

学習データに着目した画像超解像モデルの性能向上

2025 年度

大谷 豪

学位論文 博士（工学）

学習データに着目した画像超解像モデルの性能向上

2025 年度

慶應義塾大学大学院理工学研究科

大谷 豪

概要

画像超解像 (Image Super-Resolution; SR) は, 単一の低解像度画像から対応する高解像度画像を復元するタスクであり, 医療や衛星画像など多様な分野で広く利用されている. 深層学習の導入以降, 特に CNN ベースのアーキテクチャが性能向上を牽引し, SR 分野における研究の中心は, モデル構造の設計に置かれてきた.

一方で, SR はピクセル単位での精緻な再構成が求められるため, 従来から学習データの質が性能に大きく影響することも認識されていた. 実際, 多くの CNN ベース手法では, 比較的良質な小規模データセットを用いることで, 十分な性能を実現していた.

しかし近年では, CNN ベース手法よりも高い表現力を有する Transformer ベースのモデルが登場し, その能力を十分に引き出すためには, 従来以上に大規模かつ高品質な学習データが求められている. このように, モデルの高性能化に伴い, 学習データの質と多様性の両立が新たな研究課題として顕在化している.

そこで本研究では, 学習データの質と多様性に着目したアプローチとして, 損失関数とデータ構築の両面から SR 学習を再考し, 実用性と性能の両立を目指す. 具体的には, 基盤モデルを活用した Perceptual loss の再設計, Web 画像を用いた高品質データセットの自動構築, および JPEG 圧縮画像を再活用する前処理パイプラインを提案した. これらの手法を通じて, 学習データの特性に基づく損失関数設計や学習データ品質の最適化が, アーチファクトの抑制や汎化性能の向上といった SR モデルの性能改善に寄与することを示し, 学習データを中心とした新たな SR 学習の枠組みを提示した.

第 1 章では, 本研究の背景と目的を述べる. 第 2 章では, 画像超解像の基礎知識, 評価指標, モデルアーキテクチャ, ならびに従来用いられてきたデータセットについて概説し, 研

究の立ち位置を明確にする．第3章では, Perceptual loss に用いる事前学習モデルに着目し, 大規模データによる事前学習が知覚品質の向上に与える影響を検証する．これにより, 損失関数の構造とその背後にある学習データの特性との関係を明らかにする．第4章では, 損失関数による改善には限界があることを踏まえ, 学習データの質と多様性に注目したアプローチとして, Web 画像からの自動選別によって高品質なサブセットを構築する手法を提案する．第5章では, 従来, 品質の観点から SR モデルの学習には利用が難しいとされていた JPEG 圧縮画像を, 未知の劣化に対応可能な SR モデルで補正し学習に使用することで, 質と多様性を両立する手法を提案する．第6章では, 本研究全体を総括し, 損失関数とデータ構築の両面から得られた知見を整理するとともに, 今後の展望について述べる．

目次

第1章	序論	1
1.1	研究背景	1
1.2	問題設定	3
1.3	本論文の構成	5
第2章	予備知識・関連研究	6
2.1	画像超解像の概要	6
2.2	歪み品質を重視した設定（既知劣化）	11
2.2.1	SRモデルの基本構造	12
2.2.2	学習方法	14
2.2.3	評価方法	15
2.3	知覚品質を重視した設定（既知劣化）	16
2.3.1	学習方法	17
2.3.2	評価方法	18
2.4	知覚品質を重視した設定（未知劣化）	18
2.5	学習データセット	19
2.6	評価データセット	23
第3章	Perceptual lossの再考	26

3.1	導入	26
3.2	事前知識	27
3.2.1	Contrastive Language-Image Pre-training (CLIP)	27
3.3	提案手法	28
3.3.1	手法の流れ	28
3.3.2	ViT を用いた損失関数	29
3.4	実験	30
3.4.1	実験設定	30
3.4.2	従来の Perceptual loss との比較	31
3.4.3	SR モデルによる比較	34
3.4.4	CLIP モデルを使用した損失の重み付けによる比較	34
3.4.5	CLIP モデルのアーキテクチャによる比較	37
3.5	まとめ	37
第 4 章	SR 学習データ構築の再考	39
4.1	導入	39
4.2	DiverSeg データセット	42
4.2.1	ソース選定	43
4.2.2	オブジェクトベースのフィルタリング	45
4.2.3	データセット統計	46
4.3	実験	47
4.3.1	実験設定	47
4.3.2	主な結果	48

4.3.3	SOTA 手法との比較	49
4.3.4	分析 1: フィルタリングの効果	50
4.3.5	分析 2: 画像品質の影響	53
4.3.6	分析 3: 視覚的比較	55
4.4	議論	57
4.4.1	高解像度データセットへの適用	57
4.4.2	セグメント数と周波数成分の関連性	58
4.5	まとめ	59
第 5 章	JPEG 圧縮画像の有効活用	60
5.1	導入	60
5.2	前提知識	62
5.2.1	JPEG 圧縮アルゴリズム	62
5.3	提案手法	65
5.3.1	Compression Noise Estimation (CNE) モジュール	67
5.3.2	Quality Enhancement (QE) モジュール	68
5.3.3	Artifact Suppression (AS) モジュール	68
5.3.4	データセット統計	69
5.4	実験	70
5.4.1	実験設定	70
5.4.2	フィルタリング手法との比較	72
5.4.3	RECA の有効性	73
5.4.4	CNE モジュールの閾値	75

5.4.5	AS モジュールの有効性	77
5.4.6	QE モジュール内における未知劣化 SR モデルの比較	78
5.4.7	デノイジングモデルとの比較	78
5.5	議論	79
5.5.1	学習データと評価データの特徴	79
5.5.2	画像復元タスクへの応用	80
5.5.3	未知劣化への拡張	81
5.5.4	学習可能な設計の欠如	81
5.5.5	計算コストの課題	81
5.5.6	感度分析の制約	81
5.6	まとめ	82
第 6 章 結論		83
謝辞		85
参考文献		85

目 次

1.1	SR の概要	2
2.1	複数の高解像度画像をダウンサンプリングしたときに, 同一の低解像度画像になる例	6
2.2	歪み品質と知覚品質の関係性	8
2.3	歪み品質と知覚品質がトレードオフであることを示す定性的な比較例	8
2.4	SR 問題設定	9
2.5	SR モデルの基本構成例	13
2.6	既知劣化設定における基本的な学習過程	15
2.7	DIV2K を構成する画像の例	21
2.8	Flickr2K を構成する画像の例	21
2.9	LSDIR を構成する画像の例	22
2.10	ImageNet を構成する画像の例	22
2.11	Set5 を構成する全画像	23
2.12	Set14 に含まれる全画像	23
2.13	BSD100 を構成する画像の例	24
2.14	Urban100 を構成する画像の例	24
2.15	Manga109 を構成する画像の例	25
2.16	DIV2K Validation を構成する画像の例	25

