CI-001

符号化特徴と復元画像の相互情報量最大化に基づく GAN ベース画像符号化方式の検討

工藤 忍 † Shinobu Kudo　折橋 翔太 † Shota Orihashi　谷田 隆一 † Ryuichi Tanida　清水 淳 † Atsushi Shimizu

1. はじめに

近年，画像データの高精細化や映像サービスの普及により映像トラフィックが増加し，データ圧縮率向上の需要が高まっている．H.265/HEVC[1] に代表されるブロック単位の予測と変換に基づく符号化方式では，高周波数成分を多く含む複雑なテクスチャに対しては効率的な予測が行えないため，低ビットレートにおいて主観画質の低下が問題となっている．これに対してニューラルネットワークを用いた学習ベースの非線形予測・変換により，複雑なテクスチャに対する復元精度を改善する手法が多数提案されている．その中で主観画質改善を目指した手法の 1 つに，敵対的生成ネットワーク (GAN)[2] と呼ばれる枠組みを用いて，復元画像の分布が自然画像の分布に近づくように最適化する手法 [3][4] が報告されている．しかし，これらの手法は復元画像が主観的に自然であるか否かに重点を置いて最適化されるため，エンコーダから得られる符号化特徴に原画像とは無相関な成分が混入し，見た目は自然であっても原画像と別の物体に見えたり，印象が変わって見えたりする場合がある．

本稿では符号化特徴と復元画像の相互情報量に着目し，その最大化を正則化に導入することでエンコーダから得られる符号化特徴が明示的に原画像の属性情報として抽出されるように最適化を図り，主観的に自然で且つ見た目の印象変化を低減する画像符号化方式を提案する．

2. 従来技術

2.1. GAN ベース画像符号化

GAN ベースの画像符号化手法 [3][4] はオートエンコーダに基づく画像符号化手法に GAN[2] の枠組みを導入したもので，エンコーダ E，デコーダ G，ディスクリミネータ D の 3 つのネットワークにより構成される．エンコーダは入力画像 x を符号化特徴 $w(=E(x))$ にマッピングし，量子化演算 Q を施し，ビットストリーム $z(=Q(w))$ を出力する．デコーダはビットストリームから復元画像 $\hat{x}(=G(z))$ を得る．ディスクリミネータは入力画像と復元画像を判別し，エンコーダ及びデコーダと敵対的に学習する．本手法は下記の目的関数を最適化する min-max 問題として定式化される．

$$\min_{E,G}\max_{D} L = \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))] + \lambda_d \mathbb{E}[d(x, G(z))] + \lambda_r r(z). \quad (1)$$

ここで $d(x,\hat{x})$ は入力画像 x と復元画像 $\hat{x}$ の再構成誤差を表す関数，λ_d は再構成誤差の重み係数，$r(z)$ は符号量を表す関数，λ_r は符号量の重み係数を表す．また f と g はスカラ関数で様々な提案がされているが，$f(y)=\log(y)$, $g(y)=\log(1-y)$ としたものは Vanilla GAN[2] と呼ばれ，x と $G(z)$ の KL ダイバージェンスの最小化に対応する．第 1 項及び第 2 項が敵対的生成誤差を表し，これらが加わることによって復元画像が自然な画像であるか否かが考慮され，主観的に自然な出力となるように最適化される．

2.2. 従来技術の問題点

従来技術は復元画像が主観的に自然な画像であるか否かに重点を置いて最適化されるため，敵対的生成誤差を考慮せずに最適化を行なった場合と比較して，エンコーダから得られる符号化特徴に原画像とは無相関な成分が混入する．これにより見た目は自然であっても原画像と別の物体に見えたり，印象が変わって見えたりする場合がある．特に人間の顔や文字などの認識に関わる画像に対して，人の目は変化に敏感であると考えられるため，例えば人間の顔を入力とした場合，顔の表情や目の形，髪型などが微かに変化しただけでも，別人に見えたり印象が変わったりしてしまう．これは符号化という本来の機能が失われてしまっているため好ましくない．

