深層学習の内部表現に基づく画像認識を指向した情報源圧縮の研究

鈴 木　　聡 志

2021 年度 博士論文

深層学習の内部表現に基づく画像認識を指向した情報源圧縮の研究

Image Recognition-Aware Data Compression Based on Intermediate Representation of Deep Learning

2022 年 2 月

指導教員　庄野 逸 教授

電気通信大学 大学院
情報理工学研究科 情報学専攻

鈴木　聡志
Satoshi Suzuki

Abstract

Image recognition applications that use deep neural networks (DNNs) have achieved high performance. This achievement sparks the demand for DNN-based practical applications on front-end devices, such as self-driving vehicles, autonomous drones, and surveillance systems. However, because DNNs have many parameters, processing is too computationally expensive to perform on such front-end devices. Therefore, image recognition applications on front-end devices often use cloud servers to mitigate the computational budget of the DNNs on the front-end ones.

Communication ways for image recognition applications to use cloud servers are roughly divided into two ways, called "cloud-based intelligence" and "collaborative intelligence". In the former, the image signal captured on the front-end device is transmitted to the cloud server. Then, a DNN in the cloud server recognizes this image. The latter splits and distributes the DNN model between the front-end and cloud. Therefore, the process of the DNN relies on not only the cloud server but also the front-end device. Furthermore, in the collaborative intelligence, the front-end device transmits "deep features", which are the outputs of the DNN on the front-end device, to the cloud server.

Both of the above two ways need to transmit a large number of image signals or deep features to the cloud server. Therefore, to improve transmission efficacy, the bitrate reduction using the data compression technique is an important factor. The data compression technique, such as image compression, is widely studied as an application field of information theory, signal processing, and so on. However, since most of these studies are intended for television broadcasting, they assume that compressed signal is viewed by humans ultimately. Therefore, the existing data compression methods lack the viewpoint of image recognition of the DNN. Therefore, we have the possibility to obtain higher compression efficiency by introducing the viewpoint of image recognition to existing methods. In this thesis, we propose new data compression methods in two types of communication ways, cloud-based and collaborative intelligence, by introducing the viewpoint of image recognition. We try to propose data compression methods that maintain the image recognition accuracy of DNN as much as possible when the bitrate is largely reduced.

Conventional data compression methods are designed with insights into the property of the image signal and the property of human perception. Similarly, we need to design

the data compression to maintain the accuracy of the DNN with understanding the property of the DNN. To achieve this goal, we focus on the studies on the intermediate feature representation of the DNN, which has been remarkably studied in recent years. These studies mainly aim to understand the DNN, such as "why does the feature representation extracted by DNN work well?". Therefore, they did not intend for the data compression method. However, several studies are expected to provide a good insight into the image recognition performed by the DNN. In this thesis, we present a methodology by utilizing the outcome of these studies for designing data compression.

First, we studied image data compression aimed at image recognition in the cloud-based intelligence. In Chapter 3, we proposed an image pre-transformation method that appropriately transforms images before they are compressed. The proposed image pre-transformation method maintains only the signals that are important for DNN recognition in the input image and transforms the other ones so that the bitrate becomes low during compression. Our method showed a bitrate reduction effect while maintaining the recognition accuracy in all the tested compression standards and obtained a bitrate reduction effect of up to 20.5 %.

Second, we studied data compression aimed at image recognition in the collaborative intelligence. In Chapter 4, we proposed a novel spatio-temporal arrangement method that spatially arranges the deep features as images and temporally arranges them as a video. The spatio-temporal arranged deep features are compressed by the video compression method. This spatio-temporal arrangement method utilizes the redundancy of the deep features that could not be removed by the previous methods and realizes high compression efficiency. As a result of the evaluation, we found that the proposed method can achieve higher compression efficiency than the previous methods.

Based on the above results, we have succeeded in improving the compression efficiency by introducing the viewpoint of image recognition of DNN into data compression. This enables us to obtain efficient communication between front-end devices and cloud servers in image recognition applications based on DNNs. We also showed that image recognition-aware data compression can be developed based on the studies on the intermediate representation of DNN. This methodology may accelerate the development of recognition-aware data compression in the future.

要旨

深層学習モデルの一種である Deep Neural Network (DNN) の発展によって，画像認識技術は飛躍的な精度向上を果たした．この精度向上に起因して，自動運転車や自動運転ドローン，監視システムなどのフロントエンドデバイスで動作する画像認識アプリケーションの実社会応用への期待が高まっている．しかし，DNN は一般に多くのパラメータを持ち，計算負荷が非常に大きいため，利用できる計算資源が限られているフロントエンドデバイスのみで DNN を動作させることは現実的ではない．このため，フロントエンドデバイスで動作する多くの画像認識アプリケーションでは，クラウドサーバを利用し，豊富な計算資源の下で認識を行うことで，フロントエンドデバイスの計算負荷を低減させている．

DNN を用いた画像認識アプリケーションがクラウドサーバを利用するための通信の方式は，クラウド型知能方式と協調型知能方式に大別される．前者は，フロントエンドデバイス上で撮像した画像信号をクラウドサーバに伝送し，DNN の認識を全てクラウドサーバで完結させる．後者は，DNN を 2 つに分割し片方をフロントエンドデバイスに配置することで，フロントエンドデバイスでも DNN の一部の処理を行う．後者の場合，フロントエンドデバイスからクラウドサーバに伝送されるのは画像信号ではなく，フロントエンドデバイス側の DNN の出力である深層特徴となる．上記の二種類の方式は，いずれも，逐次撮像される大量の画像信号もしくはそこから抽出した深層特徴をクラウドサーバに伝送する必要がある．これらの大量のデータを効率良く伝送するために，情報源圧縮技術を用いて通信に要するビットレートを低減させることは画像認識アプリケーションの実用化において重要な要素である．ビットレート低減のために，一般に，クラウド型知能方式では撮像した画像信号に対して画像圧縮を行っている．また，協調型知能方式でも，深層特徴を擬似的に画像もしくは映像信号に変換して圧縮を行うことが一般的である．

画像の情報源圧縮は，情報理論や信号処理における応用研究という位置づけで，多くの研究がなされてきた．しかし，それらの研究のほとんどはテレビジョン放送への応用を考えているため，自然画像を圧縮し，最終的には人間が視聴することを前提とした設計がなされている．したがって，クラウド型知能方式や協調型知能方式が想定している DNN の画像認識という観点を導入することで，既存の情報源圧縮手法よりも高い圧縮効率を得られる可能性がある．上記のような背景に鑑みて，本論文では，クラウド型知能方式および協調型知能方式の二種類の通信方式において，通信に要するビットレートを低減しても DNN の認識精度を可能な限り保持する観点を導入した新たな情報源圧縮の研究を行う．

従来の情報源圧縮手法は，圧縮対象である画像の信号特性と視聴する対象である人間の知覚特性に対する深い洞察の上で設計されている．本論文が対象とする DNN の精度保持を目的とした情報源圧縮も，ちょうど同じように DNN の認識に対する特性を十分に把握した上で設計する必要がある．この目的を達成するために，本論文では，近年著しく研究が進んでいる DNN の内部の特徴表現 (内部表現) に関する研究に着目した．DNN の内部表現に関する研究は，“DNN が抽出する特徴表現はなぜ上手く働くのか” といった DNN の挙動の理解を中心に考えて行われている研究であり，情報源圧縮の開発のために行われているわけではない．しかしながら，いくつかの研究成果は DNN の行う画像認識に対して深い洞察を与えることが期待される．本論文では，これらの研究成果を応用し，DNN の認識精度を保持するという観点を導入した情報源圧縮を設計する方法論を提案する．

第 3 章では，クラウド型知能方式における画像認識を指向した情報源圧縮技術について研究し，画像を圧縮する前に適切に変換する画像プレ変換手法を提案した．提案する画像プレ変換手法は，入力画像の信号の中で DNN の認識に対して重要な情報のみを保持し，それ以外の信号を圧縮時にビットレートが低くなるように変換する．提案手法は，実験を行った全ての圧縮標準で認識精度を保持しつつビットレート低減効果を示し，最大 20.5% のビットレート低減効果を得た．

第 4 章では，協調型知能方式における画像認識を指向した情報源圧縮技術について研究し，深層特徴を構成する要素である特徴マップを複数枚の画像から成る映像のように配置する時空間的配置法を新たに提案した．時空間的配置法は，従来手法で除去することができなかった冗長性を除去し，高い圧縮効率を実現する．評価実験の結果，提案する時空間的配置法は，従来手法と比較して高い圧縮効率を達成できることが分かった．

以上の研究成果より，DNN の画像認識精度を保持するという新たな概念を情報源圧縮に導入し，圧縮効率を向上させることに成功した．これによって，深層学習による画像認識アプリケーションにおけるフロントエンドデバイスとクラウドサーバ間の効率的な通信が可能になったと考えられる．また，DNN の内部表現に関する研究によって得た知見を基にすることで，画像認識を指向した情報源圧縮が開発できることを示した．この方法論の確立によって，今後の本論文が対象とする研究分野において，DNN が抽出する特徴表現の特性に基づいた情報源圧縮の実現が加速し，継続的かつ加速度的に発展することが期待される．

目次

第 1 章

序論

1.1 研究背景

深層学習モデルの一種である Deep Neural Network (DNN) [1–7] の発展によって，画像認識技術は飛躍的な精度向上を果たした．DNN は，入力画像から処理に必要な特徴表現を学習の過程で獲得する機構を有している．DNN が抽出する特徴表現は，タスクごとの最適化学習の結果として得られた表現であるため，Scale Invariant Feature Transforms (SIFT) [8,9] 等の人間が設計した特徴表現と比較して良好な性能を示し，多くの画像認識タスクでデファクトスタンダードとなりつつある [1,10–12]．この DNN の発展に起因して，自動運転車や自動運転ドローン，監視システムなどのフロントエンドデバイスで動作する画像認識アプリケーションの実社会応用への期待が高まっている．しかしながら，一般に DNN は多くのパラメータを持ち，計算負荷が非常に大きい [13,14]．このため，一般に利用できる計算資源が貧弱なフロントエンドデバイスのみで DNN の処理を完結させることは現実的ではない．このような背景と，近年の通信インフラの整備に伴って，Amazon Web Services [15] や Google Cloud Platform [16]，Microsoft Azure [17] といったクラウドサーバを利用した画像認識アプリケーションが増加している．一般に，クラウドサーバはフロントエンドデバイスと比較して潤沢な計算資源を有している．したがって，クラウドサーバを利用することでフロントエンドデバイスの計算負荷を低減させることができ，画像認識アプリケーションの実現可能性が高まることが期待される．

クラウドサーバは，一般に，フロントエンドデバイスと物理的に離れた場所に存在している．このため，画像認識アプリケーションがクラウドサーバを利用するにはインターネット等を介した通信が必要になる．画像認識アプリケーションがクラウドサーバを利用するための通信方式は，大まかに二種類の方式に分類される．図 1.1 にその概要を示

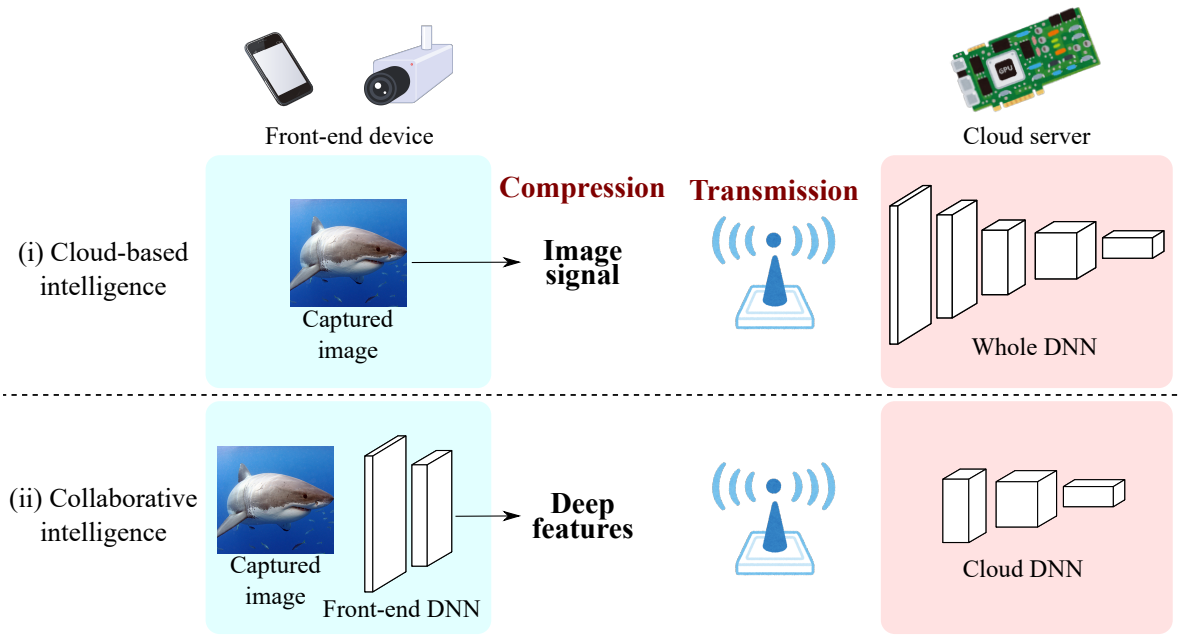

図 1.1: DNN を用いた画像認識アプリケーションのための二種類の通信方式についての概要図．(i) はフロントエンドデバイス上で撮像した画像信号をクラウドサーバに伝送し，DNN の認識を全てクラウドサーバで完結させるクラウド型知能方式を示している．また，(ii) は DNN を 2 つに分割しそれぞれをフロントエンドデバイスとクラウドサーバに配置することで，フロントエンドデバイスでも DNN の一部の処理を行う協調型知能方式を示している．

す．(i) に示す一つ目の方式では，撮像した画像をクラウドへ伝送し，クラウドサーバで DNN を用いた画像認識を行う．本論文では，この方式を “クラウド型知能 (Cloud-based Intelligence)” 方式と呼ぶ．取得した信号をクラウドへ伝送する通信方式は，画像認識に限らず，多くのクラウドサーバを利用したアプリケーションで採用されているため，比較的一般的な方式であると言える．(ii) に示す二つ目の通信方式では，DNN を 2 つに分割し，フロントエンドデバイスとクラウドサーバにそれぞれを配置する．“協調型知能 (Collaborative Intelligence)” 方式と呼ばれるこの方式では，画像信号の代わりに，フロントエンドデバイスに配置した DNN の出力である “深層特徴” をクラウドサーバに配置した DNN へ伝送することで画像を認識する．協調型知能方式は，クラウド型知能方式と比較して一般的ではないが，高い伝送効率を示し，結果として低遅延性やフロントエンドデバイスの消費電力の面でクラウド型知能方式よりも優れていることが明らかになってきている [18–20]．

上記の二種類の方式は，いずれも，逐次撮像される大量の画像信号もしくはそこから抽出した深層特徴をクラウドサーバに伝送する必要がある．これらの大量のデータを効率良

く伝送するために，情報源圧縮技術を用いて通信に要するビットレート[*1]を低減させることは画像認識アプリケーションの実用化において重要な要素である．例えば，圧縮によって非圧縮の状態からビットレートが半減した場合，クラウドサーバとの通信にかかる時間も同様に半減する．また，ビットレートの低減が不十分で，インターネット回線上に大量のパケットが流入すると，輻輳と呼ばれる現象が発生する恐れがある[*2]．輻輳が発生すると，通信速度の大幅な低下，あるいは通信システムそのものがダウンしてしまう等の弊害が発生し，画像認識システムそのものが動作しなくなる可能性がある．したがって，通信に要するビットレートを可能な限り圧縮によって低減させた上で，伝送することが重要となる．ビットレートを低減させるために，クラウド型知能方式では一般に撮像した画像信号に対して画像圧縮を行う．また，協調型知能方式でも，深層特徴を擬似的に画像もしくは映像信号に変換して圧縮処理を行うことが一般的である [21]．ただし，ビットレートを大きく低減させるためには，画像信号や深層特徴から情報を削減する非可逆の圧縮処理を行う必要があり，この情報削減は DNN の精度に悪影響を及ぼす．つまり，ビットレートの低減と DNN の認識精度保持はトレードオフの関係にあると言える．したがって，ビットレートを低減させても認識精度を可能な限り保持するような情報源圧縮技術を開発することでクラウドを利用する画像認識アプリケーションの実現可能性は大きく向上すると考えられる．

本論文は，上記のような背景に鑑み，クラウド型知能方式 · 協調型知能方式の 2 つの通信方式において，通信に要するビットレートを低減しても DNN の認識精度を可能な限り保持する情報源圧縮技術を提案し，画像認識アプリケーションの実現可能性および実用性を高めることを目的とする．

1.2 従来の情報源圧縮技術

画像を情報源としてその画質をできるだけ保ちつつコンパクトに表現する情報源圧縮技術，つまり画像圧縮技術は 1950 年代のデジタル映像黎明期から今日に至るまで，長年にわたって研究がなされてきた [22–25]．画像圧縮技術は大きく分けて，原画像を完全に復元できる可逆圧縮と完全に復元することはできない非可逆圧縮の 2 種類に分けられる．本節では，可逆圧縮と非可逆圧縮を題材に，画像圧縮の基本構成を概説する．

本節では，可逆/非可逆圧縮を図 1.2 に示す大まかな処理系に基づいて説明する．なお，さらに詳細な各処理系の説明は 2.5 節で行う．可逆圧縮は，入力となる画像信号から情報を抽出 (Extraction of information) し，抽出した情報をできる限り短い符号長で表現

[*1] 圧縮処理後のデータ容量を指す．

[*2] 例えば，電気通信大学の半期成績開示日に学務情報システムに対してアクセスが集中する際に良く見られる現象である．

図 1.2: 情報源圧縮技術の基本的な構成．図中中央の量子化 (Quantization) 処理を担う処理ブロックは非可逆圧縮の場合に利用される．可逆圧縮の場合はこの処理ブロックは利用されない．

するためのエントロピー符号化 (Entropy coding) を行う．非可逆圧縮では，上記の 2 つの処理に加えて抽出した情報に対する量子化処理 (Quantization) を行う．情報抽出処理は，画像信号に含まれている冗長な情報をできるだけ取り除いて信号を無相関化し，量子化/エントロピー符号化すべき情報を抽出する処理である．一般に画像信号は，近傍画素では画素値は類似していることが多い，といったような冗長性を多く有する．このような冗長性を削減することで，圧縮した際のビットレートを低減しつつも入力信号を可逆的に復元することが可能になる．非可逆圧縮では，情報抽出処理の後に冗長性を削減しながら抽出した信号に対して，数値精度を下げる量子化処理を行う．量子化処理では量子化誤差が発生するため，その誤差に含まれる情報は非可逆的に削減される．非可逆圧縮ではこの処理を導入することで，画像を完全には復元できなくなる反面，可逆圧縮よりも高いビットレート低減を実現できる．可逆圧縮の場合，その圧縮率の理論的な限界が示されている [26][*3]ため，達成すべきターゲットとなるビットレートがある場合，非可逆圧縮を用いる方が実用上好ましいと考えられる．また，非可逆的に削減される情報は，可能な限り画質に影響を及ぼさないことが望ましい．このため，量子化処理では人間が差分を知覚しにくい高周波の信号から積極的に情報削減を行うことが一般的である．エントロピー符号化は上記のような処理を行った後の信号を符号化対象として，できる限り短い符号長で表現できるように二進符号を割り当て，二値化されたビットストリームとして出力する．このビットストリームのデータ容量をビットレートと呼ぶ．

上記のように，従来の画像圧縮は情報抽出 · 量子化 · エントロピー符号化の 3 つの処理によって圧縮を実現しているが，この中でも，とりわけ，情報抽出と量子化は画像信号とその適用対象に対する深い洞察の上で設計されている．具体的にはそれぞれ，

1. 情報抽出処理では，入力である画像信号の特性を十分に把握した上で，冗長性除去を行っている．
2. 量子化処理では，圧縮した画像信号を人間が視聴することを前提に，人間が変化を知覚しにくい情報から優先的に情報削減を行っている．

[*3] Shannon の情報源符号化定理もしくは第一基本定理として広く知られている．

といったような設計がなされている．しかし，上述の通り，本論文はクラウドサーバを利用する画像認識アプリケーションが主な対象であり，必ずしも入力は画像信号とは限らない上に，圧縮される信号は DNN の認識に用いられることを前提としている．したがって，従来の画像圧縮技術は，本論文で目的とするような “深層学習を用いた画像認識を指向する情報源圧縮” とはなっていないと考えられる．

以上のような研究背景に基づいて，本論文では，DNN の認識精度保持という観点を新たに導入することで，既存の情報源圧縮手法と比較して高い圧縮効率を得られるのではないか，と考えた．この基本的なアイデアをベースに，本論文では，クラウド型知能方式および協調型知能方式のそれぞれにおいて既存の圧縮手法よりも高い圧縮効率を得ることを目的とした圧縮手法を提案する．

1.3 研究の位置づけ

本論文では，クラウド型知能方式および協調型知能方式の二種類の通信方式において，通信に要するビットレートを低減しても DNN の認識精度を可能な限り保持するという観点を導入した新たな情報源圧縮の研究を行う．また，本論文の提案手法では，新たな圧縮技術を一から設計するのでは無く，既存の情報源圧縮技術をそのまま利用しつつ，DNN の認識精度を保持するような観点を導入するアプローチを採用する．これは，多くのフロントエンドデバイスには，既存の情報源圧縮技術に基づくエンコーダ (圧縮を担う処理系) が LSI (Large Scale Integration) のような形で既にハードウェア実装されている [27] ことに起因する．各フロントエンドデバイスのハードウェア上に実装されているエンコーダを DNN の認識のために全てリプレースすることは現実的では無く，実用上，大きな課題となる可能性が高い．本論文で採用するようなアプローチによって，フロントエンドデバイスに既に実装されている圧縮技術を効率的に再利用でき，実用性が高い技術となることが期待される．

また，1.2 節で述べたように，従来の圧縮技術は，情報抽出と量子化において画像信号とその適用対象である人間の知覚特性に対する深い洞察の上で設計されている．DNN の精度保持を目的とした情報源圧縮でも同様に，DNN の特性を十分に把握する必要があると考えられる．具体的には，それぞれの通信方式で以下のような特性を把握することが重要であると考えられる．

クラウド型知能方式 情報抽出を行う入力信号は従来の画像圧縮と同様に画像信号である．しかし，圧縮後の画像信号の用途は主に DNN が行う認識であり，人間が視聴することは主な用途ではない．このために，DNN が，どのような画像信号のどのような変化を知覚しやすい/しにくいのか，といった特性に則って情報削減を行う

必要がある．

協調型知能方式 情報抽出を行う入力信号はもはや画像信号ではなく，フロントエンドデバイスに配置されている DNN から出力される深層特徴である．情報源圧縮に際して，冗長性を除去し，高い圧縮効率を実現するためには，DNN が抽出した特徴表現である深層特徴がどのような性質を持った信号なのかを十分に把握する必要がある．

しかし，DNN が特徴表現を抽出する機構はブラックボックス化されており [28]，上記のような DNN の特性を具体的に把握することは容易ではない．このような，DNN のブラックボックスな振る舞いは，DNN の認識精度を保持するという観点を導入した情報源圧縮の研究に対して大きなボトルネックとなっている．

このボトルネックの解消を目指して，本論文では，近年著しく研究が進んでいる DNN の内部の特徴表現 (内部表現) に関する研究に着目した．DNN の内部表現に関する研究は，“DNN が抽出する特徴表現はなぜ上手く働くのか” といった DNN の挙動の理解を中心に考えながら行われている研究であり，本論文で取り組む情報源圧縮のために行われている研究ではない．しかしながら，いくつかの研究成果は DNN の抽出する特徴表現に関する深い洞察を与えており [29–31]，情報源圧縮の研究開発にも応用が可能であると考えられる．本論文では，これらの研究成果を上手く応用することで，DNN の認識精度を保持するという観点を導入した情報源圧縮を設計する方法論を提案する．この方法論は，従来の動画像圧縮技術が人間の視覚特性に学んで人間の視覚に沿った情報源圧縮へと成長したのと同様に，DNN の画像認識のおける特性を学ぶことで，画像認識を指向する情報源圧縮技術を実現することができるのではないか，というナイーブな発想に基づいている．この方法論が確立できれば，今後の本論文が対象とする研究分野において，DNN が抽出する特徴表現の特性に基づいた情報源圧縮の実現が加速し，継続的かつ加速度的に発展することが期待される．

1.4 本論文の構成

本論文は，以下の 5 つの章からなる．本章は序論であり，本論文の研究背景や従来技術の概要，研究の位置付けについて述べた．第 2 章では，本論文で取り扱う研究テーマに関連する先行研究について述べる．具体的には，深層学習による画像認識技術とその内部表現に関する研究について概要を説明した後，クラウド型知能方式 · 協調型知能方式と，それらに応用するための情報源圧縮の従来技術について説明する．第 3 章では，クラウド型知能方式への応用を念頭に，画像を圧縮しても可能な限り DNN の認識精度を保持するように設計した画像プレ変換手法を提案する．第 4 章では，情報源圧縮の協調型知能方式へ

の応用を念頭に，深層特徴に存在する冗長性を除去し，圧縮効率を高めるための新たな手法を提案する．第 5 章は本論文の結論であり，各章で得られた知見から本論文をまとめ，今後の課題と展望について述べる．

第 2 章

関連研究

本章では，本研究の主題である画像認識を指向した情報源圧縮に関する研究を紹介する．具体的には，2.1 節および 2.2 節で深層学習による画像認識技術について概説し，2.3 節でその内部表現について述べる．2.4 節ではその実応用のための通信方式，つまりクラウド型知能方式と協調型知能方式について述べる．次いで，2.5 節で従来の情報源圧縮技術について，広く用いられている圧縮標準を題材に概説し，2.6 節および 2.7 節でそれぞれ，クラウド型知能方式と協調型知能方式における情報源圧縮の先行研究について述べる．

2.1 深層学習による画像認識

近年，画像認識技術は，深層学習モデルの一種である，Deep Neural Network (DNN) によって飛躍的な精度向上を遂げた．画像認識を行う DNN において，最も一般的に用いられるモデルは畳み込み演算を用いたモデルである [32, 33]．本節では，この畳み込み演算を用いた DNN を中心に，深層学習を用いた画像認識技術に関して概説する．

畳み込み演算を用いた DNN による画像認識は，Fukushima の提唱した Neocognitron [34, 35] にルーツを持つ．Neocognitron は哺乳類の視覚システムの仕組みを模倣し工学的に応用する，という発想から発明された Neural Network (神経回路) モデルである．図 2.1 に Neocognitron の処理の概要を示す．Neocognitron は Hubel と Wiesel [36] が発見したネコ 18 野における，単純型細胞 (Simple Cell) と複雑型細胞 (Complex Cell) を階層型 Neural Network の枠組みで実現したモデルである．単純型細胞は細胞の反応に影響を与える空間領域である受容野 (Receptive Field) の特定の位置に特定の方位の入力が存在する場合に反応を示す細胞である．一方，複雑型細胞は単純型細胞に対して入力の位置には寛容的で，受容野内に特定の方位が入力されれば，反応を示す細胞である．Neocognitron ではこれらの細胞をそれぞれ畳み込み (Convolution) 演算と空間プーリン

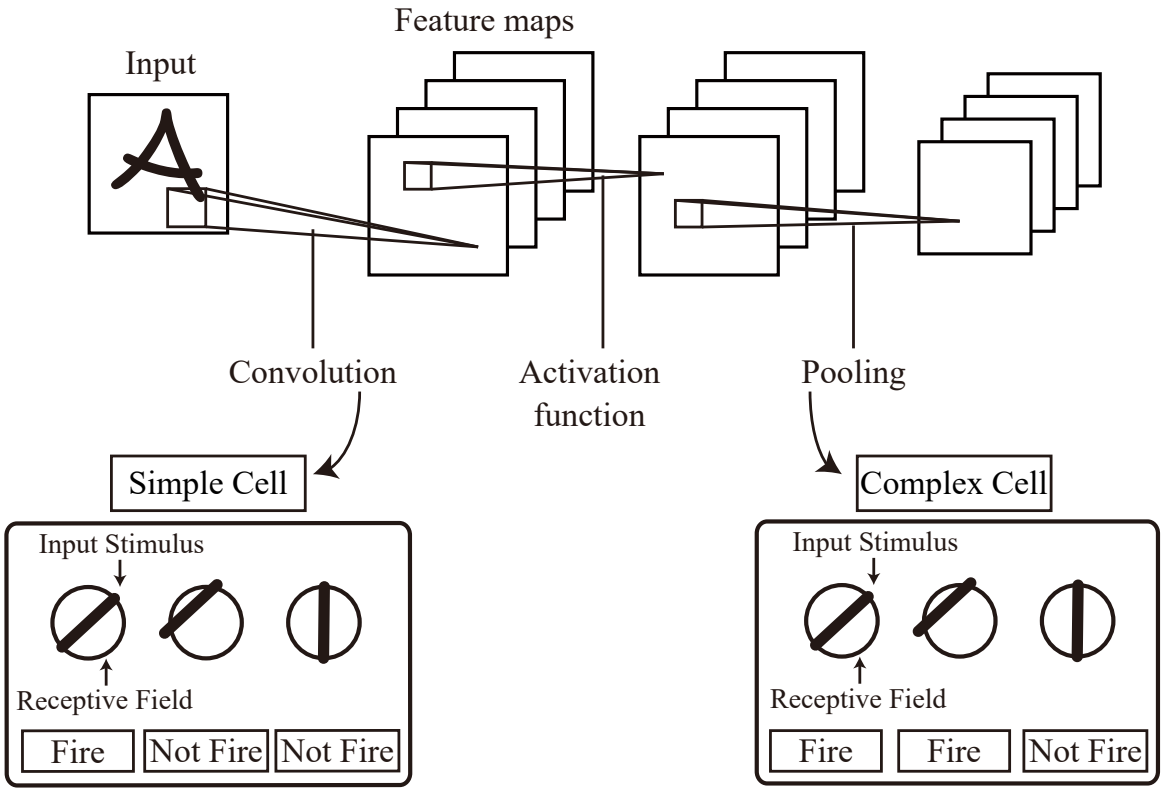

図 2.1: Neocognitron の構成要素である畳み込み · 空間プーリング演算と単純 · 複雑型細胞との関係．畳み込み演算が単純型細胞，空間プーリング演算が複雑型細胞をそれぞれ実装している．

グ (Pooling) 演算を用いて実現している．畳み込み演算では，入力に対してエッジや線分などの局所的な特徴をフィルタ演算によって抽出し，それぞれの特徴の位置を示す特徴マップ (Feature maps) を出力として得る．空間プーリング演算では，得られた特徴マップに対して空間的近傍の出力をまとめる演算を施す．Neocognitron では，畳み込み · 空間プーリングを繰り返し適用することで入力から特徴表現を抽出している．このような，畳み込み · 空間プーリングを利用したモデル構造は，現在の DNN においても一般的な構造となっている [1]．また，Neocognitron も含め，多くの DNN モデルは畳み込み演算の後に活性化関数 (Activation function) による処理を導入している．活性化関数に関しては 2.2.1 節で説明する．

Neocognitron は，現代的な DNN モデルが持つモデル構造の設計要素のほとんどを含んでいたが，その学習方法に関しては教師なし学習を採用していた [34, 35, 37]．LeCun ら [32, 33] は，Neocognitron と多層パーセプトロン [38] に相当する全結合型の Neural Network を組み合わせたアーキテクチャを誤差逆伝播法 [38, 39] を用いて学習させる

手法を提案した．現在，画像認識で用いられる DNN でも，Neocognitron に類似した畳み込み演算を持つモデルに対して誤差逆伝播法を用いて学習を行う．DNN は，画像や音声信号そのものを入力として扱い，従来のパターン認識システムにおける特徴抽出器に相当する機能を学習によって獲得する [40]．DNN が抽出する特徴量は，タスクごとの最適化学習の結果として得られた表現であるため，Scale Invariant Feature Transforms (SIFT) [8,9] や Histograms of Oriented Gradients (HOG) [41] 等の人間が設計した特徴量と比較して良好な性能を示し，多くの画像認識タスクでデファクトスタンダードとなっている [1,10–12,42]．

Fukushima や LeCun らの研究は主に，1980-90 年代頃に行われており，現代的な DNN の多くの構成要素は既に 20 世紀の段階で完成していた．しかし，画像認識分野で飛躍的な精度向上を実現した Krizhevsky らによる AlexNet [1] は 2012 年に提案されたものである．この間，画像認識分野においては，Neural Network を利用する手法が主流であったとは言い難い状況で，主に，上記で触れた SIFT や HOG を画像からの特徴抽出として用い，Support Vector Machine [43] によって認識を行う手法がデファクトスタンダードであった．各要素技術に大きなブレイクスルーがあったわけではないのにも関わらず深層学習が勃興した要因として，GPGPU (General-Purpose computing on Graphics Processing Units) 等の計算機の発達と Web や Social Networking Service (SNS) の普及によるデータセットの整備が進んだことが挙げられる [37,40]．多量のパラメータを持つ DNN を十分に学習するためには多量の学習画像が必要となり，また，その学習を現実的な時間内に行う場合には，高速な計算機が必要となる．要素技術以外にも上記のような背景も深層学習の発展には重要であり，特に学習データの整備は深層学習の適用に対して非常に重要であることが実験的に明らかになっている [44]．

2.2 画像認識を行う深層学習モデルの構成要素

前述の通り，画像認識を行う DNN の基本的な構造は畳み込み・空間プーリング演算であり，そのような構造を持つモデルを誤差逆伝播法によって学習する．以下では，本論文の実験で用いる VGG [2] と ResNet [4] に関連するものを中心に，本論文の実験を理解するために必要ないくつかの要素を詳細に説明する．

2.2.1 活性化関数

一般に，DNN の各階層の処理結果には活性化関数と呼ばれる非線形関数を用いた処理が施される．DNN における畳み込みや全結合型の Neural Network による処理は線形演算であり，非線形性は存在しない．活性化関数を導入することで，DNN の処理に非線形