3.1	画像と言語のペア学習によりゼロショット分類を実現する CLIP の構成 . . .	27
3.2	提案手法の概要	29
3.3	学習に使用した損失関数による定性比較	32
3.4	学習に使用した損失関数による知覚-歪み曲線の比較	33
3.5	CLIP モデルのアーキテクチャによる知覚-歪み曲線の比較	35
3.6	ResNet ベースの CLIP モデルによる Perceptual loss の重み係数の値による 定性比較	36
3.7	ViT ベースの CLIP モデルによる Perceptual loss の重み係数の値による定 性比較	36
4.1	SR 学習データセットを構築する提案パイプライン	41
4.2	DiverSeg の画像とそのセグメンテーションマスク	42
4.3	JPEG 品質による画像劣化の比較	44
4.4	Blockiness 分布の比較	45
4.5	パワースペクトル画像における周波数領域	52
4.6	JPEG 圧縮による学習プロセスの比較	54
4.7	ImageNet-1k, PASS および Places365 における学習プロセスの比較	54
4.8	JPEG 圧縮によって生じたアーチファクトと, 異なる JPEG 品質値で学習 された SR モデルによるアーチファクトの比較	55
4.9	DF2K, LSDIR, および DiverSeg で学習した SR モデルの定性比較	56
4.10	セグメント数と高周波含有率の分布	59
5.1	従来のフィルタリング手法と提案手法との比較	61
5.2	提案手法 RECA パイプラインの概要とその有効性	66

5.3	ベースライン, ブロックノイズによるフィルタリング, および提案手法 RECA の視覚的比較	73
5.4	AS モジュールの効果を示す視覚的な例	76

表 目 次

3.1	異なる損失関数の組み合わせによる定量評価	33
3.2	様々な SR モデルによる定量評価	34
3.3	ResNet ベースの CLIP モデルによる Perceptual loss の重み係数 λ による定 量評価	35
3.4	ViT ベースの CLIP モデルによる Perceptual loss の重み係数 λ による定量 評価	35
3.5	CLIP モデルのアーキテクチャによる定量評価	37
4.1	データセットの統計情報	46
4.2	既存の高解像度 SR データセットと DiverSeg の比較	49
4.3	最先端手法との性能比較	50
4.4	Blockiness によるフィルタリング	50
4.5	周波数成分によるフィルタリング	52
4.6	フィルタリング手法の比較	53
4.7	異なる閾値 θ における性能比較	53
4.8	JPEG 品質の効果の分析	53
4.9	LSDIR に対してオブジェクトベースのフィルタリングを適用した時の性能 比較	58

5.1	データセットの統計情報	70
5.2	提案手法 RECA の性能比較	72
5.3	異なるアーキテクチャを持つ SR モデルの定量的比較 (iCartoonFace) . .	74
5.4	異なるアーキテクチャを持つ SR モデルの定量的比較 (Food-101)	75
5.5	異なる閾値 θ における性能比較	76
5.6	AS モジュールの効果	76
5.7	QE モジュールにおける実世界 SR モデルの性能比較	77
5.8	QE モジュールと AS モジュールの組み合わせと Denoising モデルの性能比較	78
5.9	QE モジュールと AS モジュールの組み合わせと Denoising モデルの性能比較	79
5.10	DIV2K と iCartoonFace で学習された SR モデルの性能比較	79

第1章 序論

1.1 研究背景

画像超解像 (Image Super-Resolution; SR) は, 図 1.1 に示すように, 単一の低解像度画像から対応する高解像度画像¹を復元する, コンピュータビジョン分野の基本的なタスクである. SR は, 医療画像や防犯カメラ映像, 衛星画像の解析など, 多様な応用分野において不可欠な役割を担っている [1–3].

深層学習に基づく SR 研究は, 2014 年に提案された SRCNN [4] を契機として急速に発展した. 当初, 畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) を用いた手法が主流であり, 比較的小規模なデータセットでも高い学習効果を得られたため, 研究の焦点は主にモデル設計に置かれていた. 実際, 多くの研究は 800 枚の高解像度画像からなる DIV2K [5] や, その拡張版である 3,450 枚の DF2K [5,6] を用いており, 学習データセットの構築手法への注目度は限定的だった.

しかしながら, 画像認識や画像生成の分野では, こうしたモデル設計中心のアプローチとは対照的に, 学習データがモデル性能に与える影響の大きさが近年広く認識されている. 特に Transformer [7] の登場により, モデルの表現力が著しく向上し, 学習に用いるデータの規模が一層重要な要素となった. この流れの中で, 大規模データと強力なモデルを組み合わせる汎用的な表現を学習する基盤モデルが登場し, 多様なタスクで高い適応性を示して

¹本論文では, HD 解像度 (1280×720 ピクセル) を基準として, これよりも高い解像度の画像を高解像画像, 低い解像度の画像を低解像度画像と定義する.

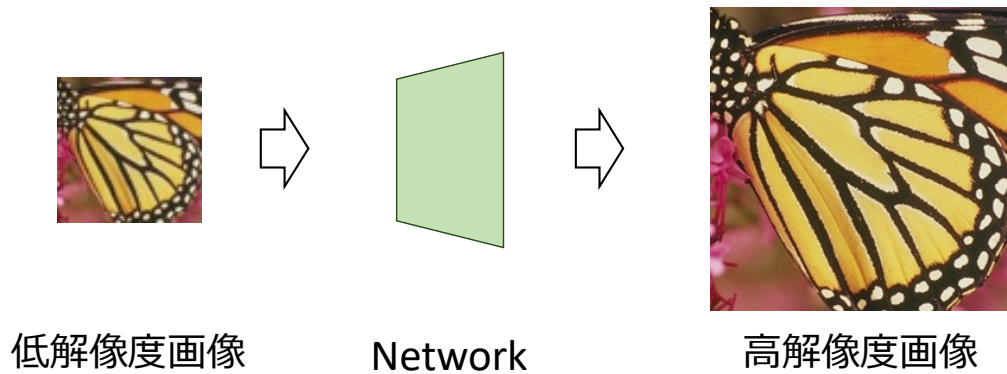


図 1.1: SR の概要.

いる. 例えば, 視覚と言語を統合した CLIP [8] は, 約 4 億対の画像と言語ペアを用いて学習され, 幅広い下流タスクへの転移能力を持つ. また, テキストから高品質な画像を生成する Stable Diffusion [9] は, 約 4 億枚の画像とキャプションのペアからなる LAION-400M [10] を用いて学習され, 多様な視覚概念に対応した生成能力を獲得している.

このように, 学習データがモデル性能を左右するという認識が広がる中, その影響はモデル設計だけでなく損失関数の設計にも及び得る. 特に, 知覚品質の向上を目的とした Perceptual loss は, SR を含む多くの画像生成・変換タスクで広く用いられているが, その特徴抽出器としては依然として古典的な CNN である VGG ネットワーク [11] が主流である. VGG は ImageNet [12] で事前学習された分類器であり, 局所的なエッジやテクスチャの抽出には優れる一方で, 画像全体の文脈的關係や意味的整合性を十分に捉えることが難しいという課題がある. したがって, SR における知覚品質をより適切に評価するためには, より高次の意味的特徴を獲得した基盤モデルを Perceptual loss の構成要素として活用することが望ましい.

