Paper

Masked Window-based Attentionを用いた深層学習によるRGBA画像符号化

稲津慶紀，木全英明

透明性を確保するためのアルファチャンネルを含む抽象的なRGBA画像は、実世界のアプリケーションで一般的である。従来のRGBA圧縮法は、RGBとアルファチャネルの両方に同じ方法を適用しているが、特性が異なるため、最適でない結果を導く可能性がある。本論文では、RGB信号とアルファチャエネルに適した注意モジュールを個別に導入したディープニューラルネットワークを提案する。提案手法は、RGB信号とαチャンネルの2つのネットワークから構成され、それぞれに適切な注意モジュールが適用される。特に、αチャンネルのマスクされていない領域に焦点を当てた新しい注意モジュールが適用される。評価では、提案手法を、入力層と出力層を3チャンネルから4チャンネルに拡張した単純なディープニューラルネットワークと、古典的なRGBA画像圧縮手法と比較する。

キーワード：画像符号化，深層学習，アルファチャネル，RGBA，masked window-based attention

1. はじめに

画像は様々なアプリケーションで使用されており、画像圧縮は基本的な技術であり、性能向上のための多くの研究が進行中である。近年、ディープラーニングを用いた画像圧縮が研究されている。Ballélaらは変分オートエンコーダ型CNNベースの手法[1]を研究し、ZouらはSwin Transformer[3]を含む変換器ベースの手法[2]を研究している。さらに、LiuらはCNNとSwin Transformerを組み合わせたネットワーク[4]を研究している。これらの研究により、ディープラーニングを用いない古典的な手法を凌駕する画像圧縮性能が示された。しかし、これらの研究の範囲はRGB画像に限定されている。

このようなRGB信号の研究は数多くあるが、画像編集などの実用的なアプリケーションで一般的に使用されているため、透明性を確保するためにαチャンネルを含むRGBA画像の効率的な圧縮符号化が必要である。RGBA画像の古典的な圧縮方法の例としては、ロスレス圧縮を用いたPNG[5]と、HEVCビデオエンコーディング[7]のイントラフレームを用いたBPG[6]がある。

これらの方法では、RGB 信号に対して直接圧縮法を適用することで、アルファチャンネルを単一チャンネルの色信号として圧縮する。しかし、αチャンネルとRGBカラー信号では信号の特性が異なるため、古典的な圧縮手法では高い圧縮性能が得られない可能性がある。例えば、我々の計算によると、本論文のセクション4で使用した一般的な画像圧縮で使用したKodakデータセットに手動でアルファチャンネルを追加したRGBA画像データセットでは、自己相関係数はRGBチャンネルで約0.886、アルファチャンネルで約0.973であることがわかった。また、RGBA画像データセットP3M-10kを搭載した評価データセットP3M-500-NPは、人間の画像に着目し、RGBチャンネルで平均約0.956、αチャンネルで平均約0.995となった。なお、自己相関係数は、x、y方向に1画素ずつずらして測定している。P3M-10kの自己相関係数の差がKodakデータセットよりも小さい理由は、P3M-10kが人の写真に焦点を当てているためであり、肌や衣服のような滑らかな色遷移を特徴とすることが多いからである。これらの観察から、アルファチャンネル信号はRGB信号よりも類似性が高いことが示唆される。これらの事実は、αチャンネル信号がRGB信号よりも類似性が高いことを意味する。そこで、RGB信号とαチャンネルのサブネットワークで構成されるネットワーク全体を最適化することで、全体的な圧縮効率を向上させる学習ベースのアプローチを研究する。各サブネットワークは各信号の特徴を学習する。

Received September 18, 2024; Revised November 27, 2024; Accepted December 19, 2024
†Graduate School of Engineering, Kogakuin University
(Tokyo, Japan)

Paper

Deep Learning-based RGBA Image Compression with Masked Window-based Attention

Yoshiki Inazu† and Hideaki Kimata†

Abstract RGBA image that includes an alpha channel for transparency is common in real-world applications. Traditional RGBA compression methods apply the same methods to both RGB and alpha channel, but potentially leading to suboptimal results due to their different characteristics. This paper proposes a deep neural network that introduces attention modules individually suitable for RGB signals and alpha channel. The proposed method consists of two networks, one for the RGB signal and one for the alpha channel, with an appropriate attention module applied in each. In particular, a new attention module that focuses on the unmasked regions of the alpha channel is applied. In the evaluation, the proposed method is compared with a simple deep neural network with input and output layers extended from three to four channels and classical RGBA image compression methods.

Keywords: image compression, deep learning, alpha channel, RGBA, masked window-based attention.

1. Introduction

Images are used in a variety of applications, and image compression is a fundamental technology, and a large amount of research for improving the performance is ongoing. In recent years, deep learning-based image compression has been studied. Ballé et al. have studied the variational autoencoder type CNN-based method [1] and Zou et al. have researched the transformer-based method [2], which include Swin Transformer [3]. Moreover, Liu et al. have studied the network [4] combining CNN and Swin Transformer. These studies have shown image compression performance that outperforms classical methods not using deep learning. However, the scope of these studies is limited to RGB images.

While there are many studies of such RGB signals, efficient compression coding of RGBA images, which include alpha channels for transparency is needed because they are commonly used in practical applications such as image editing. The examples of the classical compression methods of RGBA images are PNG [5] with lossless compression and BPG [6] that uses an intraframe of HEVC video encoding [7]. In these methods, the alpha channel is compressed as a single-channel color signal by directly applying the compression method for RGB signals. However, since the characteristics of the signal are different between the alpha channel and RGB color signals, high compression performance may not be achieved by classical compression methods. For example, according to our calculations, in the RGBA image dataset with an alpha channel manually added to the Kodak dataset used in the general image compression used in Section 4 of this paper, the autocorrelation coefficient is found to be about 0.886 for the RGB channel and about 0.973 for the alpha channel. In addition, the evaluation dataset P3M-500-NP, which comes with the RGBA image dataset P3M-10k, focusing on human images, resulted in an average of about 0.956 for the RGB channel and about 0.995 for the alpha channel. Note that the autocorrelation coefficient is measured by shifting 1 pixel in the x and y directions. The reason why the difference in the autocorrelation coefficients of P3M-10k is smaller than that of the Kodak dataset is that P3M-10k focuses on photographs of people, which often feature smooth color transitions, such as those in skin and clothing. These observations suggest that alpha channel signals exhibit greater similarity than RGB signals. These facts mean that the alpha channel signals have more similarity than the RGB signals. Therefore, we study a learning-based approach to improve the overall compression efficiency by optimizing the whole network consisted of sub-networks for RGB signals and alpha channel. Each sub-network learns the features of each signal. In addition,

Received September 18, 2024; Revised November 27, 2024; Accepted December 19, 2024
†Graduate School of Engineering, Kogakuin University (Tokyo, Japan)

また、RGB信号のネットワークは、αチャンネルの情報を取り込んで学習する。

本研究では、学習ベースのRGBA画像圧縮に関する我々の先行研究を拡張する[8] 。我々の以前の研究では、RGB信号には3チャンネルの入力エンコーダを、1チャンネルの入力アルファチャンネルエンコーダを使用した。対応するデコーダは、それぞれRGB画像とアルファ画像を再構成するように設計されている。RGB信号を学習する際に、損失関数にアルファチャンネルの情報を追加した。圧縮効率は、前回の研究では十分ではなかった。本研究では、RGBおよびαチャネルのエンコーダとデコーダの性能を向上させるための新しい注意モジュールを提案する。具体的には、RGB信号のネットワークに対して、αチャンネルからのマスクされていない領域の注意を、αチャンネルのネットワークに対して、簡略化した注意モジュールを紹介する。

本論文では、提案手法が、RGB信号の圧縮ネットワークを単純に拡張したRGBA信号の4チャンネルネットワークに加えて、BPG[6]やAVIF[9]のようなRGBAをサポートする既存の古典的手法を凌駕することを実証する。また、αチャネルの学習データセットへの依存性を評価し、セマンティックセグメンテーションタスクでよく使われるCOCOデータセットと、画像マットタスクで使われるP3M-10kデータセットを用いることで、αチャネルの圧縮性能が大幅に向上することを示す。

本論文の貢献は以下の通りである：

1. RGB信号を処理するネットワークに対して、αチャネルを用いた新しい注意モジュールを提案する。
2. 4チャンネルで処理するネットワークや他の古典的な手法と比較することで、提案手法の性能優位性を実証する。
3. また、αチャネルの適切な学習データセットを用いることで、さらなる改善が見られることを示す。

2 関連研究

2.1 深層学習を用いたRGB画像圧縮

一般に、深層学習を用いた画像圧縮[10,11]は、VAE型ネットワークと4つの変換モジュールを用いる。エンコーダ g_a (x; Φ_g) は、元画像 x を入力画像とし、複数の畳み込み層と非線形関数を用いて潜在特徴変数 y に変換する。ハイパーエンコーダ h_a (y; Φ_h) は潜在特徴変数 y から潜在表現 z に変換する。次に、ハイパーデコーダh_s (z^; θ_h)を用いて、量子化潜在表現z^ = Q(z)からエントロピーモデル p_{y^|z^} (y^*z^)のパラメータを推定する。最後に、デコーダg_s (y^; θ_g)を用いて、量子化された潜在特徴変数y^ = Q(y)から再構成画像x^をデコードする。Φ_g、Φ_h、θ_h、θ_gは各モジュールの最適化パラメータである。これらの方法に加えて、Minnenら[12]によって提案されたチャネル単位のエントロピーモデルを用いる。エントロピーモデルでは、[y]の代わりに、各[y μ]を丸め、算術コーダーを用いて符号化し、分散σを持つ単一のガウス分布としてモデル化された[y μ] + μをy^として復号し、復号器g_s (y^; θ_g) に送る。さらに、このエントロピーモデルは、yをs個のスライス{y_0, y_1, ・・・, y_{s-1} }に分割することで、符号化効率を向上させる。zについては事前分布がないので、因数分解密度モデル ～を用いて式(1)のように符号化する。

$$p_{\hat{z}|\psi}(\hat{z}\mid\psi)=\prod_j\left(p_{z_j|\psi}(\psi)*\mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)\right)(\hat{z}_j) \quad (1)$$

式(1)において、z_jはzのj番目の要素を表し、jは各要素または各信号の位置を指定する。z^はそれぞれ専用のh_sを用いて復号され、2つの潜在特徴μ'とσ'が得られる。これらは以下のスライスネットワークSN$_i$の入力として使用される。

$$r_i,\mu_i,\sigma_i=SN_i(\mu',\sigma',\bar{y}_{<i},y_i) \quad (2)$$

$$\bar{y}_{<i}=\{\bar{y}_0,\bar{y}_1,\ldots,\bar{y}_{i-2},\bar{y}_{i-1}\} \quad (3)$$

これらの過程はp_{y^|z^} (y^*z^)~N(μ, σ^2) と仮定できる。潜在残差予測の出力結果r_iは、量子化によって生じる量子化誤差(y-y^)を低減するために用いられる。このr_iの値を用いて式(4)で Δ_i を求める。

$$\bar{y}_i=r_i+\hat{y}_i \quad (4)$$

出力結果 Δ はデコーダ g_s (y^; θ_g) への入力である。ディープラーニングを用いた画像圧縮法における損失関数は、以下の式で定義される。

$$L=H(\hat{y})+H(\hat{z})+\lambda\cdot D(x,\hat{x})=\mathbb{E}\left[-log_2\left(p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z})\right)\right]+\mathbb{E}\left[-log_2\left(p_{\hat{z}|\psi}(\hat{z}\mid\psi)\right)\right]+\lambda\cdot\mathcal{D}(x,\hat{x}) \quad (5)$$

式(5)において、H(y^)とH(z^)は画素あたりのビット数、Dは入力画像xと再構成画像x^の誤差である歪みである。λはRDトレードオフを操作するためのハイパーパラメータである。

CNNを用いたROI(関心領域)画像圧縮法は、Akutsuらによって発表された[13] 。ROI画像符号化法は、特定の画像領域の効率的な圧縮強調画質を達成することを目的とする。ROIに基づく圧縮では、領域を示すマスクデータをエンコーダネットワークに導入し、さらに、マスクデータは出力しないが、

the network for RGB signals learns by incorporating the information of the alpha channel.