性を与えることができる．

画像認識を行う DNN において，最も一般的に用いられる活性化関数として，Rectified Linear Unit (ReLU) [45] がある．ReLU は，式 (2.1) のように入力値 x が 0 以上の場合，入力値そのものを線形に出力する恒等関数である．

$$\mathrm{ReLU}(x) = \max(0, x). \tag{2.1}$$

DNN では，式 (2.1) のような処理が，各階層の処理結果である数値すべてに施される．Krizhevsky らによる AlexNet [1] で採用された ReLU は，活性化関数のデファクトスタンダードの一つであり，本論文の実験で用いる VGG や ResNet でも採用されている．また，ReLU から派生した活性化関数も存在しており，これらも DNN に良い性能をもたらすことが経験的に知られている [46, 47]．

2.2.2 正規化

DNN は各階層で畳み込み演算等の線形な演算を施し，活性化関数による処理を経て処理結果をより深い階層へ伝播する．DNN の各階層は処理を行いながらパラメータを学習によって更新するため，更新の過程で処理結果の分布が変化する可能性がある．この変化は各階層で増幅するため，階層数の多い DNN の学習において，適切な学習を阻害する要因となっていた．最も単純な解決策は，誤差逆伝播法の学習率を小さい値とすることで処理結果の変化分を小さく抑える方法であるが，i) 学習の進捗が遅くなる，ii) 局所最適解に陥りやすくなる，といった課題があった．

現在，多くの DNN で広く用いられている Batch Normalization (BN) [48] は，各階層の処理結果に対して正規化処理を導入することで上記の問題を解決しようとする手法である．なお，BN は DNN の学習で広く用いられているミニバッチ学習を前提としている．ミニバッチ学習とは，データ数 M のデータセットから m 個のデータをミニバッチとしてランダムに抽出し，ミニバッチごとにモデルのパラメータを学習する学習方式である．BN は，式 (2.2) のように特定の各階層の処理結果である x をミニバッチ単位で平均 β，分散 γ^2 となるように正規化し，正規化処理後の x' をより深い階層へ伝播する．

$$x' = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta. \tag{2.2}$$

$$\text{Where, } \mu = \frac{1}{m}\sum_{i}^{m} x_i,\ \sigma^2 = \frac{1}{m}\sum_{i}^{m}(x_i - \mu). \tag{2.3}$$

ここで，ϵ はゼロ除算を防ぐパラメータであり，非常に小さい値を設定することが一般的である．また，平均 β および分散 γ^2 は学習によって適切なパラメータを決定する．DNN の各階層の処理に対して活性化関数の前に BN を導入するのみで，高い認識精度

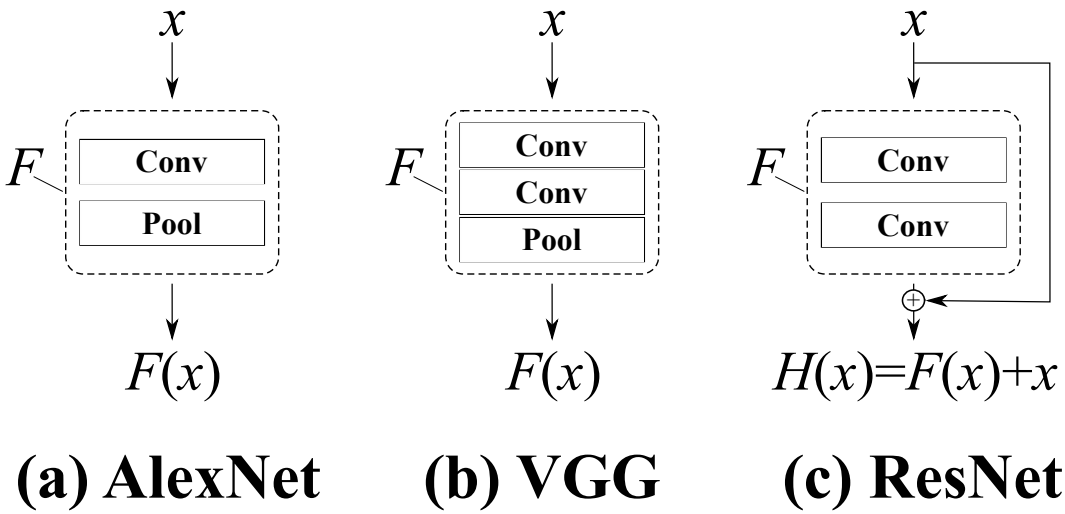

図 2.2: 画像認識を行う代表的な DNN モデルの畳み込み演算等の概要図．(a) は AlexNet [1]，(b) は VGG [2]，(c) は ResNet [4] それぞれの DNN モデルに対応する概要図を示している．

を得ることが可能になるため，ResNet 等のモデルで幅広く用いられている [4]．なお，VGG には BN が用いられていない[*1]が，PyTorch [49] 等のフレームワークでは，VGG に対して BN を適用したモデルが公開されており，従来の VGG よりも高い精度を示すことが知られている．

2.2.3 モデル構造

VGG のモデル構造

VGG [2] は，図 2.2 (b) に示すように，2 層の畳み込み演算と空間プーリング演算をまとめて一つの処理ブロックとして，このブロックを繰り返すモデル構造を有する．この畳み込み演算では，AlexNet よりも小さい畳み込みカーネルを用いることで，AlexNet よりも深い構造ながらモデルパラメータを可能な限り少なくするように設計されている．このようなカーネルを持つ畳み込み処理は，ResNet をはじめとした様々な DNN において一般的に用いられている．また，VGG は階層数の違うモデルがいくつか提案されている [2] が，その中でも高い精度を示す 16 層もしくは 19 層から成る DNN モデルを用いることが一般的である．本論文の実験においても 16 層から成る VGG モデルである VGG-16 を利用する．

[*1] VGG が提案された 2013 年頃には BN はまだ提案されていなかった．

ResNet のモデル構造

ResNet [4] は，残差学習を取り入れた DNN モデルであり，従来のモデルと比較して非常に階層数が多い (すなわち，“深い”) 構造が特徴のモデルである．残差学習は，図 2.2 (c) に示すようなバイパス構造を取り入れている．従来の AlexNet や VGG は，図 2.2 (a,b) に示すように，入力 x に対して DNN の各階層での処理から得られる出力を $F(x)$ とした際に，最適な $F(x)$ を学習する．一方で，ResNet 等の残差学習を採用している DNN モデルでは，入力 x と $F(x)$ の和を $H(x) = F(x) + x$ とした際に，最適な $H(x)$ を学習するような構造となっている．DNN は，階層を増やすほどに学習に必要な誤差勾配情報が指数的に消失してしまう課題が存在した [4] が，この残差学習ではバイパス構造の導入によって，誤差勾配情報の消失が発生しにくい構造となっている．ResNet は，残差学習に BN やパラメータの初期化 [50] 等の工夫を組み合わせることにより，100 層を超えるような，非常に深いモデルを学習することを可能にしたモデルであり，VGG 等の従来の DNN と比較して高い認識精度を示す．残差学習は，単純な構造ながら，深い DNN モデルの学習には不可欠な要素技術として，ResNet 以降に提案された様々な DNN モデルで採用されている [5–7].

2.3 深層学習の内部表現

深層学習によって画像等のメディア処理の精度が高まるにつれて，“深層学習がなぜ画像等のメディアを適切に処理できるのか” という疑問を解明する研究が活発になされるようになった．このような研究は様々な観点でなされており，例えば，数理的な枠組みで DNN の学習手法や要素技術を解析する研究 [51–55] や，Neural Network の層の数が多い場合と少ない場合とで表現能力にどのような差があるのかを理論的に解析した研究 [56–58] 等がある．本論文では，序論にて述べた通り，

1. DNN がどのような画像信号のどのような変化を知覚しやすい/しにくいのか
2. DNN が抽出した特徴表現である深層特徴がどのような性質を持った信号なのか

といった 2 つの点を把握することで，DNN の画像認識を指向した情報源圧縮手法を実現しようとしている．したがって，以下では数多くある深層学習の解析に関する研究から DNN がどのような画像特徴をどのようにその内部で表現しているのか，という点について行われた研究について説明する．

AlexNet の登場以降，人間と同等程度の画像認識精度を示す DNN が人間とどの程度同じように画像を見るのかについては多くの研究がなされてきた．Le ら [59] は，教師なし学習によって得られた巨大な DNN モデル内部のニューロンの出力を最大化させる画像信号を擬似的に作り出し，人間やネコといった概念に特異的に反応するニューロンが存在

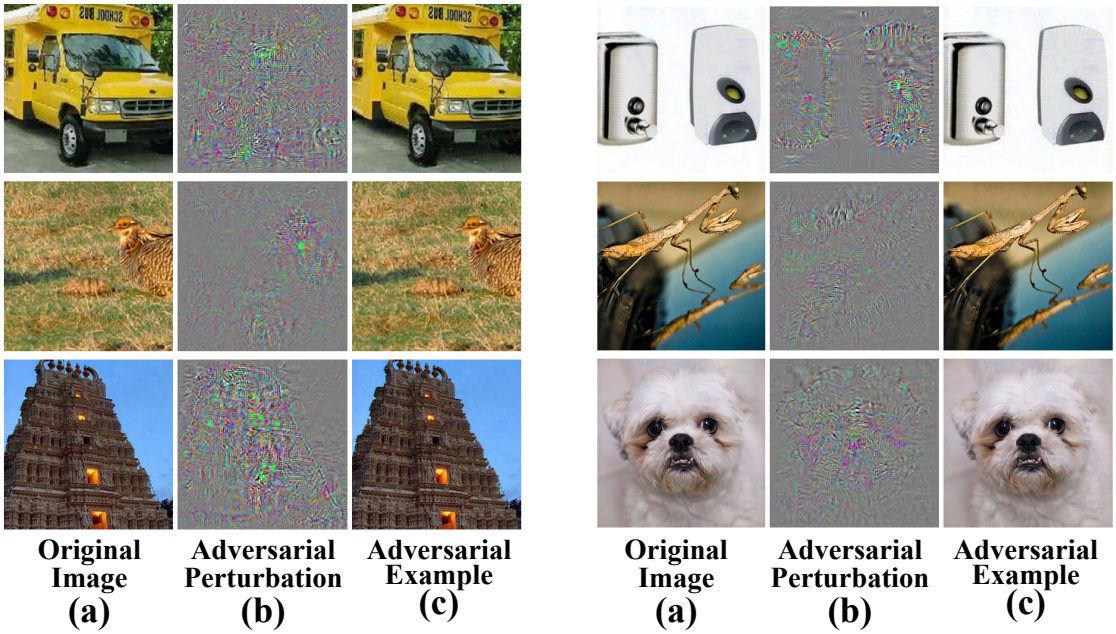

図 2.3: 敵対的摂動 (Adversarial Perturbation) の具体的な例．画像は Szegedy ら [62] から引用したものである．図中 (a) は認識を行う原画像，(b) は敵対的摂動，(c) は敵対的摂動を重畳した結果の画像である．DNN は図中 (a) は正しく認識を行えるが，(c) の画像は全てダチョウと認識してしまう [62].

することを示した．また，Yamins ら [60] は，DNN は人間の脳内に存在する神経細胞の活性パターンを良い精度で予測できることを明らかにした．これらの知見は，人間の視覚システムと DNN の類似性を指摘している研究成果である．しかし，人間と DNN の画像を認識する機構は必ずしも同一ではない，という知見も得られてきている．その中で最も代表的な研究は敵対的摂動 (Adversarial Perturbation) に関するものである [61–63]. 敵対的摂動とは，認識対象の画像に重畳すると DNN の認識結果を大幅に変化させるノイズ状の摂動を指す．敵対的摂動の例を図 2.3 に示す．なお，この図は Szegedy ら [62] から引用したものである．図中 (a) は認識を行う原画像，(b) は敵対的摂動，(c) は敵対的摂動を重畳した結果の画像を示す．DNN は図中 (a) に示す画像は正しく認識を行えるが，(c) の画像は全てダチョウと認識してしまう [62]．一般に敵対的摂動は，非常に微小な摂動であるために，我々人間は知覚できないが DNN には大きな影響を及ぼす．実際に，図中 (a) と (c) は，ほとんど差分を知覚することはできない．敵対的摂動に関する研究によって人間と DNN の画像認識に関する機構には異なる点が存在することが明らかになった [64].

上記のような敵対的摂動に関する研究に鑑みて，近年では，人間と DNN の画像認識

における差異はなぜ発生するのか，という点に関する定性的 · 定量的な解析がなされている．Jo ら [65] は，画像に写っている物体のクラス (コンセプト) を人間が視認できる状態を保持したまま周波数空間で画像信号に摂動を与え，DNN の認識精度を調査した．その結果，DNN による画像認識は大きな精度低下を示した．この研究成果は，DNN は画像のコンセプトに対応する画像信号を認識に活用しているのではなく，何か別の画像信号を用いて認識を行っていることを示唆している．Geirhos ら [29] は，DNN が認識に対して活用している信号を物体形状とテクスチャに分割して検討した．人間の被験者を用いた画像認識実験と比較すると，DNN は形状よりもテクスチャを認識に対して活用していることが明らかになった．このような，DNN が認識に用いる画像信号に関する研究は現在でも主要な研究テーマであり，人間と DNN ではやはり認識に対して異なる画像信号を用いるという成果が多く得られている [66–71].

DNN が抽出している特徴表現に関する研究も本論文の主題に関連する重要な研究テーマである．ただし，図 2.1 に示す通り，DNN が抽出する特徴表現は畳み込み演算のフィルタリング結果に対して再度畳み込み演算を行った結果であるために，深い階層の特徴表現は具体的な特徴量を見るだけでは適切な解析を行うことは難しい．Zeiler ら [30] は，Deconvolutional Network [72] と呼ばれる手法を用いて，大規模自然画像データセットで学習された AlexNet の内部のニューロンが特異的に反応する画像特徴を可視化し，人間が理解できる形で表現する手法を提案した．この Zeiler らの手法は，DNN の抽出した特徴表現の可視化という新たな分野を切り開き，多くの後続研究がなされている [73–75]. 例えば，Simonyan ら [73] は上記の Zeiler らの手法と DNN の学習に用いる誤差逆伝播法との関連性を指摘し，誤差逆伝播法で入力まで伝播された誤差信号が類似の可視化結果であることを示した．また，Springenberg ら [74] は，Zeiler らの手法では可視化が成功しない DNN モデルが存在することを示し，安定した可視化結果をもたらす手法を提案している．これらのいずれの手法においても，DNN は浅い層ではエッジや線分などの局所的な画像特徴に反応し，階層が深くなるにつれて反応する画像特徴が抽象化されることが明らかになっている．上記のような誤差逆伝播法に基づく可視化手法以外にも，DNN が注視している領域をヒートマップの形式で可視化する手法も開発されている [76–78]．これらの手法は，DNN が所望の画像領域に対して正しく反応を示しているかどうかの確認や特定の物体の位置推定 [76] などに広く用いられている．

2.4 クラウド型知能方式と協調型知能方式

フロントエンドデバイスで動作する画像認識アプリケーションが DNN を適用するには潤沢な計算資源を有するクラウドサーバを利用することは重要である．序論で述べた通り，クラウドサーバの使い方には大まかに，クラウド型知能方式と協調型知能方式の 2 つ

の伝送方式がある．本節では，この 2 つの通信方式に関して概説する．

クラウド型知能方式は，取得した画像信号をクラウドへそのまま伝送する通信方式である．この通信方式は，Amazon Web Services [15] や Google Cloud Platform [16]，Microsoft Azure [17] 等，多くのクラウドサーバを利用したアプリケーションで採用されており，一般的な方式であると言える．多くの画像認識アプリケーションはクラウド型知能方式を前提としているが，近年，フロントエンドデバイスの電力消費量やアプリケーション全体の低遅延性でより優れた性能を示す協調型知能方式 [18–20] が注目されている．協調型知能方式では，DNN を 2 つに分割し，フロントエンドデバイスとクラウドサーバにそれぞれを配置する．画像信号の代わりに，フロントエンドデバイスに配置した DNN の出力である深層特徴をクラウドサーバに配置した DNN へ伝送することで画像を認識する．Kang ら [18] は，この協調型知能方式の枠組みにおいて，フロントエンドデバイスの電力消費量やシステム全体の低遅延性の観点で最適な DNN の分割点を検証した．この結果，多くの DNN モデルおよび通信状況 (WiFi や LTE 等) で，最適な分割点は DNN モデルの中間の階層であることが経験的に明らかになった．また，Kang らの検証によれば，中間層の中でも比較的深い層が最適な分割点である傾向が見られている．なお，この検証においては最も浅い層である入力層の前で分割する場合，つまりクラウド型知能方式，および最も深い層である出力層の後で分割する場合，つまり全ての DNN の処理がフロントエンドデバイスで完結する方式も比較対象に入っている．これらに対しても協調型知能方式が高い伝送効率を示しているということは，DNN の最適な分割点は入力層や出力層のような極端な箇所ではなく中間部分である，という事実を示唆していることに他ならない．分かりやすさのため，上記の議論の概要図を図 2.4 に示す．

上記のような研究によって，協調型知能方式はクラウド型知能方式と比較して伝送効率が高いことが明らかになったが，この事実のみで協調型知能方式があらゆる実用上のユースケースでクラウド型知能方式よりも優れていると結論付けることはできない．例えば，協調型知能方式では深層特徴を伝送するため，入力となった画像そのものを人間が見ることはできない[*2]．これは，DNN が行った画像認識が適切かどうか人間が最終的な判断を行ったり，学習データとして伝送された画像を蓄積したい場合には好ましい性質とは言えない．したがって，クラウド型知能方式や協調型知能方式はそれぞれ求められるユースケースが異なり，いずれの通信方式においても，より圧縮効率を高めるための情報源圧縮技術を提供することは重要である．このような観点に鑑みて，本論文では，それぞれの通信方式において圧縮効率を高める情報源圧縮技術を提案する．

[*2] 深層特徴のような特徴表現から入力画像を推定する手法 [79–81] はいくつか提案されているが，深い層から抽出した特徴表現を完全に復元することは困難である．これは，DNN による階層的な処理の過程で入力画像の情報の多くが取り除かれていることが原因であると考えられる．

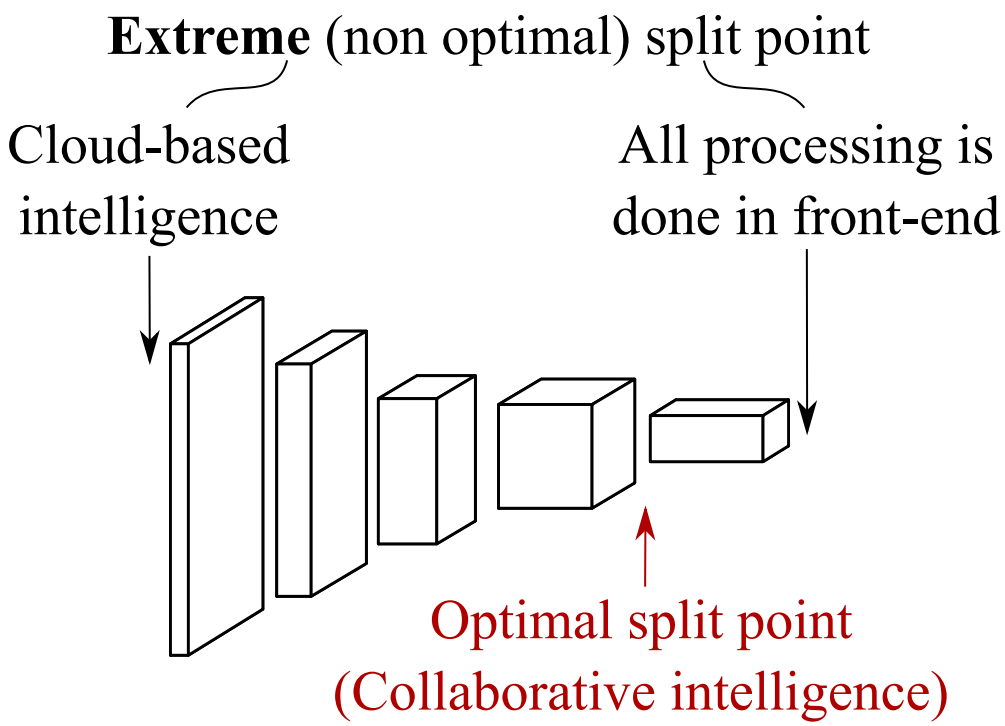

図 2.4: Kang ら [18] の検証によって得られた結果の概念図．DNN の入出力に相当する階層での分割は協調型知能方式の極端なケースであり，伝送効率の点で最適ではない．多くの場合，DNN モデルの深い層で分割することで低遅延性や電力効率の点で最適な伝送が実現できる．

2.5 画像 · 映像の圧縮に関する従来技術

序論では，従来の情報源圧縮の基本的な考え方を画像信号の圧縮を題材に解説した．本節では，画像圧縮のみならず，その対象を映像信号に拡張した動画像圧縮技術も含めて概要を説明する．具体的には，JPEG [22]，JPEG2000 [23]，H.265/HEVC (High Efficiency Video Coding) [24] および最新動画像圧縮標準の H.266/VVC (Versatile Video Coding) [82, 83] などの既存の圧縮技術において，動画像の圧縮を担う処理系であるエンコーダについて説明する．なお，圧縮された信号の解凍処理を担う処理系であるデコーダに関しては，本節で説明するエンコーダと逆の処理を行うものである．

JPEG や JPEG2000 のような画像圧縮標準は大まかに，周波数変換 [84, 85]· 量子化 [86]· エントロピー符号化 [87, 88] の 3 つの処理機構を有する．現在，広く用いられている映像圧縮標準である H.265/HEVC [24] や最新の映像圧縮標準の H.266/VVC [83] では，上記の処理機構に加えて，入力信号の時空間方向の冗長性を利用した予測処理であるフレーム内予測 [89] とフレーム間予測 [90] を導入している．ここで，フレームとは，

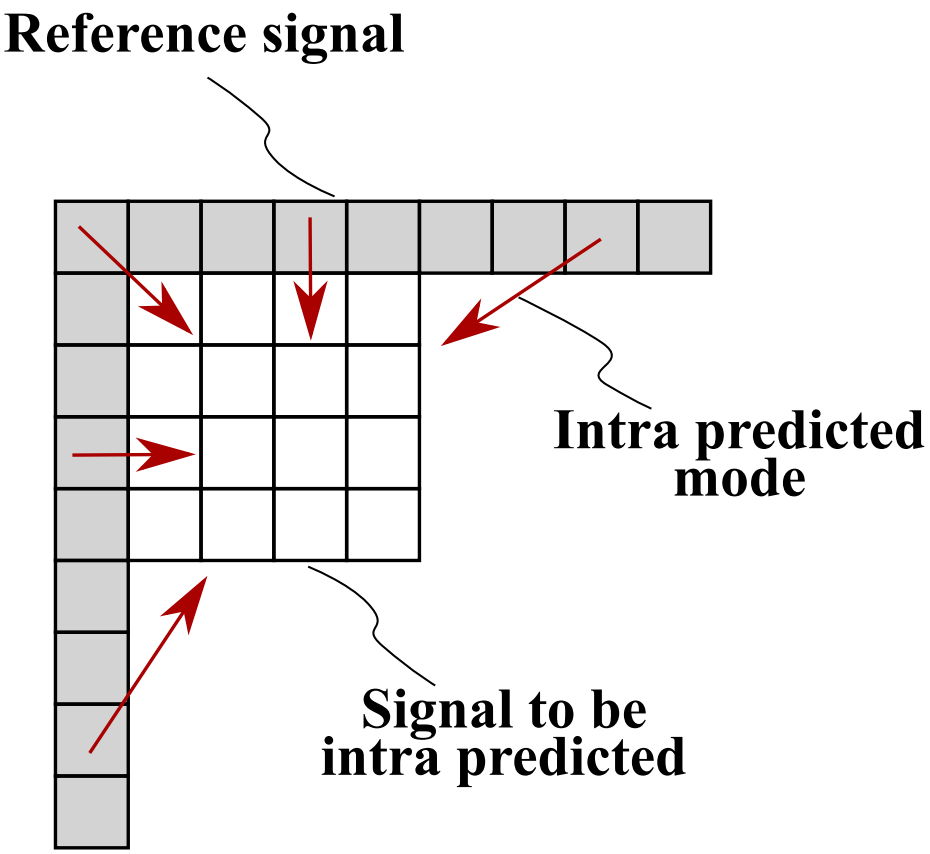

図 2.5: HEVC のフレーム内予測の概要．図中の灰色の箇所は既に復号済みの予測の参照信号，白色の箇所は予測対象の信号をそれぞれ示す．赤矢印は方向性予測の予測方向をいくつか例示しているものである．

静止画像のことを指す[*3]．フレーム内部での空間方向の予測をフレーム内予測，フレーム間の時間方向の予測をフレーム間予測とそれぞれ呼ぶ．HEVC におけるフレーム内予測およびフレーム間予測の処理の概要を図 2.5 と 2.6 にそれぞれ示す．図 2.5 に示すフレーム内予測では，予測対象の信号 (Signal to be intra predicted) に対して 33 方向から選択された予測方向の参照信号 (Reference signal) を予測値としてコピーする方向性予測，参照信号の平均を予測値とする直流 (DC) 予測，参照信号に距離に応じて重み付けした値を予測値とするプレーナー (Planar) 予測の 3 つの予測方式が定義されている．画像信号は，一般に空間的に近傍の画素値は似通っていることが多いため，フレーム内予測によって空間的な冗長性を除去できる．また，図 2.6 に示すフレーム間予測は，入力が映

[*3] 画像圧縮では，圧縮対象となるフレームは 1 つのみであり，映像圧縮では，複数のフレームが圧縮対象となる．

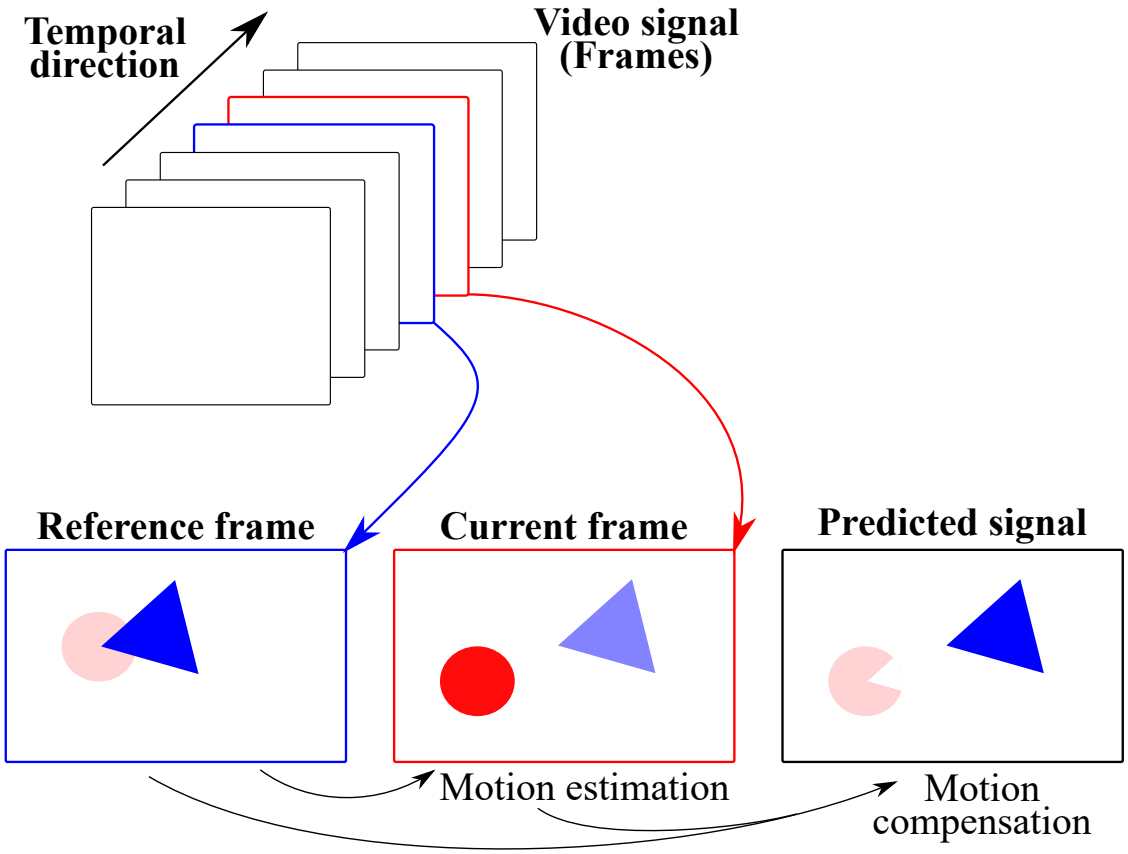

図 2.6: HEVC のフレーム間予測の概要図．図中上部は映像信号を構成するフレームの概要を示し，図中下左部のフレームは予測の参照フレーム，下中央部は予測対象となる現在のフレーム，下右部はフレーム間予測の結果導出された予測信号をそれぞれ示す．

像信号である場合に用いられる．図に示す通り，参照フレーム (Reference frame) と予測対象の現フレーム (Current frame) の間で平行移動のブロックマッチングによる動き予測 (Motion estimate) を行い，動き補償 (Motion compensation) により予測信号を生成する．映像信号は，一般に時間的に近接しているフレームは同じような被写体が継続して写っていることが多く，フレーム間予測によってそのような映像信号の時間的な冗長性を除去できる．HEVC や VVC では，これらの予測処理によって，空間的 · 時間的な冗長性を除去した結果である予測残差信号を圧縮する．したがって，予測精度を高めて冗長性を除去すればするほど，残差信号がスパースになり，圧縮効率が向上する．

次に，図 2.7 に HEVC の処理の全体概要を示す．まず，入力となる動画像信号は Coding Tree Unit (CTU) と呼ばれる 64×64 画素のブロックに分割される．基本的に圧縮処理は，この CTU ブロックごとに行われ，入力信号の左上からラスタスキャン順に処理される．CTU に分割された入力信号は，フレーム内予測 (Intra-frame prediction) お

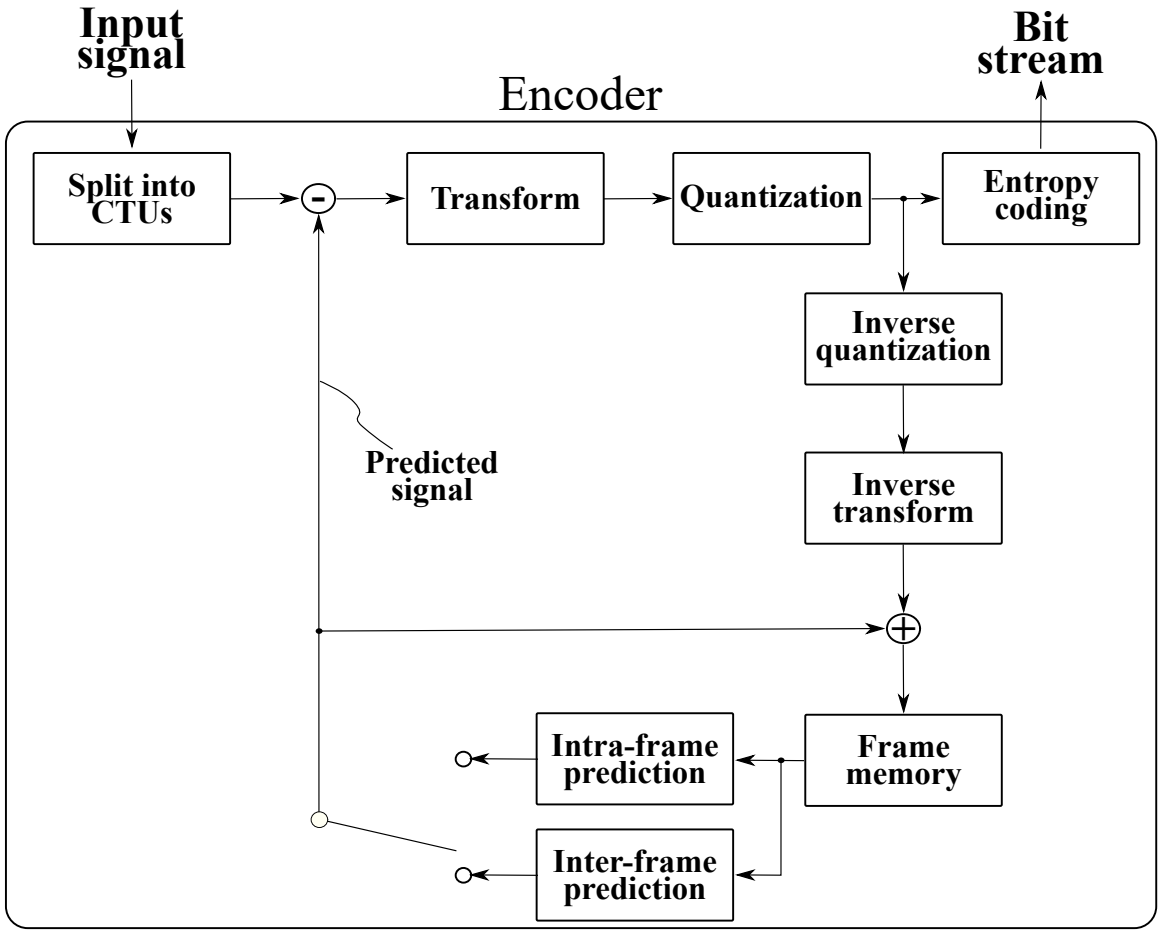

図 2.7: HEVC の圧縮を担う処理系 (エンコーダ) の処理に関するブロック図．入力となる動画像信号をエンコーダに入力し，予測 · 変換 · 量子化等の処理を経て最終的にバイナリ表現であるビットストリームの形式で圧縮された結果が出力される．なお，圧縮された信号の解凍処理を担う処理系であるデコーダに関しては，図示するエンコーダと逆の処理を行うものである．

よびフレーム間予測 (Inter-frame prediction) によって生成された予測信号を除去され，予測残差信号となる．予測残差信号は二次元直交変換処理 (Transform) によって周波数空間へ変換される．直交変換後の変換係数は量子化処理を経てエントロピー符号化によってビットストリームとして出力される．2020 年に最終標準規格が完成した H.266/VVC においても，新規ツールは導入されているが大まかな方式は HEVC の処理方式を踏襲している．