一方で, 学習データセットの構築そのものにも, SR 特有の課題が存在する. SR タスクでは, 高解像度かつ劣化の少ない画像が必要とされるため, 品質・多様性・解像度という複数の条件を同時に満たすデータセットの構築は容易ではない. Web 上には大量の画像

が存在するものの、多くは JPEG 圧縮などの劣化を含んでおり、そのまま利用することは困難である。結果として、大規模かつ高品質な学習データを得ることが SR モデルの性能向上における深刻なボトルネックとなっている。さらに、衛星画像や監視カメラ画像など、現実環境で取得される画像は、一般的な自然画像とは異なる特性を持つことが多い。これらのドメイン特有の画像では、品質を厳密に担保しようとする使用可能な枚数が大幅に制限され、十分な学習量を確保することが難しい。そのため、多少の劣化を含む画像も適切に活用し、質と多様性の両面を考慮した学習データの構築が、応用性の高い SR モデルの構築において重要な課題となっている。

このような背景から、本研究では学習データの質と多様性の両面から SR 性能向上を目指す新たなアプローチを提案する。まず、歪み品質と知覚品質の両立を目的として、基盤モデル CLIP の特徴空間を導入した Perceptual loss を設計し、アーチファクトを抑制した SR を実現する。次に、高解像度画像の大量収集が困難であることを踏まえ、品質と多様性の双方を満たす低解像度画像を活用し、効果的な学習データセットを構築する。さらに、JPEG 圧縮ノイズを含む画像が除外されることによる学習データの規模不足に対処するため、圧縮による画質劣化を補正するパイプラインを導入し、これらの画像を学習データとして再活用する。本研究は、モデルアーキテクチャの改良が主流となっている SR 研究の中で、データ構成の観点から性能向上を実現し、実環境への適用も視野に入れた新たなアプローチを提案している。

1.2 問題設定

SR の性能を最大限に引き出すには、モデルアーキテクチャの設計だけでなく、学習に用いるデータと損失関数の設計も極めて重要となる。特に、Perceptual loss のようなニュー

ラルネットワークベースの損失関数においては、特徴抽出器の事前学習データに起因するバイアスが、出力画像の知覚品質や自然さに直接的に影響する。また、多様性を備えた高品質な学習データをいかに構築・活用するかもモデルの汎化性能に大きな影響を与える。本研究では、これら損失関数と学習データの両面から SR 性能を支えるデータ設計を見直し、性能向上を目指す。

まず第 3 章では、基盤モデルの 1 つである CLIP を用いて、ニューラルネットワークをベースとした損失関数である Perceptual loss を再考する。従来の Perceptual loss は、ImageNet で学習された VGG ネットワークの中間特徴を用いることが一般的であった。しかし、これは画像分類に適した局所的な特徴に偏っており、SR 画像に不自然なアーチファクトを引き起こす要因となっていた。本研究では、自然画像とテキストのペアを用いて事前学習された CLIP の ViT [13] エンコーダを特徴抽出器として利用し、大域的な文脈と意味情報を損失関数に導入することで、アーチファクトを抑制し、歪み品質と知覚品質の両立を実現することを目指す。

続く第 4 章および第 5 章では、学習データの品質と多様性を向上させるためのデータセット構築手法を提案する。第 4 章では、大規模な画像群から SR 学習に適した画像を自動的に選別するため、JPEG 圧縮ノイズの定量的評価と構造的多様性の指標を組み合わせたフィルタリング手法を開発し、高品質で多様な学習データセットを効率的に構築する。さらに第 5 章では、従来、圧縮ノイズが多いため学習から除外されていた低品質画像群に対し、多様な劣化に対応した SR モデルを用いて補正を施し、学習可能なデータとして再活用する方法を提案する。圧縮ノイズの推定と補正プロセスを組み合わせることで、従来は活用困難とされていた画像群を学習に取り込み、データセットの規模と多様性を両立する。以上のように、本研究では SR の学習性能を支えるデータ設計に焦点を当て、損失関数と学習データの両面から、高性能な SR モデルの学習を可能にする。

1.3 本論文の構成

第 1 章では, 本研究の背景と目的を述べ, 解決すべき課題を整理した. 第 2 章では, 画像超解像に関連する基礎知識, 評価指標, モデルアーキテクチャ, およびデータセットについて概説する. 第 3 章では, Perceptual loss に用いる事前学習モデルを再考し, 新たに基盤モデルである CLIP を活用することで歪み品質と知覚品質の両立を図る. 本章の主な内容は, 文献 [14] で発表したものである. 第 4 章では, 損失関数による性能改善の限界を踏まえ, 学習データの品質と多様性を再考し, 2 段階のフィルタリングを適用することで, 高性能な SR データセットを構築する手法を提案する. この主な内容は, 文献 [15] で発表したものである. 第 5 章では, これまで利用困難だった JPEG 圧縮ノイズが含まれる低品質画像を, 未知の劣化にも対応できる SR モデルによって補正し, 再活用可能とすることで, 学習データのスケールを維持しつつ品質を向上させるアプローチを示す. この主な内容は, 文献 [16] で発表したものである. 最後に, 第 6 章では, 本研究全体の総括と今後の展望について述べる.

第2章 予備知識・関連研究

2.1 画像超解像の概要

画像超解像 (Image Super-Resolution; SR) は, 単一の低解像度画像から失われた高周波成分を復元し, 対応する高解像度画像を生成するタスクである. このタスクは, 図 2.1 に示すように, 同一の低解像度画像に対して複数の高解像度画像が正解となり得る不良設定問題 (ill-posed problem) であるため, 高品質な復元を実現するには学習ベースのアプローチが不可欠である.



図 2.1: 複数の高解像度画像をダウンサンプリングしたときに, 同一の低解像度画像になる例. [17] から引用.

一般的に SR では, 入力となる低解像度画像 I_{LR} は, 正解 (Ground Truth; GT) となる高解像度画像 I_{HR} に対して, 縮小やブラー, ノイズ付加, 圧縮などの処理を施す劣化関数 Φ を適用することで生成されると仮定される. すなわち,

$$I_{LR} = \Phi(I_{HR}; \theta_L), \quad (2.1)$$

と表され, この θ_L には縮小倍率, ブラーの強度, ノイズ量, JPEG 圧縮率などのパラメータが含まれる. 劣化関数 Φ は通常, 非可逆かつ未知であるため, SR の目的はこの劣化処理に対する逆変換 Φ^{-1} を推定し, I_{LR} から対応する高解像度画像 (SR 画像) I_{SR} を復元することである:

$$I_{SR} = \Phi^{-1}(I_{LR}; \theta_S). \quad (2.2)$$

ここで, θ_S は SR 処理におけるパラメータを表す. 近年では, この逆変換 Φ^{-1} をニューラルネットワークで実現するアプローチが主流となっている.

SR では, 劣化関数 Φ の設定に基づき, 大きく既知劣化 (**non-blind SR**) と未知劣化 (**blind SR**) の 2 つに分類される. 既知劣化では, 劣化関数を明示的に仮定するため, GT 画像との比較が可能であり, モデルの純粋な復元能力を評価するのに適している. 一方, 未知劣化では, 劣化過程が不明な場合を想定していることから GT 画像が存在せず, 多様な実画像に対しても適応可能な高い汎用性が求められる.

