様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １５ 日現在

機関番号：３４４０６
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１３０９

研究課題名（和文）ディープラーニングに対する電子透かし埋め込みに関する研究

研究課題名（英文）Study on Digital Watermarking into Deep Learning Model

研究代表者
酒澤　茂之（Sakazawa, Shigeyuki）

大阪工業大学・情報科学部・教授

研究者番号：８０５３０８２３

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：学習済みディープラーニングモデルの著作権保護のために、電子透かしを埋め込む技術を開発した。画像認識に用いられる学習モデルに対して、著作権者の情報が視覚的なロゴとして表示できる方式を国際学会IEEE MIPRおよび国内の研究会・シンポジウムで発表し、PCSJ/IMPS優秀論文賞とAVM賞優秀賞の2件の表彰を受けた。また、研究事例のまだ乏しいRNN型学習モデルについても、電子透かし埋め込みが可能であることを明らかにし、国際学会IWAITおよび国内の研究会で発表した。これらの成果を広く社会で活用できるように、プロトタイプソフトをGithubにおいて公開している。

研究成果の学術的意義や社会的意義
ディープラーニングは、様々な画像認識・合成の分野や文章やコンピュータプログラムの作成に無くてはならない技術である。その開発には、大量の整理されたデータや技術者の人件費がかかっている一方で、学習済みモデルのコピーや再利用は容易である。したがって、ビデオや音楽コンテンツと同様に、その著作権者の情報を明らかにする技術「電子透かし」を導入することによって、学習済みモデル自体が一種のコンテンツとして流通する新たなビジネスの枠組みが可能となる。本研究の成果は、その端緒となるものであり、学術界・産業界における研究を促進させた。

研究成果の概要（英文）：We have developed a technique for embedding a watermark for copyright protection of learned deep learning models. We presented our method, which allows copyright holders' information to be displayed as a visual logo on the learned models used for image recognition, at the international conference IEEE MIPR and at research conferences and symposiums in Japan, and received two awards: the PCSJ/IMPS Best Paper Award and the AVM Award for Excellence. We also clarified the feasibility of watermark embedding for RNN-type learning models, for which there are still few research examples, and presented the results at IWAIT, an international conference, and a domestic workshop. We are releasing a prototype software on Github so that these results can be widely used in society.

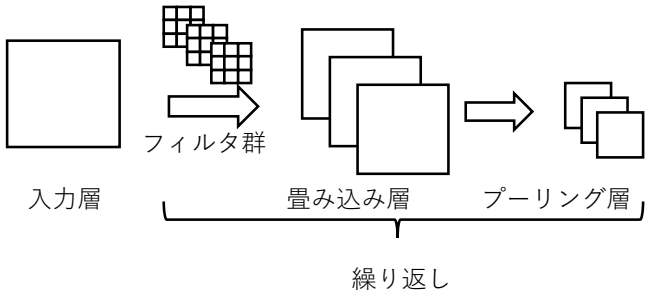
研究分野：コンテンツセキュリティ

キーワード：深層学習　知的財産保護　電子透かし

## １．研究開始当初の背景

ディープラーニング(深層学習)は、入力されたデータからクラス分類や回帰を行う機械学習の手法の一つで、画像認識、自然言語処理等において、高い性能を示すことから注目を集めている。高い性能を達成する上で、非常に大量のデータセットと、膨大な学習計算量を必要とする。一方で、学習済みの深層学習モデルを利用することによる新規システムの開発は比較的容易であり、学習済みのモデルは知的財産的な価値を持つと考えられる。