上記のような圧縮標準は，“波形符号化” と呼ばれる圧縮技術に分類される．波形符号化とは，「圧縮対象の動画像信号はどのような信号で何が映っているか」といった意味レベルの高度な内容は勘案せず，動画像を単なる波形信号として捉え，可能な限り少ないビットレートで表現する手法である．そのため，圧縮効率は Peak Signal-to-Noise

Ratio (PSNR) 等の客観画質で評価することが一般的である．PSNR 以外では，人間が動画像を鑑賞した際の主観画質での評価を行うこともあるが，本論文で対象とする DNN の認識精度保持等を評価指標とした動画像圧縮標準は現在のところ存在しない．

圧縮標準に代わる有望な動画像圧縮手法として，学習ベースの動画像圧縮手法が近年提案されている．学習ベースの動画像圧縮は DNN を用いてエンコーダとデコーダを定義し，図 2.7 のような処理フローに分けた形ではなく，最適化学習の結果として入力動画像のビットレート低減を図るものである．学習ベースの動画像圧縮手法の中で最先端の手法は HEVC に肉薄する高い圧縮性能を示している [91,92]．しかし，これらの手法では，単一モデルでは適応的な情報削減処理を行うことができず，ビットレートを状況に応じて制御することができない [91,92]．このため，ビットレートを適応的に制御するためには複数のモデルを独立して学習および用意する必要がある．本論文が主眼とする画像認識アプリケーションへの応用を考えると，計算資源が限られているフロントエンドデバイスに複数の DNN モデルを導入することは現実的ではないと考えられる．さらに，学習ベースの圧縮手法まだ標準化された技術ではなく，相互運用性を保証できないという欠点も存在する．以上のような理由から本論文では，学習ベースの手法ではなく既存の圧縮標準に対して画像認識を指向するという観点で評価を行っていく．ただし，本論文の第 3 章および第 4 章で提案する手法の適用先は圧縮標準に限ったものではなく，学習ベースの圧縮手法にも適用可能なものである．

2.6 クラウド型知能方式に向けた圧縮技術

2.5 節で述べた通り，DNN の認識精度の保持を評価指標とした動画像圧縮標準は存在しない．しかし，いくつかの先行研究は，既存の画像 · 映像圧縮標準を認識精度を保持するように改良することで，改良前と比較して精度保持に必要なビットレートの削減に成功している [93,94]．本節では，クラウド型知能方式へ適用できる圧縮技術としてこれらの手法について概要を説明する．

先行研究で提案されている手法は，既存圧縮標準の量子化処理の改良によって精度保持の観点を導入している．これらの手法は基本的に “画像中の認識対象となるオブジェクトが量子化処理によって劣化しなければ DNN の精度は保持できる” というコンセプトに基づくものである．したがって，画像中の認識対象となるオブジェクトの位置を推定し，その位置での量子化はなるべく細かく (劣化を少なく) して圧縮を行う．具体的には，Choi ら [93] は，YOLO9000 [11] の中間層の特徴マップの反応から認識に対して重要な画像領域を推定した．得られた推定結果に基づいて，HEVC の量子化制御モデルである R-λ モデル [95] の制御結果を修正することで認識精度を保持したままビットレートを削減することに成功した．また，Galteri ら [94] は，複数の顕著性マップの値と画像パッチを用い

て，Support Vector Machine [43] を学習し，画像中の重要領域を推定するモデルを構築した．Galteri らの結果によれば，重要でないと判断された画像領域を極めて粗く量子化しても DNN の画像認識精度にはほとんど影響を与えず，結果的に認識精度を保持したままビットレートを削減することに成功した．これらの手法は動画像圧縮の研究分野に DNN の精度保持という観点を取り入れた先進的な研究と言えるが，そのアプローチは量子化処理の改良のみに留まっている．

量子化制御に基づく手法は，圧縮処理を担うエンコーダの内部構造に依存するため，異なるエンコーダには直接適用できない可能性がある．例えば，JPEG2000 は HEVC とエンコーダの機構が大きく異なるため，HEVC への適用をを前提に開発された量子化処理の改良手法はそのまま適用することができない．また，JPEG のように適応量子化機構を有さない圧縮標準に対しては，上記のような量子化処理の改良に基づく手法は適用できない．

2.7 協調型知能方式に向けた圧縮技術

協調型知能方式においても，伝送効率を向上させるためには，DNN の中間層の出力値である深層特徴の圧縮が不可欠である．本節では，協調型知能方式の高度化のために行われている既存の深層特徴圧縮手法について概要を説明する．

既存の深層特徴圧縮手法は，以下でそれぞれ説明する通り，可逆圧縮 · ニアロスレス圧縮 · 非可逆圧縮の 3 種類に大別される．可逆圧縮では，ファイル圧縮等に一般的に用いられる GZIP [96] 等の可逆圧縮手法を用いて，深層特徴を圧縮する．ニアロスレス圧縮や非可逆圧縮においては，擬似的に深層特徴を画像や映像とみなして圧縮を行う．可逆圧縮は，原理的に深層特徴の有する情報を一切削減することなく圧縮することができる．このような可逆圧縮の特性は好ましい特性であるものの，Chen ら [97] は，可逆圧縮では，深層特徴の圧縮時にビットレートを大きく低減させることは困難であることを実験的に示した．そのため，現在は深層特徴を画像もしくは映像と見なして，ニアロスレス圧縮や非可逆圧縮によって圧縮する手法が主流となっている．

前述の通り，画像認識を行う DNN は畳み込み演算を利用しているため，深層特徴を構成する特徴マップを画像として見ると自然画像と同様に空間的な冗長性を持つ．GZIP 等のファイル圧縮では，このような冗長性を圧縮に利用できないが，ニアロスレス圧縮や非可逆圧縮はこのような冗長性を利用することで高い圧縮効率を実現する．非可逆圧縮を用いた深層特徴圧縮において，Choi ら [21] は深層特徴を特徴マップを全て空間的に画像として配置する手法を提案した．“空間的配置法” と呼ばれるこの手法では，それぞれの特徴マップが画像化され，1 枚の画像にまとめられる (第 4 章の図 4.2 (a) 参照)．Choi らは，空間的に配置された深層特徴を，グレースケール (YUV 4:0:0) フォーマットの画像

信号を圧縮できる HEVC Range extension (REext) [98] を用いて圧縮している．HEVC REext のような画像圧縮手法は，フレーム内予測を用いて深層特徴に隠された空間的な冗長性を除去することで圧縮効果を高めている．また，空間配置法によって配置された深層特徴を，HEVC REext の可逆モードや可逆画像圧縮 (例えば PNG など) を用いて圧縮することもできる．これは，深層特徴を画像信号として扱うためには 8-bit 程度のビット深度を持つ信号に量子化[*4]する必要があり，厳密には可逆圧縮ではないことから，“ニア” ロスレス圧縮 [99] と呼ぶ．ただし，ビット深度に対する量子化処理を行った後の深層特徴に関しては可逆な圧縮処理となっている．

深層特徴圧縮に関する既存研究の多くは，この空間的配置法を利用して他の課題を解決する技術の確立に取り組んでいる．Eshratifar ら [100] は BottleNet と呼ばれるアーキテクチャを提案した．BottleNet は Auto-Encoder [101] 状の DNN モデルであり，深層特徴を抽出する階層に BottleNet を取り付けて DNN モデルを再学習することで，伝送する深層特徴データサイズを減少させることができる．類似の目的のために Alvar ら [102] は DNN が HEVC 等の圧縮標準が圧縮し易くするような深層特徴を出力する新たな学習手法を提案した．Choi ら [103] も，深層特徴の特徴マップの数を減らす学習手法を提案している．Alvar ら [104, 105] は複数の深層特徴を伝送する際の最適なビットレート割り当てを数理的な枠組みによってモデル化した．Alvar らの手法は，DNN モデル内部で複数の出力を持つ，例えば ResNet [4] や DenseNet [7] のようなモデルで深層特徴を抽出する場合に適用できる．このように，空間的配置法は深層特徴圧縮の基盤技術の一つとして，この研究分野の進展に大きく寄与していると言える．しかしながら，深層特徴圧縮という研究分野はまだ初期段階にあり [106, 107]，空間的配置法以外にも圧縮効率を高める手法を新たに開発することは今後の研究分野のさらなる発展を促進するという観点で重要である．

空間的配置法以外の深層特徴を圧縮するアプローチとして，いくつかの先行研究は深層特徴の各特徴マップをビデオのフレームと解釈し，時間的な方向に沿って配置した “時間的配置法” が挙げられる (第 4 章の図 4.2 (b) 参照)．空間的配置法では画像圧縮手法が使われるため，フレーム内予測のみが冗長性除去に対して用いられていた．前述の通り，フレーム内予測は，方向性予測等を用いているため，比較的単純なエッジや塗りつぶしなどの冗長性を除去している．時間的配置法では，時間的に深層特徴を配置することで，フレーム間予測を冗長性除去に用いることができるようになるため，空間的配置法と比較してより複雑な深層特徴を時間的冗長性として除去することが期待される．しかし，Choi ら [21] は，空間的配置法と時間的配置法の圧縮効率を比較し，空間的配置法がより高い

[*4] エンコーダ内部の量子化処理は変換係数を量子化テーブルにしたがってと量子化するが，ここでの量子化は深層特徴信号そのものを量子化する異なる処理であることに注意されたい．

効率を示したことを報告している．Chen ら [108] は，空間的配置法と時間的配置法の圧縮効率を DNN の様々な階層から抽出した深層特徴で比較した．Chen らの報告によると，浅い層では両者には差がないものの，深い層から抽出した深層特徴では空間的配置法が時間的配置法よりも高い圧縮効率を得ることが明らかになった．このような時間的配置法の性質は，協調型知能方式では伝送効率の良い分割点が DNN の比較的深い階層である [18] ことを考えると，協調型知能方式には適していない可能性が高いと考えられる．また，Chen らは，時間的配置法に対して，時間的な冗長性を高めるための配置順序探索アルゴリズムを提案している [108]．しかし，配置順序探索アルゴリズムを用いても，時間的配置法では時間的配置法には圧縮効率が大きく劣っているのが現状である．

第 3 章

圧縮による画像認識の精度劣化を抑制する画像プレ変換とその解析

本章では，クラウド型知能方式 (図 1.1 (i)) での，画像認識を指向した情報源圧縮技術について考える．クラウド型知能方式は，フロントエンドデバイスで撮像した画像信号を圧縮した上で伝送し，クラウドサーバに配置されている DNN で認識を行う通信および認識の方式である．DNN を用いた画像認識では，認識対象となる入力画像から DNN が獲得した機構に基づいて特徴抽出を行う．したがって，入力画像は，DNN が十分に特徴抽出を行うことができるように画像品質が保たれている必要があり，品質が損なわれていない原画像を入力として利用することが望ましい [109]．しかし，クラウド型知能方式においては，多くの場合，圧縮効率向上のために非可逆の画像圧縮処理が施されるため，原画像を DNN の入力に利用することができない．2.5 節で述べた通り，画像の非可逆圧縮では，人間の知覚特性に基づいた情報削減処理が導入されており，客観画質や人間の主観画質に沿って圧縮性能の評価がなされている．つまり，DNN の精度保持という観点で見ると必ずしも適切ではなく，圧縮の結果生じる画像品質の劣化は DNN の特徴抽出を阻害し，認識精度の低下を招くと考えられる．本章で行う研究の目的は，クラウド型知能方式の高度化のために，DNN の認識精度を保持するという新たな観点を既存の画像圧縮に導入することである．まず，3.1 節では第 2 章で述べた関連研究に対する本研究の位置付けについて整理し，本研究で着目する深層学習の内部表現に関する研究について説明する．その後，3.2 節で提案手法について詳細を述べ，3.3 節と 3.4 節で提案手法の有効性を評価するための実験とその解析についてそれぞれ述べる．最後に，3.5 節で本章をまとめる．

3.1 問題設定

2.5 節で述べたように，既存の動画像圧縮標準は，人間の鑑賞を前提に PSNR 等の客観画質や人間の主観画質で評価を行っていた．このため，DNN の認識精度を評価指標とした圧縮標準は現在のところ存在しない．ただし，圧縮標準としては存在しないものの，幾つかの先行研究は既存圧縮標準の量子化処理を改良することで，改良前と比較して精度保持に必要なビットレートの低減に成功している [93, 94] (2.6 節参照)．これらの手法は，DNN が認識対象とするオブジェクトの位置を推定し，推定結果に基づいてオブジェクトが存在する領域を細かく，存在しない領域を粗く量子化する．この結果，ビットレートを低減させても認識精度を保持することが可能になる．上記のような研究は動画像圧縮の研究分野に DNN の精度保持という観点を取り入れた先進的な研究と言える．しかし，DNN が抽出する画像特徴に関する知見を導入したものではなく “認識対象となるオブジェクトが劣化しなければ DNN の精度は保持できる” というプリミティブな知見に基づいているため，その具体的なアプローチは量子化処理の改良のみに留まっている．量子化処理は 2.5 節で述べた通り，圧縮を担う処理系であるエンコーダに依存してその機構が変わってくるため，上記のようなアプローチには汎用性の面でいくつかの制約が存在する．例えば，JPEG2000 や HEVC は量子化処理の機構が大きく異なるため，HEVC のエンコーダに対して設計された量子化制御をそのまま適用することができず，互換性を持たない．また，JPEG のようにエンコーダ内部に適応量子化機構を有さない圧縮標準では，そもそも上記のような量子化処理の改良手法は適用できない．非可逆圧縮によって品質が劣化した画像データを用いた Fine-tuning に基づく手法 [110, 111] も，DNN の精度低下を防ぐ有望なアプローチとして考えられる．しかし，Dodge ら [112] の報告によるとこれらの手法は Fine-tuning 時のデータに存在しないタイプの歪みには汎化しないことが明らかになっている．したがって，量子化処理の改良手法 [93, 94] と同様に，Fine-tuning の際に用いるエンコーダへの依存性が高く，未知の圧縮手法とその歪みに対してはロバスト性が保証されないという制約が存在する．そこで本章では，上記のアプローチとは異なり，単一の手法でエンコーダによらず効果を示すアプローチの確立を目指す．

エンコーダによらずに非可逆圧縮時の精度低下を防ぐため，本章では，画像を圧縮する前に適切に変換する画像プレ変換手法を提案する．提案する画像プレ変換手法は，入力画像の信号の中で DNN の認識に対して重要な情報のみを保持し，それ以外の信号を圧縮時にビットレートが低くなるように変換する．これにより，同一ビットレートに圧縮した場合，提案手法で変換した画像の方が原画像よりも圧縮の影響を受けにくく，高い認識精度を示すことが期待される．提案手法は，2.3 節で述べた Geirhos ら [29] によって行われた研究に基づいている．Geirhos らの内部表現に関する研究では，人間が認識にあたって重視する画像特徴と DNN が重視する画像特徴は異なることを示している．したがっ

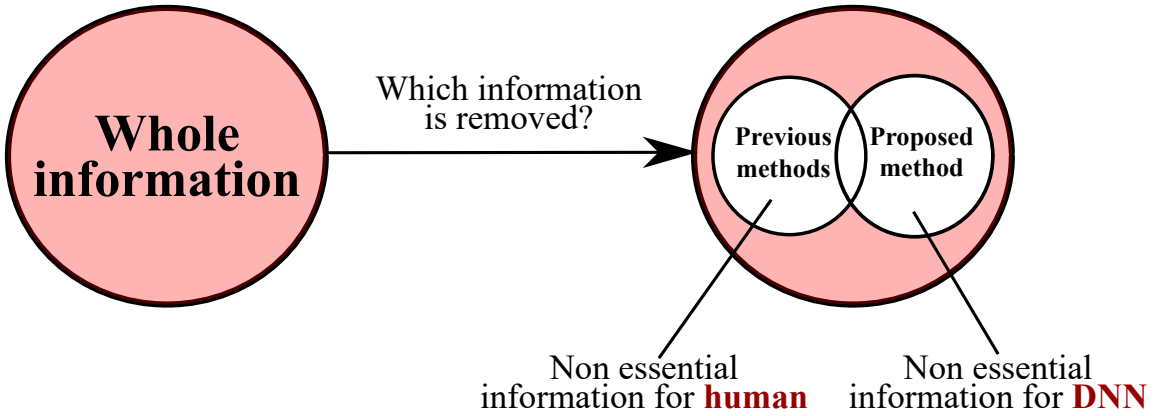

図 3.1: 本章の提案手法と従来手法の違いについての概念図．従来手法は人間の認識にとって重要でない情報を積極的に削減していたが，提案手法は DNN の認識に対して重要でない情報を削減する．

て，DNN が認識にあたって重視する情報以外の画像信号を積極的に削減することで，認識精度を保持しつつも大幅なビットレート低減が実現できると考えられる．提案手法と従来手法の違いを，図 3.1 に示す．本章で提案する手法は，このようなコンセプトを画像プレ変換の枠組みで実現したものである．近年，画像プレ変換処理を用いた手法が，画像圧縮におけるビットレート低減や DNN の認識精度向上に寄与することが明らかになってきた [113, 114]．Palacio ら [113] は Encoder-Decoder 型の DNN モデル (ED モデル) の一種である SegNet [115, 116] を画像プレ変換モデルとして採用し，認識を行う DNN モデルから伝播される損失で学習する手法を提案した．Palacio らの手法で学習された SegNet は DNN の画像認識に対して有利な変換を獲得し，多くの DNN モデルで原画像を上回る精度を示す変換を実現した．Shaham ら [114] は，オプティカルフローの枠組みで画像に微小な変形を与え圧縮しやすい画像へプレ変換することで，主観画質を保持しつつ大幅なビットレートの低減を実現した．また，Shaham らの報告は，画像プレ変換の枠組みを用いれば，エンコーダによらずビットレート低減効果を得られることも明らかにしている．画像プレ変換に関する上記の先行研究は，DNN が認識にあたって重視する情報を獲得できる，画像信号を圧縮した際のビットレートが低くなるように変換できる，といった事実を明らかにした．本研究では，この事実に着目し，画像プレ変換を用いて認識精度を保持しつつも圧縮時のビットレートを低減する手法を提案する．

提案手法は，Palacio ら [113] と同様に ED モデルに基づいた画像変換手法であり，認識を行う DNN モデルから伝播される損失とビットレートを低減させる損失の線形和で学習される．本研究では，さらに，Palacio ら [113] が利用していた ED モデルの構造の改良，および精度を高める損失の強度を決定するハイパーパラメータの適応的な調整手法

を提案する．ImageNet 2012 [117] データセットを用いた検証において，提案手法の変換画像は原画像と比較して，実験した全ての圧縮標準で，精度を保持しつつビットレートを低減する効果を得た．

本章で行った研究の主な貢献は以下の 3 点である:

1. 様々な圧縮標準で，認識精度を保持しつつビットレートを削減するための画像プレ変換手法を提案した．提案手法は，精度向上と圧縮時のビットレート低減を同時に考慮する画像プレ変換であり，それぞれを単独で考慮する手法では，得られないビットレート低減効果を確認した．
2. 先行研究 [113] で用いられていた ED モデル構造の改良と，精度向上の強度を適応的に調整する手法を導入し，その効果を確認した．
3. 提案手法が変換画像に与える影響について解析を行い，認識に重要な信号が圧縮後も保たれている等，提案手法が有効に作用する要因を発見した．

3.2 画像認識の精度劣化を抑制する画像プレ変換

本研究では，エンコーダによらず非可逆圧縮による DNN の精度低下を防ぐため，圧縮処理の画像プレ変換手法を提案する．Palacio ら [113] の手法と同様，本研究では，変換モデルとして DNN モデルの一種である ED モデルを採用した．ED モデルを，DNN の認識誤差を表現する損失 (以降 “認識損失” と呼ぶ) と圧縮時のビットレート低減効果を表現する損失の線形和で学習し，認識精度を保持しつつ，圧縮しやすい画像へ変換を実現する．

図 3.2 に提案手法の概要を示す．図中左部 (a) は ED モデルのアーキテクチャを示している．図中右部 (b) で認識を行う DNN (Recognition network) を通じて認識損失を算出し，(c) ではビットレート低減を促進させるため Total Variation [118] に基づく損失を算出する．(b)·(c) で算出した損失の勾配を ED モデルへ伝播することで学習を行う．本節では，提案手法の各構成要素の詳細を述べる．

3.2.1 Total Variation と認識損失に基づく学習

DNN モデルの学習に導入する損失は，微分可能な関数で構成されている必要があるため，ビットレート低減を表現する損失も微分可能な関数であることが求められる．しかし，非可逆圧縮の処理プロセスには微分不可能な処理が複数存在するため，ビットレートを低減させる損失を微分可能な形式で直接表現することは難しい．そこで，本研究では圧縮アルゴリズムの特性から画像信号の平滑化によるビットレート低減に着目し，平滑化を促進させる損失をビットレートを低減させる代理の損失として用いる．画像信号が平滑化

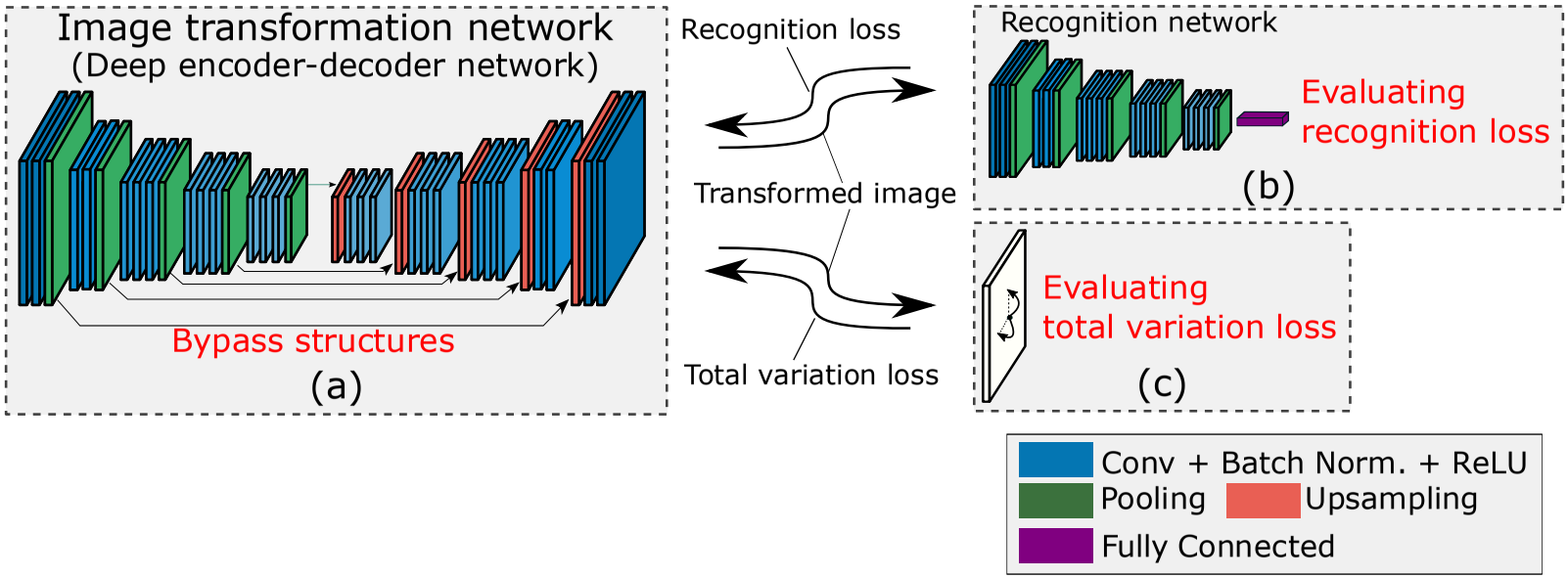

図 3.2: 提案手法の概要図．図中左部 (a) は画像プレ変換を行う ED モデルのアーキテクチャを示す．図中右部 (b) で認識を行う DNN を通じて認識損失を，(c) でビットレート低減を促進する Total Variation 損失をそれぞれ算出し，その勾配を (a) の ED モデルへ伝播して学習を行う．

されると，

1. 画像の空間方向の冗長性を除去する予測処理であるフレーム内予測 [89] の精度が向上する
2. 画像に含まれる信号のうち，高周波成分が消失することで情報量が削減される

といった理由から，圧縮した際にビットレートが低減する効果が期待される．提案手法では，画像信号を平滑化させる損失として，Total Variation (TV) [118] を採用した．TV は隣接画素間の差分に基づいて算出される微分可能なパラメータであり，画像信号の平滑化度合いを評価できる．例えば，TV を利用した画像信号分解手法である TV-L1 画像分解 [119] において，TV が小さくなるような制約条件の下で分解された Cartoon 成分は画像信号が平滑化されたような成分となることが知られている．

ED モデルの学習には，TV に基づいた損失と認識損失の組み合わせを用いる．TV に基づく損失関数 L_{TV} は以下のように表される:

$$L_{\mathrm{TV}}(T) = \sum_{i,j}(|T_{i+1,j} - T_{i,j}|^2 + |T_{i,j+1} - T_{i,j}|^2). \tag{3.1}$$

ここで，T は ED モデルが変換した画像を示し，i, j は画素の位置を表すインデックスである．なお，上記の定式化は，隣接画素間の二乗を取っている点で Rudin ら [118] が提唱した TV の定式化とは若干異なるものである．しかし，上記の定式化を利用することによって微分計算が容易になるため，画像の画風変換 (Neural Style Transfer) [120] のような深層学習による画像変換タスクでは式 (3.1) で定義される TV が一般的に用いられている [79]．また，認識損失 $L_{\mathrm{Recog.}}$ は確率分布間の距離を最小化させる効果がある交差エントロピー損失関数 [121] 等が一般に用いられる．交差エントロピー損失関数は以下のように定義される:

$$L_{\mathrm{Recog.}}(\mathbf{x}, \mathbf{y}) = \sum_{q} y_q \log(x_q). \tag{3.2}$$

ここで，$\mathbf{x}$ は入力画像に対する DNN の推定結果ベクトルを示す．$\mathbf{y}$ は正解ラベルのベクトルであり，通常，正解ラベルに相当する要素が 1，それ以外の要素は 0 となる．また，q はベクトルのラベルインデックスを示す．ED モデルは，上記の $L_{\mathrm{TV}}, L_{\mathrm{Recog.}}$ の線形和:

$$L_{\mathrm{Final}}(\mathbf{x}, \mathbf{y}, T) = \lambda_{\mathrm{Recog.}} \times L_{\mathrm{Recog.}} + \lambda_{\mathrm{TV}} \times L_{\mathrm{TV}}, \tag{3.3}$$

で学習される．$\lambda_{\mathrm{Recog.}}$，$\lambda_{\mathrm{TV}}$ は損失の強度を調整するハイパーパラメータである．

3.2.2　原画像情報を用いたハイパーパラメータ調整

ED モデルを学習する式 (3.3) の損失関数は，$\lambda_{\text{Recog.}}$ と λ_{TV} の 2 つのハイパーパラメータを有する．これらのハイパーパラメータは変換画像の精度向上とビットレート低減という，提案手法が有する 2 つ効果のトレードオフを調整する重要なパラメータである．本研究では，原画像を認識した際の情報を活用し，$\lambda_{\text{Recog.}}$ を学習過程で適応的に変動させる手法を提案する．

$\lambda_{\text{Recog.}}$ は，認識損失 $L_{\text{Recog.}}$ の強さを調整するパラメータであり，大きすぎるとビットレートを十分に低減できず，小さすぎると認識精度が原画像と比較して大きく低下してしまうため，適切な値を設定する必要がある．提案手法は，原画像と同等の認識精度を保持しつつビットレートの低減を実現する手法であるため，原画像の認識精度を基準として，その精度を下回らない中で，よりビットレートの低減を促進するような学習をすることが望ましいと考えられる．

本節で提案するハイパーパラメータ調整法では，$\lambda_{\text{Recog.}}$ が上記のような挙動を示すように，原画像をそのまま DNN で認識した場合と，変換画像を DNN で認識した場合の認識結果を活用することで，$\lambda_{\text{Recog.}}$ を学習過程で適応的に変動させる．具体的には，原画像よりも変換画像の認識精度が低い場合は，精度向上をもたらすために，$\lambda_{\text{Recog.}}$ は相対的に大きい値を取り，原画像よりも変換画像の認識精度が高い場合，よりビットレート低減効果を大きくするため，$\lambda_{\text{Recog.}}$ は小さい値を取るように変動させる．提案手法では，認識結果を示すパラメータとして，式 (3.2) で定義したような交差エントロピー損失を用いて，以下のように $\lambda_{\text{Recog.}}$ を定義した:

$$\lambda_{\text{Recog.}} = \frac{L_{\text{Recog.}}}{L_{\text{Recog.\_org}} + \epsilon} \ . \tag{3.4}$$

ここで，$L_{\text{Recog.\_org}}$ は原画像を DNN で認識した際の交差エントロピー損失である．また，ϵ はゼロ除算を防ぐパラメータであり $\epsilon = 0.0001$ とした．交差エントロピーは推定結果の負の対数尤度である [121] ため，式 (3.4) は変換画像よりも原画像の認識精度が高い場合，1 より大きい数値を示し，原画像よりも変換画像の認識精度が高い場合，1 より小さい数値を示す．

3.2.3　ED モデルのバイパス構造

Palacio らの手法 [113] では，バイパス構造を有さない ED モデルである SegNet を用いていたが，本章での提案手法では，Encoder の特徴マップを Decoder の特徴マップにバイパスするモデルを画像プレ変換モデルとして利用する．バイパス構造の導入には，2 つの利点が存在する．1 つ目は，階層数の多い DNN モデルを学習する際に生じる勾配消

失 [4] による精度低下を緩和できる点である．2 つ目は，局所的な画像構造を活用することで，モデルの表現能力が向上する点である．一般に，DNN の高次階層では抽象的な画像特徴を抽出しており [30]，また，高次階層においては空間的な情報の多くは失われている [122]．したがって，バイパス構造を用いない ED モデルの場合，限られた画像特徴しか利用できず，表現能力が不十分となってしまう恐れがある．上記のような要因から，バイパス構造を ED モデルに導入することで，変換性能が向上することが期待される．

バイパス構造を持たない ED モデルの処理は Encoder/Decoder でそれぞれ以下のように定式化できる:

$$\begin{aligned} E_m(X) &= \mathcal{F}_{E_m}(E_{m-1}(X)), \\ D_n(X) &= \mathcal{F}_{D_n}(D_{n-1}(X)). \end{aligned} \tag{3.5}$$

ここで $\mathcal{F}_{E_m}$ は Encoder の m 層目の処理を示す関数であり，$\mathcal{F}_{D_n}$ は Decoder の n 層目の処理を示す関数である．これらの関数は前の層の出力を入力として処理する．特に，$E_0(X)$ は入力画像 X を示し，$D_0(X)$ は Encoder ネットワークの出力を示す．バイパス構造は Decoder の関数を以下の $D_n^{\text{bypass}}(X)$ に修正することで実現できる:

$$D_n^{\text{bypass}}(X) = \mathcal{F}_{D_n}(D_{n-1}(X) + E_{n'}(X)). \tag{3.6}$$

ここで n' はバイパス元となる Encoder の階層を示す．式 (3.6) に示すバイパス構造は U-Net [123] と類似しているが，バイパス処理が加算演算であり，concatenate 演算では無い．本研究では，3.3.2 節で後述する ED モデルの学習において，事前学習と提案手法による学習の 2 段階の学習を採用している．そのため，事前学習と提案手法による学習の際に Decoder の次元が変更されないように加算演算を用いている．

3.3 評価実験

本節では，提案手法の有効性を検証するために，自然画像データの識別タスクにおいて提案手法を既存の圧縮方式のプレ変換として用い，効果を検証する．

3.3.1 データセットの詳細

本研究では，検証のために，大規模自然画像データセットである ImageNet 2012 [117] を用いる．ImageNet 2012 は，一般に画像識別タスクで用いられるデータセットであり，計 1000 カテゴリ，約 128 万枚の訓練画像と 5 万枚の検証画像を有する．本研究では，検証画像を {1 万, 5 千, 3 万 5 千} 枚のサブセットにランダムに分割し，それぞれをテスト用，パラメータチューニング用，事前学習検証用として利用する．また，各実験では，画像サイズを事前に 256×256 にリサイズした画像を用いる．なお，ImageNet の画像は，

表 3.1: 提案手法の学習に用いたパラメータ．

	parameters
Learning rate (LR)	$\eta = 0.0001$
Batch size	16
LR decay param.	$\eta = 0.1\eta$
LR decay timing	each $100,000$ iterations
Max iterations	$400,000$ iterations

すべて JPEG によって非可逆圧縮済みの画像であるが，本研究では，JPEG によって圧縮された画像を原画像とみなして検証を進める．

3.3.2　学習の詳細

本実験では比較を容易にするために，ED モデルの中でも，Palacio ら [113] の手法と同様に SegNet を用いた．提案手法を用いた学習を行う前に，SegNet に対して自然画像の再構成損失 (Mean Squared Error: MSE) を用いて教師なし事前学習を行った．事前学習では ImageNet 2012 の学習用サブセットを利用し，事前学習検証用サブセットで学習率を調整した．なお，この際，前述のバイパス構造は導入しない．同様の事前学習は，Palacio ら [113] も行っており，学習パラメータは先行研究の記載に従った．学習終了後の最終的な MSE は 0.000866 であり，Palacio らの報告とほぼ一致している [113].

事前学習後の SegNet に対して，バイパス構造を導入し，3.2.1 節で定義した損失関数 L_{final} および 3.2.2 節の手法によって調整されるハイパーパラメータで学習を行う．バイパス構造は計 4 箇所に導入し，式 (3.6) の表記に従うと，導入した位置は $(n', n) \in \{(2, 12), (4, 10), (7, 7), (10, 4)\}$ のように示される．画像認識モデルは，ImageNet 2012 の識別に対して有効性が確認されている VGG-16 [2] を用いた．提案手法の学習の際に用いたパラメータを表 3.1 に示す．また，式 (3.3) の λ_{TV} は $\{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$ のうち，パラメータチューニング用サブセットを用いた検証で最も高い性能を示した $\lambda_{\text{TV}} = 0.75$ を用いた．本章における，全ての実験において，DNN の利用にあたっては Caffe [124] と呼ばれるフレームワークを使用した．また，再現性を高めるために，学習済みモデルには Caffe がオンラインに公開しているモデルを使用した．

表 3.2: 検証で用いた圧縮用パラメータ.

