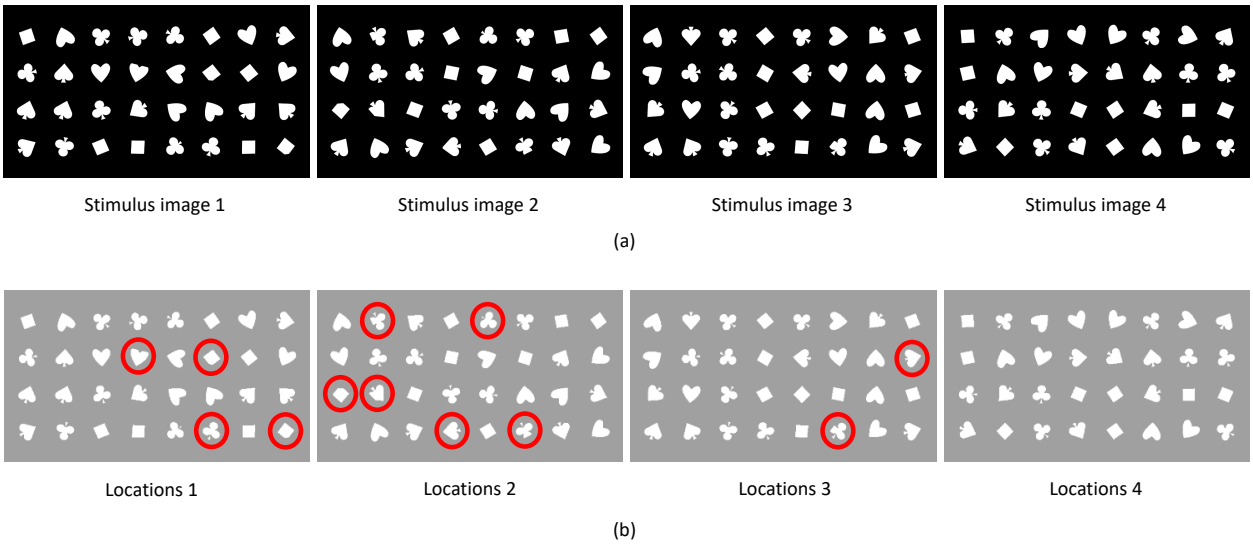


Fig. **3** Examples of stimulus images of the simple task of the marks (club, spade, diamond, and heart). The red circles indicate the locations of anomalous symbols.

S2. シンボルを生成するためのパラメータを決定した．パラメータとしてシンボルのクラス，回転，欠損の大きさ，欠損の向きを用いた．
- クラスとして，Fig. **1** (a) のクラブ，スペード，ダイヤ，ハートの 4 つの中からランダムに 1 つを選択した．
- 回転角度として，1 度刻みで 0 度から 359 度の範囲でランダムに設定した．その候補の例を Fig. **1** (b) に示す．
- シンボルを欠損させるか否かについて，ランダムに決定した．ただし，最初に決定された異常シンボルの個数に達していた場合は欠損させなかった．
- 欠損させる場合に，欠損の向きと大きさを以下のように設定した．
  - 向きについて，上，下，左，右の 4 つの中からランダムに 1 つを決定した．その候補の例を Fig. **1** (c) に示す．
  - 大きさについて，4，8，12，16% からランダムに 1 つを決定した．その候補の例を Fig. **1** (d) に示す．

S3. 決定されたパラメータを用いて，正常シンボルまたは異常シンボルを生成した．

S4. 生成されたシンボルを格子上の点に配置した．

S5. シンボルが規定数の $4 \times 8 = 32$ 個に達するまで S2 から S4 を繰り返した．

生成された刺激画像の例を Fig. **3** (a) に示す．これらの例に含まれる異常シンボルを図中 (b) の赤丸で示す．生成される刺激画像の総パターン数は約 7.6×10^{118} 通り存在するため，完全に一致する刺激画像ができる可能性は極めて低い．以下，3. では生成した刺激画像を用いて人間による異常検出の精度の調査結果を述べ，4. では深層学習による異常検出の精度の調査結果を述べる．

3. 実験協力者による単純タスクの検出精度の調査

3.1 実験環境

実験協力者が単純タスクの目視検査に取り組むことで，異常検出の精度を調査した．実験協力者の人数は，日本人学生 20 名（男性 15 名，女性 5 名）で，その平均年齢は 22.2 ± 1.0 歳であった．刺激画像を表示するために，16 インチの液晶タブレット (Cintiq Pro 16; Wacom) を配置した．その画面解像度は 1920×1080 画素とした．実験協力者が検出した異常シンボルを記録するために，液晶タブレットに付属するペンを用いた．Fig. **4** に実験環境の配置と風景を示す．実験協力者は，液晶タブレットで作業するため椅子に着席した．照明条件を統制するため，照度が 360 ± 5 lx の遮光室を用いた．

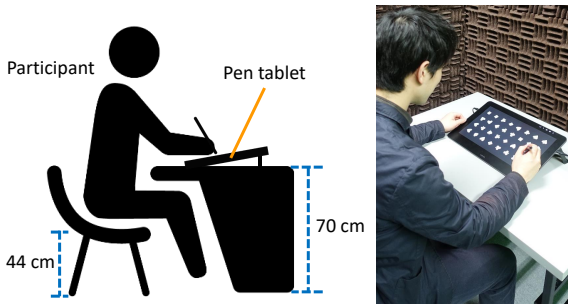


Fig. **4** Setting of the investigation using the simple task for the participants.