そうした背景のもとで、研究代表者は、電子透かし分野の研究やコピー画像検索技術の開発経験を踏まえて、急速に重要性の高まりつつある深層学習モデルに着目し、その著作権保護に向けて電子透かしを挿入する研究を行っていた。研究開始時点までに、画像認識の分野で一般的に用いられる畳み込みニューラルネットワーク(Convolutional Neural Network: CNN)に対して、その学習プロセスを操作することで、深層学習モデルの性能を保ちつつ、電子透かしを挿入できることを確認していた。その模式的説明図を図１に示す。CNN は、ネットワーク中に畳み込み演算を行うための多数のフィルタ係数を保持しているが、提案技術は係数の集合に対して透かしの埋め込みを行っているため、深層学習モデル特有の改変処理であるファインチューニングに対する耐性を持っている。しかし、同研究は基本的な特性評価を実施した段階であり、なぜ電子透かしの埋め込みが可能か、埋め込み可能な電子透かしの量はどれだけか、攻撃耐性や不正な読み取りなど安全性はどこまで保証できるか、という検証と評価は未着手であった。さらに、CNN 以外の深層学習モデルに対する電子透かしの埋め込みの可否については、未踏の領域であった。




CNNでは畳み込み層と縮小(プーリング)層の繰り返し処理が行われ、多数のフィルタ係数を持つ。電子透かし埋め込みでは、フィルタ係数群と秘密鍵となる行列と掛け算した結果が特定の値となるように、フィルタ係数値を改変する。

図 1 CNN の構造(一部)と電子透かしの対象

## ２．研究の目的

本研究は、深層学習モデルに対する電子透かしの実用化のために、様々なタイプの深層学習器に対する適用試験を行い、電子透かしの読み取りを妨害する攻撃への耐性、電子透かしの情報の改ざんへの耐性を明らかにすることを目的とし、以下の内容の研究を進める。

① 対象とする深層学習器の種類を増加させる。深層学習器は画像認識を対象とするもの、系列データとなる自然言語処理や音声認識を対象とするものに大別される。それぞれに対する電子透かし埋め込みと、埋め込んだ透かしの検出能力、および深層学習モデルの認識性能に与える影響の評価を行う。

② 電子透かしの検出を妨げる攻撃への耐性を分析し、適用範囲を明らかにする。特に、深層学習モデルに特有の改変手法について検証と考察を行う。

③ 埋め込まれた電子透かしの情報分析や改ざんに対する耐性を分析し、リスクを明らかにする。これまでの画像・テキスト・音声等に対する電子透かしの研究と同様に、不正な電子透かしの埋め込み等の攻撃と評価を行う。

④ プロトタイプの開発と公開を行う。再利用可能なソフトウェアとして電子透かし埋め込み機能と検出機能を持つプロトタイプを開発し、公開する。

## ３．研究の方法

本研究で取り扱う深層学習モデルは、教師有り学習によるものとする。すなわち、入力と正解ラベルの組からなる大量のデータを用いて、入力に対する推論結果と正解ラベルとの誤差が小さくなるように、モデル内部のパラメータ群を調整していくプロセスによって学習される。

深層学習モデルは、対象とするタスクによって大雑把に以下のように大別される。画像分類には畳み込みフィルタ演算をその内部に持つ CNN モデルが、時系列信号やテキスト情報などの系列データにはフィードバック系を内部に含む RNN モデルがそれぞれ用いられる。したがって、本