SR の評価軸としては, 歪み品質 (**Distortion**) と 知覚品質 (**Perception**) の 2 つがある. 歪み品質は, GT 画像との画素単位の一緻度を重視し, Peak Signal-to-Noise Ratio (PSNR) や Structural Similarity Index (SSIM) [18] などの指標を用いて定量的に評価される. 一方, 知覚品質は, 出力画像が人間の視覚にとってどれだけ自然かつ高品質に感じられ

るかを評価するものであり、その決定要素には、テクスチャの質感や構造の自然さ、ブラーの有無など、画像の質感や鮮明さといった知覚的に重要な視覚特徴が含まれる。これらを総合した知覚品質を定量的に評価するために、Learned Perceptual Image Patch Similarity (LPIPS) [19] などのニューラルネットワークに基づく指標が用いられる。

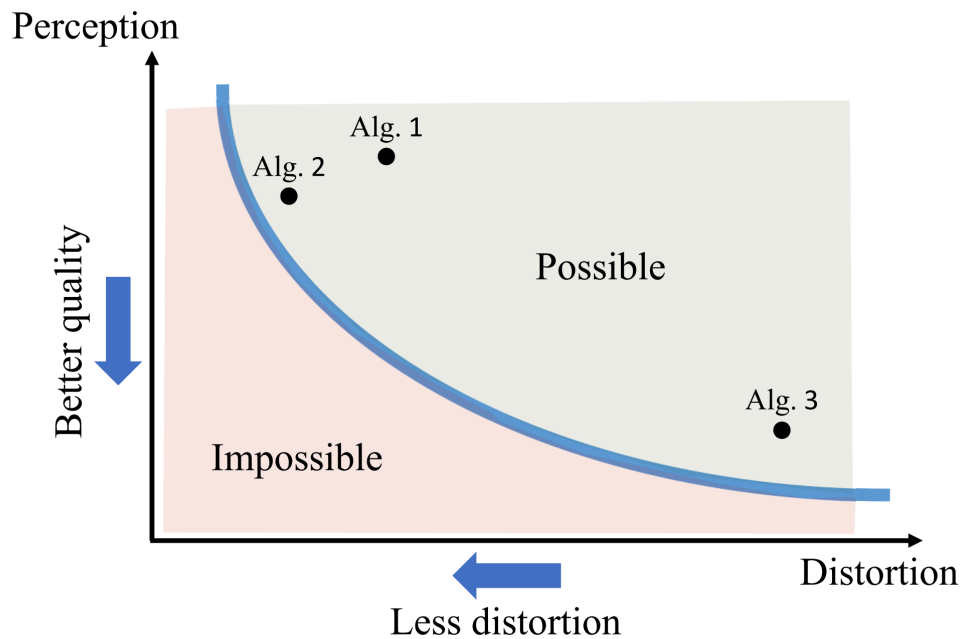


図 2.2: 歪み品質と知覚品質の関係性. [20] より引用.



図 2.3: 歪み品質と知覚品質がトレードオフである定性的な比較例. [21] より引用.

劣化条件	評価軸	評価指標	損失関数	目的
既知 (Bicubic補間)	歪み品質 を重視	PSNR, SSIM (GTあり)	L1 loss	モデルの ポテンシャル検証
	知覚品質 を重視	LPIPS, FIDなど (GTあり)	L1 loss + Perceptual loss Adversarial lossなど	歪み品質との バランス最適化
未知 (実世界の劣化 過程を再現)	知覚品質 を重視	NIQE, MUSIQなど (GTなし)	L1 loss + Perceptual loss Adversarial lossなど	汎用性向上・実応 用

図 2.4: SR 問題設定.

また, 図 2.2 に示すように, これら 2 つの品質はトレードオフの関係にあることが報告されている [20]. 歪み品質を重視したモデルでは, 複数の正解候補を平均化した結果としてディテールが失われた出力となる傾向があり, 必ずしも視覚的な評価と一致するとは限らない. 一方で, 知覚品質を重視したモデルでは, テクスチャの強調や局所的なノイズを伴いやすく, 画素単位での整合性が損なわれやすい (図 2.3 参照). このように, SR では両者のバランスを考慮した設計が求められる.

劣化条件 (既知/未知) と評価軸 (歪み品質/知覚品質) は, SR モデルの目的や設計方針を大きく左右する主要な視点であり, これら 2 軸を組み合わせることで, SR の代表的な 3 つの設定²に分類することができる (図 2.4).

本章では, 図 2.4 に示した 2 軸の問題設定に基づき, 以下の 3 つの代表的な設定に沿って, 各研究動向と手法の特徴を整理する.

- 歪み品質を重視した設定 (既知劣化): 劣化関数が既知であり, GT 画像との画素単位の一貫性を重視する設定である. 画素ごとの差分を評価する PSNR や, 構造の類

²未知劣化下で歪み品質を評価するのは困難であるため, この組み合わせは通常想定されない.

似性を評価する SSIM といった指標を用いた定量的な評価が一般的であり、ベンチマークでも最も広く採用されてきた。学習には L1 loss などの画素ベースの損失関数が用いられ、GT 画像との画素単位の整合性に優れる一方、出力画像は平均化されやすく、必ずしも視覚的な評価と一致するとは限らない。本設定に関する関連研究、学習手法、評価指標については 2.2 節で説明する。

- 知覚品質を重視した設定 (既知劣化): 劣化関数が既知である一方で、歪み品質を重視した設定では再現が難しい、GT との視覚的な自然さや質感の一致度を重視する設定である。Perceptual loss や Adversarial loss [22] など、出力画像の意味的・視覚的整合性を重視した損失関数が導入されており、L1 loss のみの場合と比較して、ブラーの低減やエッジの鮮鋭化、テクスチャの復元といった高周波成分の再構成が可能となる。評価には、GT 画像との視覚的類似度を深層特徴空間上で測定する LPIPS や、生成画像の自然さを統計的に評価する FID [23] などが用いられる。本設定に関する関連研究、学習手法、評価指標については 2.3 節で説明する。
- 知覚品質を重視した設定 (未知劣化): 実環境における複雑な劣化を前提とし、劣化へのロバスト性と視覚的な自然さを重視する設定である。対応する GT 画像が存在しないため、GT との一致度を測定する指標は適用できず、NIQE [24] や MUSIQ [25] などの画像単体で評価可能な指標や主観評価に基づく実用性重視の評価が行われる。本設定に関する関連研究については 2.4 節で述べる。なお、本論文ではこの設定における学習手法および評価指標は使用していないため、詳細は割愛する。

また、本研究がそれぞれの設定においてどのようなアプローチを提案するかについては、各節末において明確に位置付けを示す。

2.2 歪み品質を重視した設定（既知劣化）

既知劣化条件下の SR 研究では, PSNR や SSIM といった歪み指標の向上を目的とする手法が主流である. これらは, SR 画像と GT の画素単位での一致度を評価し, 定量的な整合性を保証するための指標として広く採用されている. この枠組みの下で, より高精度な再構成を実現するために, 様々な CNN ベースのモデルが提案されてきた.

