線形化した知覚品質予測と Stein の不偏リスク推定による画像復元手法の研究

代表研究者　　宮田　高道　　千葉工業大学　先進工学部知能メディア工学科　教授

1 はじめに

本研究の目的は，多様な劣化を受けた画像の知覚品質を単一の学習済みモデルで向上させる手法を開発することである．この目的を達成するために当初予定していた研究の第 1 段階は，深層学習によって獲得された静止画像品質の評価手法である LPIPS (Learned Perceptual Image Patch Similarity) [1]の線形近似と，Stein の不偏リスク推定 (SURE) [2]を組み合わることで知覚品質の高い画像ノイズ除去を実現することであった．これを元に，続く段階２では，学習済みの画像復元手法と交互方向乗数法 (ADMM) を組みあわせることにより，ノイズ除去手法の提供範囲を一般の画像復元へと拡張することを目標としていた．

このうちの段階 1 については，2024 年度に基礎的な検討を行い，線形化した LPIPS が知覚品質をある程度の精度で推定できることを明らかにした (**2-1 節**)．この検討と並行して，知覚品質の予測と知覚品質を考慮した画像復元を実現するための様々な研究を行った．本報告書では，それらの成果についても述べる．まず 2023 年の前半に，ビジョン言語モデルの一種である CLIP (Contrastive Language-Image Pretraining)[3] を用いて，原画像を参照することなく画像の知覚品質を評価する手法を開発することに成功した．この手法を更に発展させ，2024 年には IEEE ACCESS (IF=3.9)に論文を発表した．これらの手法の開発を通し，画像の知覚品質を原画像なしで評価するという当初の第 1 段階の目的を達成したといえる．これらの成果については，**2-2 節**で述べる．続く第２段階については，まず JPEG 画像を対象とした符号化ノイズの除去において，知覚品質を向上させることを目的とした手法を開発し，国際会議である KJCCS 2024 にて発表した他，その手法を発展させて IEICE NoLTA Journal に論文を発表した (**2-3 節**)．並行して，ノイズの付加された画像に対する構造-テクスチャ分離を実現する手法を開発し，その成果を IEEE ICAIIC 2024 にて発表した (**2-4 節**)．さらに，近年急速に研究が進んでいる拡散画像生成モデルを核して様々な画像復元を実現する手法を潜在拡散モデルへと拡張することにより，知覚品質の尺度の一つである Fréchet Inception Distance (FID)を改善することにも成功した．この手法は，IEEE ICAIIC 2025 にて発表した (**2-5 節**)．以上の取り組みによって，研究の第２段階についても，当初予定とはやや異なる経路を辿ったものの，その目的は十分に達成できたといえる．

2 研究成果の概要

2-1 線形化した知覚品質予測の精度の検証

知覚品質の推定手法として知られる Learned Perceptual Image Patch Similarity (LPIPS)の線形近似手法，および線形化した LPIPS （ここでは Linearized Relative Perceptual Similarity, LRPS とよぶ）を様々な劣化要因を含む画像に適用し，その知覚品質推定精度を検証した．

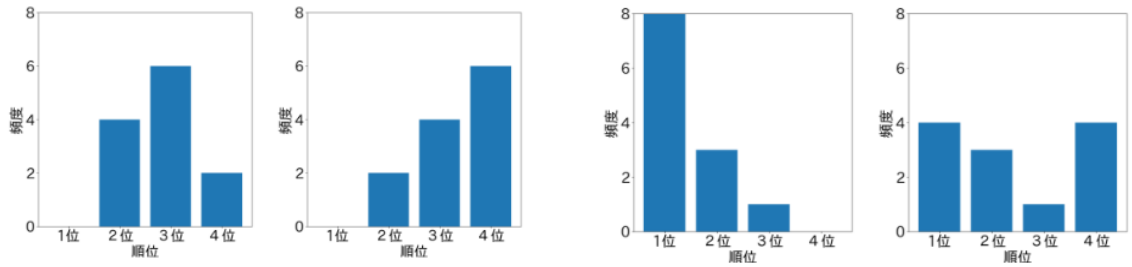
LPRS は，LPIPS が深層特徴量の抽出に用いている VGG-16 の最初の畳み込み層を取り出し，バイアスおよび活性化関数を除去したうえで，チャンネルごとに（LPIPS で学習済みの）重みをかけたものである．一般的な知覚品質データセットである TID2008, TID2013, LIVE, CSIQ で，ベースラインである PSNR, SSIM, LPIPS と LRPS の予測精度を比較した．評価尺度としてはピアソン相関係数，スピアマン順位相関，ケンドール順位相関の 3 つを用い，各データセットおよび各評価尺度で LRPS の精度が何位となったかを調査した．その結果を図 1 に示す．

これらの結果より，提案する LRPS は 1 位を獲得する確率が LPIPS についで高く，このことから高い知覚品質予測精度を実現しているといえる．その一方で，予測精度が最下位となる確率が SSIM についで高かったことからは，性能が大幅に劣化するデータセット/評価指標の組合せが存在したことを示唆している．

この問題を検証するため，最も予測精度の低かった LIVE データセットにおいて，ヒトの知覚品質の平均値である MOS と，LRPS およびベースラインの手法である PSNR, SSIM, LPIPS の品質予測値の散布図を図 2 から図５に示す．図 5 より，LRPS では一部の画像で大幅に MOS と予測値とが大きく乖離しており，このことが予

公益財団法人電気通信普及財団
研究調査助成報告書　第40号　2025年度

測性能の低下の原因であることがわかる.