	JPEG	JPEG2000	HEVC	VVC
Encoder software	libjpeg	OpenJPEG	HM-16.0	VTM-5.0
Bit depth	8 bit	8 bit	8 bit	8 bit
YUV format	4:2:0	4:2:0	4:2:0	4:2:0
Compression rate	QL ∈ {85,70,40, 25,15,10,5}	CL ∈ {5,7,10,20, 30,40,60}	QP ∈ {15,20, 22,27,32,37,40}	QP ∈ {15,20, 22,27,32,37,40}

3.3.3 評価実験の詳細

提案手法の有効性を検証するため，3.3.2 節の学習を行った ED モデルを，圧縮処理のプレ変換として適用し，認識精度とビットレートの関係を調査する．圧縮方式には，JPEG [22]，JPEG2000 [23]，H.265/HEVC [24] および最新映像圧縮標準の H.266/VVC [83] を用いた．なお，圧縮対象は静止画であるため，映像圧縮標準である HEVC，VVC を用いた実験では，静止画用の圧縮構造である Intra-only 構造を適用する．その他の圧縮に用いたパラメータを表 3.2 に示す.

ビットレートについて定量的に議論するため，Bjøntegaard Delta Bitrate (BD-Rate) [125, 126] を用いて圧縮性能を比較する．BD-Rate は，2 種類の圧縮方式の性能を比較する際に用いられる指標であり，画像品質が同一となる際に，どの程度ビットレートを低減できるかを評価する．一般的に，BD-Rate は PSNR 等を画像品質の指標としていることが多いが，どのような評価指標を用いても BD-Rate は算出可能である．本研究では，認識精度を保持しつつビットレートを低減することを目的としているため，認識精度を評価指標として設定した．また，BD-Rate は，ある画像品質を得るために必要なビットレートのペア 4 組から算出する．この 4 組の選び方は任意であるが，本研究では，標準化策定作業で一般に用いられている，量子化パラメータ (Quantization Parameter: QP) ∈ {22, 27, 32, 37} の 4 組を用いた．JPEG および JPEG2000 では，表 3.2 で規定した各圧縮率のうち HEVC の QP ∈ {22, 27, 32, 37} と概ね同程度のビットレートを示す 4 組を利用した.

3.3.4 ImageNet 2012 における効果検証

図 3.3 に，JPEG，JPEG2000，HEVC，VVC のそれぞれの圧縮標準において，提案手法を画像プレ変換モデルとして用いた際のビットレートと認識精度の関係を示す．なお，精度の指標として，上位 1 位識別結果 [1] を用いた．赤線は提案手法による変換画像，黒線は原画像を圧縮した場合の結果を示している．また，緑線と青線は比較手法の結果であり，それぞれ TV-L1 画像分解 [119]，Palacio らの手法 [113] の結果を示している.

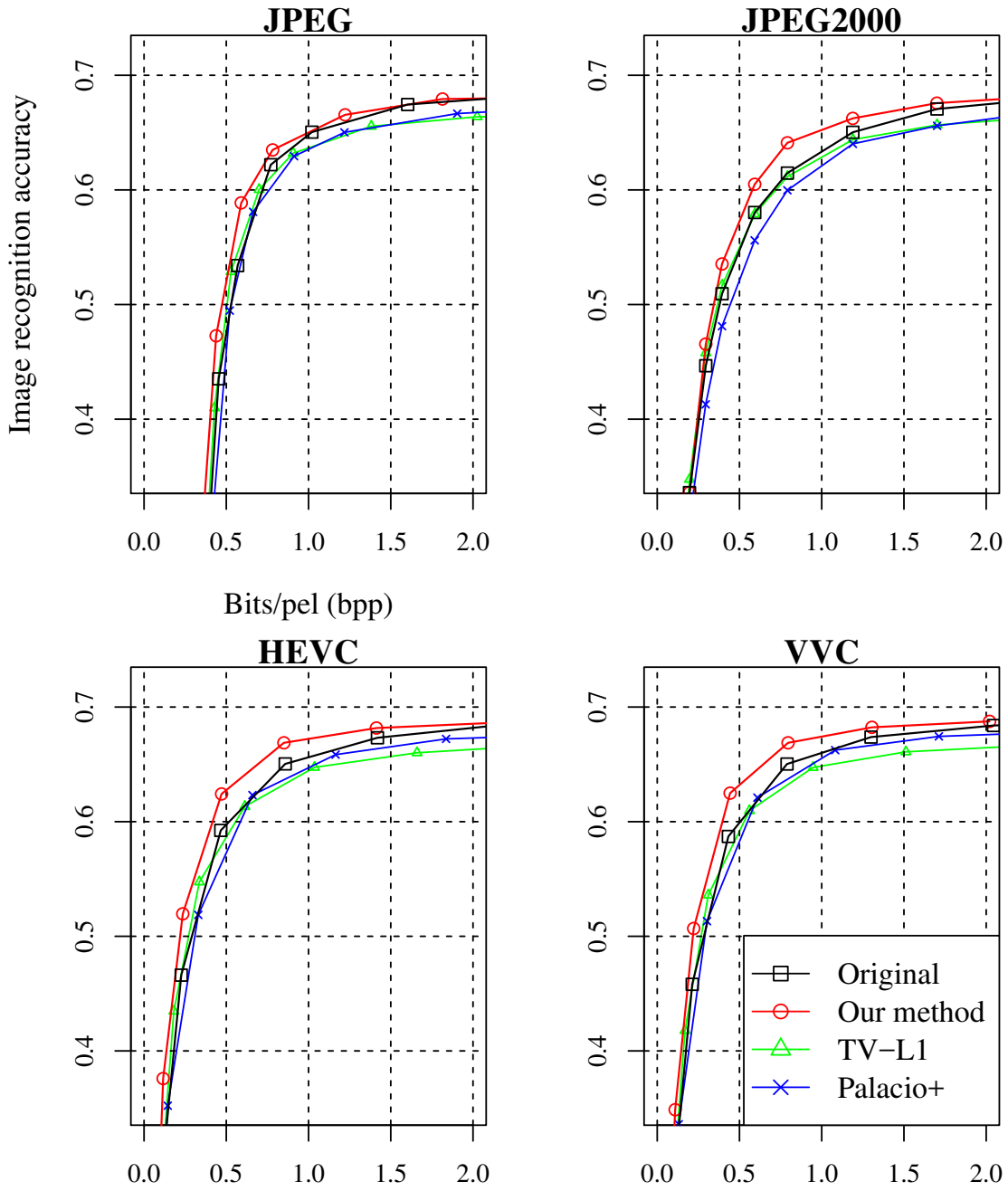

図 3.3: 自然画像識別タスク (ImageNet 2012) での，変換画像と原画像のビットレートと認識精度の関係．それぞれ，JPEG (左上)，JPEG2000 (右上)，HEVC (左下)，VVC (右下) での結果を示す．各線はそれぞれ黒線: 原画像，赤線: 提案手法による変換画像，緑線: TV-L1 画像分解による変換画像，青線: Palacio らの手法による変換画像の結果である．

表 3.3: 自然画像識別タスク (ImageNet 2012) での，各画像プレ変換手法 · 圧縮標準における BD-Rate の比較．各列 (圧縮標準) で最も低い BD-Rate を示している数値を太字で示している．

	JPEG	JPEG2000	H.265/HEVC	H.266/VVC
Palacio+ [113]	+8.9%	+16.2%	+13.1%	+14.1%
TV-L1 [119]	+2.3%	+2.0%	+14.4%	+10.9%
Our method	**−8.6%**	**−16.5%**	**−20.8%**	**−19.5%**

TV-L1 画像分解[*1]は，前述の通り TV を正則化項とした画像分解手法であり，変換画像の平滑化のみを考慮している手法と言える．一方，Palacio らの手法は認識損失のみで SegNet を学習したものであり，変換画像の精度向上のみを考慮している手法と言える．

図 3.3 の結果が示す通り，提案手法の結果 (赤線) は全ての圧縮標準において，原画像 · 比較手法による変換画像よりも低ビットレート帯で高い認識精度を示している．これは，提案手法が様々な圧縮標準およびそのエンコーダにおいて圧縮時の精度低下を防ぐために有効である，という我々の仮説の妥当性を示唆していると考えられる．一方，TV-L1 画像分解と Palacio らの手法の結果は，図 3.3 からは原画像との大きな差分は見て取れない．以下で，BD-Rate を用いて，定量的にビットレート低減効果を検証する．なお，3.3.3 節で述べた通り，BD-Rate は画像品質が同一となる際に，どの程度ビットレートを削減できるかを評価する指標であり，図 3.3 の認識精度 (縦軸) を同一にした際のビットレート (横軸) の変化量を評価するものに他ならない．

表 3.3 に，提案手法と比較手法の BD-Rate を示す．なお，比較対象となるアンカーには原画像の結果を用いている．表 3.3 に示す通り，Palacio らの手法および TV-L1 画像分解で変換した画像は，いずれの圧縮標準においても BD-Rate は正の値を示しており，精度保持に必要なビットレートが増大していることが分かる．一方で，提案手法は，すべての圧縮標準に対して，負の BD-Rate が得られており，精度保持に必要なビットレートを低減できていることが分かる．提案手法の BD-Rate は最大で −20.8% (HEVC)，最も小さいものでも −8.6% (JPEG) であり，圧縮標準ごとに最適化を行っていないにも関わらず，各標準で大幅なビットレート低減に成功している．上記の結果は，ビットレート低減と認識精度向上を同時に考慮して変換を行う提案手法の有効性を示していると考えられる．

表 3.3 より，提案手法は HEVC で −20.8%，VVC で −19.5% の BD-Rate 利得を得たが，JPEG，JPEG2000 では HEVC，VVC で示した程のビットレート低減効果は得ら

[*1] 本研究では，Cartoon 成分を変換画像として用いる．

れなかった．これは，JPEG，JPEG2000 と HEVC，VVC の圧縮方式の違いに起因していると考えられる．提案手法に導入した TV 損失は，

(i) フレーム内予測の精度向上

(ii) 高周波成分の減少

の 2 つの効能からビットレートを低減させる効果がある．しかし，JPEG および JPEG2000 にはフレーム内予測機構が存在せず，上記のうち 2 つ目のビットレート低減効果のみが作用したため，HEVC，VVC に比肩する BD-Rate が得られなかったと考えられる．圧縮標準およびそのエンコーダごとに差分を持たず，最大のビットレート低減効果を実現する損失の設計は今後の課題である．また，表 3.3 より，JPEG と JPEG2000 においても約 8 ポイント程度の圧縮効率の差分が存在していることが分かる．これは，本章の実験で用いている原画像が JPEG で既に圧縮された画像であることが要因の一つである可能性がある．提案手法は，入力画像の高周波成分を削減するために平滑化を進めるが，本章の実験において，この入力画像は JPEG で既に圧縮された画像となっており，高周波成分の一部は既に削減されている．その結果，JPEG では，類似する高周波成分を二度削減するような状況となり，結果として提案手法によるビットレート低減効果を十分に発揮できなかったと考えられる．その他の圧縮標準は，JPEG とは圧縮の機構に差異があるため，JPEG が削減していない高周波成分を削減することができ[*2]，提案手法のビットレート低減効果が比較的得やすかったと考えられる．したがって，JPEG やその他の圧縮標準においても，検証を非圧縮の原信号で行った場合，さらに高い圧縮効率を得られる可能性がある．しかし，上記の仮説の検証には，非圧縮の原信号を用いて提案手法の学習を行う必要がある．大規模画像認識データセットにおいて非圧縮の画像データを利用しているものは現在のところ存在せず，実際の検証は難しい．非圧縮の画像データセットの構築も含めて，今後，検証を進めていきたい．

3.4　提案手法の解析

上記の評価実験の結果，提案する画像プレ変換手法は最大で −20.8% という大幅なビットレート低減効果を得た．しかし，提案手法が当初の狙い通り認識に重要な信号を優先的に保持するような変換を獲得できているか，あるいは提案した 3 つの手法それぞれがビットレート低減のために上手く作用しているか等の検証はできていない．そこで，本節では，提案手法の有効性を，

i) 原画像と変換画像の比較

[*2] 実際に，どの周波数帯の信号をどのように削減するかは，圧縮の処理機構および量子化の粒度を決定する量子化テーブルによって決まる．

ii) 提案 3 手法の効果検証
iii) TV 損失のビットレート低減効果
iv) 主観画質に基づく歪み低減効果の調査

の 4 軸から検証する．

3.4.1 原画像と変換画像の比較検証

まず，原画像と提案手法による変換画像を比較するため，それぞれの圧縮前および HEVC による圧縮後の画像を比較する．2 つの画像を例として，図 3.4 と図 3.5 に，それぞれの画像とその識別結果を示す．なお，識別結果は DNN の出力を指し，認識対象の画像に各カテゴリが含まれている確率を示している．図 3.4 と図 3.5 には，上位 3 位の確率を示すカテゴリと対応する確率値を記す．

まず，図 3.4 について説明する．原画像と変換画像における圧縮画像は同程度のビットレートを示しているが, 変換画像の識別結果が原画像を上回っており (**47.2%** v.s. 21.9%), 提案手法が有効に作用していることが分かる．原画像および変換画像の識別結果の差異要因を調査するため，圧縮画像の特性を赤枠と青枠の 2 つの領域で比較する．赤枠で示した領域は，認識対象である犬がほとんど写っておらず，認識にはあまり重要でない画像領域であると考えられ，一方，青枠で示した領域は認識対象である犬が写っている画像領域であるため，認識に重要な画像領域であると考えられる．原画像の圧縮画像では，赤枠領域は歪みが少なく画像構造を保持しているが，青枠の領域では犬の毛並みのテクスチャが滲んでしまい歪みが大きくなっていることが分かる．変換画像の圧縮画像では，赤枠の木のテクスチャ等が滲んでいる一方で，青枠領域の毛並みは比較的歪みが少なく，テクスチャが保持されていることが見て取れる．これは，認識精度向上とビットレートの低減を同時に考慮した提案手法の効果であり，上記のような現象の結果として低ビットレート帯においても認識に重要な信号を保持することに成功したと考えられる．

図 3.4 では，提案手法が有効に作用する例を示したが，図 3.5 のように提案手法が認識精度およびビットレートに悪影響を及ぼす例も存在する．図 3.5 においても，圧縮後の画像は同程度のビットレートを示すが，図 3.4 とは逆に，変換画像の識別結果が原画像を下回っている (**32.0%** v.s. 68.5%)．赤枠と青枠の領域の比較においても，変換画像の圧縮画像は原画像よりも歪みが目立つ結果となっている．例えば，赤枠で示した領域は，赤破線で囲っている部分にブロック歪みが見られ，また，青枠で示した領域は原画像を圧縮した場合よりもモスキートノイズが生じている．図 3.5 の原画像は，認識対象の物体以外は輝度変化が少ない背景であり，画像プレ変換を行わなくてもフレーム内予測が当たりやすく，かつ高周波成分が少ないという点で圧縮しやすい信号であると考えられる．このよう

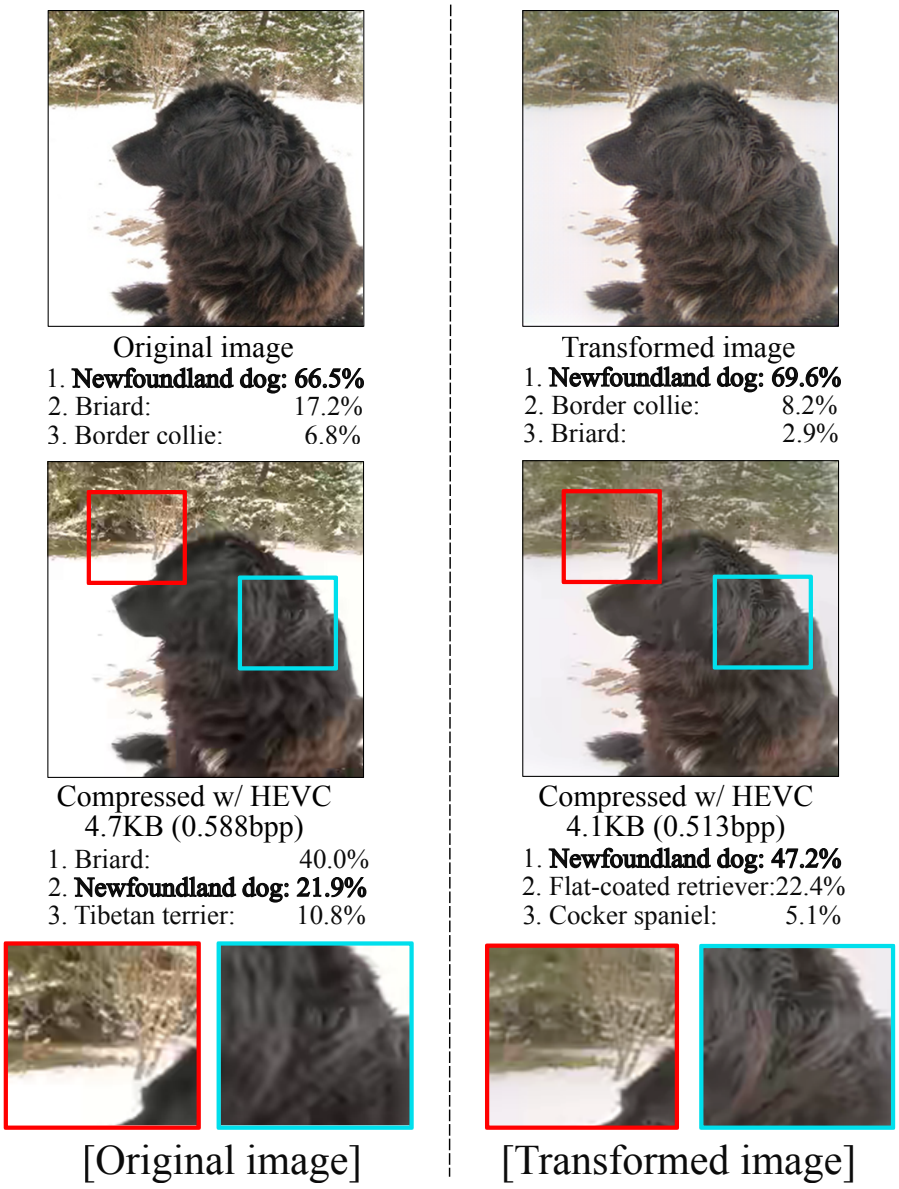

図 3.4: ImageNet 2012 データセットにおける原画像と変換画像の比較．ここでは，提案手法が有効に作用している例を示す．図中 [Original image] は，原画像およびその圧縮結果，[Transformed image] は変換画像およびその圧縮結果を示している．各画像の下部は，VGG-16 の識別結果であり，太字は正解ラベルを示す．図中下部の赤枠および青枠は圧縮後の画像の対応する領域の拡大画像である．

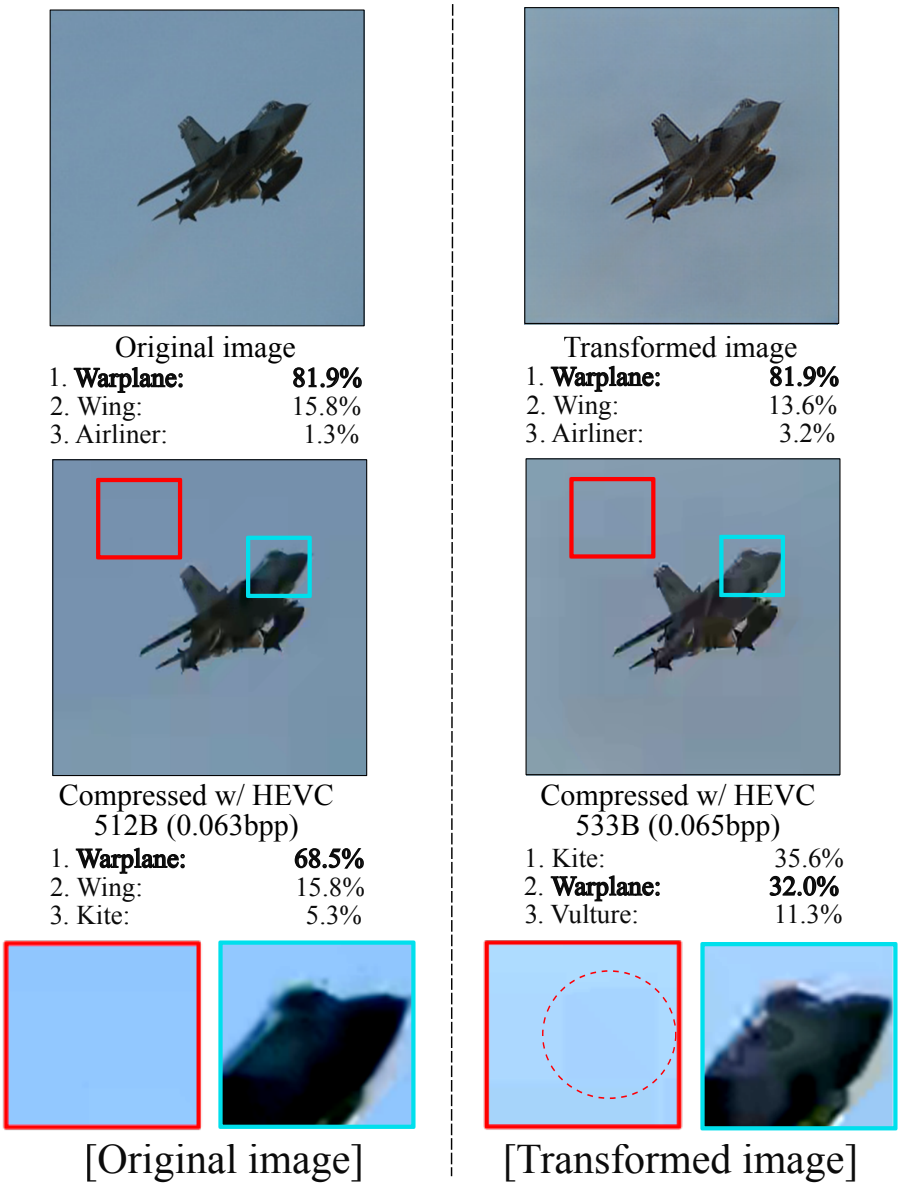

図 3.5: ImageNet 2012 データセットにおける原画像と変換画像の比較．ここでは，提案手法が悪影響を及ぼしている例を示す．図中 [Original image] は，原画像およびその圧縮結果，[Transformed image] は変換画像およびその圧縮結果を示している．各画像の下部は，VGG-16 の識別結果であり，太字は正解ラベルを示す．図中下部の赤枠および青枠は圧縮後の画像の対応する領域の拡大画像である．なお，わかり易さのため，拡大画像の明度とコントラストを変更している．

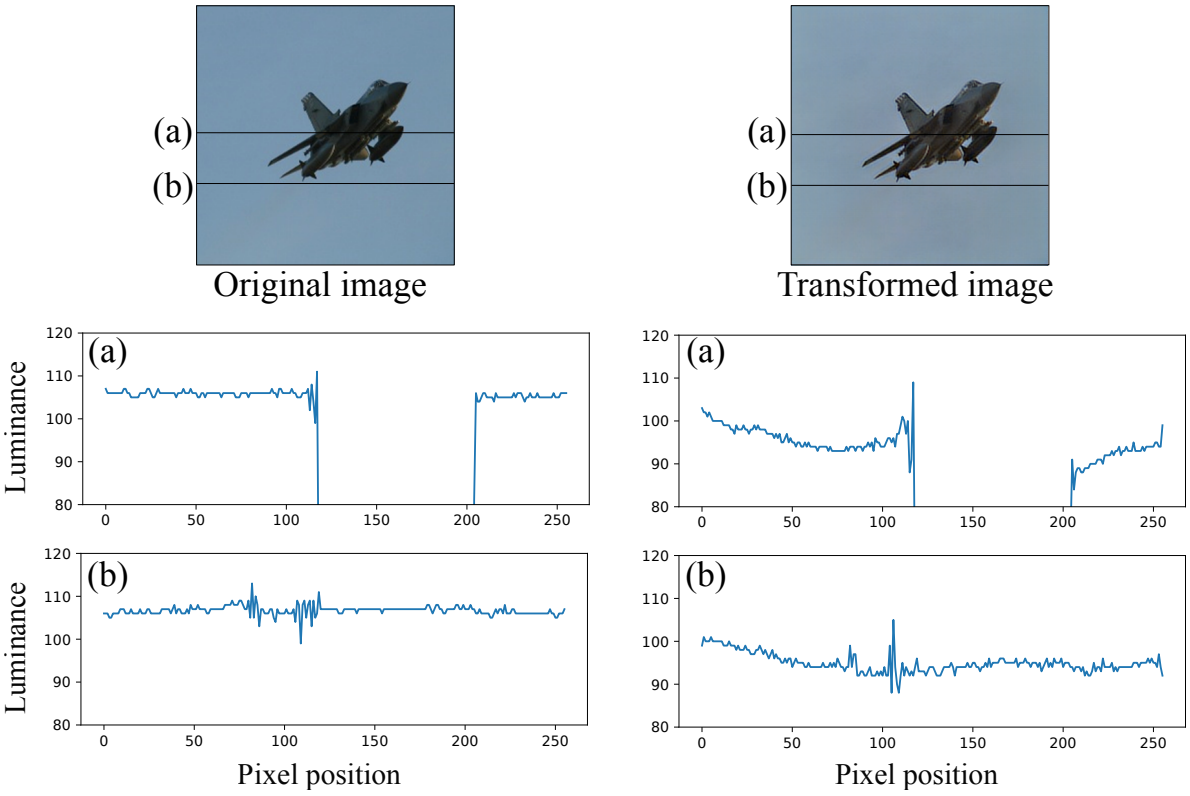

図 3.6: 図 3.5 の画像において，水平方向の輝度値を可視化した結果．グラフの横軸は画素の位置，縦軸は輝度値を示す．

な画像の場合，TV 損失を最小化しようとする提案手法が，本来変換する必要が無い輝度変化が少ない信号まで変換してしまい，図 3.5 のような悪影響を及ぼす結果となったと考えられる．図 3.6 に原画像および，変換画像の水平方向の輝度信号を可視化した結果を示す．図中左部の原画像の信号では，青色の領域の輝度信号にほとんど変化は見られない一方で，変換画像では輝度信号に大きな変化が見られる．特に，物体 (Warplane) 付近で輝度値の変化が大きくなっており，TV による平滑化によって物体領域内部の画素が滲んでしまったものと考えられる．その結果として，背景領域が原画像よりも圧縮しにくい信号となってしまい，同程度のビットレートを持つ圧縮画像において識別性能が低下してしまったものと考えられる．

以上の比較検証から，提案手法は平滑化作用によって圧縮時に認識に重要でない画像信号から優先的に情報を削減するような画像へ変換することが分かった．しかし，原画像がすでに十分に平滑化されたような信号である場合，変換する必要が無い信号まで変換してしまう可能性があることも併せて示唆された．ImageNet 2012 は自然画像識別データセットであり，その多くはデジタルカメラ等で日常の風景を撮影したものである．したがって，データセット全体では豊富にテクスチャを含んだ画像が多く，図 3.5 のような画

像は少数であるため，平均を取ると，提案手法は表 3.3 に示すような顕著なビットレート低減効果を得られたと考えられる．また，図 3.5 の結果に対する考察から，画像中のエッジ付近で TV 損失の影響を小さくするなどの改良方針も考えられる．ただし，多くの自然画像データでは図 3.4 の画像のように認識に対して重要でないエッジやテクスチャを多く含むため，認識に重要なエッジ付近でのみ適用的に TV 損失の影響を低減する必要がある．そのような新たな損失関数の開発は今後の課題である．

3.4.2 提案手法の効果検証

本研究における画像プレ変換モデルは，以下の 3 つの提案手法で構成されている．

1. TV に基づいた損失関数の導入 (3.2.1 節),
2. 原画像の認識結果を用いたハイパーパラメータ調整法 (3.2.2 節),
3. ED モデルへのバイパス構造の導入 (3.2.3 節).

ここでは，上記の 3 手法が認識精度を保持するための画像圧縮に，どのような影響をもたらすのか検証する．

表 3.4 に，HEVC での圧縮における提案 3 手法の組み合わせと BD-Rate の関係を示す．ハイパーパラメータ調整法を用いない場合は，$\lambda_{\text{Recog.}} = 1.0$ とし，λ_{TV} は $\{0.25,\ 0.5,\ 0.75,\ 1.0,\ 1.25,\ 1.5\}$ のうち，パラメータチューニング用サブセットで最も高い性能を示した，$\lambda_{\text{TV}} = 1.0$ を用いた．提案 3 手法の全ての手法を導入しない場合は，Palacio らの手法 [113] と同値である (表 3.4: 2 列)．表 3.4 に示す通り．Palacio らの手法と比較して，バイパスのみ/TV のみを導入した場合でも BD-Rate の改善が見られる．特に TV に基づく損失関数の導入は大きな効果を有しており，約 19 ポイント ($+8.1\% \rightarrow -11.2\%$) の改善を示している．そして，これら両者を組み合わせ，ハイパーパラメータ調整法を導入した提案手法が最も大きいビットレート低減効果 (-20.8%) を得ている．これは，提案 3 手法が，それぞれ認識精度を保持するための画像圧縮の実現に対して有効であることを示唆している．

また，表 3.4 から，ハイパーパラメータ調整法については，他の 2 手法と比較して，ビットレート低減効果の改善が小さいことが分かる (-0.5〜-0.9 ポイント)．これは，調整法を用いない場合に適用されるハイパーパラメータ $\lambda_{\text{Recog.}} = 1.0$ が，提案する調整法が導く値と近しくなっていることに起因していると考えられる．式 (3.4) で定義したハイパーパラメータ $\lambda_{\text{Recog.}}$ は，変換画像と原画像の精度が同一の場合，1.0 に近似するように設計されている[*3]．変換画像の認識精度は，画像単位ではバラつきがあるが，データセット

[*3] 式 (3.4) において分母に ϵ を導入しているため，1.0 に漸近する．

表 3.4: 提案 3 手法が BD-Rate へ与える影響の解析．提案 3 手法を全て用いない場合は Palacio ら [113] の手法と同値であり (2 列目)，全て用いる場合は 3.3.4 節で検証されている提案手法と同値である (7 列目)．

	Palacio+ [113]	(i)	(ii)	(iii)	(iv)	Ours
Bypass structures		✓			✓	✓
Total variation loss			✓	✓	✓	✓
$\lambda_{\text{Recog.}}$ adjustment				✓		✓
BD-Rate	+8.1%	−1.3%	−11.2%	−12.1%	−20.3%	**−20.8**%

全体の平均では原画像の精度と大きく変わらないと考えられる．そのため，式 (3.4) で定義されるハイパーパラメータ $\lambda_{\text{Recog.}}$ も平均するとおおよそ 1.0 に近い値を取ると考えられ，$\lambda_{\text{Recog.}} = 1.0$ で設定した比較手法 (表 3.4 (iii)) と大きな差分が生じなかったと考えられる．しかしながら，大きな改善幅ではないものの，BD-Rate が改善していることから，表 3.4 は適応的に $\lambda_{\text{Recog.}}$ を調整する提案手法の有効性を示していると考えられる．

3.4.3　TV 損失のビットレート低減効果に関する検証

3.4.2 節の検証によって，TV 損失の導入は BD-Rate の改善に大きく貢献していることが明らかになった．前述の通り，TV 損失は隣接画素間の差分を小さくすることで画像信号を平滑化し，変換画像のビットレート低減を促進する働きを持つ．ここでは，この TV 損失のビットレート低減効果について検証する．

TV 損失のビットレート低減効果を検証するため，まず原画像と変換画像のエントロピーを比較する．ImageNet 2012 のテスト用サブセットからランダムに抽出した 10 枚の画像およびその変換画像の輝度成分における，画像信号と HEVC のフレーム内予測の残差信号のエントロピーを表 3.5 に示す．エントロピーは画像の画素値分布の偏りを評価する指標であり，小さい値であるほど，冗長性が高く圧縮しやすい信号であることを示している．画像信号におけるエントロピーは画素値の分布そのものが持つ平均情報量であり，主に JPEG や JPEG2000 のような画像信号をそのまま変換する圧縮標準への有効性を表す．一方，予測残差信号のエントロピーは HEVC や VVC 等のフレーム内予測の結果の残差信号を変換する方式への有効性を表している．表 3.5 より，提案手法の変換画像は画像信号・フレーム内予測の残差信号のいずれにおいても原画像よりもエントロピーが減少しており，提案手法のビットレート低減効果を示唆している．また，画像信号におけるエントロピーの低減が約 5% (6.88 bits → 6.51 bits) なのに対して，予測残差信号のエントロピーの低減は約 15% (5.11 bits → 4.40 bits) である．これは，TV 損失のビット

表 3.5: 原画像，提案手法の変換画像における画像信号とフレーム内予測の残差信号のエントロピー．画像は ImageNet 2012 のテスト用サブセットからランダムに 10 枚抽出し，エントロピーは輝度値から算出した．

	Raw Images		Pred. Residuals	
	Original	Transformed	Original	Transformed
1	5.31	5.34	4.74	4.18
2	7.02	6.85	5.19	4.55
3	7.11	6.53	5.36	4.63
4	6.95	6.58	4.28	3.77
5	7.27	6.81	5.16	4.50
6	6.98	6.41	6.40	5.12
7	7.01	6.36	5.91	5.09
8	7.17	6.77	4.83	4.22
9	7.21	6.95	4.96	4.41
10	6.78	6.45	4.23	3.56
Average	6.88 bits	**6.51** bits	5.11 bits	**4.40** bits

レート低減効果がフレーム内予測機構を有する圧縮標準に対して，特に大きいことを示しており，HEVC と VVC に対する有効性が JPEG や JPEG2000 に対するものよりも高かったという表 3.3 の結果と一致する．

次に，提案手法による変換画像と原画像でフレーム内予測にどのような差分が生じるかについて検証する．図 3.7 に，表 3.5 で調査した 10 枚の画像に対してフレーム内予測を適用した際の各予測モードの選択確率を示す．図 3.7 より，提案手法による変換画像と原画像では，ほとんどの予測モードで大きな差分は見られないが，方向性予測モードの 10·26 番では，原画像と変換画像との間で選択確率に差分が生じている．方向性予測のモードの選択確率に差分が発生することは直接的にビットレート低減には関係しないものの，図 3.7 に示す現象は TV 損失による効果であることを示唆していると考えられる．10 番は，ちょうど水平方向の画素値を予測画素として外挿し，26 番は，ちょうど垂直方向の画素値を予測画素とする方向性予測モードである．式 (3.1) で定義した通り，TV 損失は垂直 · 水平方向に隣接する画素の画素値との差分を最小化し，垂直および水平方向に類似する画素値が増加するような損失である．つまり，図 3.7 の解析結果によれば，TV 損失の影響で変換画像の 10 番と 26 番の方向性予測の予測効率が増幅し，そのモード選択率が高まったことでフレーム内予測の残差が減少したと考えられる．

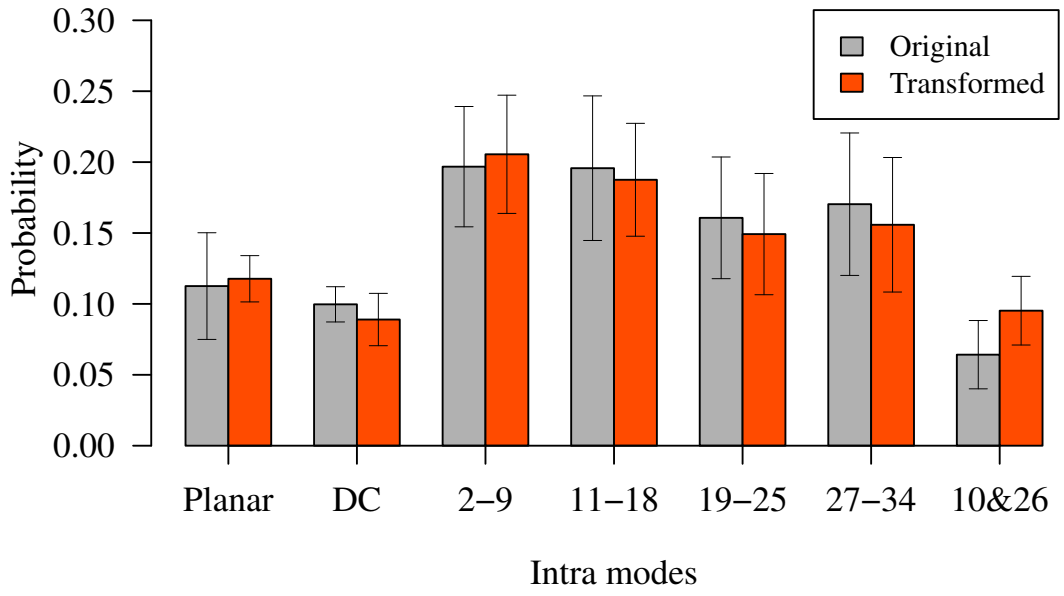

図 3.7: 提案手法による変換画像と原画像のそれぞれの圧縮処理で選択されたフレーム内予測の予測モードの比較.