3.2 実験手順

実験協力者に以下の手順で目視検査を行わせた．

手順 1. 実験協力者をランダムに 1 人選択した．
手順 2. 実験方法を説明した上で例題を説明した．
手順 3. 実験協力者は，液晶タブレットに表示された 1 枚の刺激画像に対して目視検査を行い，異常シンボルをマーキングした．刺激画像 48 枚に対して同じ検査作業を行った．
手順 4. 実験協力者 20 名が終了するまで手順 1 から手順 3 を繰り返した．

以下では手順 2 と手順 3 の詳細を説明する．

3.2.1 手順 2: 実験方法と例題の説明

異常シンボルへマーキングするための操作方法について，実験協力者へ説明した．その後，単純タスクの例題を実験協力者へ示し，実際に目視検査を行わせた．この例題では，欠損の大きさが 4% と 16% の異常シンボルを含む刺激画像を用いた．例題の目視検査を行った後に，異常シンボルが刺激画像中の何

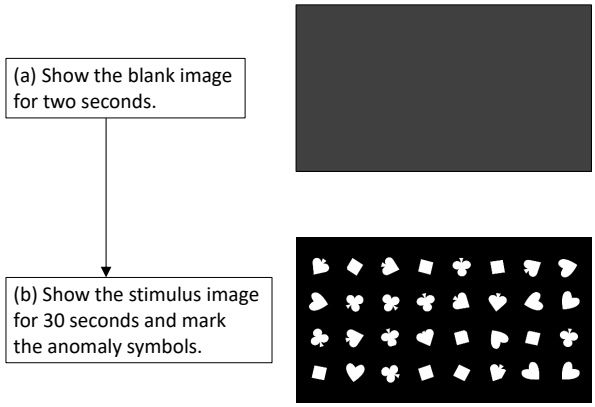


Fig. **5** Procedure of the participant in the step 3 viewing the stimulus image.

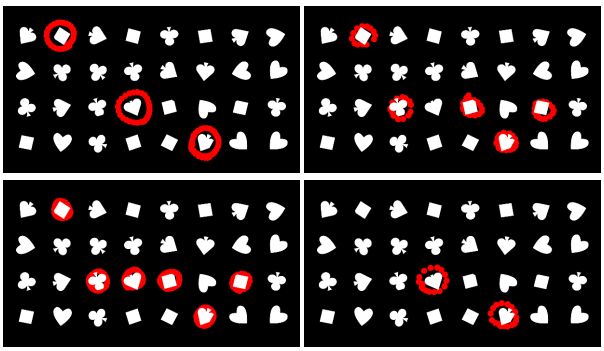


Fig. **6** Examples of symbols circled by the participants in the marking procedure.

Table **1** The inspection accuracy of participants on the simple task of the four marks (club, spade, diamond, and heart).

F-measure	Precision	Recall
0.95	0.99	0.91

処に存在したかの正解を示し，検出すべき異常シンボルについての理解を深めさせた．

3.2.2　手順 3: 刺激画像の目視検査

液晶タブレットに表示された刺激画像を目視で検査させた．実験協力者に対して，Fig. **5** (a) の挿入画像をまず表示し，(b) の刺激画像を次に表示した．異常シンボルのマーキング作業では，実験協力者にそのシンボルを，ペンを用いて丸で囲ませた．実験協力者が誤って丸をつけた場合に備え，消しゴム機能を用意した．刺激画像 1 枚あたりの目視検査の作業時間は 30 秒とし，その時間が経過すると挿入画像へ自動的に切り替わる設定とした．実験協力者がマーキングした結果の例を Fig. **6** に示す．

3.3　実験協力者の検出精度

実験協力者が検出した異常シンボルの精度評価のために，F 値 (F-measure)，適合率 (Precision)，再現率 (Recall) を用いた．実験協力者の検出結果を Table **1** に示す．F 値は 0.95 であった．適合率は 0.99 と高く，正常シンボルを異常シンボルと誤検出することは，ほぼ見られなかった．一方，再現率は 0.91 であり，見落とされている異常シンボルが多かった．

次に，異常シンボルの検出精度が，刺激画像の生成パラメータ毎でどのように変化するかを調査した．以下では，再現率を調査の指標とした．この指標を用いる理由について述べる．