(a) PSNR (b) SSIM (c) LPIPS (d) LRPS

図 1　各手法が全データセットおよび指標の組み合わせについて獲得した順位のヒストグラム

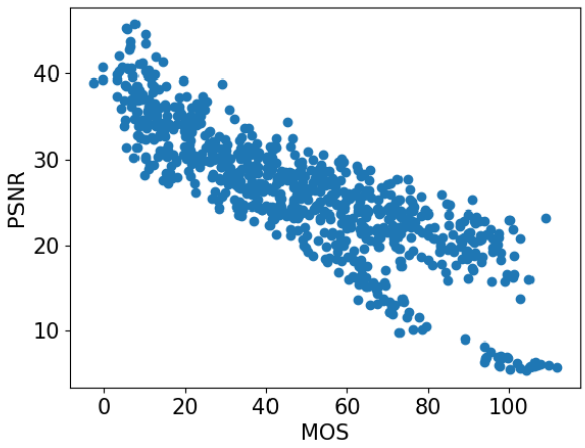


図 2: LIVE における MOS と PSNR との散布図

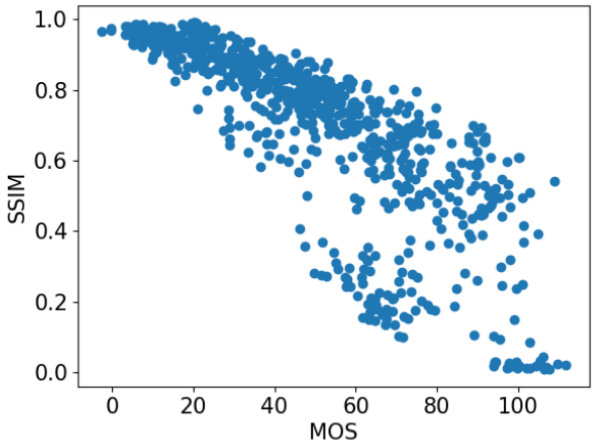


図 3: LIVE における MOS と SSIM との散布図

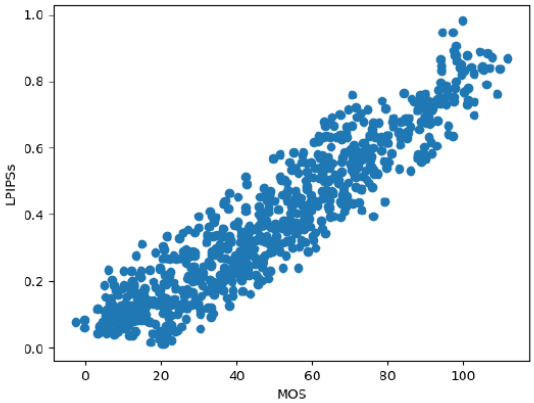


図 4: LIVE における MOS と LPIPSとの散布図.

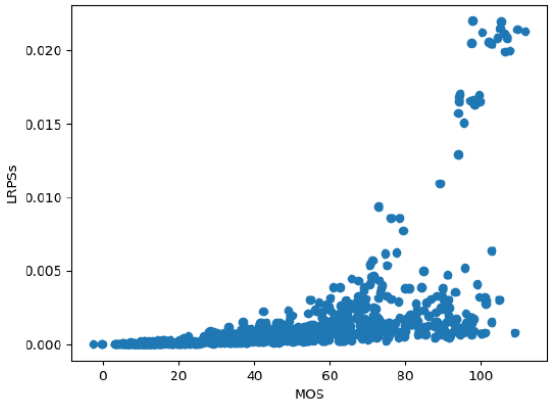


図 5: LIVE における MOS と LRPS との散布図.

2-2 ビジョン言語モデルを用いたゼロショットで参照なしの知覚品質評価手法の提案

2-2-1 概要

前節で述べたように，本研究の第 1 段階では，深層学習によって獲得された静止画像品質評価手法である LPIPS を線形近似した手法について，その性能を評価した．一方で，研究を推進する過程で，新たな技術の進歩をふまえて知覚品質評価のための異なるアプローチを模索した．まず 2023 年前半に，ビジョン言語モデルの一種である CLIP を活用したゼロショット（対象タスクやデータセットでの学習を必要としない枠組み）で説明可能な参照なし画像品質評価手法として，「Interpretable Image Quality Assessment via CLIP with Multiple Antonym-Prompt Pairs」を 2023 IEEE 13th International Conference on Consumer Electronics – Berlin (ICCE-Berlin) で発表した．この手法は，単に画像の知覚品質を評価するだけでなく，その評価の根拠となった特徴を特定できるという点で，高い説明性を有している．

具体的には，例えば「ぼやけている」と「鮮明である」，「ノイズが多い」と「ノイズがない」といった対義語のプロンプトペアを複数用意し，それぞれのペアに対する画像の類似度スコアを算出することで，画像がどのような劣化を受けているのかを言語的に説明することを可能にした．このアプローチにより，IQA の結果が単なる数値だけでなく，具体的な劣化要因を示すテキストとして提供されるため，画像処理システムのデバッグや改善に役立つ情報を提供できる点がこの手法の特徴である．実験では，様々な人工的な劣化（ガウシアンノイズ，JPEG 圧縮，ガウシアンブラーなど）や実世界における劣化（低照度，逆光など）に対して，提案手法が劣化の種類を正確に特定し，それぞれの劣化レベルに応じた品質評価を行うことを示した．また，人間の主観的知覚評価との比較においても，その説明性の高さが評価された．

この初期の成功に基づき，さらに研究を深化させ，2024 年には，この画期的な手法に関する詳細な研究成果を IEEE Access (Impact Factor: 3.9) に論文として発表した．この論文では，提案手法の理論的背景に加え，多数のベンチマークデータセットを用いた包括的な実験結果を示し，その性能が既存の最先端の参照なし IQA 手法を上回ることを明確に示した．特に，これまで評価が困難であった未知の劣化タイプに対しても，本手法が高い頑健性を示すことを明らかにした．

2-2-2 関連研究

NR-IQA の研究は，大きく手作り特徴量＋回帰モデルと深層学習ベースの二つに分かれる．前者は自然画像統計(NSS)を用いて歪み特徴を抽出し，SVR などの回帰器で品質スコアを推定する構成が一般的で，モデルの解釈性は高いものの，現実の劣化画像がもつ多様な歪みに対しては性能が頭打ちとなる [4, 5, 6, 7]．これに対して，CNN や Transformer を活用した後者の手法では，大規模データで学習することで高い相関を実現するが，訓練セット以外のデータセットへの適応性が低く，内在する判断根拠を人に示すことが困難である点が課題となる．近年，CLIP などの視覚と言語を連携させたモデルが登場し，学習済み画像-テキスト対を介してゼロショットで多様なタスクに対応可能であることが報告されている．IQA 領域においても，LIQE や CLIP-IQA [8]といった先行研究が CLIP を微調整あるいは簡易プロンプトによって品質評価を試みているが，依然として結果の説明性や各劣化因子の定量的把握には至っていない．