3.4.4　主観画質評価に基づく歪み低減効果の検証

提案する画像プレ変換モデルは認識損失と TV に基づく損失を用いて学習され，表 3.5 の結果によれば，原信号とフレーム内予測の残差信号のいずれにおいてもエントロピーを低減させる効果を持つ．このような画像変換は，画像を非可逆圧縮した際のビットレート低減効果を高め，結果的にビットレートを低減させた際の圧縮歪みを抑制することが期待される．実際に，提案手法は非可逆圧縮による DNN の精度低下を防ぐことに成功している．以下では，DNN の認識精度以外の観点からも歪みの影響を評価するため，主観画質評価に基づいて提案手法の検証を行う．

本論文では，メディア処理の専門家 30 人を対象に主観画質評価実験を行った．評価法は，主観画質評価で一般的に用いらている Absolute Category Rating (ACR) 法と Degradation Category Rating (DCR) 法に基づいており [127]，両方の評価実験を被験者全員に実施した．ACR 法は，各評価画像に対する絶対画質評価法であり，評価画像 1 枚のみを提示しその画質を被験者が評価する．一方，DCR 法は，基準となる画像と評価画像との差分の相対評価法である．DCR 法では，基準画像として原画像を用い，原画像から圧縮等の処理を行った後の画像を評価画像として提示する．各評価法それぞれで ImageNet 2012 のテスト用サブセットからランダムに抽出した 30 枚の画像を評価画像の原画像とし，計 60 枚の原画像およびその変換画像をいくつかのビットレートに圧縮した画像を評価画像とした．なお，同一ビットレートで比較するため圧縮標準には JPEG2000

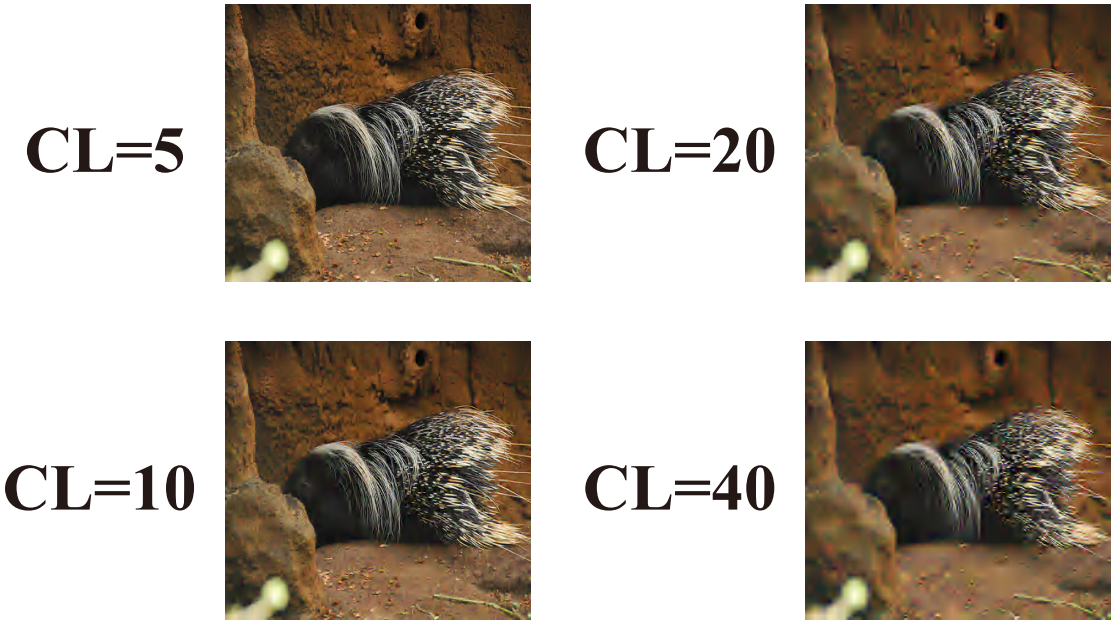

図 3.8: JPEG2000 によって圧縮された画像の例．主観画質実験に用いた各圧縮レベル，CL $\in$ {5,10,20,40}，それぞれにおける結果を示す．CL が大きくなればなるほど，圧縮歪みが大きくなっていることが分かる．

を採用し，評価は 5 段階評価とした．評価指標は，全評価者の 5 段階の評価を平均した値である Mean Opinion Score (MOS) 値を用いる．

主観画質評価の詳細な実験条件を説明する．ACR 法，DCR 法の評価に用いる原画像は ImageNet 2012 のテスト用サブセットから重複なくランダムに 30 枚ずつ抽出した．被験者に提示する評価画像は全被験者で事前に決定した順序で提示し，それぞれの圧縮レベルは CL $\in$ {5,10,20,40} からランダムに決定した．JPEG2000 によって各圧縮レベルで圧縮された画像の例を図 3.8 に示す．評価用画像は，5 秒ずつ表示した．一般に，主観画質評価では，実験の最初期に被験者に提示した画像では評価が安定しない．これは，被験者が実験にまだ慣れていないことや自身の評価値を実験の過程で安定させていくためである．そのため，本実験でも ACR 法·DCR 法ともに，最初に提示された評価画像 5 枚は用いずに画質評価結果を算出した．評価実験は，ACR 法 $\rightarrow$ DCR 法 $\rightarrow$ ACR 法の順に行い，ACR 法は二度の評価の平均値を MOS 値の算出に利用した．なお，二度の ACR 法ではそれぞれの評価画像の圧縮レベルは一致させている．

主観画質評価を行ったメディア処理の専門家は，年齢は 22 歳から 38 歳まで，性別は男性 26 名，女性 4 名であった．性別によって人数に差があるものの，それぞれの性別で主観画質評価の結果に大きな差分は生じなかった．視力や眼鏡 · コンタクトレンズ着用に関しては特に制限を設けていないが，日常生活に支障がない程度の視力がある状態で評価

表 3.6: 原画像 (O) と提案手法の変換画像 (T) における，JPEG2000 圧縮画像に対する主観画質評価の MOS 値と DNN の認識精度．各行に同一の JPEG2000 圧縮率に対応する MOS 値と認識精度を示す．

Compression level	ACR		DCR		Accuracy ([%])	
	O	T	O	T	O	T
5 (≈2.38 bpp)	**3.80**	3.46	**4.45**	4.01	68.0	**68.2**
10 (≈1.19 bpp)	3.33	**3.38**	**4.14**	3.62	65.3	**66.2**
20 (≈0.59 bpp)	2.50	**2.62**	2.67	**2.80**	58.0	**60.5**
40 (≈0.30 bpp)	1.66	**1.96**	1.63	**1.75**	44.7	**46.6**

を行った．評価を行ったデバイスは Lenovo ThinkPad X1 Carbon (5th Gen) であり，当該機種に表示される画像を基に被験者は主観画質評価を行った．当該機種のディスプレイは 1920x1080 解像度を有するが，実験に用いた画像は，認識精度の評価実験と同様のものであるため，256x256 解像度であった．

表 3.6 に，各ビットレートでの原画像と変換画像の MOS 値を ACR 法と DCR 法でそれぞれ示す．また，比較のために，3.3.4 節で算出した各ビットレートでの DNN の認識精度も併せて示す．ビットレートを決定する JPEG2000 の圧縮レベル (Compression Level: CL) は CL $\in$ {5,10,20,40} の 4 種類を用いた．表 3.6 より，ビットレートが大きい CL $= 5$ の場合，ACR 法 $\cdot$ DCR 法のいずれにおいても原画像の方が高い MOS 値を示している．また，ビットレートが低下するにつれて変換画像が原画像を逆転し，CL $= 20, 40$ では評価法によらず変換画像の方が高い MOS 値を示している．この現象は，ビットレートを低減させた際の圧縮歪みを抑制する提案手法の働きが，DNN の精度のみならず低ビットレート帯においては主観画質の保持にも寄与することを示唆している．また，CL $= 5$ では変換画像は ACR 法 $\cdot$ DCR 法のいずれにおいても原画像に比肩する主観画質を得られなかった．これは，画像プレ変換モデルが TV 損失と認識損失で学習されており，変換画像の画質を考慮する機構が存在しないことに起因すると考えられる．DNN の認識精度の観点では，CL $= 5$ において原画像 (68.0%) よりも変換画像 (68.2%) の方が高い値を示していることから，画質よりも精度保持を優先した画像変換の結果，主観画質が低下したと考えられる．人間の主観と DNN の精度が一致しない上記の事象は，

本章冒頭で述べた“人間ではなく DNN が認識にあたって重視しない情報を積極的に削減することで，認識精度を保持しつつもビットレートを低減できる”という本研究の着眼点を支持していると考えられる．

表 3.6 から ACR 法と DCR 法で結果に違いがあることも分かる．CL = 10 において，DCR 法では原画像の方が高い MOS 値を示しているが，ACR 法では変換画像の方が MOS 値が高い．CL = 20, 40 では，DCR 法でも変換画像の MOS 値が原画像を逆転するため，大きな差分は存在しないが変換画像は DCR 法に比べて ACR 法で高い評価を得る傾向があると考えられる．前述の通り，提案手法には原画像との MSE 等の画質を保持する損失が導入されていない．したがって，原画像と変換画像の画素値には乖離が生じる可能性があり，実際に図 3.6 に示すように輝度値が全体的に低下し，コントラスト等の変化が生じる．そのため，原画像との差分を評価する DCR 法では，この変化を被験者が知覚し，結果的に高い主観評価が得られなかったと考えられる．しかし，図 3.4 や 3.5 に図示する画像のように，提案手法は原画像と変換画像間で物体の形状や位置等はほとんど変化させていない．したがって変換画像は比較対象が無ければ変化を知覚しにくい画像となり，その結果，単一画像で画質を評価する ACR 法では比較的高い主観評価が得られたと考えられる．

3.5 本章のまとめ

本章では，非可逆に圧縮した画像に対しても Deep Neural Network (DNN) の認識精度を保持するため，画像圧縮処理に対する画像プレ変換に基づいた手法を提案した．提案手法は，Encoder-Decoder 型の DNN モデル (ED モデル) を利用した画像プレ変換手法であり，DNN の認識精度を保持しつつ画像信号を平滑化する処理を行うように学習されている．提案手法では，ED モデルの中でも特に SegNet と呼ばれるモデルにバイパス構造を導入し，画像認識精度を高める損失関数と画像信号を平滑化する Total Variation (TV) に基づく損失関数で学習した．先行研究で提案されている H.265/HEVC の量子化処理の改良手法はエンコーダ依存性が大きく，HEVC と量子化処理が異なる JPEG2000 や適応量子化機構が存在しない JPEG には直接適用できないという点で汎用性の制約が存在した．提案手法は，エンコーダ外部の処理であり，後段のエンコーダによらず圧縮時のビットレートを低減するため，エンコーダに依存しない形で精度を保持しながらビットレートを削減できる．ImageNet 2012 の識別タスクで，提案手法による変換画像は原画像をそのまま圧縮する場合と比較して，実験を行った，JPEG，JPEG2000，HEVC，VVC の全ての圧縮標準で認識精度を保持しつつビットレート低減効果を示した．ビットレートの低減効果は，最大で 20.5% (HEVC)，最小でも 8.6% (JPEG) であった．また，本章では提案手法が変換画像に与える影響に関しても解析を行い，圧縮後の画像において認識に

対して重要な信号を優先的に保持する働きや，フレーム内予測の予測残差を有効に低減させる働きを持つことが明らかになった．

残る課題として，ビットレート低減を促進する損失関数の改良が挙げられる．現在は TV に基づく損失関数を採用したが，この損失の場合，圧縮標準によってビットレート低減効果に差があることが明らかになった (表 3.3 参照)．この検証結果に基づき，圧縮標準によらず最大のビットレート低減効果を獲得できる損失を開発する必要があると考えられる．

第 4 章

時空間的配置法に基づく深層特徴圧縮技術とその解析

本章では，協調型知能方式 (図 1.1 (ii)) での，画像認識を指向した情報源圧縮技術について考える．協調型知能方式は，DNN を 2 つに分割し，フロントエンドデバイスとクラウドサーバにそれぞれを配置する．撮像された画像信号は，フロントエンドデバイスに配置した DNN に入力され，その出力である “深層特徴” をクラウドサーバに配置した DNN へ伝送 · 入力することで画像を認識する．協調型知能方式を提唱し，クラウド型知能方式と比較した初期の研究 [18] では，深層特徴は特に圧縮されず，そのまま伝送されることが多かった．しかし，深層特徴は信号の数値精度が 32-bit と画像信号 (通常 8-bit) よりも非常に大きく，非圧縮の場合，通信に要するビットレートが大きくなってしまう．このため，深層特徴を圧縮し，ビットレートを低減するための情報源圧縮技術が精力的に研究されている [21, 97, 99, 100, 102–105, 108, 128, 129]．図 4.1 に，深層特徴圧縮を用いた協調型知能方式の概要を示す．フロントデバイス側の DNN の出力である深層特徴は通常，2 次元配列の集合体であるテンソルであり，この 2 次元配列を画像状，もしくは映像状に変換し，圧縮 · 伝送される．これによって，非圧縮で伝送するよりも大幅な圧縮効率の向上が期待される．

DNN の抽出した特徴表現そのものである深層特徴は，従来の動画像圧縮標準や第 3 章で提案した手法が入力として想定している画像信号とは，異なる情報源である．したがって，高い圧縮効率を実現するためには，深層特徴がどのような性質を持った信号であるかを十分に把握し，冗長性を除去する必要がある．しかし，前述の通り，DNN の特徴抽出機構はブラックボックスで，深層特徴の性質は容易に把握できないため，従来の深層特徴圧縮技術は，深層特徴の信号としての性質を十分に圧縮に利用できなかった．本章で行う研究の目的は，協調型知能方式の高度化のために，深層特徴の性質に関する近年の研究成

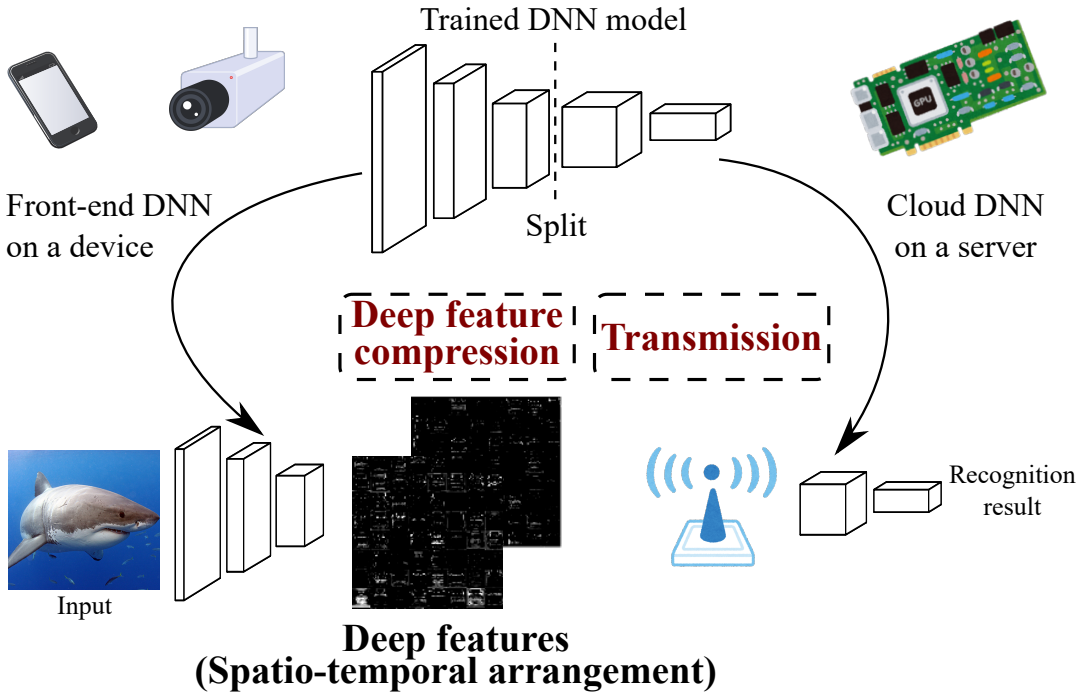

図 4.1: 深層特徴圧縮を用いた協調型知能方式による通信の基本概念．認識を行う DNN は二つに分割され，フロントエンドデバイスとクラウドサーバに配置される．深層特徴圧縮では，フロントエンドデバイス側の DNN の出力である深層特徴を圧縮し，クラウドサーバに伝送する．圧縮された深層特徴はクラウドサーバで解凍 (デコード) されクラウドサーバに配置された DNN に入力される．

果を導入することで，深層特徴の特性という観点を導入した新たな深層特徴圧縮技術を提案することである．まず，4.1 節では第 2 章で述べた関連研究に対する本研究の位置付けについて整理し，本研究で着目する深層学習の内部表現に関する研究について説明する．その後，4.2 節で提案手法について詳細を述べ，4.3 節と 4.4 節で提案手法の有効性を評価するための実験について説明する．4.5 節では，提案手法をいくつかの観点から解析する．最後に，4.6 節で本章をまとめる．

4.1 問題設定

2.7 節で述べたように，深層特徴は通常，2 次元配列の集合体であるテンソルであり，それぞれの 2 次元配列である特徴マップは画像と同様に空間的な冗長性を持つ．このため，深層特徴圧縮に関する初期の研究で，Choi ら [21, 99] は，各 2 次元配列を全て空間的に配置し，元の深層特徴と比較して空間的な冗長性を増加させた状態で一枚の画像として圧縮する手法を提案した．図 4.2 (a) に，空間的に配置された深層特徴の一例を示す．“空

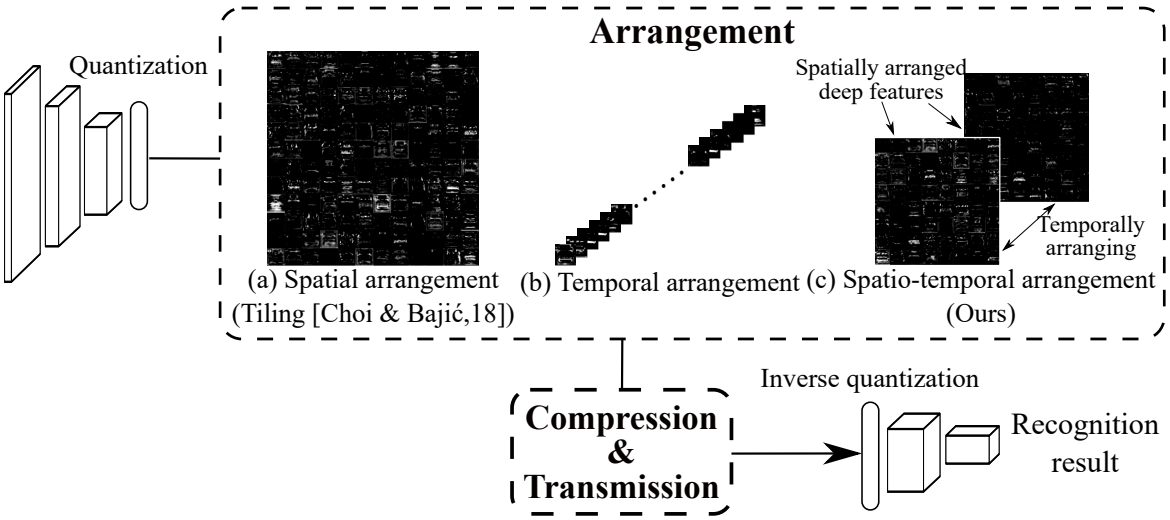

図 4.2: (a) 空間的配置法，(b) 時間的配置法，(c) 時空間的配置法で配置された深層特徴のそれぞれの例．(c) の時空間的配置法が本章における提案手法である．深層特徴は，まず，画像やビデオとして取り扱えるビット深度まで量子化される．量子化および (a) から (c) までのいずれかの手法で配置された深層特徴は，圧縮 · 伝送され，逆量子化処理の後に，クラウドサーバ側の DNN に入力される．

間的配置 (Spatial arrangement) 法” と呼ばれる Choi らの手法で画像化された深層特徴は，H.265/HEVC のような画像圧縮技術を用いて圧縮される[*1]．空間的配置法では，空間的に配置された深層特徴の間に大域的な冗長性が生じており，HEVC 等の画像圧縮手法が増加した空間的な冗長性を除去することで，圧縮効率を高めている．深層特徴圧縮における多くの研究において，空間的配置法はデファクトスタンダードとして用いられている [100, 102–105]．しかし，空間的配置法は深層特徴の特性を十分に把握した上で提案された手法ではなく，2.7 節で述べたように，ヒューリスティックに深層特徴を配置して圧縮効率を調査することでその有効性が発見された手法である．本章で行う研究の目的は，近年の内部表現に関する研究成果から明らかになってきた深層特徴の信号としての性質を深層特徴圧縮にうまく取り入れ，空間的配置法よりも高い圧縮効率を得る手法を提案することである．これは，まだ萌芽期の段階にある深層特徴圧縮という研究分野において，基盤となる新たな技術を提案することに他ならず，工学応用上重要な研究課題であると言える [106, 107].

本章で行う研究は深層学習の内部表現に関する研究において，DNN がどのような入力の画像信号に反応を示して深層特徴を作り出しているか，という観点から行われている研

[*1] 厳密には，HEVC は映像の圧縮標準であり，画像圧縮手法ではないが，先行研究では画像圧縮に用いる圧縮構造を利用している．

究に基づいている [30, 31, 73, 74]．例えば，Zeiler ら [30] は DNN の中間層に存在するニューロンが強く反応する入力の画像特徴を可視化する技術を提案した．この Zeiler らの研究は，DNN の抽出した特徴表現の可視化という新たな分野を切り開き，多くの後続研究がなされている [73–75]．これらの可視化研究では，DNN は浅い層では例えばエッジのような局所的な特徴に反応し，深い層では人の顔や体，猫の顔といったような抽象的な特徴に反応を示すことが明らかになった．ただし，これらはあくまで特定の階層の単一のニューロンの反応を調査したもので，深層特徴全体の性質の調査までは至っていない．一方，Suzuki ら [31] は，DNN が，同一階層の多数のニューロンの可視化を行った．この可視化の結果，深い層では，多くのニューロンがかなり似通った画像特徴に反応を示すことが明らかになった．つまり，DNN の抽出する特徴表現は反応する画像特徴という観点で見るとかなり冗長性が高いものであると言える．この DNN の冗長性の高さは，Pruning [14] の存在によって間接的にも証明されている．Pruning では，学習済みの DNN の結合を精度に影響を及ぼさないように取り除いており，DNN の冗長性の高さを利用しているものと言える．前述の通り，Choi らの空間的配置法は深層特徴のそれぞれの 2 次元配列に内在している空間的な冗長性を利用して画像として圧縮を行っている．画像圧縮手法は，フレーム内予測 [89] によってエッジや線分，塗りつぶしといった空間的な冗長性を除去しているが，DNN が抽出した特徴表現である 2 次元配列がそれぞれ類似しているといったような複雑な冗長性を除去することは難しいと考えられる．したがって，深層特徴に内在する，新たな冗長性を除去することで空間配置法よりも高い圧縮効率を実現することが期待される．

本章で行う研究では，深層特徴を画像ではなく擬似的に映像として扱い，圧縮する手法を提案する．ただし，深層特徴を映像として圧縮する手法は必ずしも未開拓のアイデアというわけではなく，映像圧縮に基づくアプローチの有効性を検証している研究はいくつか存在する [21, 108]．2.7 節で述べたように，“時間的配置法 (Temporal arrangement)” と呼ばれるこの配置法は，図 4.2 (b) のように深層特徴の 2 次元配列を時間軸方向に擬似的に配置する．入力が静止画像である場合，深層特徴に時間方向の次元は存在しないが，深層特徴におけるそれぞれの 2 次元配列は映像の各フレームと解釈でき，HEVC 等の映像圧縮手法を用いて圧縮できる．時間的配置法では，内部表現に関する研究成果によって存在が明らかになった前述のような冗長性を，時間的な冗長性として除去し，圧縮効率を高めることができると考えられる．これは，HEVC 等の圧縮手法が映像を圧縮する際に，フレーム内予測に加えてフレーム間予測 [90] を利用して入力の時間的な冗長性を除去するためである．2.5 節で述べた通り，フレーム間予測は隣接するフレーム (画面) の画素値をパッチレベルで予測することで冗長性を除去する．したがって，エッジや線分を予測するフレーム内予測よりも複雑な冗長性の除去が可能になる．しかし，Choi らや Chen ら [21, 108] は，時間的配置法は圧縮効率の点で空間的配置法を上回ることはできないこ

とを示している．これらの先行研究は，単純に両者の圧縮性能を比較しているため，一見深層特徴の複雑な冗長性を除去できるように見える時間的配置法がなぜ高い圧縮効率を実現できないのかについてはほとんど考察されていない．本研究では，時間的配置法が十分に圧縮効率を高められなかったのは，以下の 2 点の要因によるものと考えた:

1. 一般に，深い層から抽出した深層特徴の 2 次元配列は空間的なサイズが小さいため，各配列を 1 枚ずつ映像のフレームとすると，空間的な冗長性を十分に高められない．
2. 一般に，時間的に配置した深層特徴の近傍のフレームが類似する画像特徴に反応する特徴マップであるとは限らず，時間的な冗長性が十分に高められない．

つまり，時間的配置法は空間的および時間的な冗長性のいずれも十分に増加させることができなかったために圧縮率を十分に高められなかったものと考えられる．本章の提案手法は，これらの 2 つの要因を解消することで，圧縮効率を高める手法である．

本章では，空間的および時間的な冗長性を増加させ深層特徴圧縮の圧縮効率を向上させるため，“時空間配置法” という新たな手法を提案する．本手法では，深層特徴を時空間方向に沿って配置し，深層特徴を空間的に配置した複数のフレームからなる映像として表現する．図 4.2 (c) に時空間的に配置した深層特徴の一例を示す．提案手法では，$\mathbb{R}^{N\times M\times C}$ の要素を持つ深層特徴を，f 個のフレーム ($\in \mathbb{R}^{H\times W}$ ，ここで $H > N, W > M, f < C$) で構成された映像に配置する．また，新たに提案する配置順序探索アルゴリズムを用いて，各フレームに存在する深層特徴の 2 次元配列の配置順序を決定する．このアルゴリズムは，隣接するフレーム間の MSE を最小化する配置順序を事前に探索し，時空間的に配置された深層特徴の冗長性をより高めるような効果を持つ．以上の提案手法により，深層特徴に内在する空間的および時間的な冗長性を効果的に除去することを可能にし，従来の配置法よりも効率的な圧縮を実現する．ImageNet 2012 [117] データセットと VGG-16 [2] および ResNet-18 [4] モデルとして用いた検証において，提案手法は従来手法を凌駕する圧縮効率を示した．

本章で行った研究の主な貢献は以下の 3 点である:

1. 従来手法は，深層特徴に内在する時空間的な冗長性を圧縮に上手く活用できていなかった．提案手法は，この時空間的な冗長性を圧縮に利用することに成功した．
2. 時空間的な冗長性を高めるために，時空間的配置法と配置順序順序探索アルゴリズムを提案した．提案手法は ImageNet 2012 [117] データセットを用いた BD-Rate 評価では，空間的配置に比べて非可逆圧縮設定で 1.50% から 4.98% のビットレート低減を実現した．
3. 提案手法を解析した結果，従来手法とは異なり，提案手法はエッジやテクスチャが

豊富な画像を入力とした場合に，効果的に冗長性を高めていることを確認した．

4.2 時空間的配置に基づく深層特徴圧縮技術

深層特徴に内在する冗長性を効果的に除去するために，本章では時空間的配置法と配置順序探索アルゴリズムを提案する．本節では，これら 2 つの手法についてそれぞれ詳しく説明する．

4.2.1 深層特徴の時空間的配置

本章の提案手法では，深層特徴の各特徴マップを複数の画像状に空間的に配置し，映像の各フレームとして変換する．深層特徴は入力が静止画の場合，時間的な方向性の概念を持たないが，深層特徴を映像のフレームとして配置することで擬似的に時間的な方向性を導入することができる．提案手法は図 4.2 (c) のように，深層特徴を映像のフレームに時空間的に配置する．この際の配置順序は，あらかじめ配置順序決定アルゴリズムによって決定した配置順序である．提案する配置法は，空間的に配置された深層特徴の間に大域的な冗長性が生じるため，空間的に深層特徴を配置しない時間的配置法よりも空間的な冗長性が高くなる．さらに深層特徴は，空間的な方向だけでなく，時間的な方向にも配置されている．このような時間的な配置によって空間的配置法では存在しなかった時間的な冗長性を圧縮に利用できる．

時空間的な深層特徴の配置と圧縮処理の前に，深層特徴 $\mathbf{V} \in \mathbb{R}^{N \times M \times C}$ に対して量子化処理を行う．量子化ビット数が n-bit の信号 $\tilde{\mathbf{V}}$ は以下のような式で量子化が実現可能である:

$$\tilde{\mathbf{V}} = \mathrm{round}\left(\frac{\mathbf{V} - \min(\mathbf{V})}{\max(\mathbf{V}) - \min(\mathbf{V})} \cdot (2^n - 1)\right). \tag{4.1}$$

ここで，N，M，C は，それぞれ深層特徴の各特徴マップの高さと幅，および深層特徴の特徴マップの総数を表している．また，round(·) 関数は入力値の小数点以下の値を丸める処理を行う．本章では，従来研究の量子化処理に倣って，$n = 8$ と設定した [21, 102]. 深層特徴は一般に 32-bit の信号なので，この量子化によってビットレートを大幅に低減することが可能になる．さらに，8-bit 量子化により，深層特徴を通常の画像や映像と同様に 8-bit 信号として扱うことができる．また，上記のような量子化処理は round(·) 関数を用いて情報を非可逆に削減するため，認識精度を低下させる危険性がある．しかし，上記の 8-bit 量子化が認識精度に与える影響はごくわずかであることが実験的に明らかになっている．ビット深度の量子化が認識精度に与える影響に関する詳細な結果は，4.3.3 節で説明する．

時空間的配置法では，量子化された深層特徴 $\tilde{\mathbf{V}}$ を，高さ $H = N \times c_h$，幅 $W = M \times c_w$ のフレーム f 枚で構成される動画 F に変換する．ここで，c_h と c_w はそれぞれ垂直方向および水平方向に配置された深層特徴の特徴マップの数を示す．したがって，C は $f \times c_h \times c_w$ と同値であり，$H \geq N, W \geq M, f \leq C$ である．なお，空間的配置法と時間的配置法は，上記の定義において，それぞれ $f = 1$ と $f = C$ と解釈でき，提案する時空間的配置法の極端なケースであると考えられる．$f = C$ の場合のみ，つまり時間的配置法の場合のみ，H と W における等号が成立し，$H = N$，$W = M$ となる．

深層特徴の特徴マップの総数 C は一般に 2 のべき乗であるため，本研究では，2, 4, 8$\cdots$ のような 2 のべきの数を f に設定する．ただし，もし，C が f で割り切れない場合は，全ての値が 0 の特徴マップを擬似的に挿入することで提案手法を一般化できる．例えば，$C = 511$ かつ $f = 2$ の場合を考えると，全ての値が 0 であるチャンネルを 1 つ挿入することで $C = 512$ として深層特徴を扱うことができる．さらに，提案手法では，c_h と c_w を f を用いて以下のように一意に決定した，

$$c_h = 2^{\mathrm{ceil}\left(\frac{1}{2}\log C/f\right)}, \quad c_w = 2^{\mathrm{floor}\left(\frac{1}{2}\log C/f\right)}. \tag{4.2}$$