SRCNN [4] は, 深層学習を初めて SR に適用した先駆的手法であり, 3 層のシンプルな CNN によって構成されていた. FSRCNN [26] では, 入力画像の Bicubic 拡大を省略し, 処理の大部分を低解像度空間で行う構成とすることで, 高速化とモデルの軽量化を実現した. ESPCN [27] は Pixel Shuffle 層を導入し, 高効率なアップサンプリングを達成した.

これらの初期手法は, 低解像度空間での処理を行った後にアップサンプリングによって高解像度出力を生成する, 現代の SR モデルに共通する基本構成を確立した. その後は, より高い表現力を実現するために, 残差接続 [21, 28–30], Dense Block [31, 32], アテンション機構 [33–35] など, モジュール構成の改良やネットワークの深層化を通じて性能向上が図られてきた.

VDSR [28] は, 深層化と残差学習により勾配消失を抑制し, 高精度な復元を実現した. EDSR [29] は, バッチ正規化の除去と残差スケールリングにより学習安定性と性能を両立した. MSRResNet [30] は Residual Dense Block を導入し, 計算効率と性能のバランスを最適化した. RCAN [33] は Residual-in-Residual 構造とチャネルアテンションを組み合わせ, 特徴の重要度を動的に調整する設計を導入した.

その後, モデル設計の主軸は CNN から Transformer へと移行した. IPT [36] は ImageNet で事前学習した Transformer を活用することで, CNN ベースの性能を上回った. SwinIR [37] は, Swin Transformer [38] を基盤とし, ローカルウィンドウによるセルフアテンションに

より、高い PSNR とパラメータ効率を実現した。SwinIR は、以降の Transformer ベースのモデルにおいて、標準的なベースラインとして定着している。

さらに、SwinIR を発展させた高性能モデルとして、EDT [39], SwinFIR [40], HAT [41] などが提案されている。特に HAT は、クロスアテンションと ImageNet 事前学習を組み合わせ、Transformer の表現力を最大化することで、State-of-the-Art (SOTA) 性能を達成した。

近年では、Transformer に代わる新たなアーキテクチャとして、状態空間モデル（State Space Model; SSM）に基づく Mamba が注目されている。Mamba アーキテクチャを採用した MambaIR [42] は、Transformer ベースのモデルに匹敵する高い SR 性能を示しており、今後の有力な手法として期待されている。

このように、歪み品質を重視した既知劣化 SR の研究は、初期の CNN モデルから始まり、ネットワーク構造の改良を重ねながら進化してきた。各段階での設計上の工夫が、歪み品質の限界を着実に押し広げてきたと言える。一方で、使用される学習データセットは DIV2K や DF2K など初期に確立されたものが長らく踏襲されており、モデルの発展とは対照的にデータ側の更新はほとんど進んでいない。このような状況を踏まえ、本論文では、この設定における性能向上をさらに促進するため、第 4 章および第 5 章において、高品質かつ多様な学習データを効率的に構築・再活用するためのアプローチを提案する。

2.2.1 SR モデルの基本構造

2025 年現在、SR モデルにおける標準的なアーキテクチャは、図 2.5 に示すように、一般に **Shallow Feature Extraction**, **Deep Feature Extraction**, **Image Reconstruction** の 3 段階で構成される。この構成は RCAN [33] において明確に定式化されて以降、多くの代表的手法で踏襲されており、現在では SR モデルにおける基本設計として広く定着して

いる.

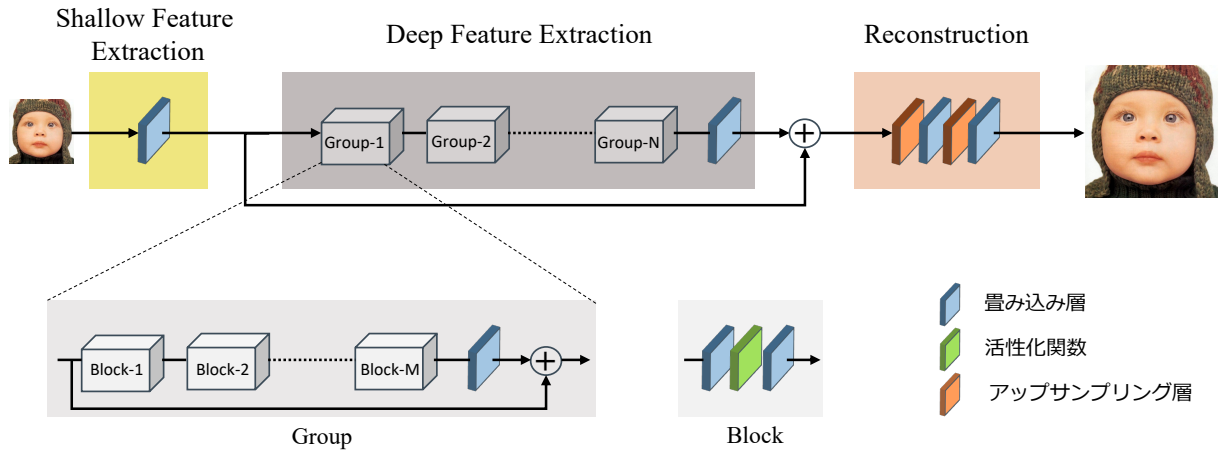


図 2.5: SR モデルの基本構成例.

Shallow Feature Extraction. 最初の段階では, 入力される低解像度画像 $I_{\text{LR}} \in \mathbb{R}^{H \times W \times 3}$ を初期の特徴空間へと変換する. ここで H と W はそれぞれ画像の高さと幅, チャンネル数 3 は RGB 画像を意味する. 通常, 1 層または少数の畳み込み層を用いて, エッジやテクスチャなどの局所的な構造を保持した特徴マップ F_{SF} を抽出する. この処理は以下のように表される:

$$F_{\text{SF}} = H_{\text{SF}}(I_{\text{LR}}). \quad (2.3)$$

ここで, H_{SF} は浅層特徴抽出を担うネットワークの処理全体を表す関数である. 得られた特徴は, 後続の深層特徴抽出の入力として用いられる.

Deep Feature Extraction. 次に, 浅層特徴 F_{SF} に対して, ブラーなどの劣化によって失われたテクスチャやエッジなどの高周波成分を復元するための変換を行う. この段階では, 残差ブロック, アテンション機構, Transformer 層など, モデルに応じた構造が用いられ, 広域な空間的文脈や特徴間の関係を学習する. この処理は次のように定式化される:

$$F_{\text{DF}} = H_{\text{DF}}(F_{\text{SF}}). \quad (2.4)$$

ここで, H_{DF} は深層特徴抽出を担うネットワークの処理全体を表す関数である. 複数の層からなる非線形変換により豊かな表現を獲得し, 高精度な復元を可能にする.