This work extends our previous work on learning-based RGBA image compression [8]. In our previous work, we used a 3-channel input encoder for RGB signals and a 1-channel input alpha channel encoder. The corresponding decoders were designed to reconstruct RGB and alpha images, respectively. Information on the alpha channel was added to the loss function when learning the RGB signal. Compression efficiency was not sufficient in the previous study. In this study, we propose a novel attention module to improve the performance of the RGB and alpha channel encoders and decoders. Specifically, we introduce the attentions of the unmasked regions from the alpha channel for the network of RGB signals, and the simplified attention module for the network of the alpha channel.

In this paper, we demonstrate the proposed method outperforms existing classical methods that support RGBA, such as BPG [6] and AVIF [9], in addition to a 4-channel network supporting RGBA signals which simply extends compression network of RGB signals. We also evaluate the dependence of the alpha channel on the training dataset and show that using the COCO dataset, which is commonly used in semantic segmentation tasks, and the P3M-10k dataset, which is used in image matting tasks, significantly improves the compression performance of the alpha channel.

The contributions of this paper are as follows:

1. We propose a new attention module using alpha channel for the network to process RGB signals.
2. We demonstrate the performance advantage of the proposed method by comparison with a network that processes with 4 channels and other classical methods.
3. We also show the further improvement by using appropriate training dataset for alpha channel.

2. Related Work

2.1 RGB image compression using deep learning

In general, deep learning-based image compression [10, 11] uses a VAE-type network and four transformation modules. Encoder $g_a(x; \Phi_g)$ takes the original image x as input image and transforms it into a latent feature variable y using multiple convolutional layers and nonlinear functions. Hyper encoder $h_a(y; \Phi_h)$ converts from latent feature variable y to latent representation z. Next, hyperdecoder $h_s(\hat{z}; \theta_h)$ is used to estimate the parameters of the entropy model $p_{\hat{y}|\hat{z}}(\hat{y}\,|\,\hat{z})$ from the quantized latent representation $\hat{z} = Q(z)$. Finally, decode the reconstructed image $\hat{x}$ from the quantized latent feature variable $\hat{y} = Q(y)$ using decoder $g_s(\hat{y}; \theta_g)$. Note that Φ_g, Φ_h, θ_h and θ_g are optimization parameters for each module. In addition to these methods, we use channel-wise entropy model proposed by Minnen et al. [12]. In the entropy model, instead of [y], each [y - μ] is rounded and encoded using an arithmetic coder, and [y - μ] + μ, modeled as a single Gaussian distribution with variance σ, is decoded as $\hat{y}$ and sent to the decoder $g_s(\hat{y}; \theta_g)$. Moreover, this entropy model improves coding efficiency by dividing y into s slices $\{y_0, y_1, ..., y_{s-1}\}$. For z, there is no prior distribution, so the factorized density model Ψ is used to encode it as in Equation (1).

$$p_{\hat{z}|\psi}(\hat{z}\,|\,\psi) = \prod_j \left(p_{z_j|\psi}(\psi) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right) \right)(\hat{z}_j) \quad (1)$$

In Equation (1), z_j represents the j-th element of z, where j specifies the location of each element or each signal. $\hat{z}$ is respectively decoded using dedicated h_s to obtain two latent features μ' and σ'. These are used as inputs to each of the following slice networks SN_i.

$$r_i, \mu_i, \sigma_i = SN_i(\mu', \sigma', \bar{y}_{<i}, y_i) \quad (2)$$

$$\bar{y}_{<i} = \{\bar{y}_0, \bar{y}_1, \ldots, \bar{y}_{i-2}, \bar{y}_{i-1}\} \quad (3)$$

These processes can be assumed to be $p_{\hat{y}|\hat{z}}(\hat{y}\,|\,\hat{z})$~N($\mu$, σ^2). The output result r_i of the latent residual prediction is used to reduce the quantization error (y–$\hat{y}$) introduced by quantization. This value of r_i is used in Equation (4) to obtain $\bar{y}_i$.

$$\bar{y}_i = r_i + \hat{y}_i \quad (4)$$

The output result $\bar{y}$ is the input to the decoder $g_s(\hat{y}; \theta_g)$. The loss function in the image compression method using deep learning is defined by the following equation.

$$L = H(\hat{y}) + H(\hat{z}) + \lambda \cdot D(x, \hat{x}) = \mathbb{E}\left[-log_2\left(p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z})\right)\right] + \mathbb{E}\left[-log_2\left(p_{\hat{z}|\psi}(\hat{z}\,|\,\psi)\right)\right] + \lambda \cdot \mathcal{D}(x, \hat{x}) \quad (5)$$

In Equation (5), H($\hat{y}$) and H($\hat{z}$) are the amount of bit per pixel, D is distortion which is the error between the input image x and the reconstructed image $\hat{x}$. λ is a hyperparameter to manipulate the RD trade-off.

There is the ROI (Region of Interest) image compression method which used CNN, presented by Akutsu et al. [13]. The ROI image encoding method aims to achieve efficient compression enhancing image quality of particular image region. In the ROI based compression, mask data was introduced to encoder network to indicate the area, and additionally the compression scheme of such mask together with RGB signals by a

4チャンネルエンコーダによるRGB信号とともに、そのようなマスクの圧縮方式を提案した。本論文では、αチャンネルを持つ画像を効率的に符号化することを目的とし、αチャンネルの異なる特徴に基づいて、それらを別々に学習・符号化する手法を提案する。

2.2 注意モジュール

ディープラーニングを用いたRGB画像圧縮では、多くの研究が注意モジュールを導入している。注意モジュールを導入することで、視覚的に重要な領域により多くのビットを割り当てることができ、他の場所に割り当てられるビットはより少なくすることができる。その結果、視覚的な品質が向上する。例えば、Chenら[14]は重要度マップを作成するために非局所的注意を適用した。Chengら[15]は、図1に示すように、より高速な処理を達成するために、非局所的注意のブロックを除去する簡略化された注意モジュールを導入した。

簡易アテンションモジュールは、特徴マップの生成、アテンションマスクの作成、残差接続の3つのブランチから構成される。メインブランチは、3つの畳み込み層からなる3つの残差ブロックを用いて特徴マップを生成する。マスクブランチは、3つの異なる残差ブロック、1つの1×1畳み込み層、非線形シグモイド活性化関数を用いてアテンションマスクを作成する。作成されたアテンションマスクは、メインブランチからの出力と要素ごとに掛け合わされる。最後に、入力データへの残差接続により、最終的な出力が作成される。

Zouら[2]は、Swin Transformer[3]に触発されたウィンドウベースの注意法を提案した。このアプローチでは、入力をM×Mの重複しないローカルウィンドウに分割し、その中で自己注意をウィンドウごとに別々に計算する。

基本構造はSW-MSAと同じである（マルチ

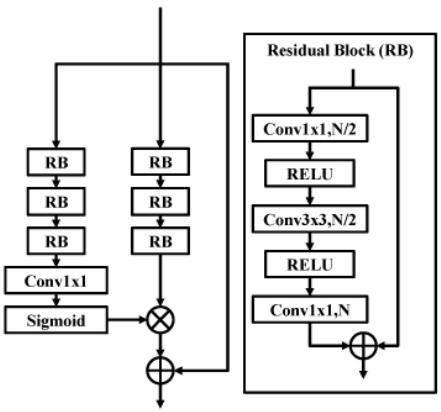

図1 簡略化された注意モジュール

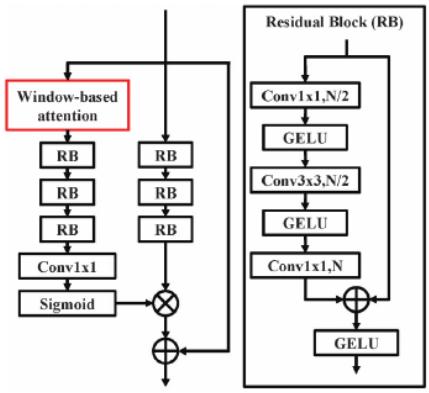

図2 ウィンドウベースの注意モジュール

ここで、$Q = xW_Q$，$K = xW_K$，$V = xW_V$ は入力画像の線形変換によって得られる。ここで、W_Q，W_K，W_Vは異なるウィンドウ間で共有される重み行列であり、xはローカルウィンドウに分割された入力画像である。一般に、Q，K，Vは次元$R^{M2\times d}$を持ち、dはクエリとキーの次元数である。最後に、注目行列は以下の式を用いて計算される。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \qquad (6)$$

ここで、Bは学習可能な相対位置バイアスである。このウィンドウベースの注意は、図2に示すウィンドウベースの注意モジュールを使用してネットワークに組み込まれる。

ウィンドウベースの注意モジュールの基本構造は、簡略化された注意モジュール[15]と同じである。その違いは、Window-based attentionの導入とResidualブロックの内容である。このモジュールでは、注意マスクが作成される分岐の最初に、ウィンドウベースの注意が導入される（図2の赤枠）。簡易注意モジュールとは異なり、このモジュールのResidualブロックは活性化関数にGELUを使用する。この活性化関数は、Residualブロックの最終出力の前にも使用される。

2.3 深層学習によるRGBA画像圧縮

学習

我々の知る限り、我々の過去の研究は、RGBA画像圧縮に深層学習を適用した唯一の研究である。我々の以前の研究[8]では、RGBA入力画像をRGBチャンネルとαチャンネルに分離し、各チャンネルを専用ネットワークに入力した。最後に、各ネットワークからの出力チャンネルを結合し、1つのRGBA画像にマージする。

four-channel encoder was proposed, although it did not output mask data. In this paper, the aim of our study is to efficiently encode an image with an alpha channel, and we propose a method to learn and encode them separately, based on the different characteristics of the alpha channel.

2.2 Attention module

In RGB image compression using deep learning, many studies introduce attention modules. By introducing attention modules, more bits can be allocated to visually important areas and fewer bits can be allocated elsewhere. As a result, visual quality is improved. For example, Chen et al. [14] applied nonlocal attention to create importance maps. Cheng et al. [15] introduced a Simplified attention module that removes blocks of nonlocal attention to achieve faster processing, as shown in Figure 1.

The Simplified attention module consists of three branches that generate feature maps, create attention masks and residual connections. The main branch generates feature maps using three Residual Blocks consisting of three convolutional layers. The mask branch creates an attention mask using three different Residual Blocks, one 1×1 convolutional layer, and a nonlinear sigmoid activation function. The created attention mask is multiplied element by element with the output from the main branch. Finally, the final output is created through a residual connection to the input data.

Zou et al. [2] proposed a window-based attention method, inspired by the Swin Transformer [3]. This approach divides the input into M×M non-overlapping local windows, within which self-attention is computed separately for each window.

The basic structure is the same as SW-MSA (multi-head self-attention using shifted window partitioning configurations) in Swin Transformer [3], and the equations are shown in Equation 6, where $Q = xW_Q$, $K = xW_K$, $V = xW_V$ are obtained by linear transformation of the input image. where W_Q, W_K, W_V are weight matrices shared among different windows and x is the input image divided into local windows. Generally, Q, K, V have dimensions $\mathbb{R}^{M^2 \times d}$, where d is the number of dimensions of queries and keys. Finally, the attention matrix is computed using the following equation.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \qquad (6)$$

where B is the learnable relative position bias. This window-based attention is incorporated into the network using the window-based attention module shown in Figure 2.

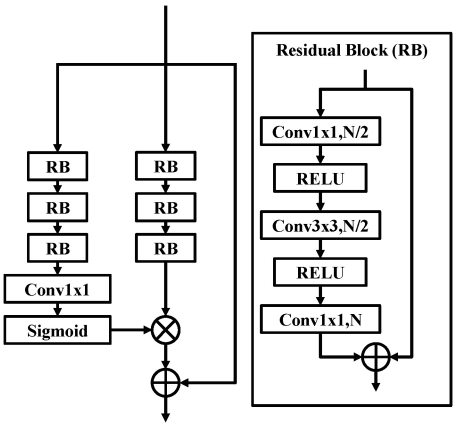

Fig. 1 Simplified attention module.

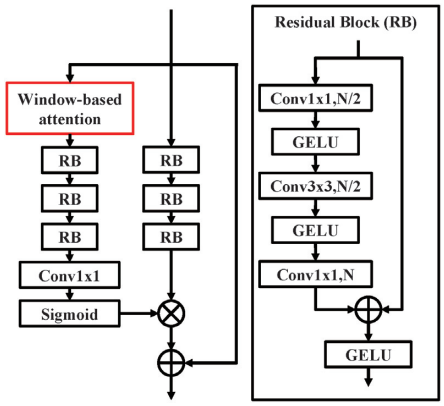

Fig. 2 Window-based attention module.