2-2-3 提案手法

提案手法である ZEN-IQA（**Z**ero-shot **E**xplainable **N**o-reference **IQA**）ではまず，画質に影響を与える要因を表現する一連のテキストプロンプトを定義する．具体的には，鮮鋭度（sharpness）やノイズの有無（noiselessness）といった対義語ペア（「sharp photo.／blurry photo.」など）および輝度の過不足を扱う対義語トリプレット（「bright photo.」「too bright, bad photo.」「too dark, bad photo.」）を用意する．次に，CLIP の画像エンコーダとテキストエンコーダでそれぞれ得られる埋め込み表現同士のコサイン類似度を計算し，特徴ごとのスコアを導出する．対義語ペアに対しては得られた類似度に softmax 関数を適用し，対義語トリプレットでは二つの softmax からより低い方の値を採用することで，過剰露光と露光不足のどちらにも対応可能としている．最後に，各特徴スコアを平均化し，0 から 100 の範囲に正規化した総合品質スコアを出力する．本手法は追加学習を必要とせず，プロンプト設計のみで

多様な歪み要因を捉えつつ，ヒトが理解しやすい説明情報を同時に提供する点が最大の特徴である．提案手法のアーキテクチャを図 6 に示す．

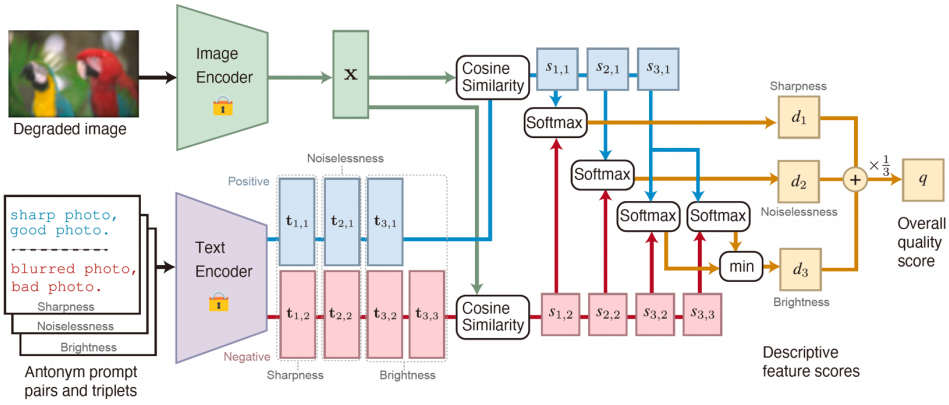


図 6.　2-2 節の提案手法 ZEN-IQA のアーキテクチャ．

2-2-3 実験結果

提案手法を KonIQ-10k，LIVE-in-the-wild，SPAQ といった大規模テストセットに適用した結果を表 1 に示す．この表より，提案手法が従来手法を上回る順位相関（SROCC）および線形相関（PLCC）を示すことがわかる．また，TID2013 や CSIQ でのクロスデータセット評価でも，深層学習モデルと同等以上の性能を保ち，学習データセット外での頑健性を実証している．さらに，各劣化要因に対応した特徴スコアは，暗転・過曝・ぼけ・ノイズなどの異なる歪みを直感的に反映し，画像処理フィルタ（例：LIME や BM3D）適用後の輝度変化やノイズ低減の効果を定量的に可視化できることを示した．これらの結果から，ZEN-IQA はゼロショット設定下での高精度かつ高い説明性を兼ね備えた NR-IQA 手法として有効であると結論付けられる．

Method	KonIQ-10k		LIVE-itW		SPAQ	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
BIQI [4]	0.559	0.616	0.364	0.447	0.591	0.549
BLIINDS-II [5]	0.585	0.598	0.090	0.107	0.317	0.326
BRISQUE [6]	0.705	0.707	0.561	0.598	0.484	0.481
NIQE [7]	0.551	0.488	0.463	0.491	0.703	0.670
CLIP-IQA [8]	0.695	0.727	0.612	0.594	0.730	0.727
ZEN-IQA (ours)	**0.776**	**0.796**	**0.672**	**0.664**	**0.757**	**0.750**

表 1.　2-2 節の提案手法 ZEN-IQA の定量評価．

2-2-4 結論

これらの成果により，当初の第 1 段階の目標であった「画像の知覚品質を原画像なしで評価する」という目的を，異なるアプローチからではあるが十分に達成したといえる．

2-3 JPEG 圧縮画像の知覚品質向上手法の提案

2-3-1 概要

本研究課題の第 2 段階として位置づけられる，知覚品質考慮型の画像復元手法については，まず JPEG 画像に特化した符号化ノイズ除去の研究をおこなった．一般的に用いられている JPEG 圧縮では，圧縮によって生じるブロックノイズやリンギングアーティファクトが，画像の知覚品質を大きく損なうことがある．これまでに，このようなブロックノイズ等の JPEG 符号化による画質劣化を改善する手法は数多く提案されてきたが，L1 ノルムや L2 ノルムを損失関数として用いる手法は知覚的品質との相関が低く，一方で GAN を用いた手法は学習の不安定性やモード崩壊を招きやすいという課題があった．本研究では，これらの問題を回避するために，複数の画像品質評価指標（IQA）を重み付きで組み合わせた損失関数を

導入し，さらにアップスケーリング層におけるチェッカーフラッグ状のアーティファクトの発生を抑制するアーキテクチャの改良を併せて提案した．この成果は，国際会議である KJCCS 2024 にて発表された後，さらに詳細な分析と改良を加え，その発展形が IEICE NoLTA Journal に論文として掲載された．この論文では，提案手法のアーキテクチャの詳細な説明に加え，複数の公開データセットを用いた広範な評価実験を行い，客観的指標（PSNR, SSIM, FID など）と主観評価の両面で，既存の最先端手法と比較して優位性があることを示した．

2-3-2 関連研究

JPEG アーティファクト除去手法には，初期の CNN 回帰モデルや DCT 領域を活用したマルチスケール CNN 手法，QF（品質係数）を推定して適応的に処理を行うブラインド JPEG アーティファクト除去手法である flexible blind convolutional neural network（FBCNN）[9]などが挙げられる．また，より高い知覚品質を狙った GAN ベースの手法（QGAC-GAN [10]等）は，テクスチャ復元に優れるものの学習の安定性に課題がある．一方，IQA を直接損失関数とする試みは，単一指標では周期的アーティファクトを誘発しやすいことが報告されている．本研究は，これらの知見を踏まえ，複数の IQA を統合することで安定かつ高品質な復元を実現しようとする点に優位性がある．

2-3-3 提案手法

提案ネットワークは FBCNN をベースに，エンコーダ，QF 推定器（Predictor），QF 情報を活用するコントローラ，デコーダの四部構成からなる（図 7）．特にデコーダのアップスケーリングには，従来の転置畳み込み層を廃し，2 倍バイリニア補間と畳み込みの組み合わせ（UpConv 層）を採用することでチェッカーボードアーティファクトを抑制している．また損失関数は，JPEG 復元後の出力と元画像との誤差を評価する再構成損失として，深層特徴距離である LPIPS と深層構造・テクスチャ評価 DISTS の重み付き和を用い，さらに QF 推定誤差を L1 ノルムで加える構成としている．このように，ネットワーク構造と知覚評価指標を組み合わせた設計により，GAN を使わずに高い知覚品質を達成する．