3. 提案法

3.1. 着眼点

従来手法の課題を解決するためにはエンコーダから得られる符号化特徴に原画像とは無相関な成分が混入しないようにする必要がある．すなわち符号化特徴と復元画像が完全に相関を持つような最適化が要求される．言い換えると，これは符号化特徴 w と復元画像 $G(z)$ の相互情報量 $H(G(z)|w)$ の最大化と等価となる．そこで本稿ではどのように符号化特徴と復元画像の相互情報量の最大化を目的関数に導入するかについて示す．提案法は符号化特徴が明示的に復元画像に対して相関のある特徴として抽出されるため，ランダム性のある処理が抑制され，復元画像が主観的に自然でありつつも見た目の印象変化を低減することができ，符号化タスクに適した符号化器の実現が期待される．

3.2. 相互情報量最大化に基づく GAN ベース画像符号化

提案法による目的関数を以下に示す．

$$\min_{E,G}\max_{D} L = \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))] + \lambda_d \mathbb{E}[d(x, G(z))] + \lambda_r r(z) - \lambda_I I(w; G(w)). \quad (2)$$

ここで λ_I は重み係数，$I(w;G(w))$ は符号化特徴と復元画像の相互情報量を表す．

$I(w;G(w))$ は事後分布 $P(w|G(w))$ によって解析的に解けないため補助分布 F を導入し，変分情報最大化

†日本電信電話株式会社, NTT メディアインテリジェンス研究所

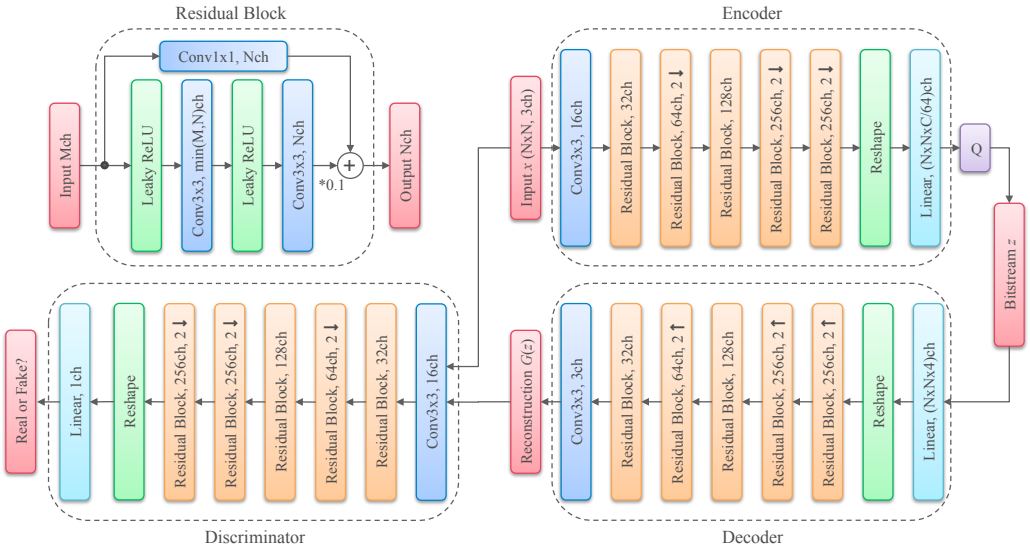

図 1 ネットワーク構成

として知られるテクニックを用いて下記のように変分下限 $L_I(G,F)$ で表す [5].

$$\begin{aligned} I(w;G(w)) &= H(G(w)) - H(G(w)|w) \\ &\geq \mathbb{E}_{x\sim G(w)}[\mathbb{E}_{w'\sim P(w|x)}[\log F(w'|x)]] + H(w) \\ &= L_I(G,F). \end{aligned} \tag{3}$$