The basic structure of the Window-based attention module is the same as that of the Simplified attention module [15]. The differences are the introduction of Window-based attention and the content of the Residual block. In this module, Window-based attention is introduced at the beginning of the branch where the attention mask is created (red frame in Figure 2). Unlike the Simplified attention module, the Residual block in this module uses GELU for the activation function. This activation function is also used before the final output of the Residual block.

2.3 RGBA image compression using deep learning

To the best of our knowledge, our previous work is the only study that applies deep learning to RGBA image compression. In our previous study [8], the RGBA input image was separated into RGB and alpha channels, and each channel were input to a dedicated network. Finally, the output channels from each network are combined

各チャンネルのネットワークアーキテクチャはCNNをベースとし、Zouら[2]の方法と同様であり、RGB信号に対してのみWindow-based attention moduleが追加される。我々の以前の研究では、学習された畳み込みネットワークを使用する手法もRGBA画像に有効であることが示された。本論文では、RGB信号のWindowベースの注意モジュールにマスクを適用したネットワークと、αチャンネルに別の注意を適用したネットワークを提案し、従来の研究よりも性能を向上させる。

3提案手法

3.1 マスク窓ベースの注意モジュール Zouら[2]が提案したRGB画像に対する窓ベースの注意モジュールは、RGBAフォーマットに対して必ずしも最適化されていない。そこで、マスクされた窓ベースの注意モジュールを提案する。マスクされたウィンドウベースの注意モジュールは、図3に示すように、ウィンドウ分割中のアルファネットワークの入力と出力のアルファ画像を参照し、アルファチャネルウィンドウ内のすべてのピクセル値がゼロの場合、同じ位置の特徴マップウィンドウは注意計算から除外される。注意の計算後、分割されたウィンドウ特徴マップは、アルファ画像を参照しながら、元の位置に再配置される。分割された窓特徴マップは、図3のオレンジ色の窓で示される注意計算で使用される窓を指すことに注意。この処理により、アルファ画像によってマスクされる画素値ゼロに対する不要な相関計算が不要となり、より効率的な学習が可能となる。例えば、図3(a)では、通常のウィンドウ分割は384個のウィンドウ分割を持つが、図3(b)の提案手法は119個のウィンドウ分割を持ち、ウィンドウ数が3分の1以下に減少し、注目度計算がより軽量になっていることを示している。図3は視覚的にわかりやすくするために32×32のウィンドウ分割を示しているが、ウィンドウが8×8のウィンドウに分割されているため、実際の計算はより細かく分割されていることに注意されたい。

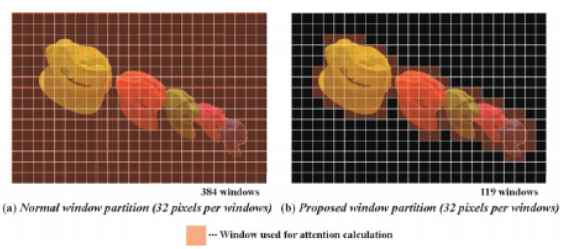

図3 ウィンドウパーティションの違い

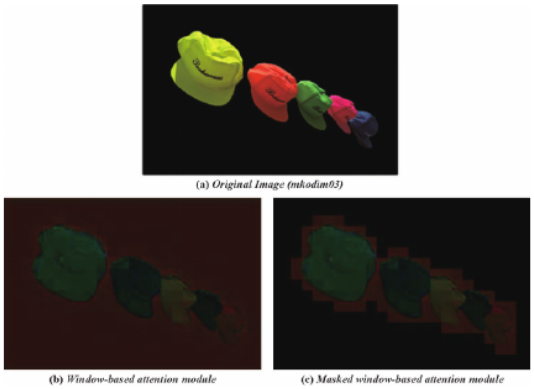

図4 注意計算後の特徴マップの違い;(c)の黒は差がないことを意味する。

注意の計算には、式(6)と同じ方法を用いる。図4は、従来手法のウィンドウベースアテンションモジュールと、提案手法計算後の特徴マップの違いを示している。図4(c)では、提案するマスク窓ベースの注意はアルファ画像によってマスクされた領域を計算しないのに対し、図4(b)の通常の窓ベースの注意は画像全体の注意マップを計算することがわかる。これらの新しいプロセスは、図2のウィンドウベースの注意の代替として使用される。

3.2 ネットワークアーキテクチャ

提案するネットワークアーキテクチャを図5に示す。前回と同様に、RGBAの入力画像はRGBチャンネルとαチャンネルに分離され、各チャンネルは異なるネットワークに入力される。各ネットワークのエンコーダ出力は、Minnenらのチャネル単位の自己回帰モデルに接続されている[12] 。
図5に示すように、RGBネットワークは前述のマスク窓ベースの注意モジュールを導入している。このモジュールはアルファ画像を必要とする。したがって、入力x_αと出力$\hat{x}_\alpha$は、それぞれカーネルサイズ3、ストライド2のmaxpool2dを用いて適切な解像度に変更され、注意モジュールで使用される。αネットワークは、Chengら[14]で使用されたSimplified attentionモジュールを使用し、計算の複雑さを軽減する。

Liuら[16]によって提案されたデコーダ側拡張モジュールは、RGBデコーダとAlphaデコーダの最終層に導入される。このモジュールは、非可逆圧縮における圧縮アーチファクトを除去する。図6に本モジュールのネットワーク構成を示す。デコーダ側拡張モジュールは、まず、点状畳み込み層を用いて、入力特徴マップのチャンネルを32チャンネルに拡張する。

and merged into a single RGBA image. The network architecture of each channel is based on CNN and similar to the method of Zou et al [2], and only for RGB signals Window-based attention module is added to it. Our previous work has shown that methods using learned convolutional networks are also effective for RGBA images. In this paper, we propose a network with a mask applied to the Window-based attention module for RGB signals and a network with another attention applied to the alpha channel to improve performance over the previous work.

3. Proposed Method

3.1 Masked window-based attention module

The window-based attention module proposed by Zou et al [2] for RGB images is not always optimized for the RGBA format. Therefore, we propose a masked window-based attention module. The masked window-based attention module refers to the input and output alpha images of the alpha network during window partitioning, as shown in Figure 3, and if all pixel values in an alpha channel window are zero, the feature map window at the same location is excluded from the attention calculation. After calculation of the attention, the partitioned window feature map is repositioned to its original position while referring to the alpha image. Note that the partitioned window feature map refers to the window used in the attention calculation, shown in orange window in Figure 3. This process eliminates unnecessary correlation calculations for pixel value zero that are masked by the alpha image, allowing for more efficient learning. For example, in Figure 3(a), the normal window partitioning has 384 window partitions, while the proposed method in Figure 3(b) has 119 window partitions, indicating that the number of windows is reduced to less than one-third, and the attention calculation is more lightweight. Note that although Figure 3 shows a 32×32 window division for visual clarity, the actual calculation is divided more finely because the window is partitioned in 8×8 windows. The same method as in Equation (6) is used for the attention calculation. Figure 4 shows the difference between the window-based attention module of the conventional method and the feature map after the calculation of the proposed method. In Figure 4(c), it can be seen that the proposed masked window-based attention does not compute the areas masked by the alpha image, whereas the normal window-based attention in Figure 4(b) computes an attention map for the entire image. These new processes are used as an alternative to the window-based attentions in Figure 2.

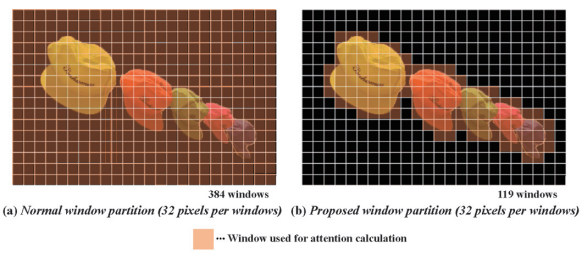

Fig. 3 Differences in window partitions.

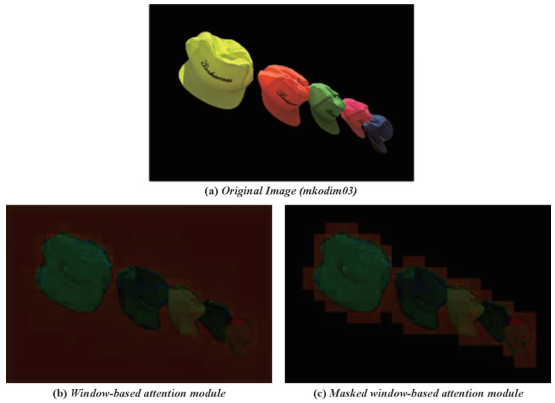

Fig. 4 Differences in feature maps after attention calculation; in (c) black means no difference.

3.2 Network architecture

The proposed network architecture is shown in Figure 5. As with our previous study, the RGBA input image is separated into RGB and alpha channels, and each channel are input to different networks. The encoder output of each network is connected to the Channel-wise Autoregressive Model of Minnen et al. [12].

As shown in Figure 5, the RGB network introduces the aforementioned Masked window-based attention module. This module requires an alpha image. Therefore, the input x_{alpha} and output $\hat{x}_{alpha}$ are changed to the appropriate resolution respectively using maxpool2d with a kernel size of 3 and a stride of 2 and used in the attention module. The alpha network uses the Simplified attention module used in Cheng et al. [14] to reduce computational complexity.

The decoder side enhancement module, which was proposed by Liu et al. [16], is introduced in the final layer of RGB decoder and Alpha decoder. This module removes compression artifacts in lossy compression. Figure 6 shows the network architecture of this module. The decoder-side enhancement module first extends the channels of the input feature map to 32 channels using the point wise convolution layer. Thus, after applying the three residual blocks, it applies them back to the

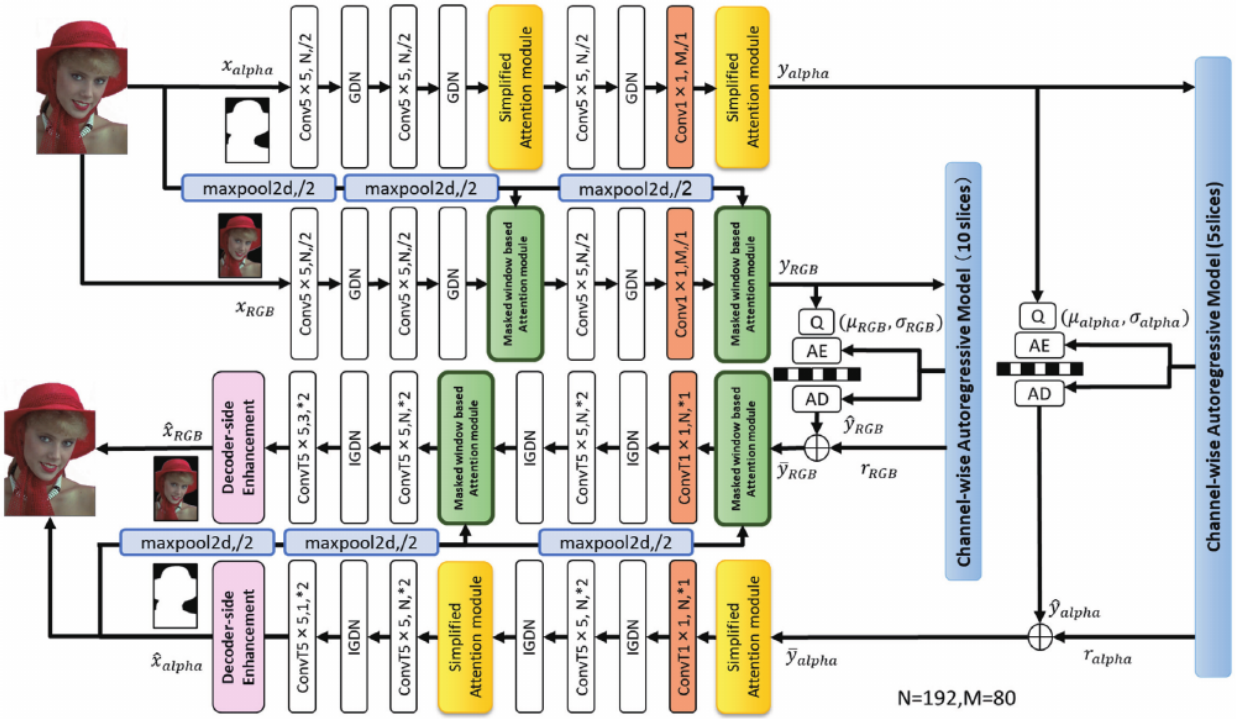

図5 提案ネットワークアーキテクチャ

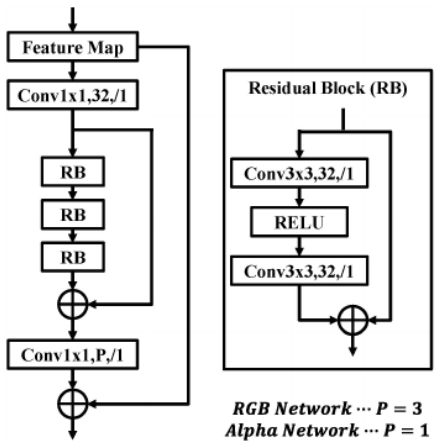

図6 デコーダ側拡張モジュール

このように、3つの残差ブロックを適用した後、最後にポイントワイズ畳み込み層を使って入力チャンネルに戻す。

図5において、GDNは一般化分割正規化[17]であり、IGDNは逆一般化分割正規化であることに注意。

3.3 損失関数のMSE

αチャンネルでマスクされたビットをマスクされていない領域のビットに割り当てることで、画像圧縮性能が向上すると仮定する。そこで、式(5)の歪みの計算で表示される画素のみの再構成誤差を計算する平均二乗誤差(MSE)を提案する。提案するMSEは、以下の計算を行うことで算出される。