Image Reconstruction. 最後に, 得られた深層特徴 F_{DF} をもとに, SR 画像 I_{SR} を出力する. ここで, 空間解像度を上げるためのアップサンプリング処理が必要となる. この時, 線形補間による補間手法ではなく, 転置畳み込みや Pixel Shuffle [27] といった学習可能な手法が使用される. 特に Pixel Shuffle は, チャンネル方向の情報を空間的に再配置することで, 高効率かつアーチファクトの少ない SR を可能にする手法であり, 現在の SR モデルにおける主流なアップサンプリング手法となっている. この処理は次のように記述される:

$$I_{\text{SR}} = H_{\text{REC}}(F_{\text{DF}}). \quad (2.5)$$

ここで, H_{REC} は画像再構成ネットワークを表す関数であり, 最終的な SR 画像の出力を担う.

2.2.2 学習方法

既知劣化設定における基本的な学習フローを図 2.6 に示す. 一般的な手法では, 正解となる高解像度画像 (GT 画像) に対して Bicubic 補間によるダウンサンプリングを行い, 低解像度画像を生成する. その後, 両画像からランダムクロップにより対応するパッチを抽出し, パッチ単位での自己教師あり学習が行われる. 具体的には, 学習パラメータ θ を持つネットワーク N_{θ} に低解像度画像 I_{LR} を入力し, 出力として SR 画像 $I_{\text{SR}} = N_{\theta}(I_{\text{LR}})$ を得る. そして, GT 画像 I_{HR} との間に定義される損失関数を最小化するように, パラメー

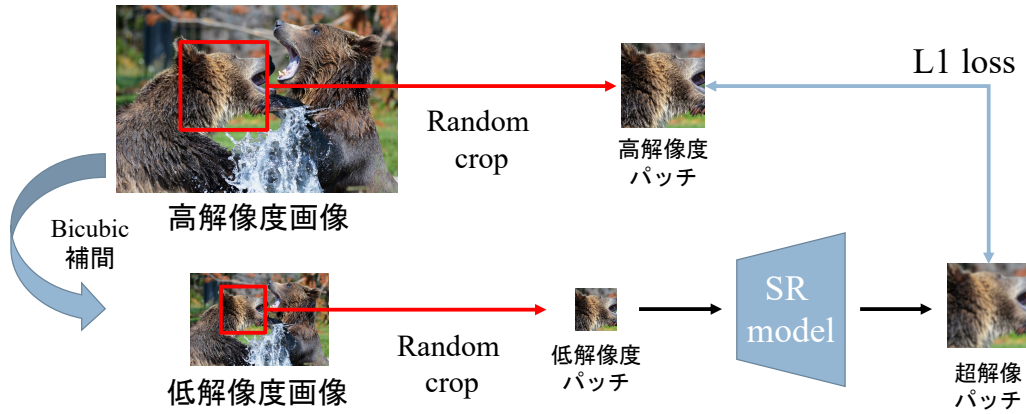


図 2.6: 既知劣化設定における基本的な学習過程.

タ θ を更新する. このとき, 多くの手法では画素単位での誤差を直接計算でき, 勾配の安定性にも優れる L1 loss が用いられる. 特に, L2 loss と比較して過度な平滑化を抑制でき, エッジ情報の保持にも優れていることから, 歪み品質の最適化において広く採用されている. L1 loss は以下のように定義される:

$$\mathcal{L}_1 = \|I_{\text{HR}} - I_{\text{SR}}\|_1. \quad (2.6)$$

2.2.3 評価方法

画素ごとの差分を評価する PSNR や, 構造の類似性を評価する SSIM [18] といった指標を用いた定量的評価が一般的である. 最も広く使用されているのは PSNR であり, 画像全体の平均二乗誤差 ((Mean Squared Error; MSE) に基づき, 次式で定義される:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\frac{1}{N} \sum_{i=1}^N (I_{\text{HR},i} - I_{\text{SR},i})^2} \right). \quad (2.7)$$

ここで MAX_I は画素値の最大値 (通常 8 ビット画像を扱うため, 255), N は画素数を表す. 一方, SSIM は輝度・コントラスト・構造の 3 要素を統合した構造的類似度指標であり,

次式で定義される：

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (2.8)$$

ここで μ_x, μ_y は平均輝度, σ_x^2, σ_y^2 は分散, σ_{xy} は共分散, C_1, C_2 は安定化定数である. これらの指標は, 2.6 項で後述する評価データセットを用いて, SR モデルの性能を比較・評価する際に広く用いられている.

2.3 知覚品質を重視した設定（既知劣化）

Johnson ら [43] は, 従来の L1 loss のみでは意味的整合性や高周波成分の再現に限界があることを指摘し, VGG ネットワークの中間特徴に基づく Perceptual loss を導入することで, SR における知覚品質を大きく改善した. この損失により, モデルは意味的整合性を保ちつつ, テクスチャや形状の再現性に優れた自然な SR 画像を生成可能となった.

この流れをさらに発展させたのが SRGAN [21] である. SRGAN は, Johnson らが導入した Perceptual loss に加え, 敵対的生成ネットワーク (Generative Adversarial Network; GAN) [22] の枠組みを統合することで, SR 画像のリアリティを一層高めた. その改良版である ESRGAN [30] は, 特徴表現の強化を目的として, 複数の残差ブロックを階層的に入れ子構造で接続した Residual-in-Residual Dense Block (RRDB) を採用し, バッチ正規化を除去することで表現力と学習安定性を両立している.

さらに近年では, より高性能な生成モデルを SR に取り入れる流れが進んでいる. GLEAN [44] は, 事前学習済みの StyleGAN [45] を復元モジュールとして活用し, エンコーダから得た特徴を StyleGAN の潜在空間にマッピングすることで, GAN の生成能力を最大限に引き出している. その後, GAN を超える新たな生成パラダイムとして, 拡散モデル [46, 47] を応

用した SR3 [48] が登場した. SR3 は, 生成過程を逐次的なノイズ除去プロセスとして定式化することで, 高精細かつ自然な SR を実現している. 現在では, 拡散モデルはその優れた知覚品質から, SR における知覚品質を重視するアプローチの中心的存在となりつつある.

このように, 既知劣化条件下における知覚品質重視のアプローチは, Perceptual loss の導入から始まり, GAN を用いた手法, さらに拡散モデルの活用へと発展しており, より自然で高品質な SR を目指して進化を続けている. 本論文では, この設定において, 基盤モデルである CLIP を活用した Perceptual loss を導入することで, 意味的・視覚的整合性に優れた出力の実現を目指す. その詳細は第 3 章で述べる.

2.3.1 学習方法

知覚品質の向上を目的としたモデルでは, 従来の L1 loss に加えて, Perceptual loss や Adversarial loss といった損失関数が導入されることが多い. Perceptual loss は, ImageNet で事前学習された VGG ネットワークの中間特徴マップを用いて, SR 画像と GT 画像の意味的・視覚的な差異を特徴空間上の距離として評価する損失関数である.