なお，c_h と c_w と圧縮効率の関係性は 4.5.6 節で詳細に解析を行っている．上記のように定義される時空間的配置法によって，提案手法は空間的および時間的な冗長性を高めることができる．まず，時空間的配置法は，フレームサイズ f が 1 である空間的配置法によって配置された深層特徴には存在しない時間的な冗長性を作り出すことができる．さらに，提案手法は，時間的に配置された深層特徴よりもはるかに高い空間的な冗長性を作り出す．例えば，ResNet-18 の 15 番目の畳み込み層では，深層特徴の要素数は $\mathbf{V} \in \mathbb{R}^{7\times7\times512}$ であるため，$f = 2$ の場合，空間的に配置されたフレームのサイズは $H \times W = 112 \times 112$ である．一方，時間的配置法を用いて配置された深層特徴では，フレームサイズはわずか $H \times W = 7 \times 7$ であるため，時空間配置法の方が遥かに大きなフレームサイズを有する (この場合は 16^2 倍)．したがって，提案する時空間的配置法では，時間的配置法と比較して，より高い空間的な冗長性を活用することが期待できる．結果として，提案手法では，従来の配置法とは対照的に，映像圧縮手法が時空間的な冗長性を有効に活用できるようになる．

時空間的に配置された深層特徴は，圧縮 · 伝送 · 解凍された後に，以下のような逆量子化処理によって $\tilde{\mathbf{V}}$ から $\hat{\mathbf{V}}$ となる，

$$\hat{\mathbf{V}} = \frac{\tilde{\mathbf{V}} \cdot (\max(\mathbf{V}) - \min(\mathbf{V}))}{2^n - 1} + \min(\mathbf{V}). \tag{4.3}$$

そして，$\hat{\mathbf{V}}$ は処理が残っている DNN，すなわちクラウド DNN (図 4.1 および 4.2 参照) に入力される．上記の逆量子化処理を行うためには，深層特徴 $\mathbf{V}$ の最大値および最

小値である $\max(\mathbf{V})$ と $\min(\mathbf{V})$ をクラウドに伝送する必要がある．ただし，本章の実験では，活性化関数に Rectified Linear Unit (ReLU) [45] を採用した DNN を使用しているため，$\min(\mathbf{V}) = 0$ とみなし $\min(\mathbf{V})$ の伝送を省略することができる．したがって，$\max(\mathbf{V})$ は 32-bit 信号であるため，本章で行う実験では圧縮した深層特徴のビットレート以外に 32 bits (4 bytes) 分のビットレートが発生する．

時空間的に配置された深層特徴である F に対する $\tilde{\mathbf{V}}$ の各特徴マップの配置順序は任意である．これは時空間的配置だけではなく空間的配置法 [21, 99] でも同様である．先行研究 [21, 99] では配置順序の詳細は具体的には記述されていないものの，特徴マップに付随するインデックスを用いて，インデックス順に F の左上からの昇順に配置していると考えられる．しかしながら，時空間的に配置された深層特徴が冗長性を持つか否かは特徴マップに付随するインデックスとは独立である．例えば，インデックスが 0 の特徴マップはヒトの顔に反応する，といったような対応関係は自明では無く，単純に昇順に特徴マップを並べるだけでは隣接するフレームに冗長性の高い (類似した画像特徴に反応する) 特徴マップが配置される保証がない．したがって，先行研究で用いられているようなインデックスに基づく配置順序は時間的冗長性を高めるには不十分である．そこで以下では，時間的冗長性を高めるための配置順序探索アルゴリズムを提案する．

4.2.2 局所探索アルゴリズムを用いた配置順序探索

特徴マップに付随するインデックスを用いて配置されている深層特徴は，前述の通り，必ずしも時間的な冗長性が高いとは限らない．似たような画像特徴に反応する特徴マップが全て同一フレームに存在する場合は，提案する時空間的配置法で時間方向に配置を行ったとしても時間的な冗長性を圧縮に利用することは難しい．そこで，各特徴マップの F への適切な配置順序探索アルゴリズムによって，時間的な冗長性を最大限高める手法を提案する．提案するアルゴリズムは，学習用のデータセットを使って事前に配置順序を探索するため，入力ごとに配置順序を探索 · 伝送する必要が無い．つまり，探索した配置順序を事前に 1 度伝送するだけで良い．

まず，提案する配置順序探索アルゴリズムの基本的な考え方を説明する．自然な映像信号では，一般に，隣り合うフレームの差分が小さいほど，時間的な冗長性が高くなり，圧縮後のビットレートが低くなると考えられる．この洞察は非常に直感的であり，さらに，Bandoh らによって実験的に確かめられている [130]．深層特徴は自然な映像信号とは信号源としての特性は異なるものの，本研究で提案する時空間的配置法はフレーム間予測を用いて時間的な冗長性を除去する構造は同様のものである．したがって，時空間的に配置された深層特徴においても，隣接するフレーム間の差分を小さくすることで，時間的な冗長性を高めることが期待できる．このようなアイデアに基づいて，隣接するフレーム間の

Algorithm 1 配置順序探索のための局所探索アルゴリズムの概要．このアルゴリズムは，量子化済みの深層特徴のサブセットを入力として受け取り，探索結果として $O_1, \cdots, O_f \in \mathbb{R}^{c_h \times c_w}$ を出力する．各 iteration において，O_1 に存在する x_1 と O_k に存在する x_{can} の 2 つのインデックスを交換しながら隣接するフレーム間の MSE を削減する配列順序を探索する．

Input: Quantized deep features subset
Output: Search results $O_1, \cdots, O_f \in \mathbb{R}^{c_h \times c_w}$

$O_1, \cdots, O_f \Leftarrow random\_assign(1, \cdots, C)$
for $i = 1$ to $c_h \times c_w$ **do**
  for $j = 1$ to $c_h \times c_w$ **do**
    for $k = 2$ to f **do**
      $x_1 \Leftarrow O_1[i],\ x'_1 \Leftarrow O_2[i]$
      $x_{can} \Leftarrow O_k[j],\ x'_{can} \Leftarrow O_{k-1}[j]$
      if $k == f$ **then**
        $cu\_mse \Leftarrow mse(x_1, x'_1) + mse(x'_{can}, x_{can})$
        $ex\_mse \Leftarrow mse(x_{can}, x'_1) + mse(x'_{can}, x_1)$
      else
        $x''_{can} \Leftarrow O_{k+1}[j]$
        $cu\_mse \Leftarrow mse(x_1, x'_1) + mse(x'_{can}, x_{can}) + mse(x_{can}, x''_{can})$
        $ex\_mse \Leftarrow mse(x_{can}, x'_1) + mse(x'_{can}, x_1) + mse(x_1, x''_{can})$
      end if
      if $ex\_mse < cu\_mse$ **then**
        $O_1[i] \Leftarrow x_{can},\ O_k[j] \Leftarrow x_1$
      end if
    end for
  end for
end for

MSE を最小化するという観点で最適な配置順序を探索する配置順序探索アルゴリズムを提案する．

提案するアルゴリズムは，局所探索アルゴリズムに基づいており，現在の配置順序よりも小さい MSE を示す配置順序を逐次的に探索する．逐次的な探索を繰り返すことで，時空間的に配置された深層特徴の MSE を最小にする配列順序の近似解となる $O_1, \cdots, O_f$ をアルゴリズムが出力する．ここで，$O_1, \cdots, O_f$ は $c_h \times c_w$ の要素を持つ行列であり，

この行列の各要素は対応する F に配置される特徴マップのインデックスを示す．これらの行列は，対応するフレームへの配列順序を決定する．例えば，O_1 上の要素は，F の第 1 フレームの順序を決定し，O_1 の 1 行 ·1 列目に 10 という要素がある場合，第 1 フレームの 1 行 ·1 列目に配置する特徴マップは 10 番のインデックスに紐づいている特徴マップとなる．したがって，本章の実験では，配置順序探索アルゴリズムを用いて $O_1, \cdots, O_f$ を得た後，これを用いて F を作り出す．

Algorithm 1 は，配置順序探索アルゴリズムの手順を示している．まず 1 行目に示すように $O_1, \cdots, O_f$ を初期化する．初期化では $random\_assign(\cdot)$ 関数を用いてインデックス $(1, \cdots, C)$ をランダムに割り当てる．このアルゴリズムでは，O_1 $(x_1 \Leftarrow O_1[i])$ の i 番目のインデックスと O_k $(x_{can} \Leftarrow O_k[j])$ の j 番目のインデックスを交換しながら，現在の配置順序よりも隣接フレーム間の MSE が小さくなるような配置順序を探索する．ここで，i，j，および k はアルゴリズムにおけるループ変数である．Algorithm 1 は，この探索処理を O_1 上のすべてのインデックスに対して繰り返し行い，最終的に繰り返しが終了した際の $O_1, \cdots, O_f$ を探索した配置順序として出力する．

図 4.3 は，Algorithm 1 に示すアルゴリズムの概要を簡略化したものである．提案するアルゴリズムの最小化対象である隣接フレーム間の MSE は，全ての隣接フレームにおいて単純に MSE を求めることによっても算出できるが，Algorithm 1 では，アルゴリズムの計算量削減のために x_1 と x_{can} のみに着目して算出する (図 4.3 の Step 1)．具体的には，交換の対象となるインデックス x_1 または x_{can} と，それに隣接するインデックスとの間の MSE を計算する (Step 2)．これによって交換に無関係なインデックスにおける MSE の計算が省略できる．x_1 の隣接するインデックスは $O_2[i]$ であり，以下では，これを x_1' と呼ぶことにする．x_{can} の隣接するインデックスの数は x_{can} が存在する k 番目のフレームが何番目のフレームであるかに依存するため，Algorithm 1 は 7 行目と 10 行目に示すように動作する．$k = f$ (7 行目) の場合，x_{can} の隣接するインデックスは $O_{k-1}[j]$ のみであり，これを x'_{can} と呼ぶことにする．これは，$k = f$ の場合，O_{f+1} が存在しないためである．その他の場合 (10 行目) では，$O_{k+1}[j]$ も隣接するインデックスとして存在する．これを x''_{can} と呼ぶ．Algorithm 1 に示す $mse(a, b)$ 関数は，インデックス a と b に対応する 2 つの特徴マップ間の MSE を計算して出力する．MSE は，入力である量子化された深層特徴のサブセットに存在するすべての深層特徴に対して計算され，$mse(a, b)$ 関数はその平均値を出力する．現在の配置順序の MSE (cu_mse) は，交換対象のインデックスである x_1 または x_{can} とその隣接するインデックスの間で計算される (8 行目または 12 行目)．配置順序を交換した場合の MSE (ex_mse) は，x_1 と x_{can} を交換した状態で cu_mse と同様に算出する (9 行目または 13 行目)．最後に，cu_mse と ex_mse を比較し，ex_mse の方が小さい場合は，交換した配置順序を新しい順序として設定する (Step 3)．

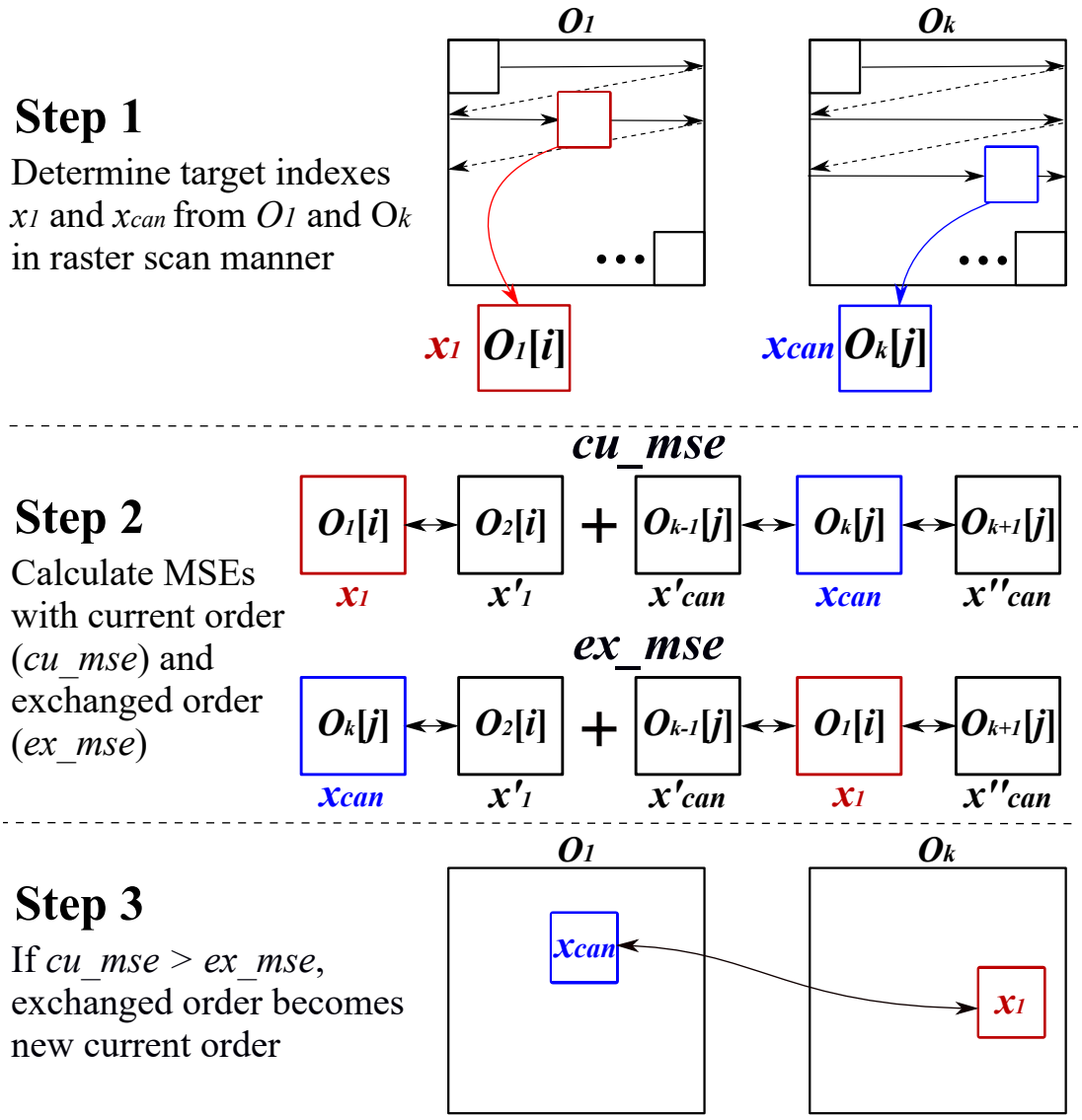

図 4.3: Algorithm 1 の処理を簡略化した図．Step 1 で，1 枚目と k 枚目のフレームからラスタスキャン順に交換対象となる 2 つのインデックス x_1 と x_{can} を選択する．Step 2 で，現在の配置順序と x_1 と x_{can} が交換された場合の配置順序での MSE をそれぞれ計算する．最後に，Step 3 で，配置順序を交換した場合の MSE が現在の配置順序のものよりも小さければ，x_1 と x_{can} を交換する．なお，この図は，$k = f$ ではない場合のアルゴリズムの処理を例示している．

このアルゴリズムでは，隣接するフレーム間の MSE 値のみを考慮しているため，フレーム間予測における動き予測 (Motion Estimate) [131] を直接的には考慮していない．したがって，提案するアルゴリズムは動き予測の冗長性除去効果を十分に引き出せていない可能性は否定できない．しかしながら，この配置順序探索アルゴリズムでは異なる映像中のフレームに類似した画像特徴に反応する特徴マップを割り当てる効果があるため，動き予測の機構も十分に活用しながら時間的な冗長性を除去できる可能性も，同様に否定できない．圧縮時の動き予測の有無によるビットレート差異を解析した結果，動き予測機構が十分に活用できていることが明らかになった．この解析の詳細な内容については，4.5.4 節で述べる．また，Algorithm 1 では，1 つのフレーム (O_1) のみから x_1 を設定し探索を行っている．O_1 以外のフレームにおいても探索を繰り返せば，MSE がより小さくなり，時空間の冗長性も高くなることが期待される．しかし，圧縮効率の向上は繰り返し探索を行ってもほとんど変化しないことが経験的に明らかになっているため，本研究では O_1 のみから x_1 を設定し探索を行っている．

4.3 ニアロスレス圧縮条件での評価実験

本節では，画像認識タスクをおいて，提案手法をニアロスレス圧縮条件 [99] の下で評価した実験について述べる．ニアロスレス圧縮条件では，式 (4.1) によって量子化された深層特徴 $\tilde{\mathbf{V}}$ を可逆に圧縮 (ロスレス圧縮) する．この圧縮条件では，深層学習の圧縮自体は可逆であるものの，n-bit の量子化処理のために深層特徴は厳密には可逆の処理とはならない．そのため，“ニア” ロスレス圧縮条件と呼ばれている [99]．本節における評価実験では，まず，事前調査として，量子化ビット数 n とフレームサイズ f が深層特徴圧縮に与える影響を調査した．その後に，提案手法と空間的および時間的配置法の圧縮効率を定量的に比較した．

4.3.1 データセットの詳細

本章では，検証のために，大規模自然画像データセットである ImageNet 2012 [117] を用いる．ImageNet 2012 は，一般に画像識別タスクで用いられるデータセットであり，計 1000 カテゴリ，約 128 万枚の訓練画像と 5 万枚の検証画像を有する．各実験では，画像サイズを事前に 256×256 にリサイズした画像を用いる．提案手法を評価するための検証用セットとして 5 万枚の検証画像から 1 万枚の画像をランダムに選択した．さらに，訓練画像からクラスごとに 5 枚の画像をランダムに選択し，Algorithm 1 に示している配置順序探索アルゴリズムの入力となる，計 5,000 個の量子化された深層特徴のサブセットを作成した．

4.3.2　評価実験の詳細

評価実験に用いる圧縮手法として，フレーム内予測とフレーム間予測の両方を利用できる映像圧縮標準である H.265/HEVC [24] を全ての実験で利用した．HEVC の中でも特に，グレイスケール (YUV 4:0:0) の信号フォーマットに対応した “HEVC RExt” [98] と呼ばれる規格を用いた．画像認識を行う DNN のアーキテクチャには，ImageNet 2012 で学習した VGG-16 [2] と ResNet-18 [4] を使用した．また，2.4 節で述べたように，協調型知能方式に関する先行研究 [18] の解析によって，伝送効率の良い分割点は DNN の比較的深い層であることが明らかになっている．そこで，各モデルの深い層の中から，深層特徴の要素の大きさが異なる 2 つの層の出力を本章の実験で用いる深層特徴として利用した．具体的には，VGG においては 8 番目および 11 番目の畳み込み層，ResNet では 13 番目および 15 番目の畳み込み層の出力を実験に用いた．本章における，全ての実験において，DNN の利用にあたっては Caffe [124] と呼ばれるフレームワークを使用した．また，再現性を高めるために，学習済みモデルには Caffe がオンラインに公開しているモデルを使用した．

空間的および時空間的に配置された深層特徴の具体的な例を，図 4.4 に示す．例示している深層特徴は，同じ入力画像 (図の右上に提示) に対応する深層特徴である．(a) VGG の第 8 層，(b) VGG の第 11 層，(c) ResNet の第 13 層，(d) ResNet の第 15 層からそれぞれ深層特徴を抽出した．時空間的に配置された深層特徴は，$f = 2$ の場合を例示している．ResNet の第 13 層から抽出した深層特徴の例が他の深層特徴よりも小さいが，これは ResNet の第 13 層の特徴マップの総数が $C = 512$ ではなく $C = 256$ であることに起因している．

提案する配置順序探索アルゴリズムに関しては使用する場合としない場合でそれぞれ圧縮効率を調査した．配置順序探索を使用しない場合，特徴マップに紐づいているインデックスによる昇順で深層特徴を配置した．探索を使用する場合は，探索された配置順序のビットレートを深層特徴の圧縮時のビットレートに含めた．ただし，4.2.2 節で述べているように，探索した配置順序は事前に学習データを用いて決定され，事前に一度だけ伝送すればよいため，ビットレートへの影響はほとんど無い．配置順序のビットレートに関する，より詳細な分析については，4.5.5 節で説明する．

提案手法と時間的配置法では，深層特徴を映像状に配置して圧縮するため，映像信号の圧縮に用いられる圧縮構造である Random-access (RA) 構造を圧縮に用いた．空間的配置法においては，圧縮対象となるのは画像状に配置した深層特徴であるため，静止画用の圧縮構造である Intra-only 構造を用いた．

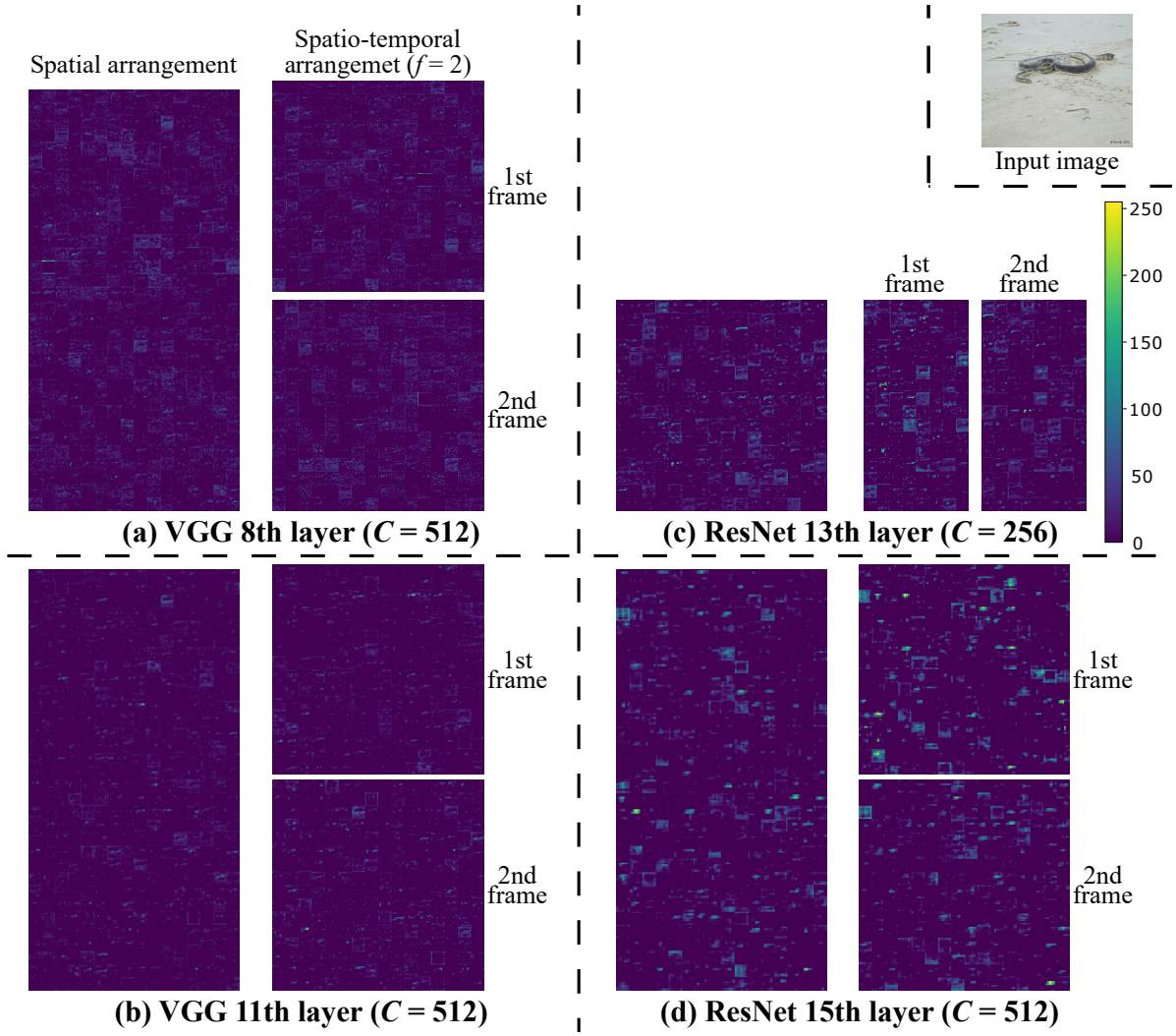

図 4.4: 空間的および時空間的に配置された深層特徴の例．(a) に VGG の第 8 層，(b) に VGG の第 11 層，(c) に ResNet の第 13 層，(d) に ResNet の第 15 層から抽出された深層特徴をそれぞれ示す．例示されている深層特徴は，全て右上に示す同じ入力画像から得られたものである．また，分かりやすさのため，深層特徴はカラーマップ状に表示している．

4.3.3 n-bit 量子化処理が認識精度へ与える影響

まず，n-bit の量子化が DNN の認識精度に与える影響を分析する．式 (4.1) に示すように，提案手法では，元来 32-bit 信号である深層特徴を画像や映像として扱うために，信号を量子化処理する必要がある．4.2.1 節で述べたように，本研究では，先行研究 [21,102] に倣って $n = 8$ として量子化を行った．本節では，8-bit 量子化が DNN の認識精度に対してほとんど影響を及ぼさないことを示し，今後の実験にて 8-bit 量子化された深層特徴を用いる妥当性について検討する．

表 4.1: 深層特徴の量子化が DNN の画像認識精度へ与える影響．認識精度の下にある数値は．本来の深層特徴の精度 (32-bit 信号の場合の精度) との差分を示している．

	layer	32-bit (Original)	16-bit	8-bit	6-bit
VGG	8th	68.89 %	68.89 % (± 0)	68.89 % (± 0)	68.89 % (± 0)
	11th	68.89 %	68.89 % (± 0)	68.89 % (± 0)	68.89 % (± 0)
ResNet	13th	68.06 %	68.06 % (± 0)	68.09 % (+ 0.3)	68.02 % (− 0.4)
	15th	68.06 %	68.06 % (± 0)	68.04 % (− 0.2)	68.04 % (− 0.2)

表 4.1 は，各量子化ビット数における DNN の認識精度を示している．この実験では，深層特徴を 32-bit，16-bit，8-bit，6-bit にそれぞれ量子化し，認識精度を算出した．深層特徴の数値精度は 32-bit であるため，32-bit 量子化時の認識精度は DNN の本来の精度を示している．表 4.1 に示す結果から，8-bit 量子化は DNN の認識精度に対してごくわずかな影響しか及ぼさないことが分かる．VGG-16 では，実験したすべての量子化ビット数で認識精度に変化がなかった．DNN は Binarized Neural Networks [132] のように，二値化された状況でもほとんどの精度を維持できるため，このような結果も妥当であると考えられる．また，6-bit の量子化においても同様の結果となっている．しかし，本研究では，以下の二つの理由により，$n = 8$ と設定した．

1. 多くの動画像圧縮では入力信号は 8-bit もしくはそれ以上の数値精度を持つ信号であることを前提としているため，8-bit を下回る量子化ビット数を設定しても圧縮率の向上への寄与は少ない．
2. ResNet は VGG とは異なり，8-bit および 6-bit に量子化した際に，若干ではあるが精度が変動している．特に 6-bit の場合，13 層目と 15 層目のいずれも認識精度が本来の精度よりも低下している．この量子化に対する精度変化は ResNet で用いられている Batch Normalization [48] の影響であると考えられる．

4.3.4　圧縮効率に対するフレームサイズ f の影響

本節では，時空間的配置法における F のフレームサイズ f が圧縮効率に与える影響を分析する．f を大きくすると，時間的な冗長性をより効率的に圧縮に利用できる可能性が

表 4.2: フレームサイズが $f = 2, 4, 8$ の場合の，ニアロスレス圧縮条件下での深層特徴圧縮における，提案手法の平均ビットレート．表中の OS は提案する配置順序探索アルゴリズムを示す．

		$f = 2$		$f = 4$		$f = 8$	
	layer	w/o OS	w/ OS	w/o OS	w/ OS	w/o OS	w/ OS
VGG	8th	159.0 KB	158.1 KB	158.8 KB	**156.7 KB**	158.8 KB	157.2 KB
	11th	34.16 KB	34.05 KB	34.26 KB	**33.49 KB**	34.49 KB	33.72 KB
ResNet	13th	22.91 KB	22.83 KB	22.94 KB	**22.60 KB**	23.15 KB	22.92 KB
	15th	10.33 KB	**10.17 KB**	10.59 KB	10.19 KB	10.87 KB	10.53 KB

ある．しかし，f が必要以上に大きくなると，空間的に配置される特徴マップの数が減少し，空間的なサイズが小さくなるため，空間的な冗長性が減少する．このトレードオフの影響を評価するために，$f = 2$, 4, 8 のそれぞれの場合で深層特徴を圧縮し，圧縮効率を比較した．

表 4.2 は，実験を行った各 f に対する深層特徴の平均ビットレートを示している．ここで，“OS” は提案の配置順序探索アルゴリズムを意味し，使用する場合と使用しない場合のそれぞれでビットレートを算出している．表 4.2 に示す結果より，$f = 4$ で深層特徴を時空間に配置した場合，ResNet-18 の第 15 層を除く全てのケースで，配置順序探索アルゴリズムを用いた提案手法が最も良い圧縮効率を得ている．一方，$f = 8$ の場合，提案手法は全てのモデルおよび階層で最も高い圧縮効率を達成することができなかった．この結果は，空間的冗長性と時間的冗長性の間に存在するトレードオフは，$f = 2$ または $f = 4$ の場合に最も優れた圧縮効率を提供することを示唆している．この結果に基づいて，以降の評価実験では，時空間的配置法に対して $f = 2$ および $f = 4$ の 2 種類のフレーム数を使用する．

また，興味深いことに，配置順序探索アルゴリズムを使用しない場合，多くのケースで，f が増加するに伴って平均ビットレートが増加していることが分かる．この結果は，f が大きくなると，同一フレーム内に類似した画像特徴に反応する特徴マップが存在することが，時間的な冗長性を高めるためのボトルネックになることを示唆している．f が大きい場合，各フレームに配置される特徴マップの数は相対的に少なくなる．そのため，類似した画像特徴に反応する特徴マップが同一フレームに複数存在すると，異なるフレームにフレーム間予測に有効な特徴マップが存在する可能性が低くなってしまう．配置順序探索アルゴリズムを用いた場合は，隣接するフレーム間の MSE を減らすような配置となるために，類似の特徴マップはなるべく異なるフレームに配置される機構が働く．したがって，上記ようなボトルネックが回避できるものと考えられる．

表 4.3: ニアロスレス圧縮条件下での深層特徴圧縮における，空間的および時間的配置法と提案手法ビットレートの比較．表中の OS は提案する配置順序探索アルゴリズムを示す．

	layer		OS	v.s. spatial [21]	v.s. temporal [108]
VGG	8th	$f=2$		−1.80 %	−13.1 %
			✓	−2.33 %	−13.6 %
		$f=4$		−1.94 %	−13.3 %
			✓	**−2.93** %	**−14.1** %
	11th	$f=2$		−1.37 %	−27.8 %
			✓	−1.69 %	−28.1 %
		$f=4$		−1.11 %	−27.6 %
			✓	**−2.66** %	**−28.8** %
ResNet	13th	$f=2$		−1.39 %	−28.2 %
			✓	−1.76 %	−28.5 %
		$f=4$		−1.27 %	−28.1 %
			✓	**−2.76** %	**−29.2** %
	15th	$f=2$		−3.57 %	−50.7 %
			✓	**−5.10** %	**−51.5** %
		$f=4$		−1.16 %	−49.4 %
			✓	−4.84 %	−51.3 %

4.3.5　空間的および時間的配置法との圧縮効率の比較

表 4.3 は，空間的および時間的配置法に対する提案手法のビットレート低減効果の比較を示している．なお，空間的 · 時間的配置法は，それぞれ $f=1$ · $f=C$ の時空間的配置法の極端なケースであると言える．

表 4.3 において，負の値は，提案手法が比較手法から平均ビットレートを低減したことを示す．提案手法は，配置順序探索アルゴリズムの有無によらず，実験を行った全ての層とモデルにおいて，空間的および時間的配置法よりも大きなビットレート低減効果を示している．さらに，表 4.2 の結果でも示している通り，配置順序探索アルゴリズムを使用することで圧縮効率はより高まっている．この結果は，提案する時空間配置法と配置順序探索アルゴリズムが時空間的な冗長性を効果的に高め，映像圧縮手法である HEVC がその冗長性を効果的に除去できることを示している．また，上記の結果は，深層特徴に内在する冗長性を圧縮に活用するための最適なフレームサイズは，$f=1$ や C といった極端な

ケースではない，という非常に重要な事実を提示していると考えられる．この事実は，深層特徴の冗長性をより効果的に引き出し，高い圧縮効率の実現を狙うための新しい研究の方向性を示唆しており，深層特徴圧縮の分野において重要な知見となる可能性がある．

さらに，時間的配置法の圧縮効率は，他の 2 つの配置法と比較して非常に低いことが実験の結果，明らかになった．この結果は，時間的配置法の圧縮効率を調査した先行研究の知見と一致している [21,108]．深い層から抽出した深層特徴の空間的なサイズは非常に小さいため，時間的に配置された深層特徴における空間的な冗長性は小さいものであると考えられる．本実験の結果は，空間的な冗長性を十分に高められていない場合，深層特徴圧縮の圧縮効率に大きな悪影響を及ぼすことを示している．本章での提案手法は，表 4.3 に示すように，この空間的な冗長性の欠如をうまく回避し，圧縮効率を向上させており，この観点からも提案手法の有効性が示されていると考えられる．