$$alpha_{input}(i,j) = \begin{cases} 1 & x_{alpha}(i,j) > 0 \\ 0 & otherwise \end{cases} \quad (7)$$

$$Input = x_{RGB} \odot alpha_{input} \quad (8)$$

$$Output = \hat{x}_{RGB} \odot alpha_{input} \quad (9)$$

$$N = \sum_{i=1}^{height} \sum_{j=1}^{width} \sum_{c\in\{r,g,b\}} alpha_{input}(i,j,c) \quad (10)$$

$$MSE_{RGB} = \frac{1}{N} \sum_{i=1}^{height} \sum_{j=1}^{width} \sum_{c\in\{r,g,b\}} \{Input(i,j,c) - Output(i,j,c)\}^2 \quad (11)$$

式(7)において、x_{α}はAlpha Encoderの入力であり、8ビット画像データであるため、この計算の結果、x_{α}は2値データとなる。次に、式(8)と式(9)の演算を組み合わせて、互いのRGB計算領域を一致させる。$alpha_{input}$は自身のピクセルをコピーし、式(8)と式(9)に代入する前に3チャンネルに展開することに注意すべきである。式(8)と式(9)の演算はハダマード積を意味する。その後、式(10)を用いてターゲット画素の総数Nを計算する。最後に、式(11)を計算することで、αチャンネルのマスクされていない領域の画素のMSEを比較することができる。この計算でアルファチャンネルの低透過領域もMSEの計算に含まれる理由は、髪や衣服のように正しく再構成されない可能性があるためである。

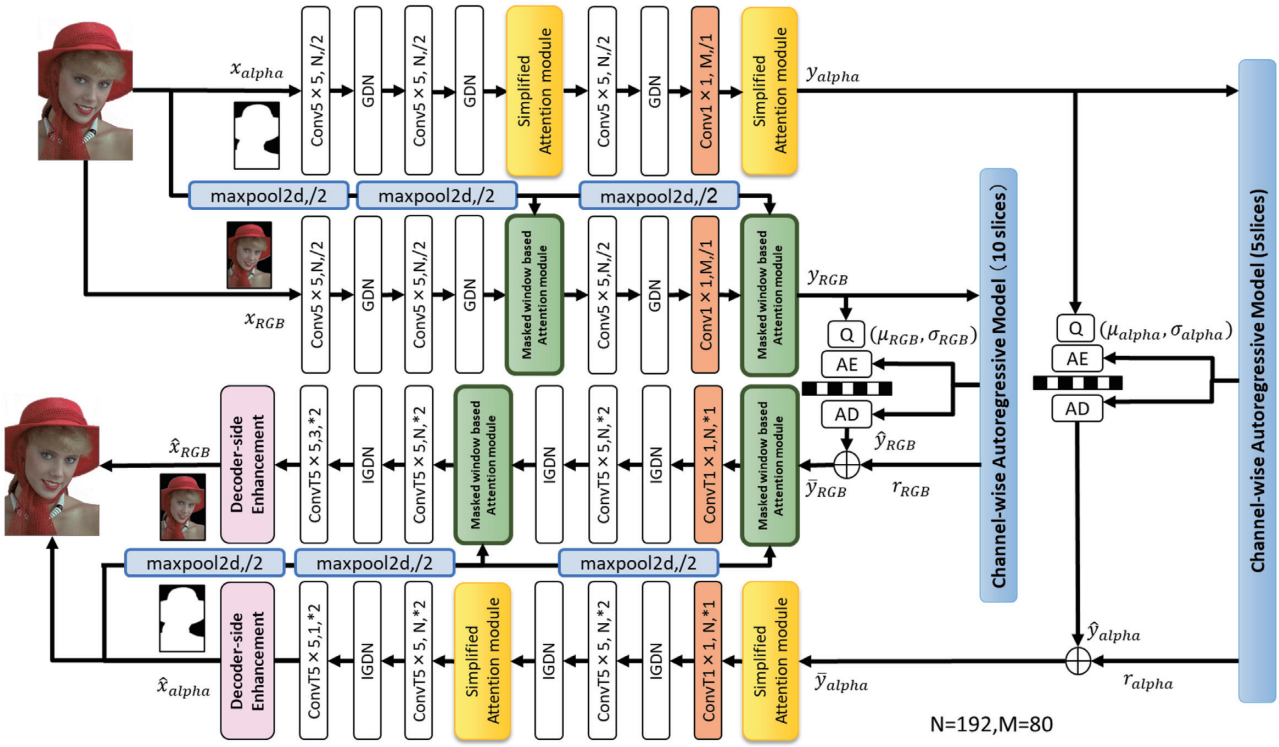

Fig. 5 Proposed Network Architecture.

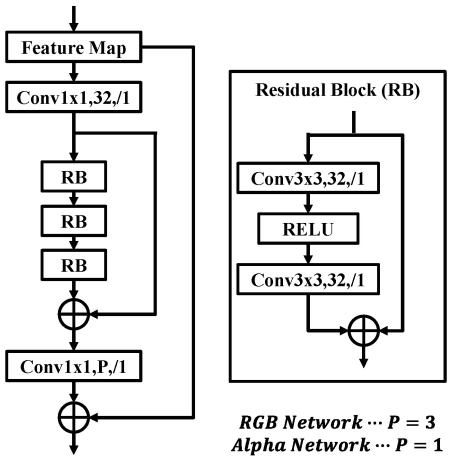

Fig. 6 Decoder-side enhancement module.

input channels using the point wise convolution layer at the end.

Note that in Figure 5, GDN is generalized division normalization [17] and IGDN is inverse generalized division normalization.

3.3 MSE for loss function

We assume that assigning the bits masked in the alpha channel to the bits in the unmasked area improves the image compression performance. Therefore, we propose a Mean Squared Error (MSE) that calculates the reconstruction error of only the pixels displayed by the calculation of distortion in Equation (5). The proposed MSE is calculated by performing the following calculations.

$$alpha_{input}(i,j) = \begin{cases} 1 & x_{alpha}(i,j) > 0 \\ 0 & otherwise \end{cases} \tag{7}$$

$$Input = x_{RGB} \odot alpha_{input} \tag{8}$$

$$Output = \hat{x}_{RGB} \odot alpha_{input} \tag{9}$$

$$N = \sum_{i=1}^{height} \sum_{j=1}^{width} \sum_{c\in\{r,g,b\}} alpha_{input}(i,j,c) \tag{10}$$

$$MSE_{RGB} = \frac{1}{N} \sum_{i=1}^{height} \sum_{j=1}^{width} \sum_{c\in\{r,g,b\}} \{Input(i,j,c) - Output(i,j,c)\}^2 \tag{11}$$

In Equation (7), x_{alpha} is the input of Alpha Encoder and it is 8-bit image data, so this calculation results in x_{alpha} being binary data. Next, the operations in Equations (8) and (9) are combined to match each other's RGB calculation regions. It should be noted that $alpha_{input}$ copies its own pixels and expands them to 3 channels before substituting them into Equations (8) and (9). The operations $\odot$ in Equations (8) and (9) mean the Hadamard product. Subsequently, the total number of target pixels, N, is calculated using Equation (10). Finally, by calculating Equation (11), the MSE of pixels in the unmasked region in the alpha channel can be compared. The reason why low-transparency regions in the alpha channel are also included in the calculation of MSE in this calculation is that they may not be

なお、αネットワークでは、式(5)の再構成誤差を計算するために通常のMSEが使用されている。また、本論文では、MSE_{RGB}はRGB信号に用いるMSE、MSE_{Alpha}はαチャネルに用いるMSEを意味する。

3.4 復号されたアルファ画像の制約条件

一般に、非可逆圧縮で復号された画像は圧縮ノイズを含むことがある。そこで、復号化されたアルファ画像に制約を導入することを提案する。0-1に正規化されたアルファ画像では、0と1はマスクされた領域の内側と外側に対応し、他の中間値はマスクされた領域とマスクされていない領域の境界を表す。提案する制約条件は、画素値1で囲まれた領域の画素値を1に設定し、画素値0で囲まれた領域の画素値を0に設定する。図7に示す制約条件により、マスクされた領域とマスクされていない領域の境界で画素値を維持したまま、圧縮ノイズを除去することを保証することができる。なお、この制約は学習過程の不安定化を避けるため、評価段階でのみ適用している。

4実験とディスカッション

提案手法の性能を評価するために、以下の2つの観点からの実験結果を示す。1つ目は、提案手法の既存手法と比較した総合的な性能である。2つ目は、4チャンネルのRGBAネットワークとの比較である。

アブレーション研究として、提案手法のマスクされた窓ベースの注意モジュールの性能をZouらの窓ベースの注意モジュールと比較する[2] 。さらに、COCOデータセット[18]のみを使用した場合と、COCOデータセットとP3M-10k[19]を使用した場合の、異なる学習データセットの性能を比較する。

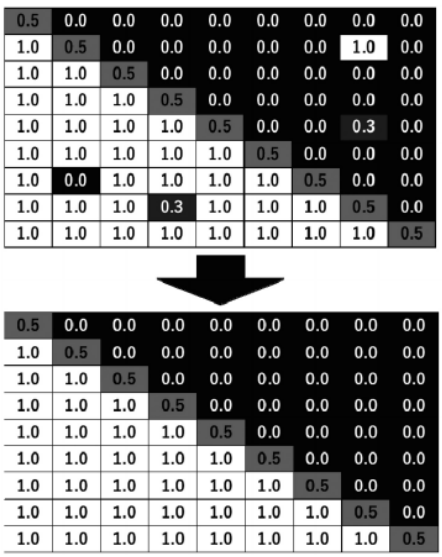

図7 提案するアルファ画像に対する制約の例。

4.1 実験条件について

ネットワークの学習には、COCOデータセット[18]から118,287枚の画像を256×256のサイズにランダムに切り出し、そのデータセットから提供されたセグメンテーション情報をアルファ画像として使用した。このデータは2値データであるため、P3M-10k[19]の9,422枚のRGBA画像を用い、1枚あたり30枚のRGBA画像を256×256の解像度に切り出した。トリミング時にすべての画素値が0または255にならないように、画像は以下の2つの条件のいずれかを満たすようにトリミングされた。

1. 画像中の少なくとも30%の画素が画素を持つ0または255の値
2. 画像中の80%以上の画素が0から255の間の値を持つ。

学習画像の多様性を考慮し、RGBネットワークの学習中に、α画像によってマスクされていない確率が25%のマスクされていない画像の混合（αチャンネルの画素値はすべて255）を使用した。ネットワークの性能を評価するために、Kodak Photo CDデータセット[20]から解像度768×512の合計24枚の写真を使用し、評価画像としてアルファ画像を手動で作成した。評価に使用した画像は[21]にある。本稿では、このデータセットをMasked Kodakデータセットと呼ぶ。Adam optimizer[22] を用いてネットワークを学習し、バッチサイズを4とした。

RGBネットワークとαネットワークは別々に学習された。学習に用いた損失関数は式(5)と同じであり、損失関数の再構成誤差Dはαネットワークでは通常のMSE、RGBネットワークでは提案手法のMSEとした。通常のPSNRとの混同を避けるため、本論文では提案手法MSE_{RGB}をPSNRに変換したものを$PSNR_{RGB}$、MSE_{Alpha}をPSNRに変換したものを$PSNR_{Alpha}$と呼ぶ。ハイパーパラメータλの値は{256, 512, 1024, 2048, 4096}とした。αネットワークは合計60万回の反復で学習された。最初の220,000反復は1×10^{-4}の学習率で学習し、残りの380,000反復は1×10^{-5}の学習率で学習した。一方、RGBネットワーク全体は1,500,000回の反復で学習された。最初の1,000,000回まで、$\lambda = 4096$のみで学習率1×10^{-4}で学習させた。

reconstructed correctly, such as hair or clothing. Note that in the alpha network, the usual MSE is used to calculate the reconstruction error in Equation (5). In addition, in this paper, MSE_{RGB} means the MSE used for RGB signals and MSE_{Alpha} means the MSE used for alpha channel.