SR 画像 $\mathbf{I}_{\text{SR}}$ と GT 画像 $\mathbf{I}_{\text{HR}}$ に対して, 特徴抽出ネットワーク ϕ の第 l 層から得られる特徴マップを用いると, Perceptual loss $\mathcal{L}_{\text{per}}$ は以下のように定義される:

$$\mathcal{L}_{\text{per}} = \|\phi^{(l)}(\mathbf{I}_{\text{SR}}) - \phi^{(l)}(\mathbf{I}_{\text{HR}})\|_2^2 \quad (2.9)$$

ここで, $\phi^{(l)}(\cdot)$ はネットワーク ϕ の l 層における出力特徴を表す. この損失により, 高周波成分や意味的整合性の保持が促進され, より視覚的に自然な出力が得られる. なお, Adversarial loss については, 本研究では使用していないため, 詳細な説明は割愛する.

2.3.2 評価方法

知覚品質の評価には, LPIPS が広く用いられている. LPIPS は, ImageNet で事前学習された深層ネットワーク (AlexNet [49] や VGG など) の中間特徴マップに基づいて, SR 画像と GT 画像の視覚的整合性を定量的に評価する指標である. PSNR や SSIM といった歪み指標とは異なり, LPIPS は人間の主観的な知覚と高い相関を示すことが報告されており, 知覚品質の評価において広く採用されている. なお, 特徴抽出器としては一般に AlexNet を用いた設定が安定して良好な結果を示すことから, 多くの研究で標準的に採用されている.

2.4 知覚品質を重視した設定（未知劣化）

実環境における画像劣化は, センサー特性や撮影条件, 圧縮処理など多様な要因が複雑に関与しており, その劣化過程を厳密に定式化することは極めて困難である. このため, 理論的にモデル化するのではなく, 現実の画像劣化を模擬的に再現することで近似的に対処する手法が主流となっている.

未知劣化下において知覚品質を高めるには, 単一の劣化では不十分であり, 現実画像に近い多様な劣化を模擬する必要がある. その代表例が BSRGAN [50] であり, ブラー, ノイズ, JPEG 圧縮など複数の劣化操作をランダムな順序および強度で適用することで, 現実的な劣化分布を近似し, 実画像に対して頑健なモデルの学習を実現している

Real-ESRGAN [51] はこのアプローチをさらに発展させ, 劣化シミュレーションの精緻化に加え, Discriminator に U-Net 構造やスペクトル正規化を導入することで, 安定した学習と高い視覚品質の両立を可能にした. その後, より実用性の高い手法が数多く提案されている [52–54].

一方で, GAN ベースの手法は写実的な出力を可能とする反面, モード崩壊やアーチファクトなどの不安定性を伴う. こうした課題を克服する新たな方向性として, SR3 [48] の流れを継承した拡散モデルをベースとしたの SR モデルが注目されている. 例えば, SeeSR [55] は, 著しく劣化した実画像に対しても意味的整合性を維持しながら高品質な復元を実現し, 拡散モデルの有効性を未知劣化環境下で実証している. このように, 未知劣化下の知覚品質重視の研究は, GAN による模擬的劣化生成から, 拡散モデルによる安定的生成へと発展している.

本論文では, この設定で学習した SR モデルを用いて, 圧縮劣化を含む低品質画像を補正することで, 学習可能なデータへと再構成する新たなデータセット構築手法を提案する. その詳細は第 5 章で述べる.

2.5 学習データセット

深層学習モデルの性能は, 学習に用いるデータに大きく依存する. SR 分野においても, 高品質かつ多様な学習データセットの整備は, モデルの精度や汎化性能の向上に直結する重要な要素である.

初期の代表的なデータセットとしては, T91 [56] および BSDS200 [57] が挙げられる. それぞれ 91 枚および 200 枚の自然画像で構成され, SRCNN や VDSR といった初期の SR モデルの学習に広く用いられていた. 当時は浅いネットワークが主流であったため, 小規模データセットであっても一定の性能が得られていた.

その後, モデルの大規模化と高性能化に伴い, より高解像度かつ多様な画像を備えたデータセットの必要性が高まり, 2017 年には DIV2K [5] が登場した. DIV2K は, 圧縮ノイズの少ない 2K 解像度の画像 800 枚からなり, 現在も標準的な学習および評価用データセット

として広く使用されている。加えて, Flickr から収集された 2, 650 枚の Flickr2K [6] と統合した DF2K も, より大規模な学習のために活用されている。

さらに大規模なモデルを学習させるためには, より大規模なデータセットが求められる。その例として, 84, 991 枚の高解像度画像からなる LSDIR [58] や, 50, 000 枚で構成される HQ-50K [59] が提案されている。これらのデータセットは, 圧縮劣化が少なく品質の高い画像のみを厳選しており, 高精度な SR 学習に適している。しかしながら, 高品質画像の収集には多大なコストを伴うため, データ規模は依然として数万枚にとどまっている。

一方で, より多様な視覚情報を取り込むために, ImageNet [12] のような大規模汎用画像データセットの活用も行われている [36, 41, 60]。ImageNet は約 128 万枚の画像を含み, 多様なカテゴリを網羅しているため, 特にテクスチャや構造の多様性を学習させる上で有用である。ただし, JPEG 圧縮された低解像度画像が大半を占めているため, 高精度な SR の学習に悪影響を及ぼす可能性も指摘されている [58]。

本研究では, 上記の多様な学習データセットを目的に応じて組み合わせて使用する。以下に標準的な学習データセットを示す。

DIV2K [5]. 800 枚の高解像度画像からなるデータセットであり, 自然風景や建物, 人物など多様なシーンを含む。主に SR タスクに使用されるが, デノイジングや JPEG 圧縮除去など他の画質復元タスクにも応用されている。各画像は慎重に選定・前処理されており, 高解像度かつ自然画像の多様性を備えた構成となっている。図 2.7 は DIV2K に含まれる画像の例である。

Flickr2K [6]. Flickr³から収集された 2, 650 枚の高解像度画像で構成されており, 日常風景から都市景観, 室内空間に至るまで比較的多様な被写体が含まれている。画像ごとの品質はばらつきがあるものの, SR モデルのスケラビリティ検証のために, DIV2K と併用さ

³<https://www.flickr.com/>



図 2.7: DIV2K [5] を構成する画像の例.



図 2.8: Flickr2K [6] を構成する画像の例.

れることが多い. 両者を統合したデータセットは, **DF2K** と呼ばれる. 図 2.8 は Flickr2K に含まれる画像の例である.