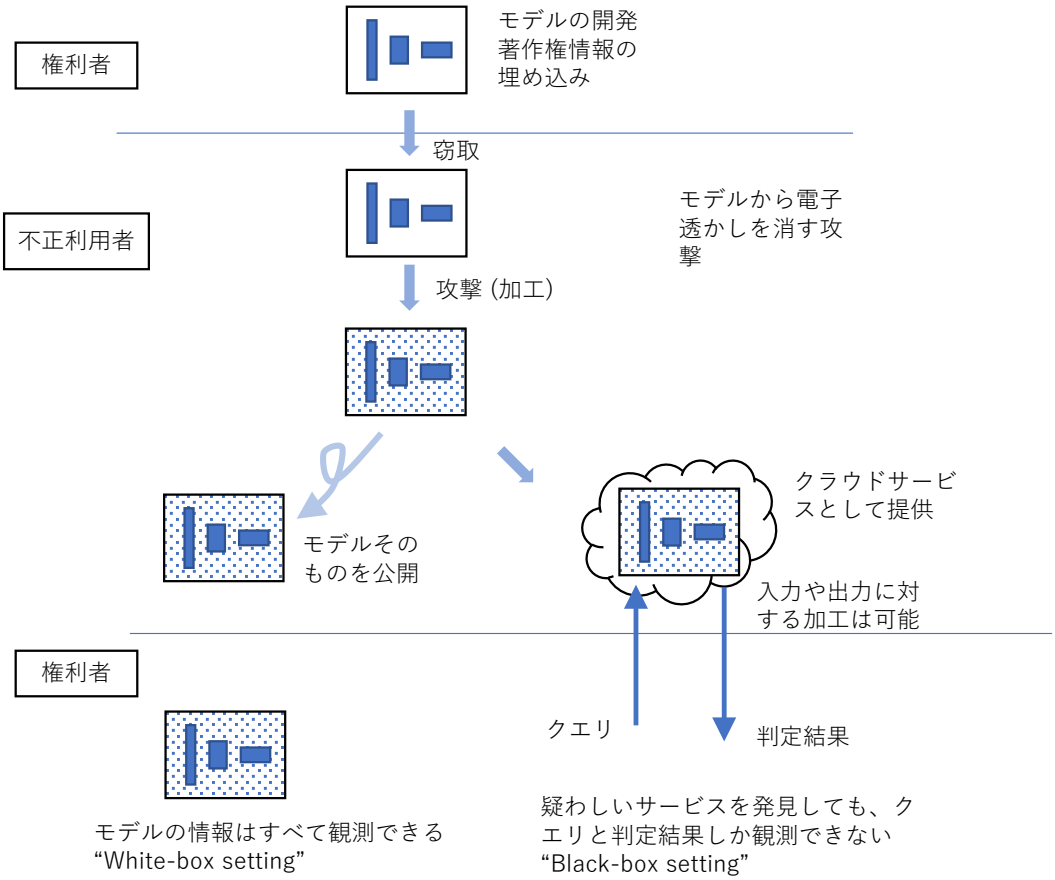



図 2　想定する脅威モデル

研究の遂行においても、この両者のモデルに対する検討が不可欠である。

次に本研究において前提としている、どのように著作権侵害が発生するかという脅威モデルについて述べる。図 2 に示すように、まず権利者が学習モデルの開発を行う。この際に、権利者を示すための情報を電子透かしとして、モデルの内部に埋め込む。そして、何らかの方法で窃取された学習済みモデルが、不正利用者による加工を経て公開される。このとき、公開の方法として、学習モデルの構造やパラメータをすべて開示する”White-box setting”と、学習モデルはクラウドサービスの内部に隠匿しておき、そのモデルへのクエリと推論結果だけが観測できる”Black-box setting”がある。

以上を踏まえて、本研究では、CNN モデルについては White-box と Black-box での電子透かしの埋め込み方法の提案を行い、攻撃環境下での検出精度について分析する。また、RNN モデルについても White-box における電子透かしの可能性を調査する。

## ４．研究成果

### 4.1 CNN モデル向け White-box 電子透かし

先行研究では、CNN モデルに電子透かしを埋め込んだのち、CNN モデルのパラメータを枝刈り処理によって 65%を削除しても電子透かしは問題なく取り出せることが示されている。本研究では、先行研究では未検証であった、量子化を行った場合、枝刈りと量子化両方を行った場合、Quantization aware training を行った場合のそれぞれについて、電子透かしに対する影響を実験的に検証した。