ここで H はエントロピー，P は確率分布である．事後分布 $P(w|G(w))$ すなわち $F(w'|x)$ は任意の分布でモデリングする必要があるが，本稿では正規分布 $N(0,1)$ と仮定し，実装上は E で代用した．

したがって，提案法の最終的な目的関数は下記の通りとなる．

$$\begin{aligned} \min_{E,G}\max_{D} L = &\mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))] \\ &+\lambda_d \mathbb{E}[d(x, G(z))] + \lambda_r r(z) \\ &-\lambda_I L_I(G,E). \end{aligned} \tag{4}$$

4. 評価実験

提案法の有効性を確認するため，シミュレーション実験を行った．

4.1. ネットワーク構成

ネットワーク構成は ResNet ベース（図 1）で，エンコーダの出力サイズは $N \times N \times C/64$（画像サイズを $N \times N$ とする）とした．各ネットワークの最終層を除く各層の出力に LeakyReLU($\alpha = 0.2$)，エンコーダ及びデコーダの出力に tanh 関数を適用し，量子化器 Q は各要素を 1bit{-1, 1} に量子化した．なお量子化後のデータに対してはエントロピー符号化等は行わず，そのままビットストリームとした．また f と g には Vanilla GAN[2] を採用した．

4.2. 実験条件

HEVC ベースの画像符号化手法である BPG[6] 及び従来手法 [4] との比較を行った．本実験では目的関数の影響のみを比較するため，従来手法のネットワーク構成は提案法と共通とした．BPG は量子化パラメータを変化させ，従来手法及び提案手法の符号量に最も近いものとした．データセットは今回，人の目が特に変化に敏感と思われる厳しい条件とするため，人間の顔のデータセット（CelebA[7]）を使用し，20 万枚の学習データと 25 枚のテストデータを抽出し，128×128 画素にリサイズして用いた．最適化は Adam[8] を用い，バッチサイズは 16，学習率は 0.0001，反復回数は 200,000 回に設定し，R_1 正則化 [9] と重み平均法 [10] を用いた．また $d(x|G(z))$ には平均二乗誤差，$P(w|x)$ は $[-1,1]$ の一様分布を使用し，その他のパラメータはそれぞれ $C = 8$，$\lambda_d = 10.0$，$\lambda_r = 0$，$\lambda_I = 10.0$ に設定した．

4.3. 評価方法

画像専門家 10 人を被験者に主観画質評価実験を行った．評価法には各評価画像に対する画質及び基準画像（原画像）との劣化度合いをそれぞれ 5 段階の評価スコアでつける Absolute Category Rating（ACR）法及び Degradation Category Rating（DCR）法 [11] を採用した．なお ACR 法では評価画像を単独で見た時の画像の自然さを表す「自然度」を，DCR 法では基準画像と比べた時の類似さを表す「類似度」を，それぞれ被験者が評価する指標とした．

4.4. 結果及び考察

図 2 に従来手法と提案手法それぞれの学習時における相互情報量の下限値 $L_I(G,E)$ の推移を示す。図 2 より，提案手法は従来手法と比べて下限値が増加し上限近く（≒ 0.5）に達しており，相互情報量の最大化が実現できていることが確認される。

図 3 に主観画質評価結果を示す．図 3 の評価スコアは評価画像 25 枚それぞれの全被験者の平均スコアを示している．この結果から BPG は自然度，類似度共に 1.2 と著しく低く，符号量が不足していることが推察される．従来手法と提案手法は自然度がそれぞれ 3.4 及び 3.6 とほぼ同程度を示したが，類似度は従来手法が