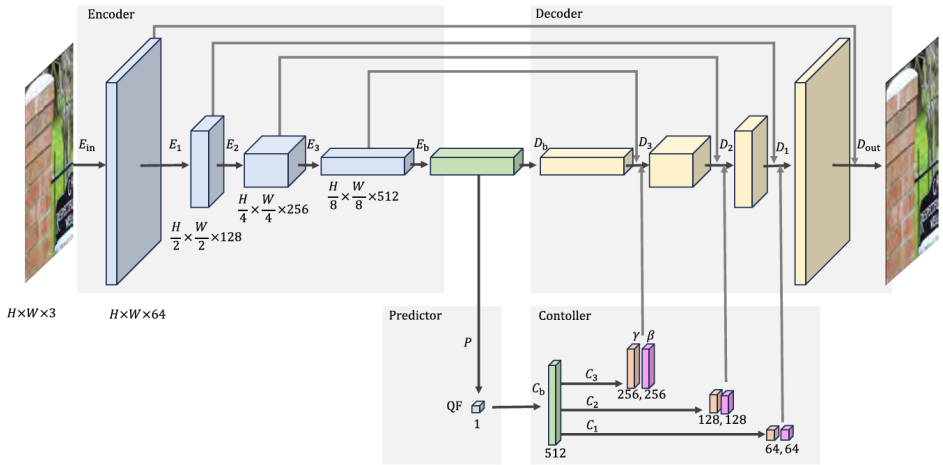


図 7.　2-3 節の提案手法のアーキテクチャ．

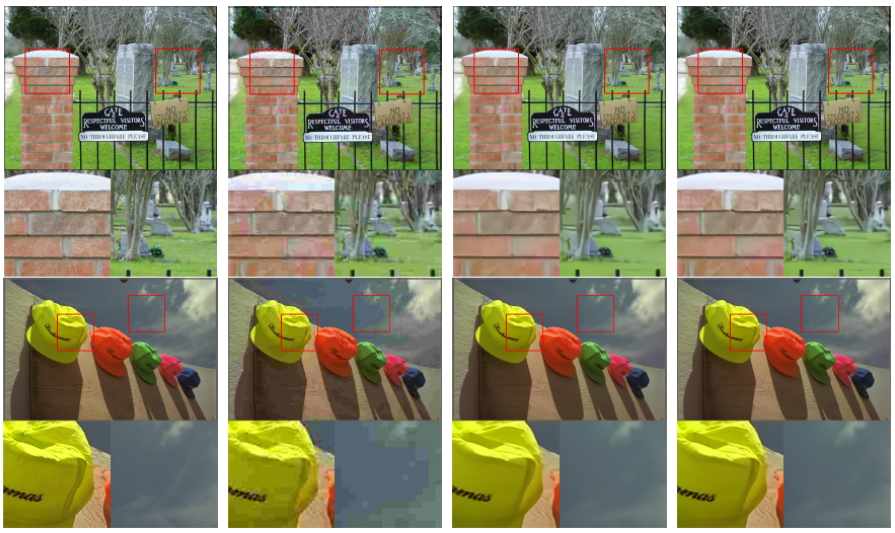
2-3-4 実験結果

学習には DIV2K および Flickr2K の約 3,450 枚，テストには LIVE1 と BSDS500 の 29 枚を用い，QF10–30 の JPEG 画像で評価を行った．比較対象は QGAC，QGAC-GAN，FBCNN，ARCRL [11]であり，評価指標には PSNR，SSIM，LPIPS，FID を採用した．定量結果（表 2）では，提案手法は LPIPS および FID で最良の性能を示し，高い知覚品質を実現した一方，PSNR/SSIM では GAN ベース手法に若干劣るものの許容範囲内であることが示された．定性評価（図 8）においても，煩雑なテクスチャの復元と平滑領域での偽色生成抑制を両立し，GAN を用いない安定した学習が有効であることを確認した．さらに，α（LPIPS と DISTS の重み）の調整や UpConv 層の有効性を検証するためのアブレーション実験を行い，提案手法の選択が有効であることを示した．

Dataset	QF	JPEG	QGAC [7]	QGAC-GAN [7]
LIVE1	10	25.66 / 0.7396 / 0.4785 / 78.83	27.67 / 0.8039 / 0.3591 / 70.25	27.42 / 0.7966 / 0.3034 / 46.04
	20	27.98 / 0.8220 / 0.3482 / 40.01	29.93 / 0.8682 / 0.2722 / 35.80	29.68 / 0.8629 / 0.2264 / 25.74
	30	29.26 / 0.8581 / 0.2812 / 27.55	31.21 / 0.8961 / 0.2290 / 24.45	30.96 / 0.8917 / 0.1892 / 17.52
BSDS500	10	25.67 / 0.7320 / 0.4601 / 97.17	27.62 / **0.7969** / 0.3372 / 96.23	27.36 / 0.7894 / 0.2816 / **59.14**
	20	27.94 / 0.8184 / 0.3205 / 55.53	29.81 / 0.8638 / 0.2498 / 53.66	29.54 / 0.8580 / 0.2026 / 36.71
	30	29.22 / 0.8561 / 0.2492 / 40.78	31.08 / 0.8926 / 0.2052 / 36.27	30.81 / 0.8879 / 0.1654 / 24.22

Dataset	QF	FBCNN [9]	ARCRL [10]	Ours
LIVE1	10	**27.79** / 0.8019 / 0.3578 / 63.91	27.78 / **0.8042** / 0.3592 / 62.50	27.12 / 0.7924 / **0.2963** / **45.72**
	20	30.09 / 0.8679 / 0.2671 / 31.63	**30.15** / **0.8704** / 0.2661 / 31.83	29.46 / 0.8599 / **0.2140** / **24.54**
	30	31.37 / 0.8962 / 0.2227 / 21.54	**31.45** / **0.8982** / 0.2217 / 21.65	30.77 / 0.8891 / **0.1760** / **16.78**
BSDS500	10	27.74 / 0.7936 / 0.3378 / 88.26	**27.75** / 0.7966 / 0.3361 / 86.43	27.08 / 0.7844 / **0.2705** / 59.22
	20	29.93 / 0.8617 / 0.2478 / 47.55	**30.01** / **0.8649** / 0.2448 / 46.49	29.30 / 0.8536 / **0.1888** / **33.76**
	30	31.15 / 0.8906 / 0.2027 / 32.09	**31.27** / **0.8935** / 0.2000 / 32.02	30.57 / 0.8838 / **0.1523** / **22.86**