10 種類の画像を見分けるための Cifar-10 データセットを学習させたモデルに対して電子透かしを埋め込む。埋め込む対象の CNN モデルは大小 2 種類の規模のものを用意した。規模の小さいモデルの総パラメータ数は約 200 万、そのうち電子透かしを埋め込む対象となるレイヤーのパラメータ数は約 9 千である。規模の大きいモデルは総パラメータ数が約 3300 万、そのうち電子透かしを埋め込む対象となるレイヤーのパラメータ数は約 3 万 7 千である。

電子透かしを埋め込んだモデルの作成後、以下の攻撃を行う。

枝刈り処理：80%のパラメータを削減。
量子化処理： 8bit 精度で整数量子化。
Quantization aware training：量子化を前提とした再学習を行った後に 8bit 精度で整数量子化する。

枝刈り処理の影響をまとめると、本研究では 256bit の電子透かしを学習モデルに対して埋め込んだが、これに対して枝刈り処理を行った場合は規模の違いによる差が見られなかった。しかし、10bit の電子透かしを埋め込んだモデルに対して枝刈り処理をした場合には、小規模モデル

の方が取り出された電子透かしの劣化が大きくなっていた。電子透かしは埋め込み対象となるレイヤーのパラメータのうち、重み係数となる部分すべてを用いて埋め込まれている。そのため、同じ電子透かしを埋め込んだ際にパラメータに対して埋め込まれている電子透かしの割合が高くなる小規模モデルのほうが、枝刈りによるパラメータ削減の影響を受けやすくなっていると考えられる。

量子化処理と Quantization Aware Training の影響は、量子化前と比較すると取り出された電子透かしの値が離散化され、見かけ上、埋め込んだ値との差分が極めて小さくなっていた。しかし、これは整数量子化によってモデル内のパラメータ値が整数化されたことによるもので、攻撃による透かしの劣化度合いが判別出来なくなる事象を引き起こすことが分かった。

### 4.2 CNN モデル向け Black-box 電子透かし

画像分類を行う CNN モデルにおいて、Black-box setting での典型的な電子透かしの手法は、分類結果が誤ることによっている。例えば、電子透かしの検出用に特別な加工が施された自動車の画像が飛行機と誤認識されることや、ある入力画像群の誤分類の確率が変わるという手法が提案されている。しかし、これらは誤分類の原因が電子透かしによるものなのか、CNN モデルそのものがある確率で誤ることによるのかが自明ではない。その誤分類が統計的に有意に電子透かし由来によることを示すことは可能ではあるが、専門家以外の人にとっては直感的とは言い難い点が課題である。そこで、本研究では、検出のプロセスと結果を DNN モデルの権利者のロゴ画像の形式で表現することができ、直観的な理解と高い納得感が得られる DNN 電子透かし方式を提案した。

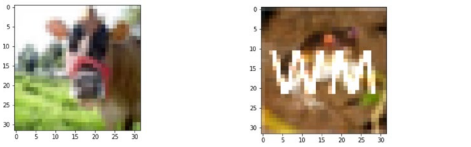



図 3　Black-box 電子透かし用鍵画像

100 種類の画像分類タスクを対象とした提案方式で説明する。画像データセット Cifar-100 は学習用のデータセット 5 万とテスト用のデータセット 1 万から構成され、それぞれ 32x32 画素の自然画像と、100 種類の画像内容のうちどれが正解であるかの正解ラベルのペアからなる。電子透かしを埋め込む対象となる DNN モデルとしては、このタスクに高い性能を示す Wide Residual Network (WRN)を用いた。電子透かしの埋め込みにあたって、本来のタスクの学習のためのデータセットとは別に電子透かし埋め込みのための鍵画像データセットを構築した。図 3 に示すように鍵画像には人間が見て意味のあるパターンを刻印している。この鍵画像とペアとなる正解ラベルに工夫を施しており、この正解ラベルの値を二次元的に並べ直して視覚化すると、権利者のロゴ画像となるようにしてある。検出の手順を図 4 に示す。