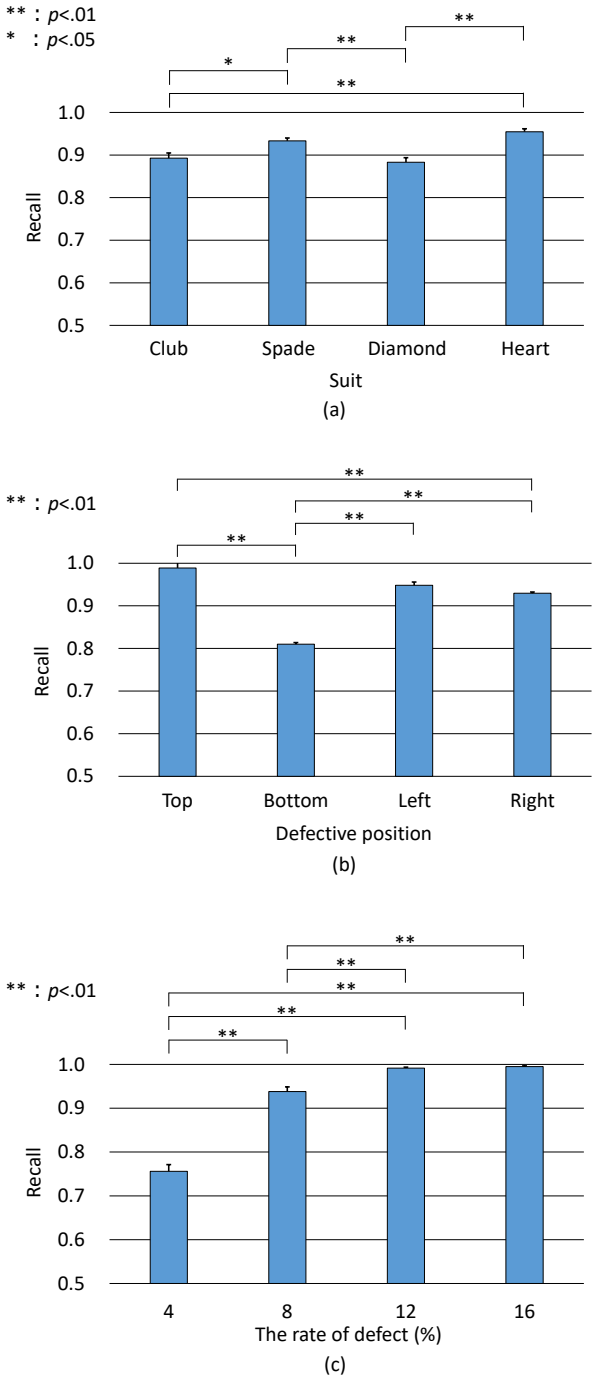


Fig. **7** Recall rates of the participants by suit, defective position, and the rate of defects in the simple task of the four marks (club, spade, diamond, and heart).

Table **1** の結果より，適合率は 1.0 にほぼ近く，一方で再現率は適合率と比べて大きく低下していた．このため，再現率が精度に強く影響を及ぼしていると判断し，再現率を指標とした．以下では，パラメータ間の精度変化を統計的に検定するために，ボンフェローニ法を用いた多重比較検定を適用した．

刺激画像を生成する際に変化させた各パラメータにおける再現率を Fig. **7** に示す．図中 (a) において，シンボルのクラス毎の再現率を調査した．ここでは，クラブとハート，スペードとダイヤ，ダイヤとハートの組み合わせで有意差があった．また，クラブとスペードの組み合わせで有意な傾向が見られた．

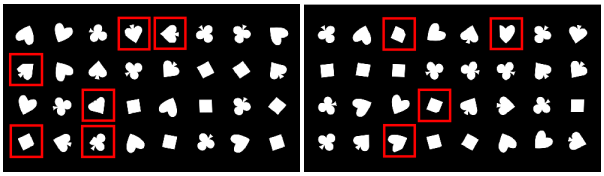


Fig. **8** Examples of outputs predicted using the SSD.

実験協力者は，シンボルのクラスがクラブやダイヤの場合に，スペードやハートに比べて，異常を検出できる可能性が低いと言える．次に (b) において，欠損の向き毎の再現率を調査した．ここでは，上と下，上と右，下と左，下と右の組み合わせで有意差があった．実験協力者は，異常シンボルの欠損の向きが下の場合に，上や左や右に比べて，異常を検出できる可能性が低いと言える．最後に (c) において，欠損の大きさ毎の再現率を調査した．ここでは，4% と 8%，4% と 12%，4% と 16%，8% と 12%，8% と 16% の組み合わせで有意差があった．実験協力者は，異常シンボルの欠損の大きさが 4% の場合に，8% から 16% と比べて，異常を検出できる可能性が非常に低いと言える．また，大きさが 8% の場合に，12% や 16% と比べて，検出の可能性が低くなると言える．一方，大きさが 12% の場合，16% と比べて同程度の精度になると言える．

4. 深層学習による単純タスクの検出精度の調査

4.1 概要

深層学習の代表的な手法として，Single-shot multibox detector (SSD) [20] と U-Net [21] を用いて，単純タスクにおける欠損の検出精度を調査した．SSD は物体検出のタスクで採用されることが多く，U-Net は領域分割のタスクで採用されることが多いが，異状検知の手法としても適用することができる．学習に用いる訓練サンプルを収集するために，2.2 の手順で刺激画像と教師信号のペアを大量に生成した．以下では，SSD の検出精度を4.2 に，U-Net の検出精度を4.3 に示す．

4.2 SSD を用いた場合の検出精度の調査

4.2.1 SSD の実験条件

刺激画像中の異常シンボルの位置を検出するために SSD を用いた．異常シンボルの位置を表すバウンディングボックスを教師信号とし，SSD のネットワークモデルの訓練を行った．学習された SSD を用いて欠損を検出した結果の例を Fig. **8** に示す．なお SSD のベースネットワークには VGG16 [22] を用いた．訓練サンプルの個数を，2,500 個から 10,000 個まで，2,500 個刻みで変更しながら，それぞれで学習を行った．学習時のランダム性を考慮するため，訓練サンプル生成を 3 回行った．それぞれの訓練サンプルでモデルを学習し，検出精度の平均を算出した．テスト画像として，訓練サンプルとは別に生成された刺激画像の 1,000 枚を用いた．学習時に与えたパラメータ間の精度変化に対して統計的な検定を行うために，ボンフェローニ法を用いた多重比較検定を適用した．

4.2.2 SSD の検出精度

SSD を用いた場合の検出精度を Fig. **9** に示す．図中 (a) において，訓練サンプルの個数毎で算出された F 値を調査した．ここでは，2,500 個と 5,000 個，2,500 個と 7,500 個，2,500 個と 10,000 個の組み合わせで有意差があった．よって SSD の検出精度は，訓練サンプルの個数が 2,500 個の場合に，5,000 個や 7,500 個や 10,000 個と比べて，低くなることが分かった．次に