4.4 非可逆圧縮条件下での評価実験

4.4.1 評価実験の詳細

以下では，深層特徴を非可逆に圧縮する非可逆圧縮条件の下で評価した実験について述べる．非可逆圧縮条件では，幾つかの量子化パラメータ (Quantization Parameter: QP) を設定し，得られたビットレートを Bjøntegaard Delta Bitrate (BD-Rate) [125,126] を用いて比較した．BD-Rate は，2 種類の圧縮方式の性能を比較する際に用いられる指標であり，画像品質が同一となる際に，どの程度ビットレートを低減できるかを評価する．一般的に，BD-Rate は PSNR 等を画像品質の指標としていることが多いが，どのような指標を用いても BD-Rate は算出可能である．本章の研究では，認識精度を保持しつつビットレートを低減することを目的としているため，認識精度を評価指標として設定した．具体的な認識精度の指標は，上位 1 位識別結果 [1] を用いている．BD-Rate は，ある画像品質を得るために必要なビットレートのペア 4 組から算出する．この 4 組の選び方は任意であるが，本研究では，標準化策定作業で一般に用いられている，QP $\in \{22, 27, 32, 37\}$ の 4 組を用いた．

提案手法における映像の圧縮では，フレーム数 f が 2 または 4 であるため ($f = 2,\ 4$)，RA 圧縮構造を以下のように変更した:

i) GOP サイズをフレーム数に揃えて 2 または 4 に設定した．

ii) 全ての QP offset を 0 に設定した．

iii) 予測フレームの QPfactor を 0.2 に設定した．

他の配置法を用いた深層特徴圧縮では，空間的配置法では Intra-only 圧縮構造，時間的配置法では RA 圧縮構造をそれぞれ用いた．本章で行った全ての実験では，ビットレー

表 4.4: 空間的配置法および時間的配置法をアンカーとした際の提案手法の BD-Rate．表中の OS は提案する配置順序探索アルゴリズムを示す．

	layer		OS	v.s. spatial [21]	v.s. temporal [108]
VGG	8th	$f = 2$		−1.31 %	−52.0 %
			✓	**−2.27** %	**−52.5** %
		$f = 4$		−0.70 %	−52.4 %
			✓	−1.05 %	−52.3 %
	11th	$f = 2$		−0.98 %	−58.5 %
			✓	**−1.50** %	**−58.7** %
		$f = 4$		−0.39 %	−58.6 %
			✓	−0.80 %	−58.0 %
ResNet	13th	$f = 2$		−1.08 %	−62.3 %
			✓	**−1.50** %	**−63.2** %
		$f = 4$		−0.22 %	−61.8 %
			✓	−0.70 %	−62.0 %
	15th	$f = 2$		−1.53 %	−74.1 %
			✓	−1.83 %	−72.5 %
		$f = 4$		−3.66 %	**−76.3** %
			✓	**−4.98** %	−75.3 %

トと非可逆圧縮に起因する圧縮歪みを最適化する RDOQ (Rate-Distortion Optimized Quantization) は使用していない．その他の圧縮パラメータは, 共通テスト条件 (Common Test Condition: CTC) [133] に従っている．

4.4.2　非可逆圧縮条件における実験結果

表 4.4 は，本章の提案手法である時空間的配置法と比較手法の BD-Rate を示している．表に示しているように，全ての BD-Rates が負の値を示している．これは，DNN の認識精度を保持するためのビットレートを提案手法が低減させたことを意味している．比較手法をそれぞれ見ていくと，時間的配置法 [108] は，提案手法が 50 % を超える BD-Rate 利得を得ていることから，深層特徴の非可逆圧縮に適していない配置法であると考えられる．空間的配置法 [21] は，提案する時空間的配置法に近い圧縮効率を示しているが，定量評価では提案手法の方が高い圧縮効率を示している．これらの結果は，ニアロスレス圧縮条件での結果とほぼ一貫しており，いずれの場合においても提案手法が優れた圧縮効率

を示している．しかしながら，非可逆圧縮条件では，$f = 4$ の場合での空間的配置法と比較した BD-Rate (例えば，VGG の 8 層目では -1.05 %) はほとんどの階層とモデルで $f = 2$ の場合 (VGG の 8 層目では -2.27 %) よりも悪化している．これはニアロスレス圧縮条件下ではほとんど見られなかった現象である．この結果は，非可逆圧縮条件下での提案手法の圧縮では，フレーム間予測の参照フレームに非可逆圧縮によるアーチファクトが発生したことが原因と考えられる．参照フレームのアーチファクトは当然，フレーム間予測の精度に悪影響を及ぼし，f が大きい値の場合において，圧縮効率を低下させる結果となったと考えられる．ニアロスレス圧縮条件の場合，このようなアーチファクトは発生しないために，f が大きい値の場合でも圧縮効率が低減しなかったものと考えられる．ただし，このような不利な状況でも，提案手法は空間的配置法に比べて BD-Rate 利得を得ることができている．図 4.5 は，定性的な評価を行うために，空間的配置法 (黒線)，時間的配置法 (緑線)，$f = 2$ かつ配置順序探索アルゴリズムを用いる提案手法 (赤線) における，深層特徴の圧縮時のビットレートと認識精度の関係を示している．図 4.5 において，青色の水平方向の直線は表 4.1 の 32-bit の列に示された DNN の本来の認識精度を示している．図に示した結果は，時間的配置法によって配置された深層特徴の圧縮効率が際立って低いことを示している．空間的配置法は本章での提案手法に匹敵する圧縮効率を示しているものの，全体的に提案手法の方が同一認識精度でのビットレートを低減できているように評価できる．これらの関係は，BD-Rate を用いた定量評価と一致しており，定量評価 · 定性評価が一貫した結果を示していることから，妥当な評価であることを示唆している．

これまで述べてきたニアロスレスおよび非可逆圧縮条件での実験結果によって，時空間配置法と配置順序探索アルゴリズムは，シンプルでありながらも協調型知能方式のための深層特徴圧縮に有効であることが明らかになった．この有効性は，提案する 2 つの手法が，時空間的冗長性を効果的に高めることに成功したために生じていると考えられる．先行研究が提案していた空間的および時間的配置法では，この時空間的な冗長性を十分に動画像圧縮手法が活用できていなかった．本章の実験結果は，既存の空間的配置法を用いた深層特徴圧縮技術の多くが，深層特徴に内在している冗長性を完全には圧縮に利用できていないという事実を示唆している．この事実は，今後，深層特徴の圧縮効率を向上させるための取り組みに対して大きな影響を与えるものと考えられる．

4.5 深層特徴の時空間的配置法の解析

ここまでに述べてきた実験結果によって，提案手法は実験的にその有効性が示されたと考えられる．しかしながら，入力信号がどのような場合に圧縮効率が高まり，あるいは，どのような場合に圧縮効率が低下するか，といった評価は不十分である．この点を明確にするために，本節では時空間配置法を解析し，その有効性について検証を行った．

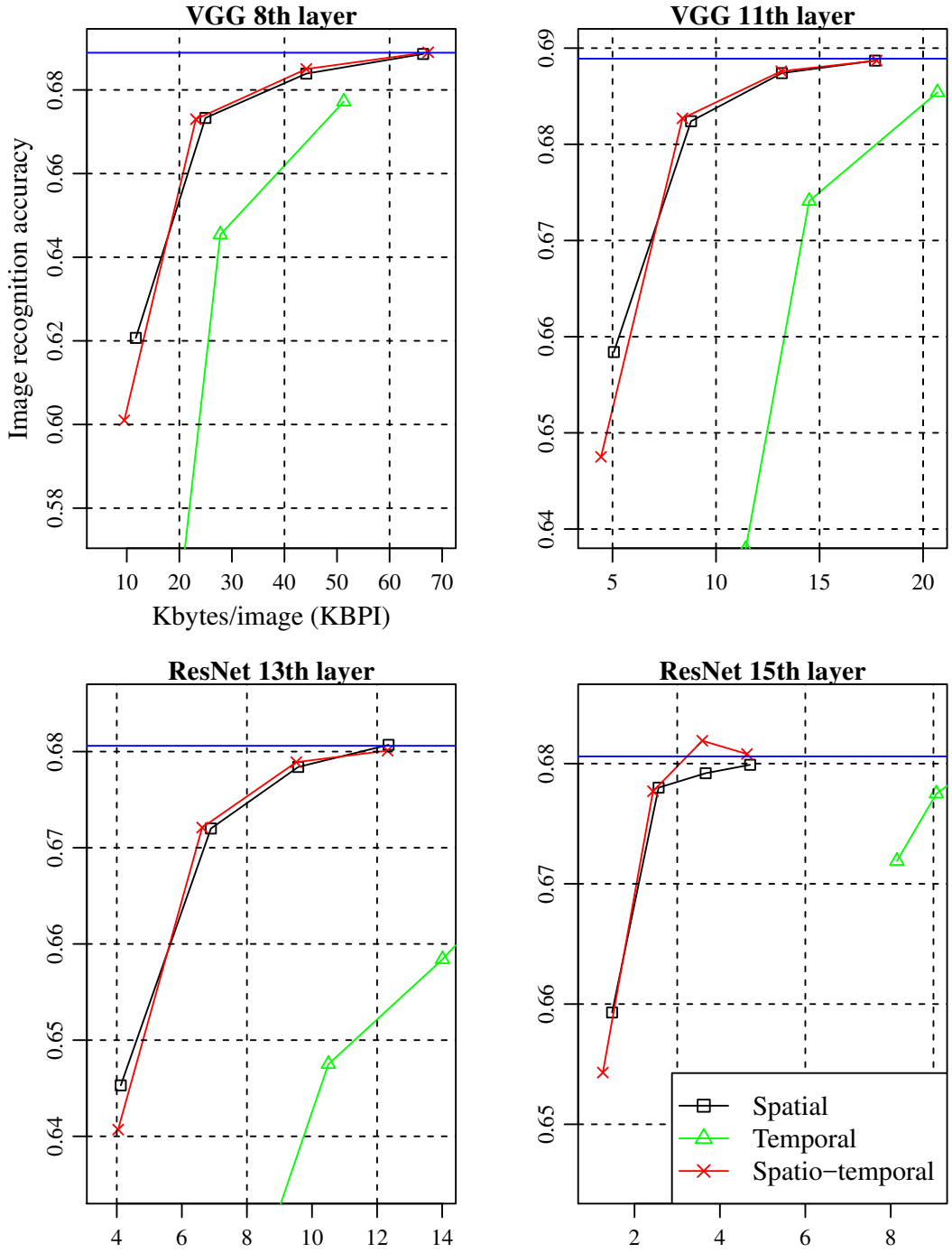

図 4.5: 各手法における，深層特徴を圧縮した際のビットレートと認識精度の関係．空間的配置法を黒線，時間的配置法を緑線，提案する時空間的配置法において配置順序探索アルゴリズムを用いた $f = 2$ の場合を赤線で示す．図中の青線は，DNN の本来の認識精度を示している．この本来の認識精度は表 4.1 の 32-bit の列の結果と同値のものである．

4.5.1 エントロピーに基づく予測残差信号の解析

本節では，まず初めに，予測残差信号の解析によって提案手法の有効性を検討する．予測残差は，フレーム内予測とフレーム間予測を利用して元の深層特徴の信号から冗長性を取り除いた結果として得られる信号である．ここでは，時空間的配置法と空間的配置法で配置した深層特徴を HEVC を用いてそれぞれ圧縮した際の予測残差信号のエントロピーを用いて評価を行った．HEVC が深層特徴から除去する冗長性が多ければ多いほど，予測残差はスパースな信号となる．スパースな信号であるほどそのエントロピーは小さくなるため，評価したエントロピーが小さい方が HEVC が深層特徴に内在する冗長性をより多く除去できていることを意味する．

検証では，ImageNet 2012 の検証画像から 20 枚の画像をランダムに選択し，実験に使用したすべての層とモデル (VGG-16 の第 8，11 層，ResNet-18 の第 13，15 層) から深層特徴を抽出した．空間的配置法と $f = 2$ の場合の提案手法における 20 個の深層特徴の予測残差のエントロピーを表 4.5 に示す．表 4.5 の最終行は各モデル · 階層 · 手法における平均エントロピー値を示している．この結果から，提案手法の平均エントロピーは，実験した全ての層とモデルにおいて，空間的配置法での平均エントロピーよりも小さい値を示していることが分かる．これは，時空間的配置法が，空間的配置法よりも冗長性を効果的に高めることに成功し，HEVC がその冗長性をフレーム内予測とフレーム間予測によって除去したため，残差信号がよりスパースでエントロピーの低い信号となったことを示唆している．この残差信号の解析結果は，提案手法のニアロスレス圧縮条件および非可逆圧縮条件下における圧縮効率の結果と一致しており，従来手法では十分に除去できなかった深層特徴に内在する冗長性を除去するという提案手法の目的が実現できていることをサポートする解析結果であると考えられる．

表 4.5 に示した数値を比較すると，深層特徴の予測残差のエントロピー値が入力画像や深層特徴を抽出する階層に依存して変化していることが分かる．このような現象がなぜ起きるのかを分析するために，深層特徴を可視化し，定性的な評価を行っていく．図 4.6 に，空間的配置法と提案手法の間でエントロピーの値の差が顕著であった入力画像とその深層特徴を示す．深層特徴は空間的に配置されたものを例示しており，分かりやすさのためにカラーマップの形で可視化している．また，各入力画像の下には，表 4.5 に対応付けられた番号とそれぞれの手法におけるエントロピー値を示している．左側の例 (a) と右側の例 (b) は，それぞれ VGG-16 の第 11 層と ResNet-18 の第 15 層から抽出した深層特徴を例示している．また，(a) と (b) のいずれも上側に空間的配置法のエントロピーが低い例，下側に時空間的配置法のエントロピーが低い例を示している．図に示す深層特徴の定性的な観察から，(a) と (b) の両方において，上側と下側の深層特徴は，それぞれ疎な反応と密な反応をする深層特徴であることが分かった．これは，それぞれの深層特徴の特

表 4.5: 空間的に配置された深層特徴と提案手法 ($f = 2$) で配置された深層特徴の HEVC による予測処理を行った後の残差信号のエントロピー値．各行は同一の入力画像における，各モデル · 階層 · 手法におけるエントロピー値を示し，一番下の行は各列の平均エントロピー値を示す．

	VGG 8th		VGG 11th		ResNet 13th		ResNet 15th	
No.	Spatial	Ours	Spatial	Ours	Spatial	Ours	Spatial	Ours
1	3.43	3.36	2.21	2.14	2.83	2.78	2.23	2.06
2	3.22	3.17	2.46	2.54	3.04	2.92	2.30	2.17
3	3.11	3.03	2.18	2.04	2.79	2.69	1.95	1.78
4	3.16	3.03	2.46	2.48	3.12	3.02	2.36	2.16
5	3.23	3.18	2.68	2.73	3.61	3.45	2.66	2.45
6	3.24	3.19	2.44	2.39	2.74	2.67	2.06	2.05
7	3.25	3.17	2.61	2.69	3.21	3.08	2.26	2.12
8	3.18	3.11	2.62	2.51	3.92	3.84	2.65	2.50
9	3.22	3.23	2.58	2.40	3.35	3.21	2.35	2.23
10	3.07	3.00	2.47	2.66	3.38	3.22	2.63	2.42
11	3.26	3.15	2.52	2.43	3.48	3.36	2.51	2.42
12	3.24	3.08	2.51	2.37	2.95	2.82	2.24	2.17
13	2.72	2.65	2.32	2.21	3.13	3.01	2.19	2.28
14	3.25	3.13	2.31	2.34	3.37	3.25	2.53	2.43
15	3.15	3.08	2.26	2.17	3.09	3.02	2.56	2.50
16	3.29	3.14	2.57	2.52	3.60	3.45	2.51	2.36
17	3.40	3.30	2.56	2.44	3.53	3.41	2.50	2.34
18	2.60	2.55	2.15	2.08	2.53	2.45	2.21	2.06
19	3.19	3.13	2.61	2.52	3.58	3.47	2.58	2.51
20	2.96	2.90	2.73	2.90	3.43	3.30	2.54	2.47
Average	3.16	**3.08**	2.46	**2.43**	3.23	**3.12**	2.39	**2.27**

定の水平方向での輝度変化 (赤い折れ線グラフ) からも見て取れる．深層特徴の反応が密になると，フレーム間予測の予測性能を向上させる参照ブロックを見つけやすくなると考えられるため，上記のような定性評価は妥当なものと考えられる．圧縮効率を評価する実験結果 (例えば，表 4.3) において，提案手法の効果はそれぞれのモデルや階層によって異なっていた．このような差異は，深層特徴の反応の疎である度合い，あるいは密である度合いに起因するものと考えられる．

さらに，図 4.6 を詳細に分析すると，入力と深層特徴の関係を定性的に評価することができる．反応が疎な深層特徴である上側の例の入力画像は，認識対象となるサルの顔等のみが大きく写った画像となっていることが分かる．一方，反応が密な深層特徴である下側の例の入力画像は，背景にぬいぐるみや家等，認識対象となる物体以外が写っている画像である．このような入力画像に対しては，DNN が認識対象以外にも映り込んでいる様々な物体に含まれる豊富なエッジやテクスチャに反応を示すため，密な深層特徴を出力すると考えられる．したがって，提案手法はエッジやテクスチャが豊富な画像が入力された場合の深層特徴の圧縮において，より高い圧縮効率を得られる可能性が高いと考えられる．

上記のような性質は，DNN を用いた画像認識アプリケーションの実用化という観点において提案手法が有効な手法であることを示唆している．例えば，自動運転車や自動運転ドローン，監視システムなどのユースケースでは，フロントエンドデバイス周囲の環境は常に変化しており，特定の認識対象物のみが写っている画像を撮像することは難しいと考えられる．そのため，入力画像は通常エッジやテクスチャが豊富で複雑な画像であり，提案手法が高い圧縮効率を得られるような入力であると言える．

4.5.2 探索アルゴリズムによる配置順序と最適な配置順序の圧縮効率の差分評価

4.3 および 4.4 節の評価実験の結果，Algorithm 1 によって得られた配置順序は，深層特徴の圧縮効率向上に寄与することが明らかになった．しかし，探索された配置順序は，あくまで隣接するフレーム間の MSE を最小化する “近似解” である．近似解ではない厳密に最適な配置順序は，非常に大きな計算時間を要するものの，取り得る全ての配置順序を総当たり的に試行し，最小の MSE を探索することで求められる．以下では，Algorithm 1 によって探索された配置順序と，上記のように総当たりで探索した最適な配置順序のそれぞれの圧縮効率を実験的に比較する．

今回の評価では，ImageNet 2012 の検証画像からランダムに 10 枚の画像を選択し，特徴マップの総数が $C = 512$ である VGG-16 の第 8 層から深層特徴を抽出した．しかしながら，最適な配置順序を総当たりで探索すると，計算量は特徴マップの総数に対して階乗オーダー ($\mathcal{O}(C!)$) となるため，全ての特徴マップで探索を行うと，現実的な時間内に計

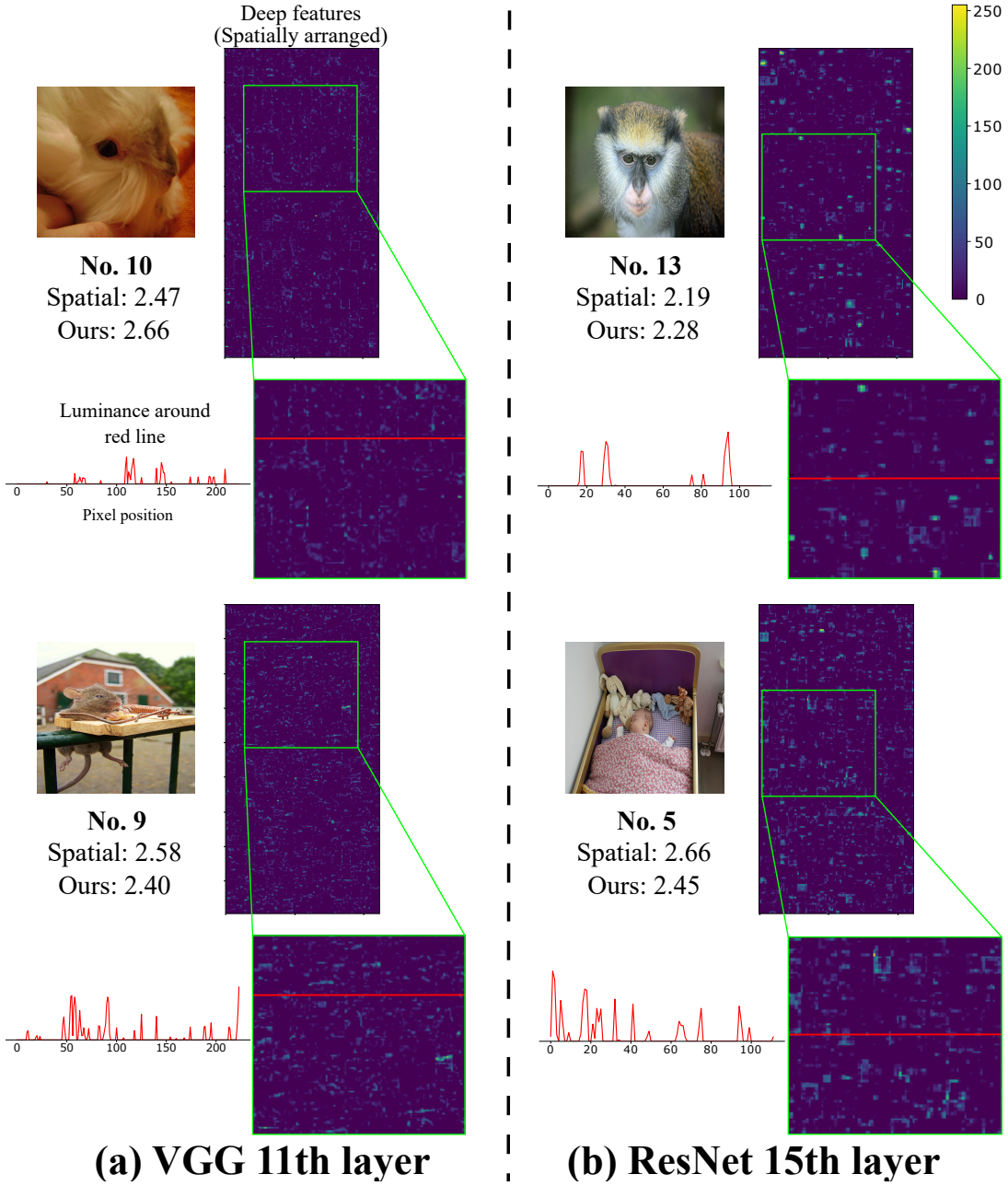

図 4.6: 空間的配置法と提案する時空間的配置法の間で予測残差信号のエントロピー値の差分が顕著だった深層特徴の例．例示する深層特徴は，空間的配置法によって配置されたものである．例示する各入力の下には，表 4.5 に関連付けられた番号とそれぞれの配置法におけるエントロピー値を記している．図中左側の例 (a) と右側の例 (b) は，それぞれ VGG の第 11 層目と ResNet の第 15 層目から抽出した深層特徴の例である．図中の緑枠は深層特徴の一部を拡大したものである．また，分かりやすさのため，深層特徴はカラーマップ状に表示している．

算が完了しない恐れがある．そこで，512 個の特徴マップから 16 個の特徴マップを抽出して計算に用いることとした．そして，Algorithm 1 および総当たり探索を用いて深層特徴の配置順序を探索した．

深層特徴を最適な配置順序またはアルゴリズムによって探索された配置順序を用いて時空間的に配置し，それぞれの圧縮効率を評価を行った．特徴マップの数を間引いている関係上，非可逆圧縮による精度への影響は評価できないため，ニアロスレス圧縮条件の下で圧縮処理を行った．最適な配置順序と探索された配置順序で配置された深層特徴のビットレートの平均値は，それぞれ 431.6 byte および 439.9 byte であった．したがって，最適な配置順序と比較して，提案するアルゴリズムによる近似解ではビットレートが約 1.92% 増加したことになる．しかしながら，最適な配置順序は入力ごとに探索する必要がある．さらに，その計算量は $\mathcal{O}(C!)$ であるため，C が大きいと探索に膨大な時間を要する．提案するアルゴリズムは，入力ごとに配置順序を探索する必要がなく，さらに，その計算量は二乗オーダー $(\mathcal{O}(C^2))$ である．したがって，提案する配置順序探索アルゴリズムは，実用性を維持しつつ，最適な配置順序の圧縮効率とほぼ同等の性能を実現できるものであると考えられる．なお，上記で示している総当たり探索や提案アルゴリズムの計算量は次節で具体的に導出する．

4.5.3 $\mathcal{O}$ 記法を用いたアルゴリズムの計算量の導出

ここでは，提案する配置順序探索アルゴリズムと総当たり探索の計算量をそれぞれ導出する．計算量はアルゴリズムの反復回数によって規定されるため，深層特徴の特徴マップの数 C に依存する．提案するアルゴリズムでは，1 枚目のフレームに配置されているインデックスと 2 枚目以降のフレームのインデックスで入れ替えを行うかどうかの評価を行う．したがって，その反復回数は 1 枚目のフレームに存在するインデックス数と 2 枚目以降のフレームに存在するインデックス数の積となる．1 枚目のフレームに存在するインデックスは $\frac{C}{f}$ であるため，提案するアルゴリズムの総反復回数は以下のように算出される:

$$\frac{C}{f} \times (C - \frac{C}{f}) = \frac{(f-1) \times C^2}{f^2}. \tag{4.4}$$

一方，総当り探索では，C 個のインデックスを全て各フレーム上の位置に配置するため，$C!$ 回の探索が通常は必要となる．しかし，本章で議論する総当り探索は，二種類の入れ替えに関して MSE 値が不変であるため，その分を除算する必要がある．一つ目は，フレーム内の配置箇所に関する不変性である．今回の総当たり探索では，隣接するフレーム間の MSE を最小化する配置順序を探索するため，隣接するフレームのインデックスが変わらなければ，フレーム内のどこにインデックスが配置されていても結果は変わらない．

表 4.6: ニアロスレス圧縮条件下における，空間的配置法と提案手法 ($f = 2$) で ME を使った場合と使わなかった場合の計 3 種類の手法の平均ビットレート．

	layer	Spatial	Ours w/o ME	Ours w/ ME
VGG-16	8th	161.9 KB	159.2 KB	**158.2** KB
	11th	34.64 KB	34.48 KB	**34.05** KB
ResNet-18	13th	23.24 KB	23.12 KB	**22.83** KB
	15th	10.71 KB	10.63 KB	**10.17** KB

したがって，各フレームにおける配置の組み合わせ数 $\left(\frac{C}{f}\right)!$ を除算する必要がある．二つ目は，フレーム順序の逆転に関する不変性である．フレーム順序が逆転した場合，隣接するインデックスの関係は変わらない．例えば，1→2→3→4，という順序で並んでいたインデックスをフレーム順序を逆転させて，4→3→2→1，としてもその隣接関係は変わらない．この組み合わせは 2 通り存在するため，2 で除算する必要がある．結果として，総当たり探索の総反復回数は以下のように算出される:

$$\frac{C!}{\left(\frac{C}{f}\right)! \times 2}. \tag{4.5}$$

上記の算出では C/f は整数である，つまり C が f で割り切れることを仮定している．また，フレームサイズ f は 2 から C までの任意の値を取れるが，表 4.3 と 4.4 の結果から，時空間的配置法では，f は通常 2 または 4 のような小さな値が適していると考えられる．したがって，提案するアルゴリズムの計算量は $\mathcal{O}$ 記法を用いると以下のように表すことができる:

$$(4.4) \in \mathcal{O}(C^2). \tag{4.6}$$

一方，総当たり探索の計算量は以下のように表せる:

$$(4.5) \in \mathcal{O}(C!). \tag{4.7}$$

4.5.4　動き予測が圧縮効率へ与える影響

前述の通り，提案する配置順序探索アルゴリズムでは，隣接フレーム間の MSE のみを考慮しているためフレーム間予測における動き予測 (Motion Estimate: ME) [131] を直接的には考慮していない．しかし，本研究において我々は，配置順序探索アルゴリズムによって求められた配置順序では，異なるフレームに類似した画像特徴に反応する特徴マップを割り当てる作用があるため，ME を直接考慮していないものの，時間的な冗長性を高

表 4.7: 配置順序探索アルゴリズムで探索された配置順序のビットレート．

	layer	$f = 2$	$f = 4$
VGG-16	8th	1.27 KB	1.64 KB
	11th	1.36 KB	1.53 KB
ResNet-18	13th	0.83 KB	1.16 KB
	15th	1.35 KB	1.63 KB

められる可能性がある．この点について検証を行うために，以下では，配置順序探索アルゴリズムがフレーム間予測の ME にどのような影響を与えるかを調査する．

ME に対するの配置順序探索アルゴリズムの効果を分析するために，ニアロスレス圧縮条件で ME を用いずに[*2]，時空間的に配置された深層特徴を圧縮した．表 4.6 は，空間的配置法，時空間的配置法の ME がある場合と無い場合の計 3 ケースにおける平均ビットレートを示している．表 4.6 に示す通り，提案する時空間的配置法において ME を用いた場合の平均ビットレートは，ME を用いない場合の平均ビットレートに比べて低い値となっていることが分かる．したがって，配置順序探索アルゴリズムは，フレーム間予測における ME を直接考慮していないにも関わらず，ME を用いて除去できる時間的な冗長性を高めることに成功していると言える．さらに，ME を使わない場合でも提案手法は，空間的配置法に比べて低いビットレートを示しており，優れた圧縮効率を実現できている．この結果から，配置順序探索アルゴリズムが，隣接するフレームの同じ位置に類似した画像特徴に反応する特徴マップを配置する効果も有していると考えられる．

4.5.5 探索された配置順序のビットレートへの影響

配置順序探索アルゴリズムを用いた提案手法においては，探索された配置順序をフロントエンドデバイスとクラウドサーバ間で共有する必要がある．4.3 節や 4.4 節で行った比較実験では，公平な比較のために，圧縮された深層特徴のビットレートを算出する際に，探索した配置順序のビットレートも含めて平均ビットレートを算出した．探索された配置順序はテキストファイルの形式で保持し，ZIP で可逆圧縮したもののビットレートを算出した．表 4.7 に，探索された配置順序のビットレートを示す．探索された配置順序のビットレートは高々 1.64 KB 程度であることが，表より見て取れる．さらに，配置順序は入力画像に依存して変化しないため，配置順序のクラウドサーバへの伝送は一度のみで良

[*2] 実装上は，Configuration ファイルには ME を用いないオプションは存在しない．そのため，HEVC テストモデル (HM) のソースコードにおいて，ME の探索範囲を 0 となるように書き換えることで ME を用いない圧縮を実現した．

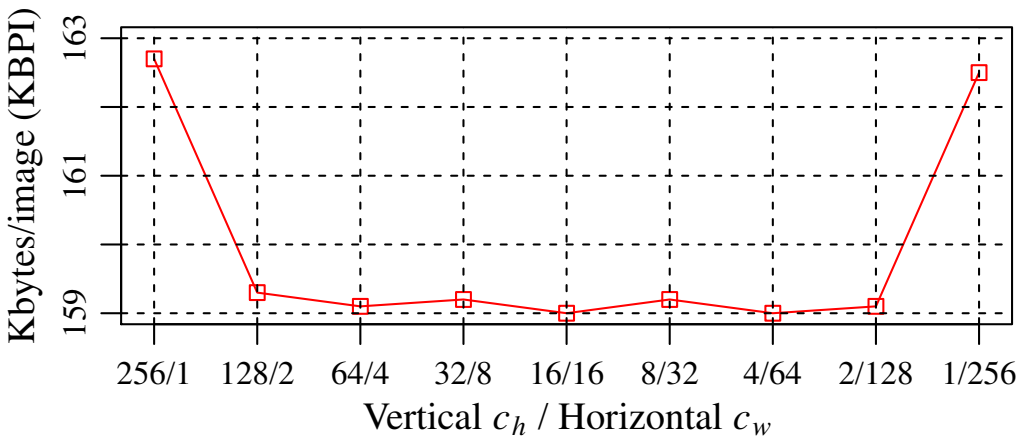

図 4.7: VGG の第 8 層から抽出した深層特徴において，時空間的配置法における c_h および c_w と圧縮時のビットレートの関係．縦軸はビットレート，横軸は縦 (c_h)/横 (c_w) に配置した特徴マップの数をそれぞれ示している．

い．したがって，1 万枚の検証画像を用いて平均ビットレートを算出する本章の検証実験では，配置順序のビットレートが与える影響はごくわずかであった．

もちろん，検証画像の画像枚数が少なければ，総ビットレートに占める配置順序のビットレートの割合が相対的に大きくなるため，配置順序が平均ビットレートを増加させる効果は高まる．しかし，定常的に動作する実用的な画像認識アプリケーションでは，認識する画像の総数は非常に大きくなると考えられる．また，探索された配置順序を事前にフロントエンドデバイスとクラウドサーバで共有できているのであれば，そもそも伝送する必要も無い．したがって，提案する配置順序探索アルゴリズムによって探索された配置順序は，ビットレートにはほとんど影響を及ぼさないと考えられる．

4.5.6　c_h と c_w がビットレートへ与える影響

上記までの全ての実験では，式 (4.2) を用いて，垂直方向に並べる特徴マップの数 c_h および水平方向に並べる特徴マップの数 c_w を決定している．しかし，c_h と c_w を決定する方法は，式 (4.2) しか存在しないわけではなく，式 (4.2) によるもの以外にも様々な c_h と c_w の組み合わせが存在する．ここでは，c_h と c_w の設定値が圧縮時のビットレートへどのような影響を与えるか調査する．

様々な c_h と c_w で時空間的に配置された深層特徴を，ニアロスレス圧縮条件の下で圧縮する．実験には VGG-16 の第 8 層の出力を深層特徴として使用し，フレーム枚数は $f = 2$ とした．また，配置順序探索アルゴリズムを使用しないで特徴マップに紐づくインデックスによる昇順で配置を行った．図 4.7 は，それぞれの c_h および c_w と対応

する圧縮時のビットレートを示している．図 4.7 に示す通り，極端なケース，すなわち $c_h/c_w = 256/1$ または $1/256$ の場合，提案手法で配置した深層特徴のビットレートは大幅に増加している．これは，垂直または水平方向の深層特徴の空間的なサイズが非常に小さなものになるために，空間的な冗長性を十分に高められないためであると考えられる．また，それ以外のケースでは，ビットレートの増減はほとんど無いことも明らかになった．これは，時空間的に配置された深層特徴は自然画像や自然映像とは異なり，c_h や c_w が多少変更されてもその統計的な性質は不変であることに起因すると考えられる．したがって，c_h と c_w が極端に小さい値とならなければ，式 (4.2) から算出した c_h と c_w 以外の値を用いても，時空間的配置法はほぼ最適な圧縮効率を得ることができる．このような特性は，提案手法は最適な c_h と c_w を探索する必要がなく，極端な値を避けるのみで良いことを示唆している．