表 2.　2-3 節の提案手法の定量評価．PSNR↑/ SSIM↑/ LPIPS↓ /FID↓．



(a) Original image　(b) JPEG　(c) QGAC [10]　(d) QGAC-GAN [10]

(e) FBCNN [9]　(f) ARCRL [11]　(g) Ours

図 8.　2-3 節の提案手法の定性評価．

2-3-5 結論

本論文では，複数の画像品質評価指標（IQA）を重み付きで組み合わせた損失関数と，従来の転置畳み込み層を廃したアップスケーリング構造（UpConv 層）の導入により，JPEG 圧縮画像の知覚品質を大幅に向上させる新しいアーティファクト除去手法を提案した．実験結果から，提案手法は従来手法（特に FBCNN）と比較して，LPIPS や FID などの知覚評価指標で定量的に優れた性能を示すとともに，定性評価においてもテクスチャの復元と偽色・チェッカーボードアーティファクトの抑制を両立できることが確認された．さらに，将来の課題として，異なるデータセットや品質係数（QF）での汎化性能検証や，ユーザーが知覚品質と PSNR のトレードオフをインタラクティブに調整できる仕組みの検討が挙げられる．

2-4 ノイズ画像に対する二段階ネットワークによる構造-テクスチャ-ノイズ分解

2-4-1 概要

画像処理における構造-テクスチャ分解は，画像からエッジや滑らかな領域からなる構造成分と，繰り返しパターンや微細なディテールからなるテクスチャ成分を分離する重要なタスクである．既存の構造-テクスチャ分解手法の多くはノイズの存在を考慮していなかったが，実画像にはノイズが含まれるのが一般的である．ノイズが存在すると，構造とテクスチャの分離が不正確になり，特にテクスチャ成分がノイズによって汚染されるという問題があった．そこで，本研究では，2 段階のネットワークを順次接続することで，ノイズ画像から構造，テクスチャ，ノイズ成分を分離できる深層学習ベースの構造-テクスチャ-ノイズ分解手法を提案した．第 1 段階のネットワークでは，入力ノイズ画像から構造成分を抽出し，それを用いてノイズ成分とテクスチャ成分が混合した残差画像を生成する．第 2 段階のネットワークでは，この残差画像と構造成分のコンテキスト情報を利用して，ノイズ成分とテクスチャ成分をさらに正確に分離する．この 2 段階のアプローチにより，構造とテクスチャの境界を明確に保ちながら，ノイズを効果的に除去し，純粋なテクスチャ成分を抽出することが可能になった．本研究は，この分解手法がノイズのある入力に対するトーンマッピング応用にも適用できることを示した．例えば，高ダイナミックレンジ（HDR）画像から低ダイナミックレンジ（LDR）画像への変換において，ノイズのある HDR 画像に対しても，提案手法で分離された構造，テクスチャ，ノイズ成分をそれぞれ適切に処理することで，視覚的に優れたトーンマッピング結果が得られることを示した．この成果は，2024 IEEE International Conference on Artificial Intelligence in Information and Communication（ICAIIC）で発表された．

2-4-2 関連研究

従来の構造–テクスチャ分解手法は，大きく分けて最適化ベースのアプローチと深層学習ベースのアプローチに分類される．最適化ベースでは，RoF（Rudin–Osher–Fatemi）モデルに代表されるように，総変動正則化を用いて構造成分を抽出し，残差をテクスチャとみなす手法が古くから研究されてきた．しかし，これらの手法は推論時に反復計算が必要であり，高解像度画像やリアルタイム処理には不向きである．また，構造とテクスチャの境界があいまいな領域では，過度な平滑化やアーティファクトが生じやすいという問題もある．一方，近年登場した深層学習を用いる手法では，事前に大規模なデータセットで学習したネットワークにより，一度の順伝播で処理を完了できるため推論速度が格段に速い．たとえば VDCNN（Very Deep Convolutional Neural Network）[12]を用いたアプローチは，入力画像と構造成分を対応づけるマッピングを直接学習することで，従来の最適化手法よりも高速かつ高精度な分解を実現している．しかしこれらのモデルはいずれもノイズ成分を独立に扱う設計にはなっておらず，入力ノイズをテクスチャに含めたまま処理するため，ノイズレベルが高い画像ではテクスチャとノイズの識別が困難になり，テクスチャ推定の品質が低下する欠点が残されている．

2-4-3 提案手法

本研究が提案する手法は，ノイズの有無にかかわらず入力画像を構造・テクスチャ・ノイズの三成分に分解する点に特徴がある．まず第一段階のネットワーク Φ_1では，入力画像をそのまま受け取りテクスチャとノイズをまとめた成分を推定する．この段階では，画像の高周波情報全体を捉えることに注力し，構造成分を差し引いた残差としてテクスチャ＋ノイズを抽出する．続いて第二段階のネットワーク

Φ_2では，第一段階で得られたテクスチャ＋ノイズ成分と，元画像から差し引いた構造成分をチャンネル方向に結合して入力することで，ネットワークに明示的に三成分分解タスクを提示する（図 9）．この二段構成により，Φ_2はテクスチャとノイズの微妙な特徴差を学習しやすくなり，特にノイズ成分の空間的・統計的性質を捉えることが可能となる．学習時には，構造成分の局所的な変動を保つための近傍一致損失，テクスチャおよびノイズの精度を高めるためのL_1損失，そして再構成画像との整合性を確保する再構成誤差という三種類の損失を組み合わせることで，各成分の分解品質を同時に最適化する．これにより，ノイズ環境下でも構造とテクスチャの分離精度を維持しつつ，ノイズ成分だけを正確に抽出することを実現している．