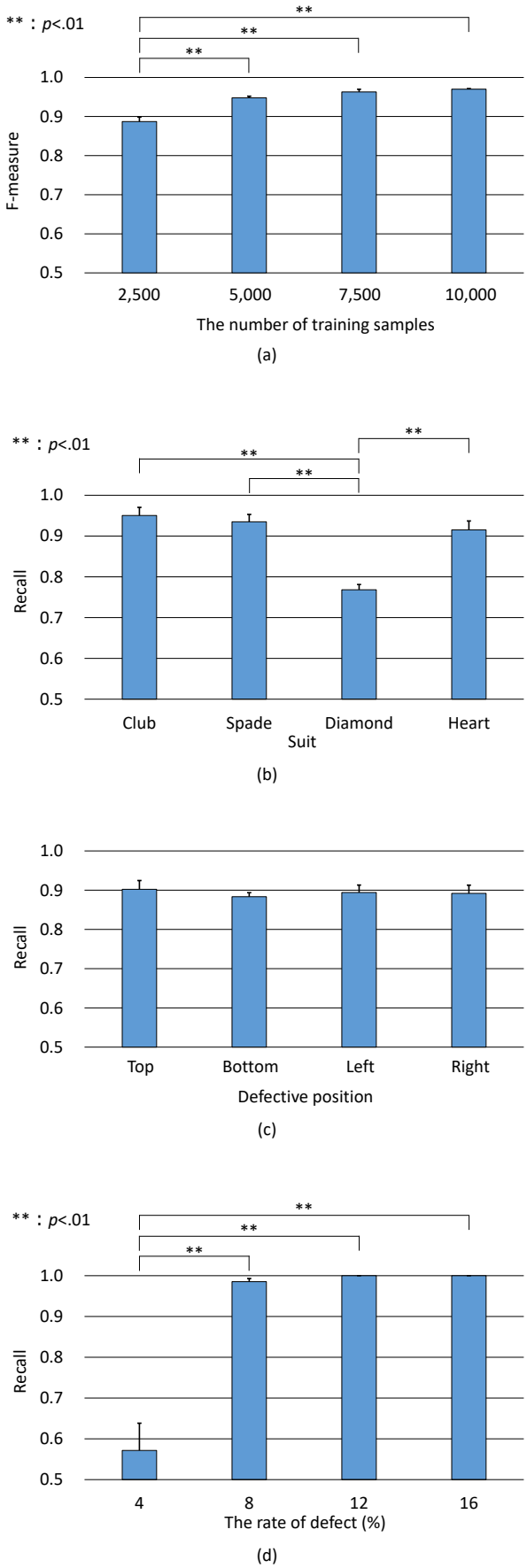


Fig. **9** Inspection accuracy using the SSD on the simple task of the four marks (club, spade, diamond, and heart).

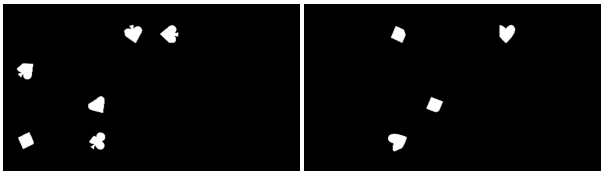


Fig. **10** Examples of outputs predicted using U-Net.

図中 (b) において，シンボルのクラス毎の再現率を調査した．ここでは，クラブとダイヤ，スペードとダイヤ，ダイヤとハートの組み合わせで有意差があった．SSD の検出精度は，シンボルのクラスがダイヤの場合に，クラブやスペードやハートと比べて，低くなることが分かった．次に図中の (c) において，欠損の向き毎の再現率を調査した．ここでは，どの組み合わせでも有意差は見られなかった．SSD の検出精度は，欠損の向きに影響を受けにくいことが分かった．最後に図中 (d) において，欠損の大きさ毎の再現率を調査した．ここでは，4% と 8%，4% と 12%，4% と 16% の組み合わせで有意差があった．SSD の検出精度は，欠損の大きさが 4% の場合に，8% や 12% や 16% と比べて，低くなることが分かった．

4.3 U-Net を用いた場合の検出精度の調査

4.3.1 U-Net の実験条件

刺激画像中の異常シンボルのみを領域分割するために U-Net を用いた．異常シンボルのみが表示された画像を教師信号とし，U-Net のネットワークモデルの訓練を行った．学習された U-Net を用いて欠損を検出した結果の例を Fig. **10** に示す．なお U-Net の構造として，ダウンサンプリング 9 層とアップサンプリング 9 層を用いた．訓練サンプルの個数を，5,000 個から 20,000 個まで，5,000 個刻みで変更しながら，それぞれで学習を行った．学習時のランダム性を考慮するため，訓練サンプル生成を 3 回行った．それぞれの訓練サンプルでモデルを学習し，検出精度の平均を算出した．テスト画像として，訓練サンプルとは別に生成された刺激画像の 1,000 枚を用いた．学習時に与えたパラメータ間の精度変化に対して統計的な検定を行うために，ボンフェローニ法を用いた多重比較検定を適用した．

4.3.2 U-Net の検出精度

U-Net を用いた場合の検出精度を Fig. **11** に示す．図中 (a) において，訓練サンプルの個数毎で算出された F 値を調査した．ここでは，5,000 個と 10,000 個，5,000 個と 15,000 個，5,000 個と 20,000 個の組み合わせで有意差があった．U-Net の検出精度は，訓練サンプルの個数が 5,000 個の場合に，10,000 個や 15,000 個や 20,000 個と比べて，低くなることが分かった．次に図中 (b) において，シンボルのクラス毎の再現率を調査した．ここでは，クラブとダイヤ，スペードとダイヤの組み合わせで有意差があった．U-Net の検出精度は，シンボルのクラスがダイヤの場合に，クラブやスペードと比べて，低くなることが分かった．次に図中の (c) において，欠損の向き毎の再現率を調査した．ここでは，どの組み合わせでも有意差は見られなかった．U-Net の検出精度は，欠損の向きに影響を受けにくいことが分かった．最後に図中 (d) において，欠損の大きさ毎の再現率を調査した．ここでは，4% と 8%，4% と 12%，4% と 16% の組み合わせで有意差があった．よって U-Net の検出精度は，欠損の大きさが 4% の場合に，8% や 12% や 16% と比べて，低くなることが分かった．ここまでの実験では，深層学習の各手法に対して精度を個別に調査していた．次節では，実