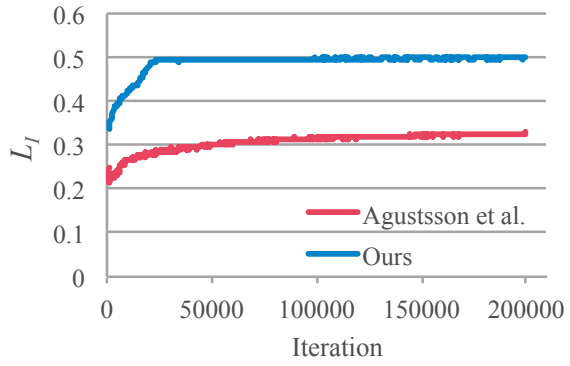

図 2 相互情報量の下限値 L_I の推移

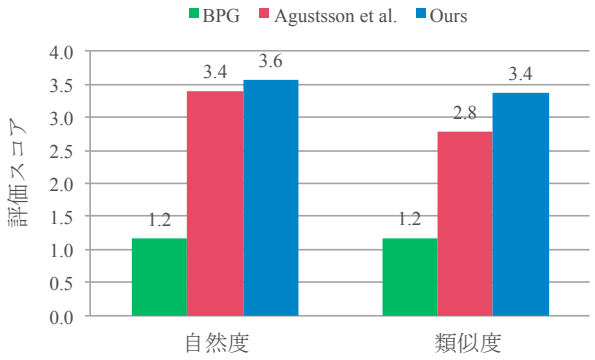

図 3 主観画質評価結果

2.8，提案手法が 3.4 と提案手法が 0.6 ポイント上回った．このことから提案手法は従来手法と同等以上の復元画質を示しながらも，原画像により近い画像を復元できていることが確認された．

図 4 に各手法による復元画像を示す．主観画質評価結果が示したように BPG は高周波数成分が失われており，著しい画質の低下が確認される．従来手法と提案手法はテクスチャが保存されており，どちらも自然な復元を実現できている．しかし従来手法は原画像と比較すると (a), (b) では目線が，(c), (d) では口元の表情が明らかに変化しており，原画像からの印象と大きく異なっていると考えられる．一方で提案手法は目線や表情が一致しており，見た目の印象変化が低減されていることが確認できる．これらが影響して主観画質評価結果の従来手法と提案手法の類似度スコアに差が表れたと考えられる．

5. まとめ

本稿では符号化特徴と復元画像の相互情報量最大化を正則化に導入した GAN ベース画像符号化方式を提案した．主観画質評価実験により提案手法は従来手法と同等以上の画質で且つ主観的に見た目の印象変化を低減できることを示した．

参考文献

[1] G. J. Sullivan J. -R. Ohm, W. J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard,” IEEE Trans. Circuits and Systems for Video Tech., Vol. 22, No. 12, pp. 1649-1668, Dec. 2012.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” Advances in Neural Information Processing Systems, pp.2672-2680, Dec. 2014.

[3] O. Rippel, and L. Bourdev, “Real-Time Adaptive Image Compression,” arXiv preprint arXiv:1705.05823v1, May 2017.

[4] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, “Generative Adversarial Networks for Extreme Learned Image Compression,” arXiv preprint arXiv:1804.02958, Apr. 2018.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,” arXiv preprint arXiv:1606.03657, June 2016.

[6] F. Bellard, “BPG Image format,” https://bellard.org/bpg/.

[7] Z. Liu. P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” Proceedings of International Conference on Computer Vision (ICCV), 2015.

[8] D. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6948, 2014.

[9] L. Mescheder, A. Geiger, and S. Nowozin, “Which Training Methods for GANs do actually Converge?,” arXiv preprint arXiv:1801.04406v4, July 2018.

[10] Y. Yazici, C. -S. Foo, S. Winkler, K. -H. Yap, G. Piliouras, and V. Chandrasekhar, “The unusual effectiveness of averaging in GAN training,” arXiv preprint arXiv:1806.04498v2, Feb. 2019.

[11] Recommendation ITU-T P. 910, “quality assessment methods for multimedia applications,” Apr. 2008.