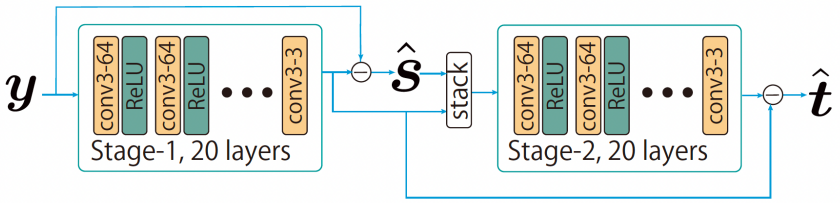


図 9.　2-4 節の提案手法のアーキテクチャ．

2-4-4 実験結果

提案手法の有効性を検証するため，BSDS500 データセットから選んだ 400 枚の画像を訓練セット，100 枚をテストセットとして使用した．ノイズレベル σ ＝0，15，25，50 の四段階で合成ノイズを付加し，従来の VDCNN および最適化型の L1 平滑化手法と比較した結果，提案手法はノイズ強度が高くなるほど従来モデルに対する優位性が顕著となった．具体的には，σ ＝25 においてテクスチャ成分の推定誤差が著しく低減し，構造成分においても精度の悪化をほとんど招いていないことが確認された（表 3）．さらに，実装を GPU 上で動作させた際の推論時間は 1 画像あたり数十ミリ秒と高速であり，CPU を用いた場合でも最適化ベース手法と比較して 1 桁以上の加速を達成した（表 4）．加えて，高ダイナミックレンジ映像のトーンマッピング応用実験では，ノイズを含む入力画像に対しても TMQI 指標が従来手法を上回る改善を示し，分解結果の利用価値が高いことを示唆している．また，視覚的評価においても，ノイズ領域とテクスチャ領域が明瞭に区別された結果が得られ，ノイズ除去とテクスチャ保存の両立が可能であることが確認された．

		structure		texture	
σ	method	wMAE	wRMSE	wMAE	wRMSE
0	VDCNN	**6.20**	**9.78**	**6.20**	**9.78**
	Ours	6.21	9.87	6.21	9.87
15	VDCNN	**7.00**	**10.55**	15.02	19.17
	Ours	7.03	10.57	**7.21**	**10.65**
25	VDCNN	7.98	11.74	20.90	26.35
	Ours	**7.85**	**11.51**	**7.87**	**11.26**
50	VDCNN	9.26	13.55	40.76	51.66
	Ours	**9.18**	**13.45**	**8.25**	**11.71**

表 3.　2-4 節の提案手法の定性評価．

method	Processor	time [s]
L1 smoothing	CPU	85
Ours	CPU	4.0
Ours	GPU	**0.08**

表 4.　2-4 節の提案手法の計算速度．

2-4-5 結論

本研究では，二段階の深層学習ネットワークによりノイズを含む入力画像を構造・テクスチャ・ノイズの三成分に高速かつ高精度に分解する手法を提案した．実験から，ノイズ混入時にもテクスチャ分離能が向上し，トーンマッピングなど他の画像処理タスクへの応用性も示された．今後はテクスチャ強調などさらなる応用領域への展開を検討する．

2-5 事前学習済み潜在拡散モデルを用いたゼロショット画像インペインティング手法

2-5-1 概要

近年急速な発展を遂げている拡散モデルを画像復元タスクに応用する研究にも取り組んだ．拡散モデルは，ノイズから画像を生成する過程を逆転させることで，画像の復元を行う強力な生成モデルとして注目されている．特に，Denoising Diffusion Null-space Model (DDNM) [13]を基盤とし，これの核となる拡散モデルを潜在拡散モデルへと拡張することで，様々な画像復元タスク，特に画像修復（Inpainting）において，従来の課題であった出力の多様性や大領域修復の課題を克服することに成功した．DDNM は，事前学習済み拡散モデルのヌル空間を利用して，画像復元制約を満たす画像を生成するゼロショット手法であり，タスク固有の学習を必要としないという利点がある．しかし，オリジナルの DDNM は，ピクセル空間で動作するため，高解像度画像や複雑な復元タスクにおいては計算コストが高いという問題があった．そこで，本研究では，潜在拡散モデル（Latent Diffusion Models, LDM）と DDNM を組み合わせることで，低次元の潜在空間で拡散プロセスと復元制約の適用を両立することを可能にした．これにより，計算効率を大幅に向上させつつ，高解像度画像の複雑な欠損領域に対しても，多様で一貫性のある画像を生成できるようになった．本研究では，このゼロショット画像修復手法が，知覚品質の重要な尺度の一つである Fréchet Inception Distance (FID) を大幅に改善できることを実験的に示した．FID は，生成された画像の分布と実際の画像の分布の類似度を測る指標であり，FID の改善は，生成された画像がよりリアルで多様であることを意味する．この研究成果は，IEEE ICAIIC 2025 にて発表された．この進展は，研究の第 2 段階の目標であった「ノイズ除去手法の提供範囲を一般の画像復元へと拡張する」という目的を，当初想定していなかった拡散モデルという強力なツールを用いることで，より高性能な形で達成できたことを意味する．

2-5-2 関連研究

画像インペインティングの従来手法として，まずジェネレーティブ・アドバーサリアル・ネットワーク (GAN) を用いたアプローチが挙げられる．Iizuka らの手法では，グローバルな文脈とローカルな文脈を同時に考慮したネットワーク構造により，高品質かつシームレスな復元を実現している．続いて，Yu らはコンテキュアル・アテンション機構を取り入れることで，画像中の遠隔領域からの情報を取り込み，自由形状マスクへの対応力を高めた手法を提案した．また，Gated Convolution を導入した DeepFill v2 は，欠損領域の境界を動的に学習し，滑らかな復元を可能にしている．一方で，GAN ベースの手法はトレーニングデータセットに依存しやすく，多様なシーンやクラスへの一般化性能に課題が残る．これに対し，拡散モデル（Diffusion Models）を利用した画像生成・復元手法が近年注目を集めている．DDPM（Denoising Diffusion Probabilistic Models）は逐次的にノイズを付加／除去するマルコフ過程を通じて高品質な生成を行い，DDIM（Denoising Diffusion Implicit Models）はその高速版として確率過程の一部を決定的更新に置き換え，推論速度を改善した．これら技術をベースに，Wang らは DDNM を提案し，先に述べた“生成画像のリアリティ向上”と“観測画像との整合性維持”を交互に実行する枠組みで，スーパーレゾリューションやカラー化，インペインティングをゼロショットで実現した ．しかし，DDNM のバックボーンである Guided Diffusion は ImageNet-1k（1,000 クラス）でのみ学習されており，馬やメロンなど一般的なクラスへの適用が困難であるという制約がある．この問題を解決するために，Rombach らが提案した Latent Diffusion Models (LDM) では，VAE によって得られる低次元潜在空間上で拡散過程を行うことで，大規模データ（LAION-5B）を用いた Stable Diffusion の学習を可能とした ．Stable Diffusion は 5 十億枚規模のデータセットと CLIP による言語条件付けを組み合わせ，多様かつ高解像度な画像生成能力を示す．本研究は，こうした LDM の強力な生成能力を活用しつつ，DDNM が備える整合性保持メカニズムを潜在空間上で再構築することで，より広範なカテゴリに対応可能なゼロショットインペインティングを実現する点に特徴がある．