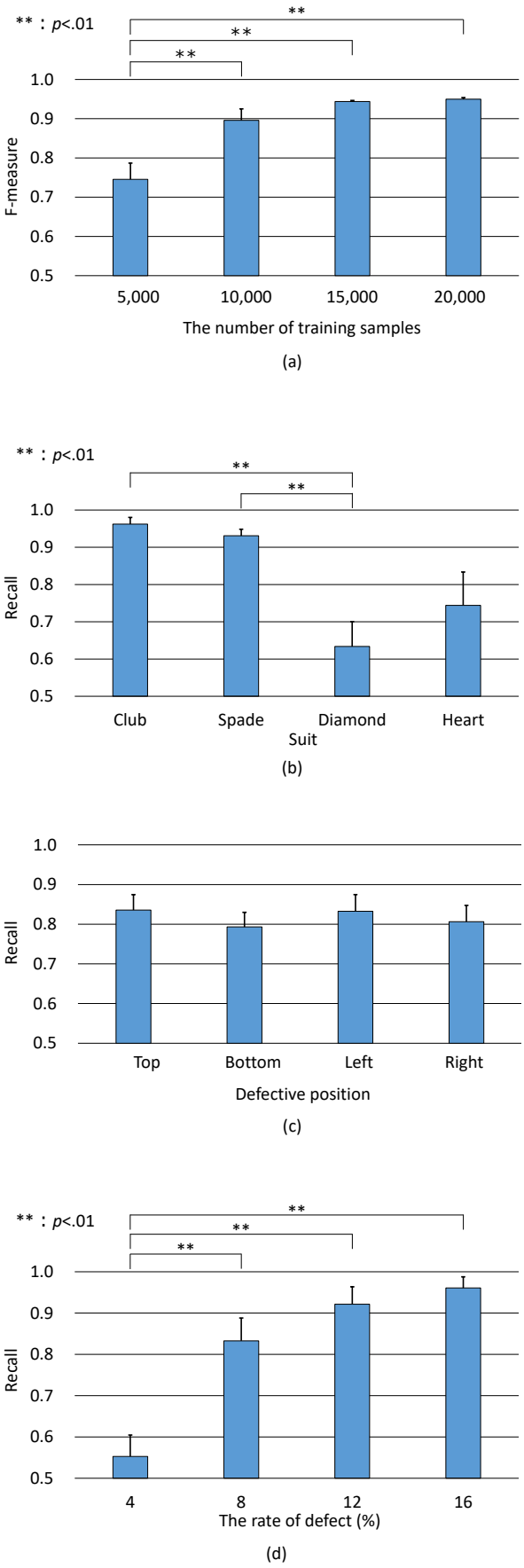


Fig. **11** Inspection accuracy using U-Net on the simple task of the four marks (club, spade, diamond, and heart).

Table **2** The label inspection accuracy of participants, SSD and U-Net on the simple task of the four marks (club, spade, diamond, and heart).

		F-measure	Precision	Recall
Participants		0.95	0.99	0.91
SSD	5,000	0.93	0.99	0.88
	7,500	0.95	0.99	0.91
U-Net	10,000	0.85	0.84	0.88
	15,000	0.92	0.98	0.87

験協力者と深層学習との間で精度を比較する．

5. 深層学習が実験協力者と同等以上の精度となるために必要な訓練サンプルの個数検証

5.1 深層学習と実験協力者との間でテストサンプルを揃えた場合

実験協力者による検出精度と，深層学習による検出精度とを比較することで，単純タスクにおける必要な訓練サンプルの個数について検証した．ここまでの実験では，実験協力者のテストサンプル数が 3. で 48 枚，深層学習のテストサンプル数が 4. で 1,000 枚を用いて精度を評価した．本実験では，両者とも同じ刺激画像のテストサンプルを用いた．その枚数は 48 枚とした．実験協力者の人数が 20 名であるため，実験協力者側の試行回数を合計で 20 回とした．なお，実験協力者 1 名あたりのラベル検査を 1 試行と数えた．SSD 側の試行回数を 8 回とし，それぞれの試行で訓練サンプルを生成しネットワークを学習した．U-Net 側の試行回数を 8 回とし，同様に学習を行った．訓練サンプルの個数として，SSD では 5,000 個と 7,500 個，U-Net では 10,000 個と 15,000 個を用いた．

実験協力者，SSD，U-Net による欠損の検出精度を Table **2** に示す．SSD の検出精度 (F 値と再現率) は，実験協力者と比べて，訓練サンプル数が 5,000 個の場合に低下していたが，7,500 個の場合に同等となっていた．一方，U-Net の精度は，15,000 個で実験協力者に近づいたものの，同等以上の結果を得ることはできなかった．この U-Net の傾向は 20,000 個でも同じであった．本実験にて，同じ訓練サンプル数を用いた場合，SSD の精度が U-Net より高かった理由について考える．単純タスクは欠損の位置検出が目的であるため，SSD のアルゴリズムが向いていると考えられる．一方，U-Net は位置検出に利用できるものの，領域分割のための手法であり，この目的には過剰性能なアルゴリズムであるからだと考えられる．

以上の結果より，2.2 で述べた条件で生成した単純タスクでは，実験協力者の目視と同等以上の精度となるために必要な訓練サンプルの個数は，SSD で 7,500 個であることが分かった．この個数は，この単純タスクで生成可能な総パターン数に対して，約 $9.9 \times 10^{-114}\%$ であった．