3.4 Constraints for the decoded alpha image

In general, the image decoded with lossy compression may contain compression noise. Therefore, we propose to introduce constraints on the decoded alpha image. In an alpha image normalized to 0-1, 0 and 1 correspond to the inside and outside of the masked region, while the other intermediate values represent the boundary between the masked and unmasked regions. The proposed constraint sets the pixel value of the region bounded by pixel value 1 to 1 and the pixel value of the region bounded by pixel value 0 to 0. The constraints shown in Figure 7 allow us to guarantee that compression noise is removed while maintaining the pixel values at the boundaries between masked and unmasked areas. It should be noted that we apply this constraint only during the evaluation phase to avoid destabilizing the learning process.

4. Experiments and Discussions

To evaluate the performance of the proposed method, experimental results on the following two perspectives are shown. The first is the overall performance of the proposed method compared to existing methods. The second is a comparison with a 4-channel RGBA network.

As ablation studies, the performance of the masked window-based attention module of the proposed method is compared with the window-based attention module of Zou et al. [2]. Furthermore, performance of different training datasets are compare between using only COCO dataset [18] and using COCO dataset and P3M-10k [19].

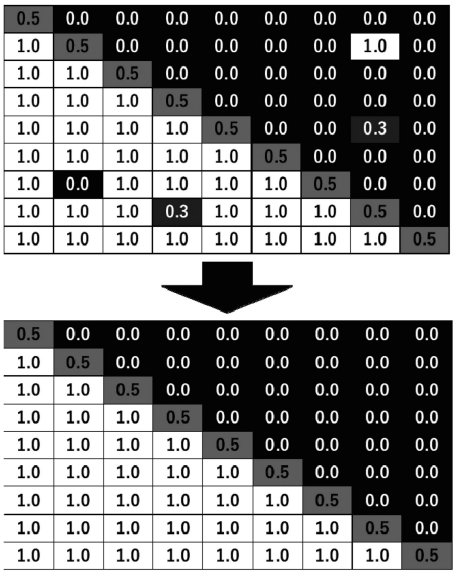

Fig. 7 Example of constraints on the proposed alpha image.

4.1 Experimental conditions

For training the network, 118,287 images from the COCO dataset [18] were randomly cropped to a size of 256 × 256 and the segmentation information provided by its dataset was used as alpha images. Since this data is binary data, we further used 9, 422 RGBA images from P3M-10k [19], cropped to a resolution of 256 × 256 with 30 RGBA images per one image. In order to prevent all pixel values from being 0 or 255 when cropping, the images were cropped to satisfy one of the following two conditions.

1. At least 30% of the pixels in the image have a pixel value of 0 or 255
2. 80% or more of the pixels in the image have values between 0 and 255

To account for the diversity of the training images, we used a mixture of unmasked images (all pixel values in the alpha channel were 255) with a 25% probability of being unmasked by the alpha image while training the RGB network. To evaluate the performance of the network, a total of 24 photos from the Kodak Photo CD dataset [20] with a resolution of 768 × 512 were used, and alpha images were manually created as evaluation images. The images used during the evaluation are available in [21]. In this paper, we refer to this dataset as the Masked Kodak dataset. Adam optimizer [22] was used to train the network and the batch size was set to 4.

The RGB network and the alpha network were trained separately. The loss function used for training was the same as in Equation (5), and the reconstruction error D of the loss function was the normal MSE for the alpha network and the MSE of the proposed method for the RGB network. To avoid confusion with ordinary PSNR, this paper refers to the proposed method MSE_{RGB} converted to $PSNR$ as $PSNR_{RGB}$ and MSE_{Alpha} converted to $PSNR$ as $PSNR_{Alpha}$. The values of hyperparameters λ were set to {256, 512, 1024, 2048, 4096}. The alpha network was trained for a total of 600, 000 iteration. The first 220,000 iterations was trained with a learning rate of 1×10^{-4}, and the remaining 380,000 iterations was trained with a learning rate of 1×10^{-5}. On the other hand, the entire RGB network was trained with 1,500,000 iterations. Up to the first 1,000,000 iterations, it was trained with only $\lambda = 4096$ with a learning rate of 1×10^{-4}. The subsequent 500,000 iterations were trained

その後の50万回の反復は、それぞれハイパーパラメータλを変化させることで、1×10^{-5}の学習率で学習された。提案されたネットワークはすべてPyTorchで実装され、実験はNVIDIA RTX A5000で行われた。

4.2 実験結果

定量的な評価として、レート歪み性能を評価した。提案手法であるBPG[6]とAVIF[9]のMasked Kodakデータセットの平均結果を図8に示す。

図8より、提案手法は他の手法と比較して、全てのビットレートにおいて画質が向上していることがわかる。このグラフにおける提案手法のbppは、αネットワークとRGBネットワークのbppの和であることに注意。グラフの作成に使用したλの値の組み合わせを表1に示す。

BPGの符号化と復号化にはWindowsの公式分散版を、AVIF[9]にはlibavif[23]を用いた。

また、MSE_{RGB}を定性的評価として最適化した場合の再構成画像として、Masked Kodak Datasetのmkodim15を図9に示す。

図9(a)の画像が入力画像であり、図9(c)、図9(d)、図9(e)は各手法で切り出した再構成画像である。図9(b)の入力画像と比較して、図9(c)と図9(d)のAVIF[9]とBPG[6]は、鼻付近で顕著な帯状ノイズを持つが、図9(e)の提案手法はノイズが少なく、視覚的に入力画像に最も近い。

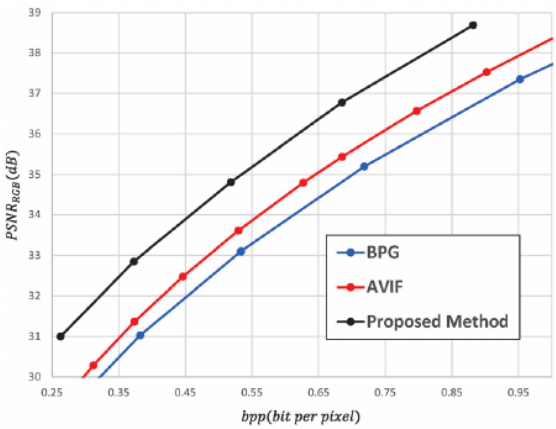

図8 Masked Kodakデータセットでの性能評価。

表 1 グラフの作成に使用したλの値の組み合わせ。

Network	Value of λ				
RGB network	4096	2048	1024	512	256
Alpha network	1024	1024	1024	512	256

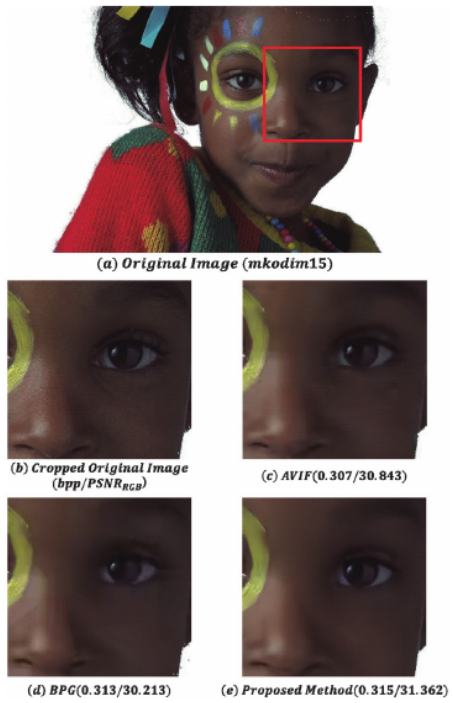

図9 再構成画像の可視化

表 2 ネットワークパラメータの比較

Method	Network Parameters
Zou et al.	75,235,779
Proposed Method	60,744,072

次に、提案手法のネットワークパラメータを表2に示す。参考として、RGB画像に対するZouら[2]ネットワークのパラメータも示す。

表2の結果は、バッチサイズ1、入力解像度256×256の入力画像に対するもので、torchinfo 1.8.0を用いて計算したものである。表2より、2つのネットワークを用いても、提案手法によりネットワークパラメータが約19%削減されていることがわかる。これは、各エンコーダとデコーダの最終層で使用されるポイントワイズコンボリューションの効果に起因する。

4.3 4チャンネルネットワークとの比較

RGBA画像は、透明性情報のためにαチャンネルを追加したRGB画像の形式であるため、通常の3チャンネルRGBネットワークの入力と出力チャンネルを4チャンネルに変更するだけで、RGBA画像に適応させることができる。そこで、RGBチャンネルとAlphaチャンネルを別々に処理する提案ネットワークと、4チャンネルネットワークを比較する。図10に比較のためのネットワークアーキテクチャを示す。また、4チャンネルネットワークでは、RGBチャンネルとαチャンネルを同時に最適化する必要がある。そこで、式(12)を以下のように修正した式(5)の損失関数D(x, x^)を使用する。

with a learning rate of 1×10^{-5} by respectively varying the hyperparameter λ. All proposed networks were implemented in PyTorch and experiments were performed on NVIDIA RTX A5000.

4.2 Experimental results

As a quantitative evaluation, rate distortion performance was evaluated. The average results of the Masked Kodak dataset for the proposed method, BPG [6] and AVIF [9] are shown in Figure 8.

From Figure 8, it can be seen that the proposed method improves the image quality at all bit rates compared to the other methods. Note that the bpp of the proposed method in this graph is the sum of the bpp of the alpha network and the RGB network. The combination of values of λ used to create the graph is shown in Table 1.

For BPG encoding and decoding, we used the officially distributed version for Windows, and for AVIF [9] we used libavif [23].

In addition, mkodim15 from the Masked Kodak Dataset is shown in Figure 9 as a reconstructed image when optimized with MSE_{RGB} as a qualitative evaluation.

The image in Figure 9(a) is the input image and Figures 9(c), 9(d), and 9(e) show the reconstructed images cropped by each method. Compared with the input image in Figure 9(b), the AVIF [9] and the BPG [6] in Figure 9(c) and Figure 9(d) have noticeable banding noise near the nose, whereas the proposed method in Figure 9(e) has less noise and is visually closest to the input image.

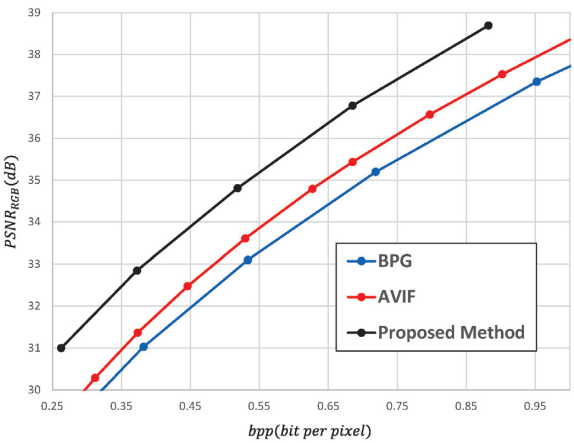

Fig. 8 Performance evaluation on Masked Kodak dataset.

Table. 1 Combination of values of λ used to create the graph.