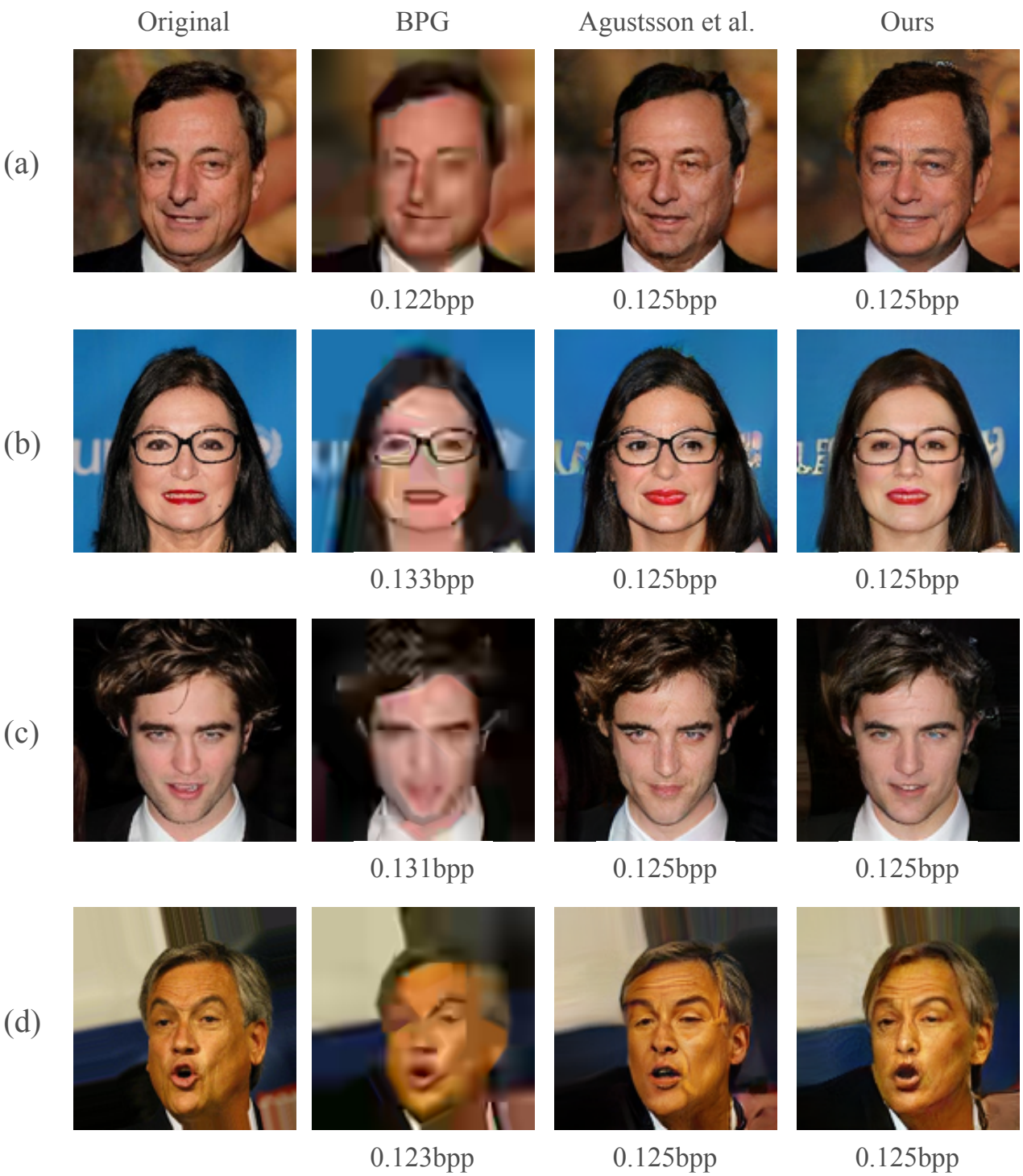

図 4 各手法による復元画像（左から原画像，BPG[6]，Agustsson[4]，提案手法）