5.2 欠損検出が難しい変動要因に注目した場合

実験協力者と深層学習が共に苦手としていた変動要因である欠損の大きさに注目し，検出に必要な訓練サンプルの個数を検証した．ここでは深層学習の手法として，5.1 で精度が高かった SSD を用いた．実験協力者の検出精度は，Fig. **7** (c) で示したように，欠損の大きさ 4% の場合が 8% と比べて著しく低かった．SSD の検出精度は，Fig. **9** (d) で示したように，実験協力者と同様の傾向がみられた．以下では，ラベル欠損の検出が困難な場合を 4%，検出が容易な場合を 8% として考察を進める．なお実験条件は，5.1 と同じとした．

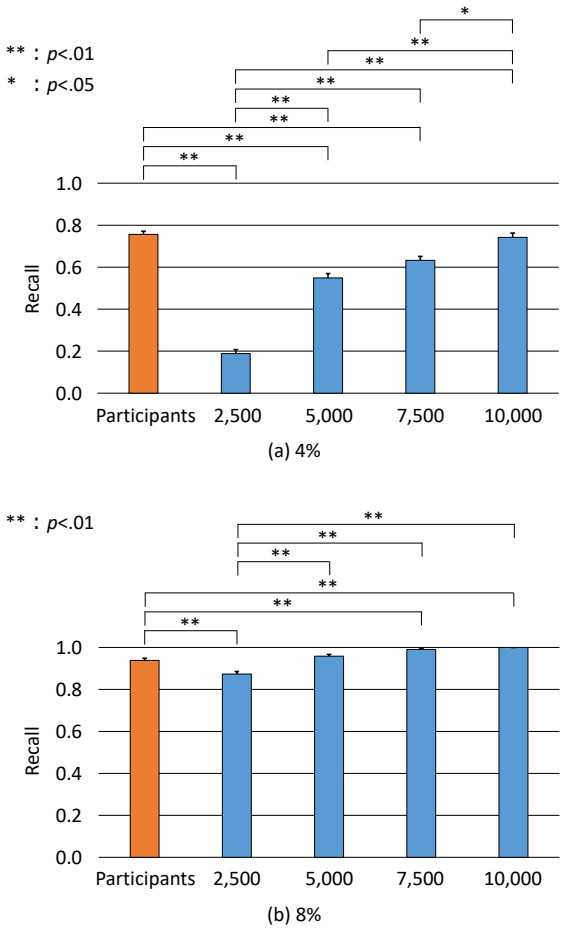
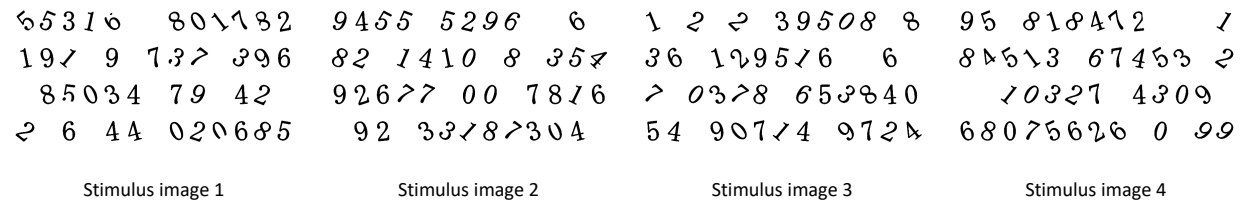


Fig. **12** Comparison of the number of training samples between SSD and participants when focusing on the rate of defect, which is a factor that causes missing detection. (a) 4% and (b) 8% in the simple task of the four marks (club, spade, diamond, and heart).

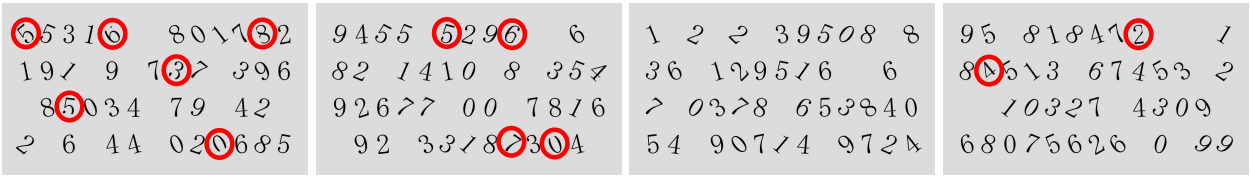
検出が困難な場合である欠損の大きさ 4% において，訓練サンプルの個数により，SSD の検出精度がどのように変化したかを Fig. **12** (a) に示す．精度変化に対して統計的な検定を行うために，ボンフェローニ法を用いた多重比較検定を適用した．検出が困難な場合の (a) において，実験協力者と SSD との間では，訓練サンプルの個数が 2,500 個と 5,000 個と 7,500 個で有意差が見られた．これらの個数では，SSD の精度は実験協力者の精度を下回ることが分かった．よって，欠損の大きさが 4% の場合，実験協力者の検出精度と同等以上となるために必要な SSD の訓練サンプルの個数は，10,000 個以上であると言える．次に，検出が容易な場合である欠損の大きさ 8% において，その精度変化を Fig. **12** (b) で示す．この場合，実験協力者と SSD との間では，訓練サンプルの個数が 2,500 個と 7,500 個と 10,000 個で有意差が見られた．よって欠損の大きさが 8% の場合，学習に必要な訓練サンプル数は 5,000 個以上と言える．

以上の結果より，実験協力者と同等以上の精度を得るために必要な SSD の訓練サンプルの個数は，検出が容易な場合には，困難な場合と比べて 2 分の 1 で済むことが分かった．



Stimulus image 1 Stimulus image 2 Stimulus image 3 Stimulus image 4

(a)



Locations 1 Locations 2 Locations 3 Locations 4

(b)

Fig. **13** Examples of stimulus images of the simple task of the digits (0, . . . , 9). The red circles indicate the locations of anomalous symbols.