Network	Value of λ				
RGB network	4096	2048	1024	512	256
Alpha network	1024	1024	1024	512	256

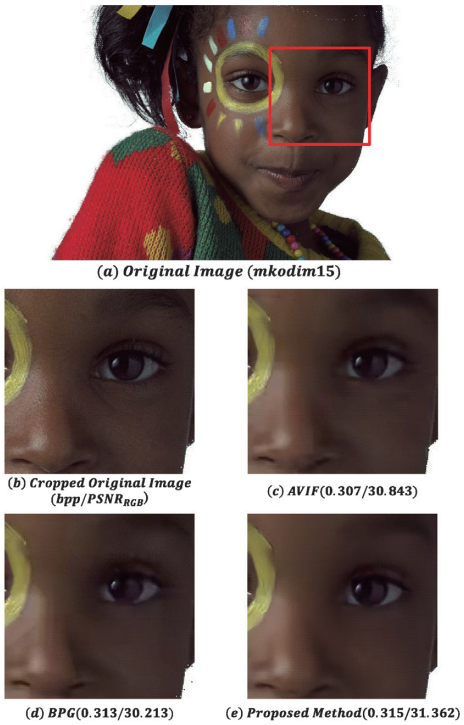

Fig. 9 Visualization of reconstructed images.

Table. 2 Comparison of network parameters.

Method	Network Parameters
Zou et al.	75,235,779
Proposed Method	60,744,072

Next, the network parameters of the proposed method are listed in Table 2. As a reference, the parameters of the Zou et al. [2] network for RGB images are also shown.

The results in Table 2 are for an input image with batch size 1 and input resolution of 256×256, calculated using torchinfo 1.8.0. From Table 2, the proposed method reduces the network parameters by about 19%, even though two networks are used. This can be attributed to the effect of pointwise convolution used in the final layer of each encoder and decoder.

4.3 Comparison with 4-channel network

Since RGBA images are in the form of RGB images with an additional alpha channel for transparency information, they can be adapted to RGBA images by simply changing the input and output channels of a normal 3-channel RGB network to 4 channels. Therefore, we compare the proposed network, which processes RGB and Alpha channels separately, with the 4-channel network. Figure 10 shows the network architecture for comparison. In addition, in a 4-channel network, the RGB and alpha channels must be optimized simultaneously. Therefore, we use the equation (5) loss function $D(x, \hat{x})$ with the following modification of the Equation (12) in the

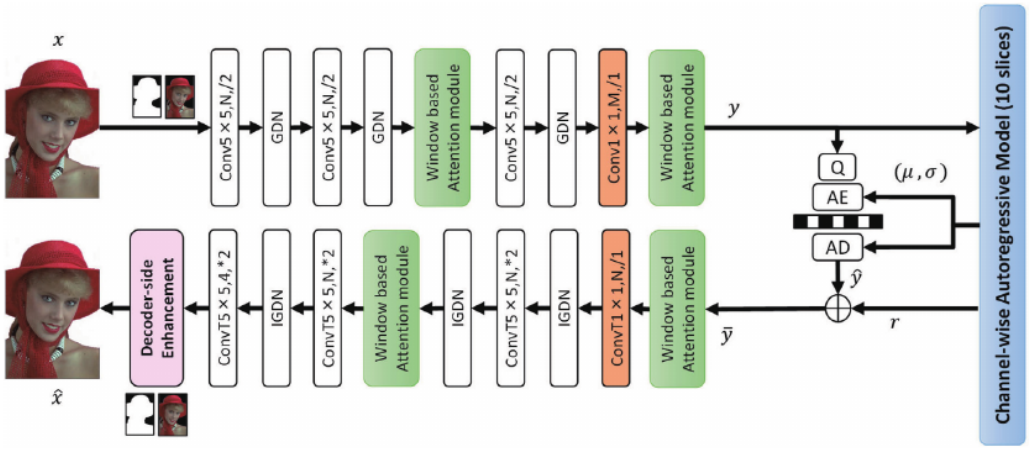

図10 比較法のための4チャンネルネットワーク

4チャンネルネットワーク

$$D(x,\hat{x}) = \left(MSE_{RGB} + MSE_{alpha}\right) \quad (12)$$

図11に$PSNR_{RGB}$を用いた比較結果のグラフを、図12に$PSNR_{alpha}$を用いた比較結果を示す。比較手法に加え、Minnen2018[11]とZou2022[2]のCNNベースとSwin Transformerベースの4チャンネル入出力層モデルの結果も示す。

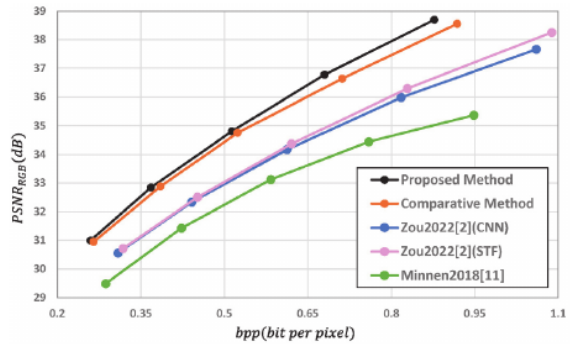

Fig. 11 提案手法との性能差について
$PSNR_{RGB}$と比較した4chネットワーク.

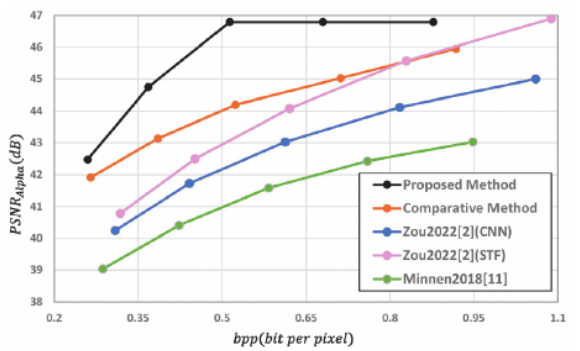

図12 提案手法と4chネットワークの性能差と$PSNR_{Alpha}$の比較
.

図11より、高ビットレートでは性能に顕著な差があることがわかる。しかし、ビットレートが低いほど性能差は小さくなる。図12の$PSNR_{Alpha}$の場合、性能差が大きい。図12に示すように、提案手法と比較手法の性能差の理由は、比較手法がRGBチャンネルとアルファチャンネル間で共有される単一のエントロピーモデルを利用しているため、最適化プロセスがより複雑になっていることに起因していると考えられる。一方、提案手法は、RGBチャンネルとαチャンネルに対して別々のエントロピーモデルを採用しているため、αチャンネルに特化したエントロピーモデルの最適化が可能である。なお、提案手法のグラフはλが固定されているため、右に移動しても平坦である。

したがって、RGB画像とα画像を別々に処理することで、学習とネットワーク構築はそれぞれの信号特性を考慮することができる。さらに、RGBチャネルとαチャネルの各ビットレートを独立に変更することで、より実用的な利用が可能となる。

4.4 アブレーション研究

4.4.1 マスク窓ベースの注意モジュールパフォーマンス

Masked window-based attention moduleの有効性を示すために、通常の window-based attention moduleとの比較を行う。この通常のウィンドウベースの注意モジュールは、我々の以前の研究[8]で使用したものと同じである。実験条件は4.1と同じである。

図13は、提案手法であるマスク窓ベースの注意モジュールを用いたネットワークが、全てのビットレートでより良い性能を発揮することを示している。

4.4.2 学習データの違いによる性能の違い

本節では、異なるデータセットにおける提案手法の性能差を評価するため、

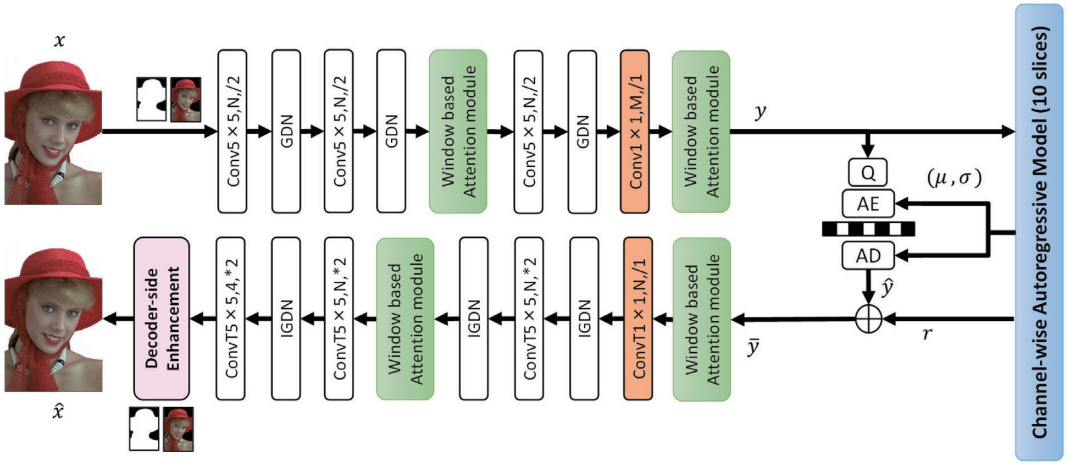

Fig. 10 4-channel network for comparative method.

4-channel network.

$$D(x,\hat{x}) = \left(MSE_{RGB} + MSE_{alpha}\right) \tag{12}$$

Figure 11 shows the graph of the results compared using $PSNR_{RGB}$ and Figure 12 shows the results compared using $PSNR_{alpha}$. In addition to the comparative method, the results of Minnen2018[11] and Zou2022[2] CNN-based and Swin Transformer-based models with 4-channel input and output layers are also shown. Figure 11 shows that there is a noticeable difference in performance at high bit rates. However, the performance difference becomes smaller at lower bitrates. In the case of $PSNR_{Alpha}$ in Figure 12, there is a large performance difference. The reason for the performance difference between the proposed method and the comparative method as shown in Figure 12, can be attributed to the fact that the comparative method utilizes a single entropy model shared between the RGB and alpha channels, making the optimization process more complex. In contrast, the proposed method employs separate entropy models for the RGB and alpha channels, allowing the entropy model to be optimized specifically for the alpha channel. Note that the graph of the proposed method is flat as one moves to the right because λ is fixed.

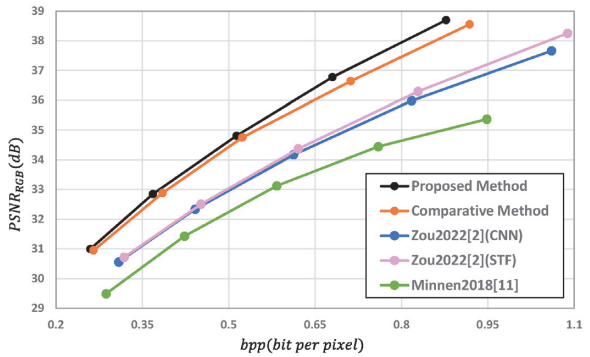

Fig. 11 Performance difference between the proposed method and 4-ch Network compared with $PSNR_{RGB}$.

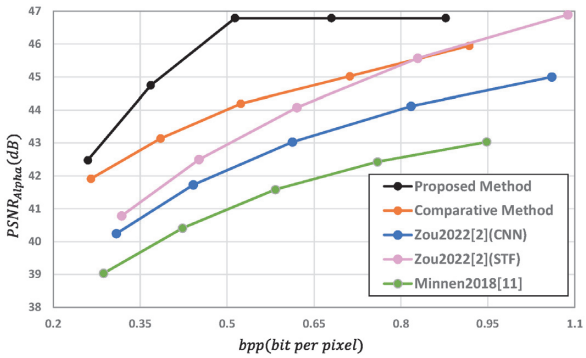

Fig. 12 Performance difference between the proposed method and 4-ch Network compared with $PSNR_{Alpha}$.

Therefore, by processing the RGB and alpha images separately, the training and network construction can consider the signal characteristics of each. Furthermore, each bit rates of the RGB and alpha channel can be changed independently, making them more practical to use.

4.4 Ablation study

4.4.1 Masked window-based attention module performance

To demonstrate the effectiveness of the Masked window-based attention module, a comparison is made with the regular window-based attention module. This regular window-based attention module is the same one used in our previous study [8]. Experimental conditions are the same as in 4.1.

Figure 13 shows that the network with the proposed method, the masked window-based attention module, performs better at all bit rates.