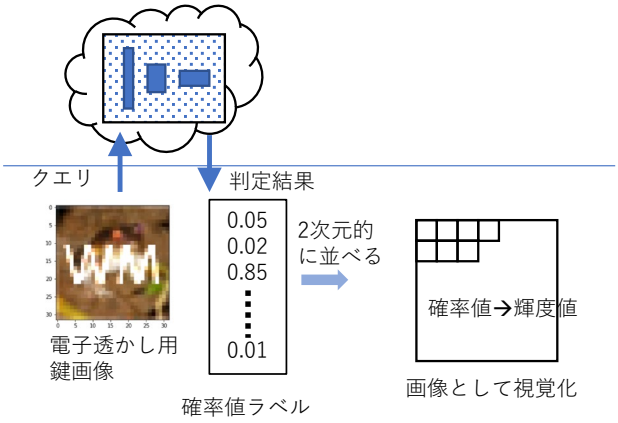



図 4　Black-box 電子透かし視覚復号

ロゴ画像としての表現にあたって、複数の鍵画像による判定結果である複数のラベルを用いることとする。このとき、鍵画像の入力枚数に応じて、判定結果のラベル値すなわちロゴ画像の輝度値を順次加算していく。その結果を図 5(a)に示す。重ねる画像枚数を増やすにつれて、権利者のロゴが浮かび上がってくることが分かる。また、学習モデルに対する攻撃やクエリ画像に対する攻撃についても検証しており、その例を図 5(b)に示す。(a)と比較してロゴの明瞭さは劣るものの、主観的にロゴが視認可能であることから攻撃耐性があるといえる。また、電子透かしを埋め込んでいない学習モデルに、鍵画像をクエリとして入力した結果を図 5 (c)に示す。結果はランダムなロゴ画像であり、鍵画像そのもの

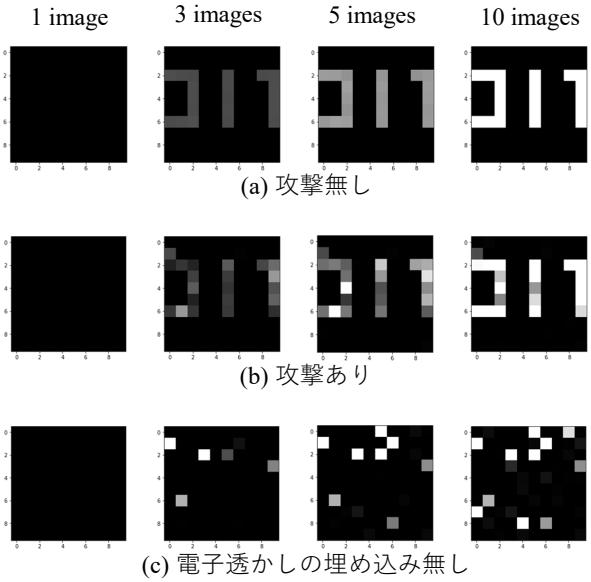


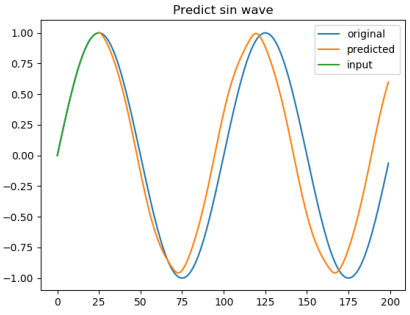

図 5　Black-box 電子透かしロゴ復号結果

がロゴを生み出すわけではないことが示された。なお、これは電子透かしシステムの観点からは、存在しない電子透かしを検出してしまう”False positive”が発生しないことを意味し、電子透かしに求められる重要な要件を満たしている。