6. 単純タスクの設計に関する考察

6.1 実際のラベル欠損との対比

ここまでに検証した単純タスクの妥当性を考えるために，その単純タスクと実際のラベル欠損とを対比しながら議論を進める．現場で実施されるラベル欠損の検査項目は，破れ，穴あき，めくれ，位置ズレ，印字スレ，印字違い，文字欠けなど多岐に渡る．本論文の単純タスクでは，文字欠けを検査項目の対象とした．なお，文字欠けとは，ラベルとして印字された文字の一部が欠損していることを指す．欠損の大きさについて，現場で要求される検出精度は，それぞれの顧客で異なると想定される．そのため本論文の単純タスクでは，検出が難しい大きさの 4% から，容易な大きさの 16% までを用いることとした．単純タスクは，実際のラベル欠損を完全に表している訳ではないが，その一部を表しているため，基礎検証の最初の一歩目としては妥当であると考えている．

さらに基礎検証を進めるために，実際のラベル欠損で扱われる文字の種類について考える．ラベルに印刷される文字の種類として，2. で用いた記号だけではなく，数字やアルファベットや漢字など様々なものが想定される．ここでは，実際のラベルで出現することが多い文字種類を用いて単純タスクを設計することを考える．具体的には，2. から5. までの実験で用いた記号の代わりに数字を用いる．次節にて，数字を用いた場合の単純タスクの詳細と評価について述べる．

6.2 数字を用いた単純タスクにおける精度の評価結果

数字を用いた単純タスクを検証するため，2.2 で述べた手続きに従い刺激画像を生成した．その刺激画像の例を Fig. **13** に示す．シンボルを 0 から 9 までの 10 クラスの数字とした．一枚の刺激画像には，空白文字を含めて 52 個のシンボルを含むこととした．シンボルを横 13 個，縦 4 個の格子状に並べて配置した．シンボルの回転角度は ±45 度の範囲でランダムとした．刺激画像の枚数を 53 枚とした．

数字を用いた単純タスクに必要な訓練サンプルの個数を探るため，実験協力者による検出精度と，深層学習による検出精度とを比較した．実験協力者と深層学習で共に同じ刺激画像のテストサンプル (53 枚) を用いた．実験協力者の人数を 8 名 (日本人学生，男性 8 名，平均年齢 22.9 ± 0.3 歳) とした．訓練サンプル生成を 8 回行い，それぞれの訓練サンプルでネットワークモデルを学習した．深層学習の手法として，5.1 で精度が高かった SSD を用いた．訓練サンプルの個数として，2,500 個，5,000 個，7,500 個，10,000 個を用いた．上記以外の実験条件は，5.1 と同じとした．

Table **3** The label inspection accuracy of participants and SSD on the simple task of the ten digits (0, . . . , 9).

		F-measure	Precision	Recall
Participants		0.89	0.99	0.80
SSD	5,000	0.88	1.00	0.78
	7,500	0.92	1.00	0.85

数字を用いた単純タスクにおける実験協力者と SSD による検出精度の比較を Table **3** に示す．SSD の検出精度 (F 値) は実験協力者の精度と比べて，訓練サンプル数が 5,000 個の場合に低くなったが，7,500 個の場合に高くなった．数字を用いた単純タスクでは，実験協力者の目視と同等以上の精度となるために必要な訓練サンプルの個数は，7,500 個であることが分かった．

次に，実験協力者と SSD が共に苦手としていた変動要因である欠損の大きさ 4% のみに注目し，検出精度を算出した．数字を用いた単純タスクにおいて，訓練サンプルの個数により，SSD の検出精度が，実験協力者の精度と比べて，どのように変化したかを Fig. **14** に示す．精度に対して統計的な検定を行うために，ボンフェローニ法を用いた多重比較検定を適用した．実験協力者と SSD との間では，訓練サンプルの個数が 2,500 個と 5,000 個で有意差が見られた．記号 (クラブ，スペード，ダイヤ，ハート) を用いた場合の Fig. **12** (a) では，訓練サンプルの個数が 7,500 個でも有意差が見られたが，数字を用いた場合の Fig. **14** には 7,500 個で有意差が見られなかった．このことにより，数字を用いた場合，訓練サンプルの個数が数千のオーダーでも，実験協力者の目視と同等の精度となる場合があると考えられる．ただし，実験協力者の目視を超える精度を SSD で得るために，訓練サンプルの個数は数万のオーダーを必要とすると予測される．実際に訓練サンプルの個数を 20,000 個とした場合，実験協力者の再現率が 0.57 に対し SSD の再現率が 0.73 であったため，実験協力者の目視精度を SSD が超えたと言え

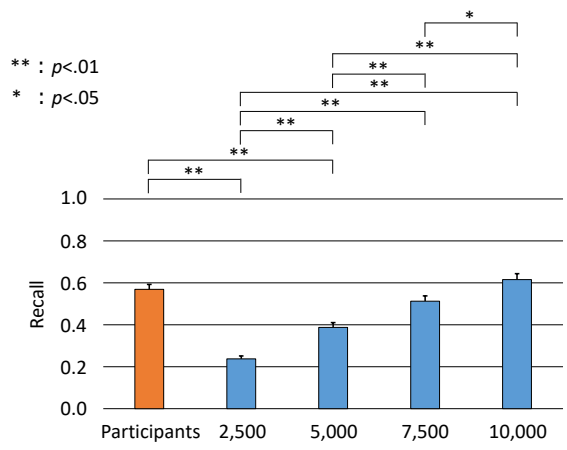


Fig. **14** Comparison of the number of training samples between SSD and participants in terms of defect rate 4% on the simple task of the ten digits $(0, \ldots, 9)$.