4.4.2 Performance differences due to differences in training data

In this section, to evaluate the performance

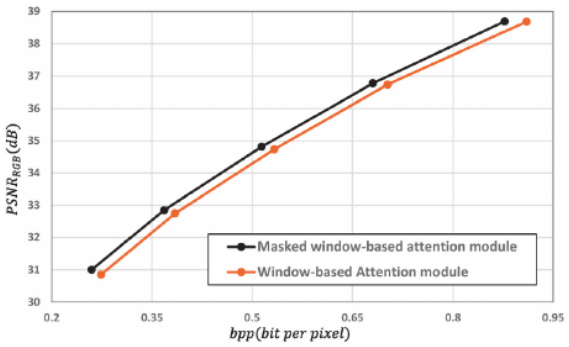

図13 RGBネットワークにおける注意モジュール間の性能差。

COCOデータセットとP3M-10kの両方で学習したネットワークと、COCOデータセットのみで学習したネットワークで実験を行った。学習方法は上記の方法と同じである。図14は、Masked Kodakデータセットに対する定量評価の実験結果である。図14の実験結果から、COCOデータセットとP3M-10kデータセットでの学習により、性能が向上することがわかる。図15は、定性的評価のイメージでもある。図15(b)、図15(c)、図15(d)の左上隅の木の枝に注目すると、図15(c)の2つのデータセットで学習したモデルは、図15(b)の元の画像と比較して、同じ量の枝を表現することができるが、図15(d)のCOCOデータセットのみで学習したモデルは、枝の量が明らかに減少していることがわかる。これらの実験結果の理由は、バイナリデータであるCOCOデータセットだけでは複雑な形状を捉えることが難しいためである。

4.4.3 透明領域の取り扱いがRGBネットワークの性能に与える影響の分析

実験では、データ量を減らすために

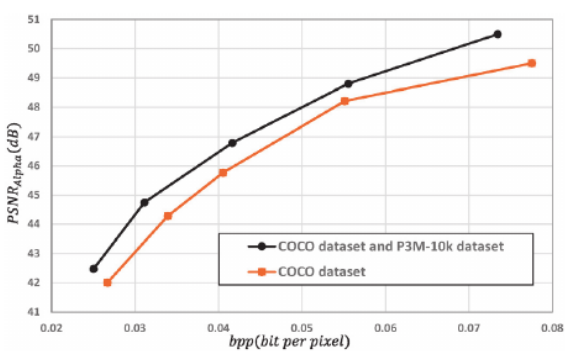

図14 学習データの違いによるアルファネットワークの性能差

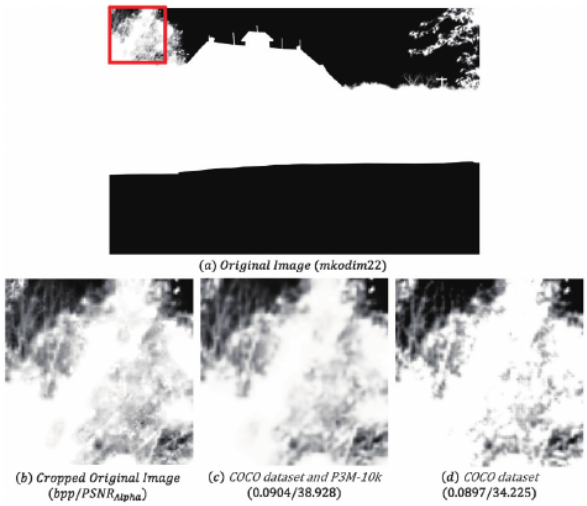

図15 異なるデータセットからの出力画像の違い。

RGBエンコーダの入力として、透明領域のRGB画像の画素値を0に設定した画像を用いて、RGBエンコーダの学習と評価を行う。しかし、提案する注意モジュールのように、学習時にα画像がRGB画像に作用するため、透明領域のRGB画像の画素値が0に設定されていない画像でRGB画像を学習することも重要であるため、透明領域のRGB画像の画素値をそのまま保持したまま学習を行い、その結果を比較・解析する。この比較により、アルファ画像でマスクされた領域の影響を考慮した圧縮方法が確立されることが期待される。実験結果を図 16 に示す。

図16において、w/ゼロ処理は、学習時に透明領域のRGB画像の画素値をゼロに設定した場合であり、w/oゼロ処理は、

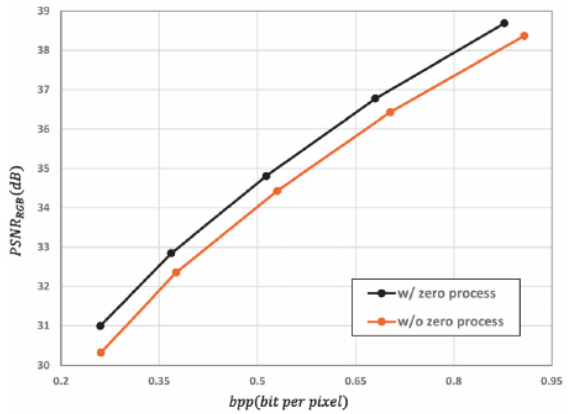

図16 RGB画像中の透明領域に対する学習方法による出力画像の違い。

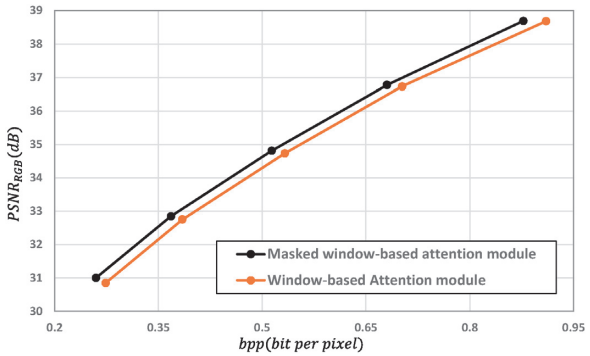

Fig. 13 Performance differences between different attention modules in RGB network.

differences of the proposed method on different datasets, experiments were conducted on networks trained on both the COCO dataset and P3M-10k, and on networks trained on the COCO dataset alone. The learning method is identical to the method described above. Figure 14 shows the experimental results in the quantitative evaluation for the Masked Kodak dataset. From the experimental results in Figure 14 show that training on the COCO dataset and the P3M-10k dataset improves performance. Figure 15 is also an image of the qualitative evaluation. Focusing on the tree branches in the upper left corner in Figures 15(b), 15(c), and 15(d) the model trained with the two datasets in Figure 15(c) can represent the same amount of branches compared to the original image in Figure 15(b), while the model trained with only the COCO dataset in Figure 15(d) clearly shows a decrease in the amount of branches. The reason for these experimental results is that it is difficult to capture complex shapes using only the COCO dataset, which is binary data.

4.4.3 Analysis of the impact of handling transparent regions on RGB network performance

In the experiments, to reduce the amount of data, the

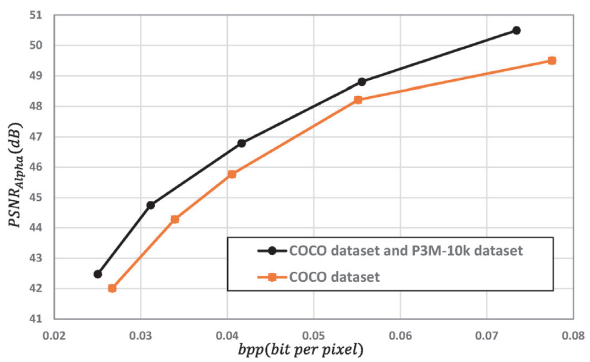

Fig. 14 Alpha network performance differences due to differences in training data.

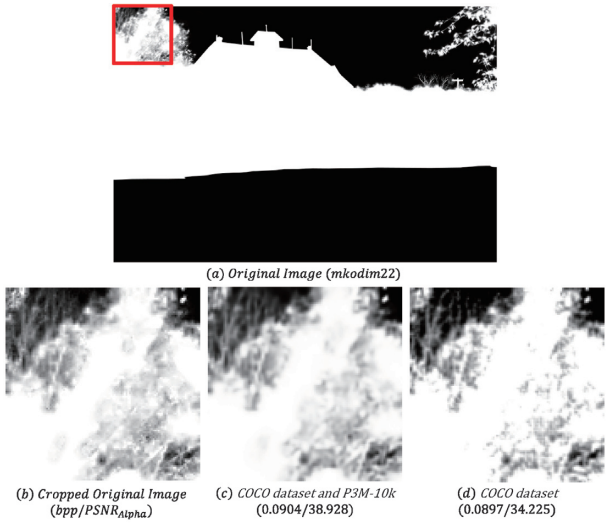

Fig. 15 Differences in output images from different datasets.

RGB encoder is trained and evaluated with images in which the pixel values of the RGB images in the transparent areas are set to 0 as input to the RGB encoder. However, since the fact that the alpha image acts on the RGB image during training, such as in the proposed attention module, it is also important to train the RGB image with images in which the pixel values of the RGB image in the transparent areas are not set to 0. Therefore, training is performed with the pixel values of the RGB image of the transparent areas is retained as it is, and the results are compared and analyzed. This comparison is expected to establish a compression method that takes into account the effect of areas masked by the alpha image. The experimental results are shown in Figure 16.

In Figure 16, the w/ zero process is the case where the pixel values of the RGB image in the transparent region are set to zero during training, and the w/o zero process is

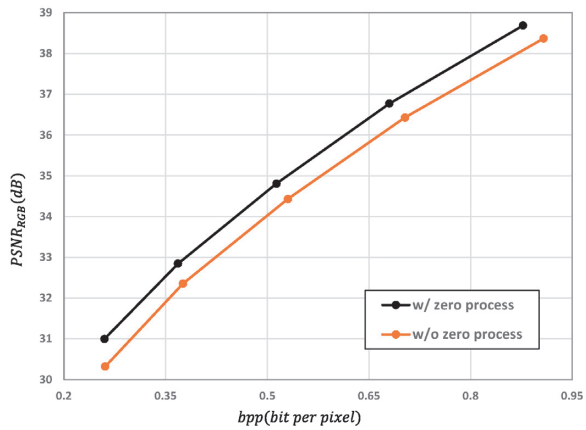

Fig. 16 Differences in output images based on training methods for transparent regions in RGB images.

ゼロに設定しない場合の結果である。図16より、透明領域のRGB画像の画素値を0にした場合の方が、ゼロにしない場合よりも、総合性能が高いことがわかる。

4.4.4 簡略化された注意モジュールがアルファネットワークの性能に与える影響の分析

本研究では、簡略化された注意モジュールを導入することで、従来のアルファネットワークの性能を向上させることを目的とする。このモジュールの性能を評価するために、簡略化した注意モジュールを利用しないモデルを用いて、同一の実験条件下で再トレーニングを行い、定量的な評価結果を比較した。図17は、アテンションモジュールの有無による定量的な評価結果である。w/簡略化された注意モジュール」とは、簡略化された注意モジュールがエンコーダとデコーダの中間層と最終層の両方に導入されるモデルを指す。一方、”w/簡略化された注意モジュールは中間層のみ ”は、注意モジュールが中間層のみに導入されたモデルを示す。実験の結果、中間層と最終層の両方にアテンションモジュールを導入したモデルは、特に中程度のビットレートで優れた性能を示すことが明らかになった。高ビットレートでは、「w/中間層のみの単純化された注意モジュール」モデルが優れた性能を示したことに注意。しかし、セクション4.2と4.3の実験では、アルファチャネルの高ビットレートデータを利用しなかったので、中層と最終層の両方で単純化された注意モジュールを組み込んだモデルが採用され、中ビットレートでより良い性能が実証された。

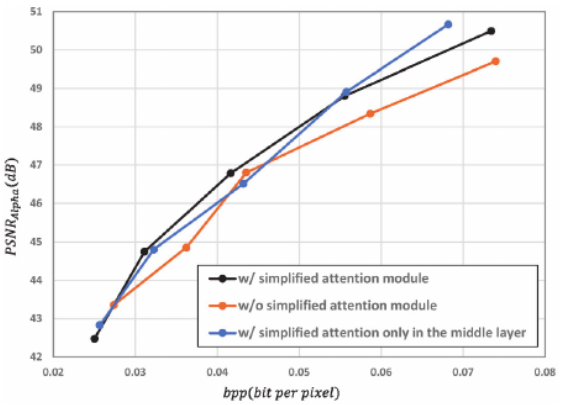

図17 単純な注意モジュールの有無による性能差。

5 おわりに

ディープラーニングを用いたRGBA形式の画像圧縮のための新しいネットワークアーキテクチャを提案する。提案するネットワークアーキテクチャは、RGB信号用とαチャンネル用の2つのネットワークから構成される。RGBネットワークのために、アルファチャンネルを用いた新しい注意モジュールMasked window-based attentionモジュールを導入する。実験結果より、提案手法は全てのビットレートにおいて、既存手法よりも優れた性能を持つことが示された。また、提案手法は、古典的なRGBA圧縮手法のように、入出力層を単純に4チャンネルに拡張する学習済み圧縮手法を凌駕することを実証する。また、マスクされたウィンドウベースのアテンションモジュールにより、従来のウィンドウベースのアテンションモジュールと比較して圧縮効率が改善されること、アルファチャネルネットワーク用の別のアテンションモジュールにより、より高いビットレートでの圧縮効率の改善に貢献できることがわかった。さらに、αネットワークの性能は学習データによって大きく異なることがわかった。今後の課題としては、損失関数に対するRGBA画像の画像評価指標の検討が挙げられる。例えば、MSEは本論文で提案した手法と同様に、視覚的に優れた復号化画像を得ることができない。したがって、マスク画像を考慮したMS-SSIMのような画像評価指標を開発する必要がある。また、本研究で採用したエントロピーモデルはマスク画像を用いずに圧縮するため、画像圧縮の効率を上げるためには、マスク領域を考慮した圧縮方法を導入することが重要である。

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22K12098.