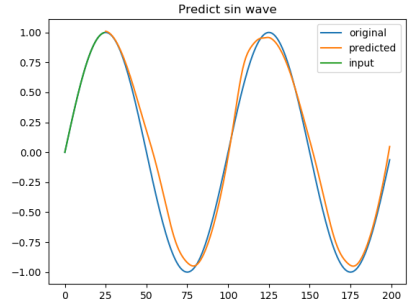
### 4.3 RNN モデル向け White-box 電子透かし

再帰型ニューラルネットワーク(Recurrent Neural Network) とはネットワーク構造の中に再帰的な構造を持つニューラルネットワークの一つである。RNN では株価の推移や音声・文章データなどの時間経過で値が変化する時系列データを扱うことが可能であり、文章解析や音声認識分野などで用いられる。時系列データを分析する上での特徴は、各データの時点で独立した状態とみなさず、特定の時点のデータが以降に発生するデータに対しても影響を及ぼすものとして考慮する点である。そのため、現在の入力と直前の状態の入力との 2 つの入力をもとに学習を行うための構造を持っている。現在の入力に対するパラメータを Kernel weight、直前の状態の入力に対するパラメータを Recurrent weight と呼ぶ。前節まで用いていた CNN モデルは現在の入力だけを用いるものであることから、RNN モデルに対する電子透かしでは、特に、Recurrent weight に対する埋め込みの可否が関心の対象となる。

まず、シンプルなタスクとして、1 周期 100 個のデータを持つ sin 波 2 周期分を学習データとして sin 波の予測を行う学習モデルを構築する。用いた RNN モデルは Simple RNN であり、Kernel weight は 20 個、Recurrent weight は 400 個のパラメータを持っている。電子透かしの埋め込みについては、内田らの方式を基にして、埋め込む対象の重み係数を上述の Kernel weight と Recurrent weight のそれぞれに対して行った。実験の結果、いずれの weight についても電子透かしの埋め込みと検出は可能であった。また、電子透かしの埋め込みが学習モデルの本来のタスクに与える影響を見るために、正弦波形の予測結果をグラフに示す。図 6 において、0 から 25 サンプルの緑色の入力に基づいて予測したオレンジ色のプロットが、正解値の青色のプロットと類似した波形を描いていることが分かる。Kernel weight の方は電子透かし埋め込み対象のパラメータ数が 20 と少ないこともあり、正解値との乖離が若干大きくなっているが、Recurrent weight の方はパラメータ数 400 で正解値との乖離が少なくなっている。




(a) kernel weight

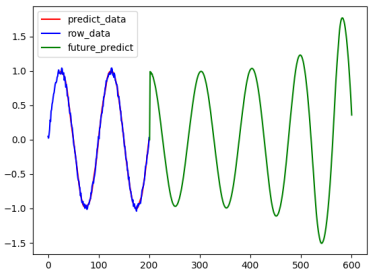



(b) recurrent weight

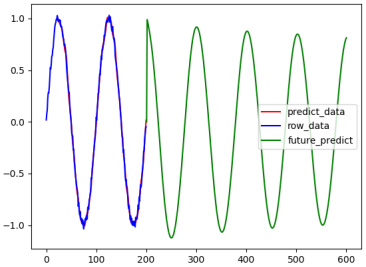
図 6 シンプルな sin 波予測

次に、より複雑なタスクとして、ノイズが重畳されている sin 波形の予測を行った。今回は、LSTM と呼ばれるより複雑な学習モデルを用いており、Kernel weight は 1200 個、Recurrent weight は 360,000 個となっている。埋め込んだ 8 ビットの電子透かしに対して、Kernel weight の方は 100%検出できたが、Recurrent weight の方は 87.5%の検出率にとどまった。本来のタスクに与える影響をグラフ化したものを図 7 に示す。青色の入力値に基づく緑色の予測値を見ると、500 サンプル以降での予測外れは大きくなっているもののおおむね妥当な予測ができていることが分かる。