る．5.2 の記号を用いた場合の単純タスクでも，20,000 個とした場合，実験協力者の目視精度を超えることを確認した (実験協力者の再現率が 0.76，SSD の再現率が 0.89)．以上より，数字，または，記号を用いた単純タスクにおいて，欠損の大きさが 4% の場合，実験協力者の目視を超える精度を SSD で得るために，訓練サンプルの個数は数万のオーダーを必要とすると考えられる．

本論文では，文字の欠けに関する検出精度において，対象とする文字種類 (記号と数字) の間で，個数のオーダーが共通する一例を実験的に示した．ただし実際のラベル欠損には多様な変動要因が含まれるため，さらなる検証が必要であると考えられる．今後の課題として，アルファベットや漢字など異なる種類の文字を用いた場合の評価，フォントタイプやサイズやレイアウトなど文字の表現方法を変化させた場合の評価，様々な年代や職種の実験協力者との精度比較，ラベルの位置ズレや印字スレなど文字欠け以外の検証項目での実験が挙げられる．

7. まとめ

本論文では，深層学習でラベル欠損を検出するために必要な訓練サンプルの個数を検証するために，単純タスクに取り組む実験協力者の精度と深層学習の精度とを比較した．訓練サンプルを自動で収集するために，ラベル欠損の検出を模擬した単純タスクを設計した．単純タスクにおける実験協力者の検出精度を求め，その精度と同等以上となるネットワークモデルを訓練するために要したサンプル個数を明らかにした．実験結果より，単純タスクにおいて実験協力者の目視と同等以上の精度を深層学習で得るためには，訓練サンプルの個数は SSD で数千のオーダーが必要であることが分かった．また，欠損の大きさにより検出が困難になった場合，SSD の訓練サンプルの個数は数万のオーダーが必要であることが分かった．深層学習の精度が人間の精度を超えるために必要となる訓練サンプルの個数は，記号や数字を用いた限定的な単純タスクではあるが，共通する傾向が見られる例があることを示した．

今後の課題として以下が挙げられる．実際にラベル欠損を検出した場合と単純タスクを用いた場合との間での精度比較が必要である．本論文の実験では欠損が含まれる割合が多い場合を想定し検証を行ったが，実際のシーンではその割合が少ない場合が考えられるため，様々な割合での検証が必要である．また，ラベルには文字だけでなく図柄も含まれる場合があるため，単純タスクの設計方法のさらなる検討が必要となる．

謝辞

本研究の一部は，JSPS 科研費 18H04114 の助成を受けたものである．

参 考 文 献

1) 村上弘記. 自動化システムの最近の動向と今後の展開. 計測と制御, Vol. 54, No. 12, pp. 889–894, 2015.
2) T. S. Newman and A. K. Jain. A survey of automated visual inspection. *Computer Vision and Image Understanding*, Vol. 61, No. 2, pp. 231–262, 1995.
3) J. Beyerer and C. Frese F. P. Leon. Machine vision: Automated visual inspection: Theory, practice and applications. *Springer*, 2015.
4) S. H. Huang and Y. C. Pan. Automated visual inspection in the semiconductor industry: A survey. *Computers in Industry*, Vol. 66, pp. 1–10, 2015.
5) F. Nagata, K. Tokuno, H. Tamano, H. Nakamura, M. Tamura, K. Kato, A. Otsuka, T. Ikeda, K. Watanabe, and M. K. Habib. Basic application of deep convolutional neural network to visual inspection. In *Proceedings of the International Conference on Industrial Application Engineering*, pp. 4–8, 2018.
6) 田中拓哉, 笠原亮介. 画像を用いた自動外観検査技術. 日本画像学会誌, Vol. 55, No. 3, pp. 348–354, 2016.
7) V. Chaudhary, I. R. Dave, and K. P. Upla. Automatic visual inspection of printed circuit board for defect detection and classification. In *Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking*, pp. 732–737, 2017.
8) D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 2019.
9) 脇迫仁, 森勇貴. 光沢のあるプラスチック部品の外観検査. 産業応用工学会論文誌, Vol. 4, No. 2, pp. 45–49, 2016.
10) 永見英臣, 水谷真也, 森田亮介, 伊藤聡. 検査工程における水栓ラベル認識システムの構築. 日本機械学会論文集, Vol. 85, No. 879, pp. 19–00158, 2019.
11) R. Ren, T. Hung, and K. C. Tan. A generic deep-learning-based approach for automated surface inspection. *IEEE transactions on cybernetics*, Vol. 48, No. 3, pp. 929–940, 2017.
12) Y. J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Buyukozturk. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 33, No. 9, pp. 731–747, 2018.
13) B. Rekabdar and C. Mousas. Dilated convolutional neural network for predicting driver's activity. In *Proceeding of the International Conference on Intelligent Transportation Systems*, pp. 3245–3250, 2018.
14) X. Wang, L. Gao, J. Song, X. Zhen, N. Sebe, and H. T. Shen. Deep appearance and motion learning for egocentric activity recognition. *Neurocomputing*, Vol. 275, pp. 438–447, 2018.
15) AI プロダクト品質保証コンソーシアム. AI プロダクト品質保証ガイドライン, May 2019.
16) 経済産業省. AI・データの利用に関する契約ガイドライン, June 2018.
17) 産業技術総合研究所サイバーフィジカルセキュリティ研究センター. 機械学習品質マネジメントガイドライン, June 2020.
18) 松崎優太, 岡安寿繁, 中村明生, 佐藤雄隆, 片岡裕雄. フラクタル幾何学を用いたデータセット構築と特性評価. 第 21 回画像の認識・理解シンポジウム, PS2-62, 2018.
19) S. Guangxin, 片岡裕雄, 佐藤雄隆. フラクタル幾何学を用いたデータセットの拡張および特性評価. ビジョン技術の実利用ワークショップ, IS2-A5, 2019.
20) W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pp. 21–37, 2016.
21) O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.
22) K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.