2-5-3 提案手法

本手法は，Stable Diffusion（SD）をバックボーンとするLDMの潜在空間上で，既知領域の情報を固定しながら欠損領域のみを更新するインペインティングアルゴリズムを実現する．まず入力画像$\boldsymbol{y}$をVAEエンコーダEを用いて潜在変数 z_tに写像し，逆拡散過程をステップ T から 1 まで繰り返す．各ステップ t において，通常のDDIM更新により得られる潜在変数$\boldsymbol{z}_{0|t}$を計算した後，既知領域マスク$\mathbf{M}_l$を用いて以下のように整合性保持を行う．

$$\hat{\mathbf{z}}_{0|t} = \mathbf{M}_l \odot E(\mathbf{y}) + (\mathbf{I} - \mathbf{M}_l) \odot \mathbf{z}_{0|t}$$

ここで，$\mathbf{M}_l$は元画像のピクセルマスク$\mathbf{M}_p$ を最近傍補間で潜在空間にリサイズ後，境界を2ピクセル拡張して滑らかにしたものである ．この処理により，VAEデコーダ D の非線形性を考慮せずとも，既知領域部分は常に観測画像と一致したまま欠損領域を生成可能となる．その後，修正された潜在変数 $\hat{\boldsymbol{z}}_{0|t}$ を用いて次ステップの潜在変数 $\boldsymbol{z}_{t-1}$ を

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{z}}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2(\eta)}\epsilon_\theta(\mathbf{z}_t, t) + \sigma_t(\eta)\epsilon_{\mathbf{z}}$$

により推定し，最終的に$D(\boldsymbol{z}_0)$を出力してインペインティング画像を得る．提案手法の流れを図10に示す．

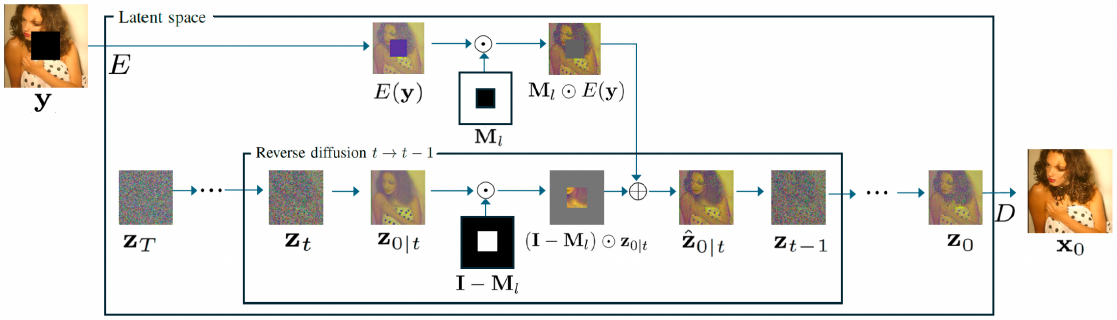


図 10.　2-5 節の提案手法のアーキテクチャ．

2-5-4 実験結果

提案手法の評価は定量評価と定性評価の両面で行った．まず定量評価においては，ベースラインとして DDNM（ゼロショット）を用い，マスクサイズ 168×168 および 136×136 ピクセルの中心マスクを BSDS500，ImageNet-1k，ImageNet-O の各テストセットで適用し，Fréchet Inception Distance（FID）を算出した（表 5）．その結果，すべてのデータセットおよびマスク条件において提案手法は DDNM を上回り，特に BSDS500 の 168×168 マスクでは 59.54→56.57，ImageNet-1k では 35.63→29.43 と顕著な改善を示した．また，BLIP による「理想的なテキストプロンプト」を導入した変種（Ours+BLIP*）では，さらなる低 FID 化（例：BSDS500 で 51.05）を達成し，テキストガイダンスの有効性を確認した．処理速度についても， 1 台の NVIDIA RTX 4080 GPU 上で約 3 秒/枚と，DDNM（約 100 秒/枚）比で約 33 倍の高速化を実現し，大規模データセットへの実用可能性を示した．定性評価では，市街地や石像，人物写真など多彩なシーンで比較を行い，DeepFill v2 ［14］ や LaMa ［15］ は局所的にぼやけや不整合を生じ，DDNM ではアーティファクトが目立ったのに対し，提案手法は欠損領域に対して高いリアリティと一貫性を保った復元が行えることが示された（図 11）．特に Ours+BLIP*では元画像から得られた語彙的手がかりにより，より精緻な復元が可能であることがわかった．以上より，提案手法は従来のゼロショットインペインティングを凌駕する品質と高速性を同時に実現し，今後の応用範囲拡大に向けた有望なアプローチであることが示された．