以上から RNN 電子透かしの可能性が示された。




(a) kernel weight




(b) recurrent weight

図 7 ノイズ入り sin 波予測

### 4.4 成果の公表

本研究の成果を第三者が検証できるようにするために、Github においてソースコードを公開している。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計15件（うち招待講演　2件／うち国際学会　3件）

| |
|---|
| １．発表者名<br>衣川晃弘，酒澤 茂之 |
| ２．発表標題<br>画像ロゴを表現できる深層学習電子透かし方式とその評価 |
| ３．学会等名<br>情報処理学会AVM研究会（招待講演） |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>山地雄大，酒澤茂之 |
| ２．発表標題<br>CNNモデル向け電子透かしの軽量化耐性 |
| ３．学会等名<br>情報処理学会AVM研究会 |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>松本幸大，酒澤茂之 |
| ２．発表標題<br>電子透かしを用いたRNN 学習モデルの保護 |
| ３．学会等名<br>情報処理学会AVM研究会 |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>衣川晃弘，酒澤 茂之 |
| ２．発表標題<br>31×31 画像ロゴを表現できる深層学習電子透かし方式の一検討 |
| ３．学会等名<br>情報処理学会AVM研究会 |
| ４．発表年<br>2021年 |

１．発表者名
酒澤茂之

２．発表標題
10x10 画素ロゴを表現可能な深層学習電子透かし方式

３．学会等名
電子情報通信学会EMM研究会(2020年度第1回)

４．発表年
2020年

１．発表者名
松本幸大、酒澤茂之

２．発表標題
電子透かしを用いたRNN 学習モデルの保護

３．学会等名
情報処理学会AVM研究会(第112回)

４．発表年
2021年

１．発表者名
山地雄大、酒澤茂之

２．発表標題
CNNモデル向け電子透かしの軽量化耐性について

３．学会等名
情報処理学会AVM研究会(第112回)

４．発表年
2021年

１．発表者名
衣川晃弘、酒澤茂之

２．発表標題
31×31 画像ロゴを表現できる深層学習電子透かし方式の一検討

３．学会等名
情報処理学会AVM研究会(第112回)

４．発表年
2021年

１．発表者名
酒澤　茂之

２．発表標題
研究テーマの開拓法：画像符号化伝送から 深層学習モデルの著作権保護まで

３．学会等名
令和元年　電気関係学会関西連合大会　講演論文集（招待講演）

４．発表年
2019年

１．発表者名
酒澤　茂之

２．発表標題
深層学習モデルに対する視覚復号可能な電子透かし方式

３．学会等名
2019年度画像符号化シンポジウム／2019年度映像メディア処理シンポジウム

４．発表年
2019年

１．発表者名
小林 栄介，酒澤 茂之

２．発表標題
DNNへの電子透かし埋め込みの特性調査

３．学会等名
第108回オーディオビジュアル複合情報処理研究発表会

４．発表年
2020年

１．発表者名
酒澤茂之

２．発表標題
深層学習モデルに対する電子透かしの要件分析

３．学会等名
2018年映像情報メディア学会年次大会

４．発表年
2018年

１．発表者名
Shigeyuki Sakazawa, Emi Myodo, Kazuyuki Tasaka, Hiromasa Yanagihara

２．発表標題
Visual Decoding of Hidden Watermark in Trained Deep Neural Network

３．学会等名
2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)（国際学会）

４．発表年
2019年

１．発表者名
Matsumoto Kota、Sakazawa Shigeyuki

２．発表標題
A feasibility study of watermark embedding in RNN models

３．学会等名
SPIE Proceedings Vol. 12177: International Workshop on Advanced Imaging Technology (IWAIT) 2022（国際学会）

４．発表年
2022年

１．発表者名
Yamaji Yudai、Sakazawa Shigeyuki

２．発表標題
Tolerance of CNN watermarking against model optimizations

３．学会等名
SPIE Proceedings Vol. 12177: International Workshop on Advanced Imaging Technology (IWAIT) 2022（国際学会）

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
| --- | --- |