References

1) J. Ballé, V. Laparra and E.P. Simoncelli: "End-to-end optimized image compression," 5th International Conference on Learning Representations (2017)
2) R. Zou, C. Song and Z. Zhang: "The devil is in the details: Window-based attention for image compression," IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
3) Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, Stephen Lin and Baining Guo: "Swin transformer: Hierarchical vision transformer using shifted windows," IEEE/CVF International Conference on Computer Vision (2021)
4) J. Liu, H. Sun and J. Katto: "Learned image compression with mixed transformer-cnn architectures," IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
5) T. Boutell: RFC 2083 - PNG (Portable Network Graphics)

the result when they are not set to zero. Figure 16 shows that the overall performance is higher when the pixel values of the RGB images in the transparent regions are set to zero than when they are not set to zero.

4.4.4 Analysis of the impact of the simplified attention module on the performance of the alpha network

This study aims to improve the performance of conventional alpha network by introducing a simplified attention module. To evaluate performance this module, we conducted re-training under identical experimental conditions with models that do not utilize the simplified attention module and compared the quantitative evaluation results. Figure 17 presents the quantitative evaluation results based on the presence or absence of the attention module. The "w/ simplified attention module" refers to models where the simplified attention module is introduced in both the intermediate and final layers of the encoder and decoder. In contrast, "w/ simplified attention module only in the middle layer" indicates models where the attention module is introduced only in the intermediate layer. Experimental results reveal that models with the attention module introduced in both the intermediate and final layers exhibit superior performance, particularly at medium bitrates. Note that at high bitrates, the model "w/ simplified attention module only in the middle layer" exhibited superior performance. However, since the experiments in Sections 4.2 and 4.3 did not utilize high-bitrate data for the alpha channel, a model incorporating simplified attention modules in both the middle and final layers, which demonstrated better performance at medium bitrates, was adopted.

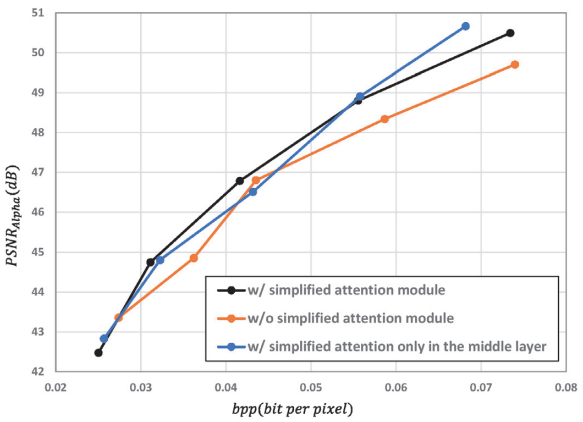

Fig. 17 Performance differences with and without simple attention module.

5. Conclusion

We propose a new network architecture for RGBA format image compression using deep learning. The proposed network architecture consists of two networks, one for RGB signals and the other for alpha channel. For the RGB network, a new attention module Masked window-based attention module using alpha channel is introduced. Experimental results show that the proposed method has better performance than existing methods at all bit rates. We also demonstrate the proposed method outperforms the learned compression method that simply extends the input and output layers to 4 channels like classical RGBA compression methods. We also found the compression efficiency can be improved by masked window-based attention module compared with the previous window-based attention module and another attention module for alpha channel network can contribute to improved compression efficiency at higher bit rates. Furthermore, it was found that the performance of the alpha network varies significantly depending on the training data. Future work includes investigation of an image evaluation metric for RGBA images for the loss function. For example, MSE, like the method proposed in this paper, cannot obtain a visually superior decoded image. Therefore, it is necessary to develop an image evaluation metric such as MS-SSIM that takes mask images into account. In addition, since the entropy model employed in this study compresses without using mask images, it is important to introduce a compression method that takes mask regions into account in order to achieve further efficiency in image compression.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22K12098.

References

1) J. Ballé, V. Laparra and E.P. Simoncelli: "End-to-end optimized image compression," 5th International Conference on Learning Representations (2017)
2) R. Zou, C. Song and Z. Zhang: "The devil is in the details: Window-based attention for image compression," IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
3) Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, Stephen Lin and Baining Guo: "Swin transformer: Hierarchical vision transformer using shifted windows," IEEE/CVF International Conference on Computer Vision (2021)
4) J. Liu, H. Sun and J. Katto: "Learned image compression with mixed transformer-cnn architectures," IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
5) T. Boutell: RFC 2083 - PNG (Portable Network Graphics)

Specification Version 1.0. Internet Engineering Task Force (1997)
6) F. Bellard: BPG image format, https://bellard.org/bpg/
7) G.J. Sullivan, J. Ohm, W. 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 (Dec. 2012)
8) Y. Inazu and H. Kimata: "Study on Learned RGBA Image Compression Using Loss Function Based on Alpha Channel," International Conference on Image, Video and Signal Processing (2024)
9) AV1 Image File Format (AVIF), https://aomediacoDec. github.io/av1-avif/
10) J. Ballé, D. Minnen, S. Singh, S.J. Hwang and N. Johnston: "Variational image compression with a scale hyperprior," International Conference on Learning Representations (2018)
11) D. Minnen, J. Ballé and G. Toderici: "Joint autoregressive and hierarchical priors for learned image compression," International Conference on Neural Information Processing Systems (2018)
12) D. Minnen and S. Singh: "Channel-wise autoregressive entropy models for learned image compression,". IEEE International Conference on Image Processing (2020)
13) H. Akutsu and N. Takahiro: "End to End Learned ROI Image Compression," IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
14) T. Chen, H. L., Z. Ma, Q. Shen, X. Cao and Y. Wang: "End-to-end learnt image compression via non-local attention optimization and improved context modeling," IEEE Transactions on Image Processing, vol. 30, pp.3179-3191 (2021)
15) Z. Cheng, H. Sun, M. Takeuchi and J. Katto: "Learned image compression with discretized gaussian mixture likelihoods and attention modules," IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
16) J. Liu, G. Lu, Z. Hu and D. Xu: "A unified end-to-end framework for efficient deep image compression," arXiv preprint arXiv:2002.03370 (2020)
17) J. Ballé, V. Laparra and E.P. Simoncelli: "Density modeling of images using a generalized normalization transformation," International Conference on Learning Representations (2016)
18) T.-Y. Lin, M. Maire, S; Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick and P. Dollár: "Microso COCO: Common Objects in Context,". European Conference on Computer Vision (2014)
19) J. Li, S. Ma, J. Zhang and D. Tao: "Privacy-Preserving Portrait Matting,"ACM International Conference on Multimedia (2021)
20) Kodak Lossless True Color Image Suite, http://r0k.us/graphics/kodak/
21) Inazu: Masked-Kodak-Dataset, https://github.com/Yoshiki172/Masked-Kodak-dataset/
22) D.P. Kingma and J. Ba: "Adam: A method for stochastic optimization," International Conference on Learning Representations (2015)
23) libavi https://github.com/AOMediaCodec/libavif

Yoshiki Inazu graduated from the Department of Information Design, School of Informatics, Kogakuin University in 2023, and has been enrolled in the Informatics Program, Graduate School of Engineering, Kogakuin University since the same year, conducting research on image compression.

Hideaki Kimata received the B.E. and M.E. degrees in applied physics, and the Ph.D. degree in electrical engineering respectively from Nagoya University, Nagoya, Japan, in 1993, 1995, and 2006. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995, and was engaged in research and development of video coding systems, realistic communication, computer vision, and machine learning. He has also been involved with the development of MPEG standards under JTC 1/SC 29. He is currently a professor at the Department of Information Design, Faculty of Informatics, Kogakuin University. His research interests include image processing, video coding, video communication, computer graphics, He is a member of IEEE, IPSJ, and IEICE.

Specification Version 1.0. Internet Engineering Task Force (1997)
6) F. Bellard: BPG image format, https://bellard.org/bpg/
7) G.J. Sullivan, J. Ohm, W. 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 (Dec. 2012)
8) Y. Inazu and H. Kimata: "Study on Learned RGBA Image Compression Using Loss Function Based on Alpha Channel," International Conference on Image, Video and Signal Processing (2024)
9) AV1 Image File Format (AVIF), https://aomediacoDec. github.io/av1-avif/
10) J. Ballé, D. Minnen, S. Singh, S.J. Hwang and N. Johnston: "Variational image compression with a scale hyperprior," International Conference on Learning Representations (2018)
11) D. Minnen, J. Ballé and G. Toderici: "Joint autoregressive and hierarchical priors for learned image compression," International Conference on Neural Information Processing Systems (2018)
12) D. Minnen and S. Singh: "Channel-wise autoregressive entropy models for learned image compression,". IEEE International Conference on Image Processing (2020)
13) H. Akutsu and N. Takahiro: "End to End Learned ROI Image Compression," IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
14) T. Chen, H. L., Z. Ma, Q. Shen, X. Cao and Y. Wang: "End-to-end learnt image compression via non-local attention optimization and improved context modeling," IEEE Transactions on Image Processing, vol. 30, pp.3179-3191 (2021)
15) Z. Cheng, H. Sun, M. Takeuchi and J. Katto: "Learned image compression with discretized gaussian mixture likelihoods and attention modules," IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
16) J. Liu, G. Lu, Z. Hu and D. Xu: "A unified end-to-end framework for efficient deep image compression," arXiv preprint arXiv:2002.03370 (2020)
17) J. Ballé, V. Laparra and E.P. Simoncelli: "Density modeling of images using a generalized normalization transformation," International Conference on Learning Representations (2016)
18) T.-Y. Lin, M. Maire, S; Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick and P. Dollár: "Microsoft COCO: Common Objects in Context,". European Conference on Computer Vision (2014)
19) J. Li, S. Ma, J. Zhang and D. Tao: "Privacy-Preserving Portrait Matting,"ACM International Conference on Multimedia (2021)
20) Kodak Lossless True Color Image Suite, http://r0k.us/graphics/kodak/
21) Y. Inazu: Masked-Kodak-Dataset, https://github.com/Yoshiki172/Masked-Kodak-dataset/
22) D.P. Kingma and J. Ba: "Adam: A method for stochastic optimization," International Conference on Learning Representations (2015)
23) libavif, https://github.com/AOMediaCodec/libavif

Yoshiki Inazu graduated from the Department of Information Design, School of Informatics, Kogakuin University in 2023, and has been enrolled in the Informatics Program, Graduate School of Engineering, Kogakuin University since the same year, conducting research on image compression.

Hideaki Kimata received the B.E. and M.E. degrees in applied physics, and the Ph.D. degree in electrical engineering respectively from Nagoya University, Nagoya, Japan, in 1993, 1995, and 2006. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995, and was engaged in research and development of video coding systems, realistic communication, computer vision, and machine learning. He has also been involved with the development of MPEG standards under JTC 1/SC 29. He is currently a professor at the Department of Information Design, Faculty of Informatics, Kogakuin University. His research interests include image processing, video coding, video communication, computer graphics, He is a member of IEEE, IPSJ, and IEICE.