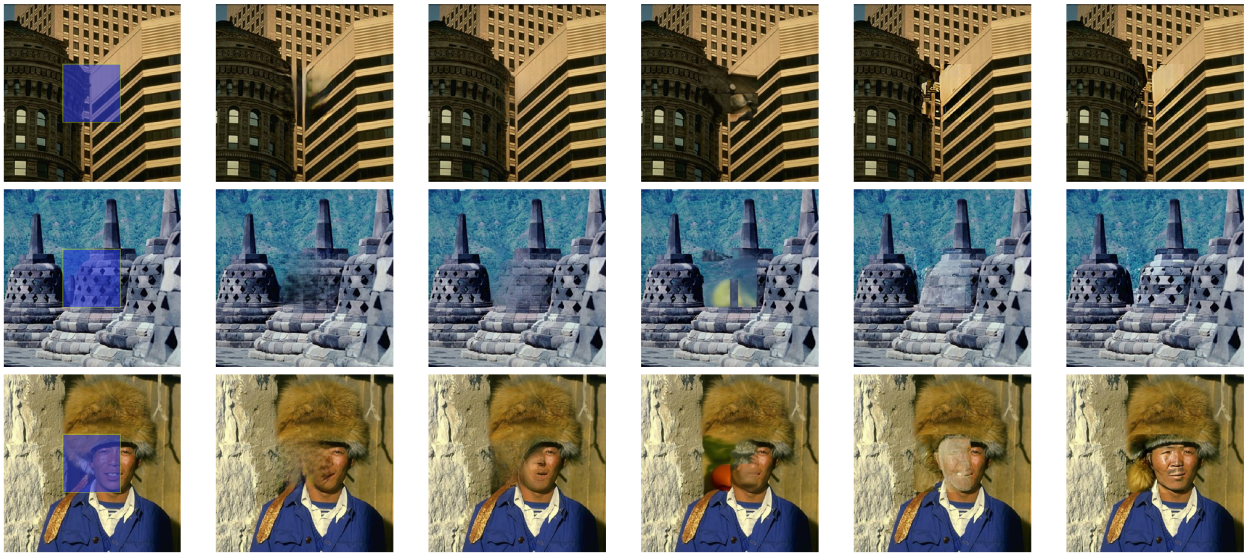
2-5-5 結論

本研究では，潜在拡散モデルの大規模事前学習の恩恵を受ける Stable Diffusion をバックボーンに採用し，従来の DDNM では困難であった多様カテゴリのゼロショットインペインティングを可能にする手法を提案した．実験結果から，従来法を上回る品質と大幅な高速化を同時に達成したことを示し，さらにテキストプロンプトを利用することで性能向上の余地も確認した．今後は，部分マスク入力から適切なプロンプトを自動生成する仕組みの検討など，さらなる性能向上を図る予定である．

.

Dataset	BSDS500		ImageNet-1k		ImageNet-O	
Mask size	168	136	168	136	168	136
DDNM [13]	59.54	43.83	35.63	22.82	36.80	25.25
Ours	**56.57**	**39.90**	**29.43**	**20.49**	**29.05**	**20.40**
Ours+BLIP* (refs. only)	51.05	37.33	27.82	21.18	32.20	23.54

表 5.　2–5 節の提案手法の定量評価.



(a) Original image $\mathbf{x}^\star$ and overlaid mask　(b) DeepFill v2 [14]　(c) LaMa [15]　(d) DDNM [13]　(e) Ours　(f) Ours+BLIP* (refs. only)

図 11.　2–5 節の提案手法の定性評価.

3 まとめ

以上に述べたように，本研究は，当初の計画であった知覚品質の高い画像復元という目標に留まらず，その過程で得られた知見と技術を応用し，画像品質評価や構造–テクスチャ分解といった多岐にわたる分野で，着実に重要な成果を創出してきた．これらの研究成果は，それぞれ国際会議や査読付き学術雑誌にて発表されており，学術コミュニティへの貢献も果たしている．本研究は，深層学習，統計的推定，最適化理論といった幅広い分野の知識を統合することで，複合的な問題に対する解決策を提供してきたといえる．今後も，これらの基盤技術をさらに深化させ，多様な実世界の問題に対する新たな応用を展開していく所存である．特に，拡散モデルのさらなる応用や，マルチモーダルデータ（画像，テキスト，3D 点群など）を統合した新たな知覚品質評価・復元手法の開発にも注力していく．

【参考文献】

[1] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2018, pp. 586-595.

[2] C. M. Stein, "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, vo. 9, no. 6, pp. 1135-1151, 1981.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, vol. 139, 2021, pp. 8748-8763.

[4] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513-516, 2010.

[5] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339-3352, 2012.

[6] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing,* vol. 21, no. 12, pp. 4695-4708, 2012.

[7] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209-212, 2013.

[8] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in Proceedings of *the AAAI Conference on Artificial Intelligence*, 2023.

[9] J. Jiang, K. Zhang, and R. Timofte, "Towards flexible blind JPEG artifacts removal," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4977-4986.

[10] M. Ehrlich, L. Davis, S.-N. Lim, and A. Shrivastava, "Quantization guided JPEG artifact correction," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 293-309.

[11] X. Wang, X. Fu, Y. Zhu, and Z.-J. Zha, "JPEG artifacts removal via contrastive representation learning," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 615-631.

[12] F. Zhu, Z. Liang, X. Jia, L. Zhang, and Y. Yu, "A benchmark for edge preserving image smoothing," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3556-3570, 2019.

[13] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," in *Proceedings of the International Conference on Learning Representations*, 2023.

[14] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[15] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2149-2159.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
A Consideration of JPEG Resistance Verification of Correlation-based Steganography	Journal of Imaging Science and Technology	2024 年 11 月
Perceptual JPEG Artifact Removal using Weighted Sum of IQAs as Loss Function	Nonlinear Theory and Its Applications	2024 年 10 月
Real Image Noise Aware Steganography with Image Denoising and Generative Adversarial Network	Nonlinear Theory and Its Applications	2024 年 10 月
ZEN-IQA: Zero-Shot Explainable and No-Reference Image Quality Assessment with Vision Language Model	IEEE Access	2024 年 5 月
PSinGAN: Single Image Inpainting by Generative Model Trained on Partial Observation	Nonlinear Theory and Its Applications	2024 年 7 月
Traffic Matrix Completion by Weighted Tensor Nuclear Norm Minimization and Time Slicing	Nonlinear Theory and Its Applications	2024 年 4 月
Zero-Shot Image Inpainting using Pretrained Latent Diffusion Models	IEEE International Conference on Artificial Intelligence in Information and Communication (ICAIIC)	2025 年 2 月
Accuracy based Rewarding for Sensors in Noisy Collaborative Point Cloud Acquisition Environments	IEEE International Conference on Artificial Intelligence in Information and Communication (ICAIIC)	2025 年 2 月
Completion of Traffic Matrix by Tensor Nuclear Norm Minus Frobenius Norm Minimization and Time Slicing	IEEE/IFIP International Workshop on Analytics for Network and Service Management (NOMS/AnNet)	2024 年 5 月
Structure-Texture-Noise Decomposition for Noisy Images with Two-Stage Network	International Conference on Artificial Intelligence in Information and Communication (ICAIIC)	2024 年 2 月
Perceptual JPEG Artifact Removal using Sum of Weighted IQA as Loss Function	Korea-Japan Joint Workshop on Complex Communication Sciences (KJCCS)	2024 年 1 月
Pseudo Real Image Noise Steganography with Image Denoising and Generative Adversarial Network	Korea-Japan Joint Workshop on Complex Communication Sciences (KJCCS)	2024 年 1 月
Completion of Traffic Matrix with Sequential Missing Values by Weighted Tensor Nuclear Norm Minimization	International Symposium on Nonlinear Theory and Its Applications (NOLTA)	2023 年 9 月
Interpretable Image Quality Assessment via CLIP with Multiple Antonym-Prompt Pairs	IEEE International Conference on Consumer Technology (ICCE) Berlin	2023 年 8 月

掲載された成果は，すべて査読あり・英文の国際論文誌および国際会議である．