4.6 本章のまとめ

本章では，協調型知能方式 [18–20] に基づく画像認識アプリケーションの高度化に向けて，フロントエンドデバイスに配置された DNN の出力である深層特徴を圧縮するための“時空間的配置法” を提案した．深層特徴は，互いに類似した反応を多数持つため，従来手法 [21, 99] では効率的な圧縮のために，深層特徴を空間的に配置して画像として圧縮することで，空間的な冗長性を除去していた．しかし，深層特徴を空間的な方向だけでなく時間的な方向に配置し，映像として圧縮した場合，空間的な冗長性のみならず，時間的な冗長性も圧縮に利用して，より効率的な圧縮が実現できる可能性がある．いくつかの先行研究 [21, 108] は，時間的な冗長性を利用するために，時間的配置法を提案したが，空間的な冗長性を十分に高めることができなかったため，空間的配置法を上回る圧縮効率は得られなかった．時空間的配置法は深層特徴を画像として空間的に配置しつつ，新たに提案する配置順序探索アルゴリズムを用いて映像として時間的な配置も行うことで，空間的および時間的な冗長性の両方を高めるような配置を行う．結果として，提案手法によって配置された深層特徴では，映像圧縮手法が空間的および時間的な冗長性を有効に活用し，高い圧縮効率を実現することが可能になった．

ImageNet 2012 の画像認識タスクを基に行った評価実験の結果，時空間的配置法は，ニアロスレス圧縮条件，非可逆圧縮条件のいずれの条件においても，従来手法と比較して高い圧縮効率を達成していることが分かった．この結果は，提案手法が深層特徴に内在する時空間的な冗長性を高め，映像圧縮手法が冗長性を除去することができたために，圧縮効率が向上することを示唆している．これは今後，深層特徴の圧縮効率をより向上させるため非常に重要な知見であると考えられる．さらに，本章では，提案手法がどのような場合に圧縮効率を高めるか，あるいは低下させるかを定性的に解析した．その結果，提案手法

は，エッジやテクスチャが豊富な画像を入力とした場合に冗長性を高め，高い圧縮効率を実現することが明らかになった．これは，従来の空間的配置法には見られなかった傾向である．

残る課題として，非可逆圧縮条件下で圧縮した場合，フレーム数を増やすと圧縮効率が低減することが挙げられる．これは，非可逆圧縮によって参照フレームに発生したアーチファクトによりフレーム間予測の予測精度が低下することに起因すると考えられる．フレーム間予測の予測精度を維持するために重要な深層特徴に対して低い QP 値を割り当てるような量子化制御を行うことで，この問題の解決の一助になると考えられる．これを実現するために，重要な深層特徴領域を推定する技術の開発に今後取り組む予定である．また，配置順序探索アルゴリズムの改良も検討している．現状のアルゴリズムは，時間的な冗長性のみを高める効果しか有さない．このアルゴリズムに対して，空間的な冗長性を高める効果も導入することができれば，より高い圧縮効率を得ることが可能になると考えられる．

第 5 章

結論

本論文は，近年目覚ましい精度向上を見せている深層学習を用いた画像認識アプリケーションの高度化を目指し，フロントエンドデバイスとクラウドサーバ間での効率的なデータ伝送を実現する情報源圧縮技術を研究した成果をまとめたものである．フロントエンドデバイスで動作する画像認識アプリケーションがクラウドサーバを利用するには，撮像した画像をそのまま伝送するクラウド型知能方式と DNN が抽出した特徴表現である深層特徴を伝送する協調型知能方式の 2 つの方式に大別される．これらの 2 つの方式では，いずれも，逐次撮像される大量の画像信号もしくはそこから抽出した深層特徴をクラウドサーバに伝送する必要がある．これらの大量のデータを効率良く伝送するために，情報源圧縮技術を用いて認識精度を保持しつつ通信に要するビットレートを低減させることは重要である．しかし，従来の情報源圧縮の枠組みでは入力は画像や映像を想定しており，また，DNN の認識精度を保持するという観点は導入されていない．したがって，本論文で目的とするような “深層学習を用いた画像認識を指向する情報源圧縮” とはなっていないという課題が存在した．そこで本論文では，既存の情報源圧縮技術に対して，DNN の精度保持 · 深層特徴の圧縮という観点を導入することで，従来の枠組みと比較して高い圧縮効率を得られるのではないかと考えた．ただし，DNN の精度保持 · 深層特徴の圧縮の観点を導入する，と言っても DNN の特徴抽出処理はブラックボックス化されており [28]，DNN の特性を具体的に把握することは容易ではない．本論文では，この問題に対処するために，近年著しく研究が進んでいる DNN の内部の特徴表現 (内部表現) に関する研究に着目した．DNN の内部表現に関する研究は，“DNN が抽出する特徴表現はなぜ上手く働くのか” といった DNN の挙動の理解を中心に考えて行われている研究であり，情報源圧縮のために行われている研究ではない．しかしながら，いくつかの研究成果は DNN の抽出する特徴表現に関する深い洞察を与えている．ちょうど，従来の動画像圧縮技術が人間の視覚特性に学んで，人間の視覚に沿った情報源圧縮へと成長したのと同様に，DNN

の画像認識のおける特性を学ぶことで，画像認識を指向する情報源圧縮技術を実現することができても不思議ではない．この着眼点のもと，DNN の精度保持 · 深層特徴の圧縮という観点を導入するための着想を内部表現に関する研究から得て，情報源圧縮技術の改良をする要素技術を開発することで問題に取り組んだ．

5.1 本論文のまとめ

第 1 章では，研究背景，従来の情報源圧縮の研究，および本研究の位置づけと方針について述べた．

第 2 章では，本研究の主題である深層学習による画像認識技術とその内部表現に関する研究動向の概要，およびその実応用のための通信方式，つまりクラウド型知能方式と協調型知能方式について述べ，本論文に関連する既存の情報源圧縮技術の概要と本論文で取り組む課題について説明した．

第 3 章では，クラウド型知能方式での，画像認識を指向した情報源圧縮技術について研究した．クラウド型知能方式では，撮像した画像信号をそのままクラウドに伝送する．したがって，この場合の画像認識を指向した情報源圧縮技術とは，画像信号の圧縮による DNN の精度劣化を抑制するための圧縮技術に他ならない．従来の情報源圧縮手法である動画像圧縮標準は，人間の鑑賞を前提に PSNR 等の客観画質や人間の主観画質で評価を行っており，DNN の認識精度を評価指標とした圧縮標準は現在のところ存在しない．ただし，圧縮標準としては存在しないものの，幾つかの先行研究は既存圧縮標準を改良することで，改良前と比較して精度保持に必要なビットレートの低減に成功している．しかし，それらの先行研究は DNN が抽出する画像特徴に関する知見を導入したものではなく“認識対象となるオブジェクトが劣化しなければ DNN の精度は保持できる”というプリミティブな知見に基づいているため，具体的なアプローチは量子化処理の改良のみに留まっている．量子化処理は，圧縮処理を担うエンコーダの内部構造に依存するため，異なるエンコーダには直接適用できない等の制約が存在する．そこで，第 3 章では，画像を圧縮する前に適切に変換する画像プレ変換手法を提案した．提案する画像プレ変換手法は，入力画像の信号の中で DNN の認識に対して重要な情報のみを保持し，それ以外の信号を圧縮時にビットレートが低くなるように変換する．提案手法は，人間が認識にあたって重視する画像特徴と DNN が重視する画像特徴が異なることを明らかにした内部表現に関する研究成果に着想を得ている．DNN が認識にあたって重視する情報以外の画像信号を積極的に削減することで，認識精度を保持しつつも大幅なビットレート低減が実現できると考えられる．具体的には，Encoder-Decoder 型の DNN モデルである SegNet [116] を画像プレ変換モデルとして採用し，DNN の画像認識精度を高めるための損失と画像を圧縮した際のビットレートを低減させる Total Variation 損失で学習を行った．提案手法を定

量的に評価するために，ImageNet 2012 の識別タスクで，提案手法による変換画像を圧縮する場合と原画像をそのまま圧縮する場合とで比較評価を行った．提案手法は，実験を行った JPEG, JPEG2000, HEVC, VVC の全ての圧縮標準で認識精度を保持しつつビットレート低減効果を示した．ビットレートの低減効果は，最大で 20.5% (HEVC)，最小でも 8.6% (JPEG) であった．また，第 3 章では，提案手法が変換画像に与える影響に関しても解析を行い，圧縮後の画像において認識に対して重要な信号を優先的に保持する働きや，フレーム内予測の予測残差を有効に低減させる働きを持つことが明らかになった．

第 4 章では，協調型知能方式での，画像認識を指向した情報源圧縮技術について研究した．協調型知能方式では，DNN を 2 つに分割し，フロントエンドデバイスとクラウドサーバにそれぞれを配置する．撮像された画像信号は，フロントエンドデバイスに配置した DNN に入力され，その出力である “深層特徴” をクラウドサーバに配置した DNN へ伝送 · 入力することで画像を認識する．DNN の抽出した特徴表現そのものである深層特徴は，従来の動画像圧縮標準や第 3 章で提案した手法が入力として想定している画像信号とは，異なる情報源である．したがって，高い圧縮効率を実現するためには，深層特徴がどのような性質を持った信号であるかを十分に把握し，冗長性を除去する必要がある．しかしながら，協調型知能方式における情報源圧縮の先行研究では，DNN が抽出する画像特徴に関する知見は導入されておらず，深層特徴の圧縮効率が高い手法をヒューリスティックに探索していたのが現状である．具体的には，図 4.2 (a) および (b) に示すように深層特徴を構成する特徴マップを空間的に配置し画像として圧縮する，もしくは，時間的に配置しビデオとして圧縮する手法を提案し，これらのうち圧縮効率が高かった空間的な配置法を採用していた．本論文では，深層学習の内部表現に関する研究のうち DNN がどのような入力の画像信号に反応を示して深層特徴を作り出しているか，という観点から行われている研究に着目した．それらの研究成果から，DNN の深い層では，多くのニューロンがかなり似通った画像特徴に反応を示すことが明らかになった．つまり，DNN の抽出する深層特徴を構成する特徴マップには反応する画像特徴が類似しているという冗長性が存在している．このような，類似する画像特徴に反応する特徴マップ間に存在する冗長性は画像圧縮手法で採用されているフレーム内予測では除去することが難しい．そこで，この冗長性を除去して圧縮効率を高めるために特徴マップを空間的に複数枚の画像として並べ，それを映像として時間的に配置する “時空間的配置法” (図 4.2 (c)) を提案した．時空間的配置法は，空間的配置法で除去することができなかった冗長性を除去することが期待され，結果として高い圧縮効率を実現できると考えられる．また，この時空間的配置法は，深層特徴を構成する特徴マップを配置する順序によって，除去できる冗長性が変化する．第 4 章では，最大限に深層特徴に内在する冗長性を除去するための配置順序を探索するアルゴリズムも提案した．ImageNet 2012 の識別タスクを基に行った評価実験の結果，提案する時空間的配置法は，従来手法と比較して高い圧縮効率 (非可逆

圧縮条件において 1.50% から 4.98%) を達成していることが分かった．さらに，提案手法がどのような場合に圧縮効率を高めるか，あるいは低下させるかを定性的に解析した結果，提案手法は，エッジやテクスチャが豊富な画像を入力とした場合に冗長性を高め，高い圧縮効率を実現することが明らかになった．これは，従来の空間的配置法には見られなかった傾向である．

5.2 今後の課題

本論文では，フロントエンドデバイスで動作する画像認識アプリケーションの高度化のために，DNN による画像認識を指向した情報源圧縮技術を DNN の内部表現に関する研究を基に開発する方法論を提案した．本論文で述べたような情報源圧縮技術の必要性は，例えば 5G (第 5 世代移動通信システム) のようなブロードバンドインターネット接続の実現によって，一見不要になっているようにも思える．しかし，画像認識アプリケーションは全国あるいは世界中津々浦々で利用される可能性があり，十分な帯域のインターネット接続が用意されていないケースも多く考えられる．そのような状況でも，画像認識アプリケーションの恩恵に預かるためには本論文で述べたような研究を続け，より高い圧縮効率を実現することが重要である．上記のような観点に鑑みて，各要素技術において圧縮効率を高めるために解決すべき課題を以下で述べる．

第 3 章では，SegNet の学習における圧縮時のビットレートを低減させる損失として Total Variation を採用したが，この損失の場合，圧縮標準によってビットレート低減効果に差があることが明らかになった (表 3.3 参照)．この検証結果に基づき，圧縮標準によらず最大のビットレート低減効果を獲得できる損失を開発する必要があると考えられる．また，画像プレ変換によるアプローチのみを採用した結果として，最大 20.5% のビットレート低減効果を得たが，エンコーダの処理を改良する手法と組み合わせることによって，より大きな圧縮効率の向上も見込めると考えられる．提案手法と相性の良いエンコーダの内部処理に関する改良手法の開発も今後の課題である．

第 4 章で提案した空間的配置法に関しては，第 4 章で述べた今後の改良手法の他に深層特徴を圧縮する全く新しい情報源圧縮の開発も検討の余地がある．深層学習の内部表現に関する研究成果に着想を得た提案手法は，5% 程度のビットレート低減効果を得た．これは，昨今の動画像圧縮標準の新規ツールと比較すると高い圧縮効率の改善幅であると言える [25, 82] が，第 3 章の提案手法と比較するとその改善幅は小さい．これは，圧縮の対象とする情報源が深層特徴であるのにも関わらず動画像圧縮標準を圧縮処理に利用しているためであると考えられる．深層特徴に適した新たな圧縮アルゴリズムを開発することでより高い圧縮効率を得られると期待される．無論，その際にも本論文で提唱している深層学習の内部表現に基づくアプローチの有効性は変わらないものであろう．

5.3 むすび

深層学習による画像認識は，1980 年代に始まった Fukushima の Neocognitron に関する研究をルーツに持っている．これらの研究は，人間等の哺乳類が持つ視覚という優れたパターン認識装置の原理を模倣し工学的に応用することを主な動機に，画像認識モデルの改良を進めてきた．深層学習の発展によって，一部の画像認識タスクでは人間の認識精度を凌駕し，ついに工学的な実用に耐えうる精度に到達しつつある．しかし，認識精度が高ければすぐに実用化が可能になるというものではない．本論文で扱った題材のように，様々な場所で人間の代わりとなるような画像認識アプリケーションを実現するためには，その通信方式とセットになる情報源圧縮手法が必要である．本論文で行った研究は，哺乳類の視覚の機構を模倣した工学的なモデルを実社会に適用させようとする深層学習による画像認識という壮大な研究分野において，情報源圧縮という観点から日本あるいは世界中の様々な場所で画像認識技術を利用できるようにするための研究に他ならない．本論文で示した研究成果および内部表現に関する研究成果に基づく研究指針が，今後の深層学習による画像認識を指向した情報源圧縮の研究開発，および実用化の一助になり，引いては深層学習に基づく画像認識の恩恵を世の中の誰もが受けられるような世界の実現に繋がると信じている．

参考文献

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

[7] G. Huang, L. Liu, Z. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, 1999.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, 2004.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR), 2016.

[11] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NIPS) Deep Learning and Representation Learning Workshop*, 2015.

[14] J. Luo, J. Wu, and W. Lin. Thinet: A filter level pruning method for deep neural network compression. In *International Conference on Computer Vision (ICCV)*, 2017.

[15] Amazon web services. https://aws.amazon.com/jp/, [Online; accessed 10-March-2022].

[16] Google cloud platform. https://console.cloud.google.com/?hl=ja, [Online; accessed 10-March-2022].

[17] Microsoft azure. https://azure.microsoft.com/ja-jp/, [Online; accessed 10-March-2022].

[18] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *SIGARCH Computer Architecture News*, Vol. 45, No. 1, pp. 615–629, 2017.

[19] P. M. Grulich and F. Nawab. Collaborative edge and cloud neural networks for real-time video processing. In *Proceedings of the Very Large Data Base (VLDB) Endowment*, 2018.

[20] A. E. Eshratifar, M. S. Abrishami, and M. Pedram. Jointdnn: an efficient training and inference engine for intelligent mobile cloud computing services. *IEEE Transactions on Mobile Computing*, Vol. 20, No. 2, pp. 565–576, 2021.

[21] H. Choi and I. V. Bajić. Deep feature compression for collaborative object detection. In *IEEE International Conference on Image Processing (ICIP)*, 2018.

[22] G. K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, Vol. 38, No. 1, pp. xviii–xxxiv, 1992.

[23] A. Skodras, C. Christopoulos, and T. Ebrahimi. The JPEG 2000 still image compression standard. *IEEE Signal Processing Magazine*, Vol. 18, No. 5, pp. 36–58, 2001.

[24] G. J. Sullivan, J.-R. Ohm, W. J. Han, and T. Wiegand. Overview of the High

Efficiency Video Coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 22, No. 12, pp. 1649–1668, 2012.

[25] B. Bross, J. Chen, S. Liu, and Y. Wang. Versatile video coding (draft 10). *Joint Video Experts Team (JVET)-S2001*, 2020.

[26] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 1948.

[27] T. Onishi, T. Sano, Y. Nishida, K. Yokohari, J. Su, K. Nakamura, K. Nitta, K. Kawashima, J. Okamoto, N. Ono, R. Kusaba, A. Sagata, H. Iwasaki, M. Ikeda, and A. Shimizu. Single-chip 4k 60fps 4:2:2 hevc video encoder lsi with 8k scalability. In *Symposium on VLSI Circuits*, 2015.

[28] V. Buhrmester, D. Münch, and M. Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv preprint, arXiv:1911.12116*, 2019.

[29] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2018.

[30] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

[31] S. Suzuki and H. Shouno. A study on visual interpretation of network in network. In *International Joint Conference on Neural Networks (IJCNN)*, 2017.

[32] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, Vol. 1, No. 4, pp. 541–551, 1989.

[33] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998.

[34] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, Vol. 36, No. 4, pp. 193–202, 1980.

[35] K. Fukushima. Neocognitron: A Hierahical Neural Network Capable of Visual Pattern Recognition. *Neural Networks*, Vol. 1, pp. 119–130, 1988.

[36] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, Vol. 106, No. 1, pp. 106–154, 1962.

[37] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, Vol. 323, pp. 533–536, 1986.

[39] S. Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, Vol. EC-16, No. 3, pp. 299–307, 1967.

[40] L. Deng and D. Yu. Deep learning: Methods and applications. Technical Report MSR-TR-2014-21, Microsoft Research, 2014.

[41] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[42] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv preprint, arXiv:1606.00915*, 2016.

[43] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[44] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *International Conference on Computer Vision (ICCV)*, 2017.

[45] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. In *International Conference on Machine Learning (ICML) Deep Learning Workshop*, 2015.

[46] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[47] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint, arXiv:1606.08415*, 2016.

[48] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[50] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, 2015.

[51] P. Baldi and P. J. Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems (NIPS)*. 2013.

[52] K. Hara, D. Saito, and H. Shouno. Analysis of function of rectified linear unit used in deep learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2015.

[53] R. Karakida, M. Okada, and S. Amari. Dynamical analysis of contrastive divergence learning: Restricted boltzmann machines with gaussian visible units. *Neural Networks*, Vol. 79, pp. 78–87, 2016.

[54] A. Camuto, M. Willetts, U. Şimşekli, S. Roberts, and C. Holmes. Explicit regularisation in gaussian noise injections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[55] L. Yuan, F. EH Tay, G. Li, T. Wang, and J. Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[56] G. F. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[57] X. Pan and V. Srikumar. Expressiveness of rectifier networks. In *International Conference on Machine Learning (ICML)*, 2016.

[58] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[59] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Machine Learning (ICML)*, 2012.

[60] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, Vol. 111, No. 23, p. 8619–8624, 2014.

[61] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in*

Databases (ECML PKDD), 2013.

[62] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[63] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[64] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Mądry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

[65] J. Jo and Y. Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint, arXiv:1711.11561*, 2017.

[66] S. Dodge and L. Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *International Conference on Computer Communication and Networks (ICCCN)*, 2017.

[67] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[68] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Mądry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[69] R. Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, 2019.

[70] Y. Hamano and H. Shouno. Analysis of texture representation in convolution neural network using wavelet based joint statistics. In *International Conference on Neural Information Processing (ICONIP)*, 2020.

[71] H. Wang, X. Wu, Z. Huang, and E. P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[72] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[73] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint, arXiv:1312.6034*, 2013.

[74] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR) Workshop*, 2015.

[75] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. In *International Conference on Machine Learning (ICML) workshop on visualization for deep learning*, 2017.

[76] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[77] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi. Automatic attribute discovery with neural activations. *arXiv preprint, arXiv:1607.07262*, 2016.

[78] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, 2017.

[79] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[80] A. Dosovitskiy and T. Brox. Inverting convolutional networks with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[81] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[82] B. Bross, J. Chen, and S. Liu. Versatile video coding (draft 2). *Joint Video Experts Team (JVET)-K1001*, 2018.

[83] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 10, pp. 3736–3764, 2021.

[84] K. R. Rao and P. Yip. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press Professional, Inc., 1990.

[85] J. J. Benedetto and M. W. Frazier (ed.). *Wavelets: mathematics and applications*. CRC Press, 1993.

[86] M. W. Marcellin, M. A. Lepley, A. Bilgin, T. Flohr, T. T. Chinen, and J. H.

Kasner. An overview of quantization in JPEG 2000. *Signal Processing: Image Communication*, Vol. 17, pp. 73–84, 2002.

[87] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, Vol. 40, No. 9, pp. 1098–1101, 1952.

[88] D. Salomon. *Data Compression: The Complete Reference*. Springer Verlag, second edition, 2000.

[89] J. Lainema, F. Bossen, J. Min, and K. Ugur. Intra coding of the HEVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 22, pp. 1792–1801, 2012.

[90] B. Bross, P. Helle, H. Lakshman, and K. Ugur. Inter-picture prediction in hevc. In *High Efficiency Video Coding (HEVC): Algorithms and Architectures*, pp. 113–140. Springer, Cham, 2014.

[91] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao. Dvc: An end-to-end deep video compression framework. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[92] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[93] H. Choi and I. V. Bajić. High efficiency compression for object detection. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[94] L. Galteri, M. Bertini, L. Seidenari, and A. D. Bimbo. Video compression for object detection algorithms. In *International Conference on Pattern Recognition (ICPR)*, 2018.

[95] B. Li, H. Li, L. Li, and J. Zhang. λ domain rate control algorithm for high efficiency video coding. *IEEE Transactions on Image Processing*, Vol. 23, No. 9, pp. 3841–3854, 2014.

[96] J. Gailly and M. Adler. Gzip. http://www.gzip.org/, [Online; accessed 10-March-2022].

[97] Z. Chen, W. Lin, S. Wang, L. Duan, and A. C. Kot. Intermediate deep feature compression: the next battlefield of intelligent sensing. *arXiv preprint, arXiv:1809.06196*, 2018.

[98] D. Flynn, D. Marpe, M. Naccari, T. Nguyen, C. Rosewarne, K. Sharman, J. Sole, and J. Xu. Overview of the range extensions for the hevc standard: Tools, profiles, and performance. *IEEE Transactions on Circuits and Systems for*

Video Technology, Vol. 26, No. 1, pp. 4–19, 2016.

[99] H. Choi and I. V. Bajić. Near-lossless deep feature compression for collaborative intelligence. In *International Workshop on Multimedia Signal Processing (MMSP)*, 2018.

[100] A. E. Eshratifar, A. Esmaili, and M. Pedram. Bottlenet: A deep learning architecture for intelligent mobile cloud computing services. In *International Symposium on Low Power Electronics and Design (ISLPED)*, 2019.

[101] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, Vol. 313, No. 5786, pp. 504–507, 2006.

[102] S. R. Alvar and I. V. Bajić. Multi-task learning with compressible features for collaborative intelligence. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[103] H. Choi, R. A. Cohen, and I. V. Bajić. Back-and-forth prediction for deep tensor compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[104] S. R. Alvar and I. V. Bajić. Bit allocation for multi-task collaborative intelligence. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[105] S. R. Alvar and I. V. Bajić. Pareto-optimal bit allocation for collaborative intelligence. *arXiv preprint, arXiv:2009.12430*, 2020.

[106] ISO/IEC. Draft call for evidence for video coding for machines. ISO/IEC JTC1/SC29/WG11 MPEG2018/w19077, 2020.

[107] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao. Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *IEEE Transactions on Image Processing*, Vol. 29, pp. 8680–8695, 2020.

[108] Z. Chen, L. Duan, S. Wang, W. Lin, and A. C. Kot. Data representation in hybrid coding framework for feature maps compression. In *IEEE International Conference on Image Processing (ICIP)*, 2020.

[109] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[110] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint, arXiv:1611.05760*, 2016.

[111] Y. Zhou, S. Song, and N. Cheung. On classification of distorted images with

deep convolutional neural networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[112] S. Dodge and L. Karam. Quality robust mixtures of deep neural networks. *IEEE Transactions on Image Processing*, Vol. 27, No. 11, pp. 5553–5562, 2018.

[113] S. Palacio, J. Folz, J. Hees, F. Raue, D. Borth, and A. Dengel. What do deep networks like to see? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[114] T. R. Shaham and T. Michaeli. Deformation aware image compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[115] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint, arXiv:1505.07293*, 2015.

[116] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 12, pp. 2481–2495, 2017.

[117] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[118] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, Vol. 60, No. 1, pp. 259 – 268, 1992.

[119] V. Le Guen. Cartoon + Texture Image Decomposition by the TV-L1 Model. *Image Processing On Line*, Vol. 4, pp. 204–219, 2014.

[120] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[121] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[122] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[123] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015.

[124] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast fea-

ture embedding. *arXiv preprint, arXiv:1408.5093*, 2014.

[125] G. Bjøntegaard. Calculation of average PSNR differences between RD-Curves. *ITU-T Video Coding Experts Group (VCEG)-M33*, 2001.

[126] J. Ström, K. Andersson, R. Sjöberg, F. Bossen, G. Sullivan, and J.-R. Ohm. Summary information on bd-rate experiment evaluation practices. *Joint Video Experts Team (JVET)-Q2016*, 2020.

[127] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. 2008.

[128] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot. Lossy intermediate deep learning feature compression and evaluation. In *ACM International Conference on Multimedia (MM)*, 2019.

[129] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot. Toward intelligent sensing: Intermediate deep feature compression. *IEEE Transactions on Image Processing*, Vol. 29, pp. 2230–2243, 2019.

[130] Y. Bandoh, K. Hayase, S. Takamura, K. Kamikura, and Y. Yashima. Theoretical modeling of inter-frame prediction error for high frame-rate video signal. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E91.A, No. 3, pp. 730–739, 2008.

[131] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, and F. Wu. Efficient parallel framework for hevc motion estimation on many-core processors. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, No. 12, pp. 2077–2089, 2014.

[132] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[133] D. Flynn. Common test conditions and software reference configurations for hevc range extensions. *Joint Collaborative Team on Video Coding (JCT-VC)-L1100*, 2013.

謝辞

まずはじめに，本論文の主査であり，日頃から多くのご指導とご鞭撻を頂いた電気通信大学 大学院 情報理工学研究科 情報学専攻 庄野逸教授に深く感謝いたします．庄野研究室の門戸を叩いた学部 4 年の頃から現在に至るまで，庄野教授には研究の面白さ，神経回路モデルの重要性等を幅広くご指導いただきました．庄野教授の存在が無ければ筆者が研究者としてのキャリアを歩むことも無かったでしょう．今後ともご指導ご鞭撻のほどよろしくお願いいたします．ご多忙の中，審査委員を務めていただいた電気通信大学の柳井啓司教授，羽田陽一教授，高橋裕樹准教授，および工学院大学の木全英明教授に感謝いたします．審査委員の先生方からは，本論文の内容や構成に関して有益な意見を多数頂戴しました．特に高橋准教授におかれましては，審査以外にも，講義履修に関しまして大変お世話になりました．本当にありがとうございました．また，筆者の研究活動を全般にわたってバックアップしてくださった庄野研究室のメンバーの皆様に感謝いたします．

本論文の一部は，筆者が日本電信電話株式会社 (NTT) で行った研究の成果をまとめたものです．NTT メディアインテリジェンス研究所および後継組織のコンピュータアンドデータサイエンス研究所におきまして，数多くの方々にご指導とご援助を賜りました．特にお世話になった方々をここに記し，深い感謝の意を表します．NTT における研究で，特に多くのご指導を頂いたのが高木基宏博士と武田翔一郎博士です．高木博士は筆者の OJT における指導者を務めていただき，筆者の企業研究者としての礎となる重要な教えを幾度となく賜りました．武田博士には，ご多忙にも関わらず，本研究を進めるに当たって非常に有益なご議論を幾度となく賜りました．また，筆者の歴代の上長 (グループリーダ) である NTT テクノクロス株式会社 清水淳氏，工学院大学 木全英明教授，日和崎祐介博士，澤田雅人氏には，本研究へのご理解を頂き，研究遂行の機会を与えていただきました．特に木全教授には本論文の審査委員も引き受けていただき，多大なご尽力を頂きました．NTT コミュニケーション科学基礎研究所の木村昭悟博士には筆者の OJT の技術アドバイザを引き受けていただき，数多くの有益なご議論を賜りました．筆者の所属した NTT メディアインテリジェンス研究所 画像メディアプロジェクト，環境情報処理プロジェクト，および NTT コンピュータアンドデータサイエンス研究所 次世代 AI 研究

プロジェクトの皆様には，本研究の遂行に際して技術面のみならず精神面からも多大なサポートを頂きました．

最後に，末尾にはなりますが，今日に至るまで筆者を育て，教育の機会を与えてくれた，父 秀典，母 美保子を始めとする親族，常に温かく励ましてくれた友人，そして，日頃から研究活動を支えてくれた妻 裕佳に心からの感謝の意を表し，本論文の結びといたします．

研究業績

本論文に関する研究業績

学術論文誌

1. 鈴木聡志，高木基宏，早瀬和也，武田翔一郎，木全英明，“圧縮による画像認識の精度劣化を抑制する画像プレ変換とその解析”，電子情報通信学会論文誌 D, 103(5), pp.483-495, 2020.

2. S. Suzuki, S. Takeda, M. Takagi, R. Tanida, H. Kimata, and H. Shouno, “Deep Feature Compression using Spatio-Temporal Arrangement toward Collaborative Intelligent World”, *IEEE Transactions on Circuits and Systems for Video Technology*, soon-to-be-published, 2021.

査読付き国際会議

1. S. Suzuki, and H. Shouno, “A Study on Visual Interpretation of Network in Network”, In *International Joint Conference on Neural Networks (IJCNN)*, 2017.

2. S. Suzuki, M. Takagi, K. Hayase, T. Onishi, and A. Shimizu, “Image Pre-transformation for Recognition-aware Image Compression”, In *IEEE International Conference on Image Processing (ICIP)*, 2019.

3. S. Suzuki, M. Takagi, S. Takeda, R. Tanida, and H. Kimata, “Deep Feature Compression with Spatio-Temporal Arranging for Collaborative Intelligence”, In *IEEE International Conference on Image Processing (ICIP)*, 2020.

研究会 · シンポジウム等

1. 鈴木聡志，庄野逸，“Network In Network の視覚システムとしての妥当性について ～方位選択性マップに関する観点から～”，電子情報通信学会 情報論的学習理論と機械学習研究会，2016.

2. 鈴木聡志， 高木基宏， 早瀬和也， 大西隆之， 清水淳，“高圧縮時の認識誤差を抑制する画像プレ変換手法”，電子情報通信学会 画像工学研究会，2019.

3. 鈴木聡志，高木基宏，渡邊真由子，谷田隆一，木全英明，“フレーム間予測を活用した深層特徴圧縮に関する一検討”, 画像符号化シンポジウム，2019.

その他の研究業績

学術論文誌

1. K. Hara, D. Saitoh, S. Suzuki, T. Kondou, and H. Shouno, “Analysis of Conventional Dropout and its Application to Group Dropout”，情報処理学会論文誌数理モデル化と応用 (TOM), 10(2), pp.25-32，2017.

2. S. Suzuki, and H. Shouno, “Support Vector Machine Histogram: New Analysis and Architecture Design Method of Deep Convolutional Neural Network”, *Neural Processing Letters*, 47, pp.767-782, 2017.

査読付き国際会議

1. H. Shouno, S. Suzuki, and S. Kido, “A Transfer Learning Method with Deep Convolutional Neural Network for Diffuse Lung Disease Classification”, In *International Conference on Neural Information Processing (ICONIP)*, 2015.

2. S. Suzuki, N. Iida, H. Shouno, and S. Kido, “Architecture Design of Deep Convolutional Neural Network for Diffuse Lung Disease Using Representation Separation Information”, In *International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, 2016.

3. S. Suzuki, and H. Shouno, “An Architecture Design Method of Deep Convolutional Neural Network”, In *International Conference on Neural Information Processing (ICONIP)*, 2016.

4. K. Hara, D. Saitoh, T. Kondo, S. Suzuki, and H. Shouno, “Group dropout inspired by ensemble learning”, In *International Conference on Neural Information Processing (ICONIP)*, 2016.

5. A. Suzuki, S. Suzuki, S. Kido, and H. Shouno, “A 2-staged transfer learning method with deep convolutional neural network for diffuse lung disease analysis”, In *International Forum on Medical Imaging in Asia (IFMIA)*, 2017.

6. S. Suzuki, S. Takeda, R. Tanida, H. Kimata, and H. Shouno, “Knowledge Transferred Fine-tuning for Anti-aliased Convolutional Neural Network in Data-limited Situation”, In *IEEE International Conference on Image Processing (ICIP)*, 2021.

7. H. Higuchi, S. Suzuki, and H. Shouno, “Measuring Shift-invariance of Convolutional Neural Network with a Probability-incorporated Metric”, In *International Conference on Neural Information Processing (ICONIP)*, 2021.