LSDIR [58]. LSDIR は, 84, 991 枚の高解像度画像からなる大規模なデータセットである. 本データセットは Flickr から自動的に画像をクロールして収集されており, 多様性を確保するために Flickr2K の構築に使用されたタグ, Flickr の人気タグ, ImageNet および ImageNet-21k のラベル群に基づいた広範なキーワード検索が用いられている. 画像の選定においては, 特にブレの検出や視覚情報に乏しいフラット領域の検出が重視され, それらに該当する画像は除外された. これに加え, 解像度, アスペクト比, 撮影日, タグ, ライセンスといったメタ情報に基づく自動フィルタリングも併用されている. フィルタリング後

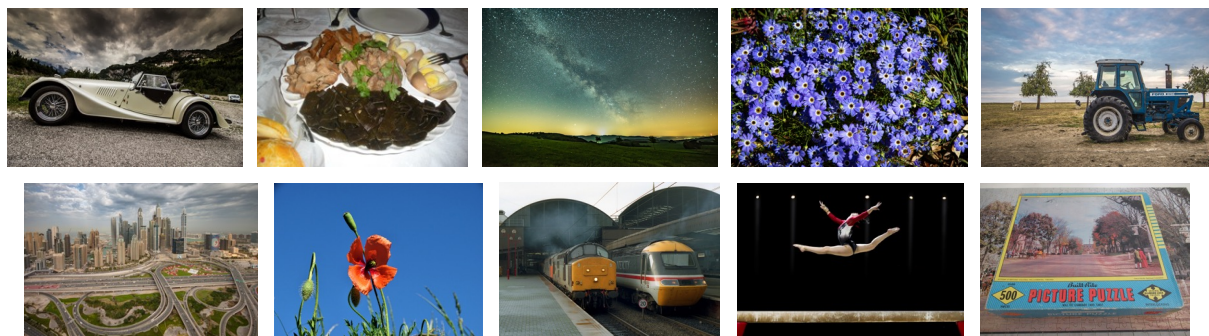


図 2.9: LSDIR [58] を構成する画像の例.



図 2.10: ImageNet [12] を構成する画像の例.

には、抽出されたサンプル画像を対象とした目視確認も行われており、画像品質の一定の保証がなされている。図 2.9 は LSDIR に含まれる画像の例である。

ImageNet [12]. ImageNet (ImageNet-21k と呼ばれる) は、約 21,000 カテゴリ・1,400 万枚以上の画像から構成される大規模な画像データセットである。その中でも、約 1,000 カテゴリ・128 万枚の画像を抜粋した ImageNet-1k が最も広く利用されており、多様な物体やシーンを含むため、汎用的な視覚表現の学習に適したデータセットである。主に VGG などの画像分類モデルの事前学習に用いられてきたほか、SR においても、Perceptual loss に用いる特徴抽出器の学習だけでなく、SR モデル本体の事前学習にも利用されている。図 2.10 は、ImageNet に含まれる画像の例である。



図 2.11: Set5 [61] を構成する全画像.



図 2.12: Set14 [62] に含まれる全画像.

2.6 評価データセット

本研究では, SR 性能を定量的・定性的に評価するために, 以下の異なる解像度や特徴を持つ評価データセットを使用する.

Set5 [61] 画像処理分野で古くから画質評価に用いられてきた 5 枚の自然画像から構成される小規模な評価データセットであり, 画像の平均解像度は約 313×336 ピクセルである.

図 2.11 に Set5 に含まれる全画像を示す.

Set14 [62] Set5 と同様, 画像処理分野で古くから画質評価に用いられてきた 14 枚の画像からなる小規模な評価データセットであり, 平均解像度は約 527×446 ピクセルである.

図 2.12 に Set14 の全画像を示す.

BSD100 [57] Berkeley Segmentation Dataset [57] から選定された 100 枚の自然画像からなる評価データセットであり, SR 分野における古典的な自然画像ベンチマークとして広

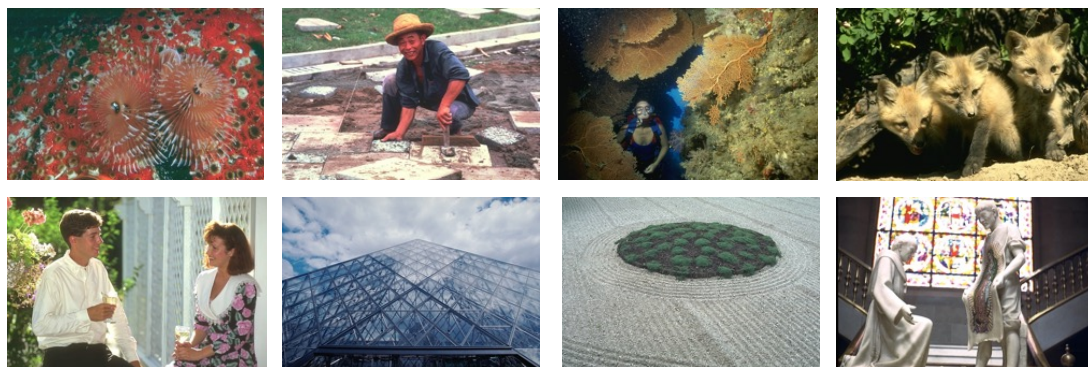


図 2.13: BSD100 [57] を構成する画像の例.



図 2.14: Urban100 [63] を構成する画像の例.

く用いられている．平均解像度は約 445×357 ピクセルで, Set5 や Set14 と同様に比較的低解像度な画像で構成されている．一般的な自然画像に対するモデルの性能評価に適している．図 2.13 は BSD100 に含まれる画像の例である．

Urban100 [63] ビルや街並みといった人工物を主題とし, 繰り返し構造や幾何学パターンを多く含む 100 枚の画像で構成される．解像度は平均 984×799 ピクセルと高く, 高周波成分の復元能力の評価に適している．図 2.14 は Urban100 に含まれる画像の例である．

Manga109 [64] 109 冊の日本の漫画作品から抽出された高解像度のページ画像で構成さ



図 2.15: Manga109 [64] を構成する画像の例.

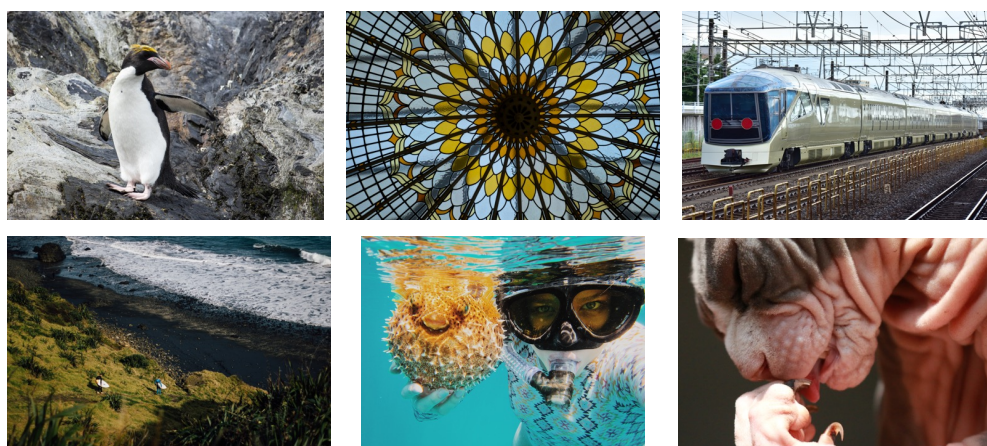


図 2.16: DIV2K Validation [5] を構成する画像の例.

れており、線画や文字など、他のデータセットとは異なる視覚的特徴を持つ。平均解像度は約 $826 \times 1,169$ ピクセルと比較的高い。図 2.15 は Manga109 に含まれる画像の例である。

DIV2K Validation [5] 高精細な 2K 解像度の画像で構成された DIV2K データセットの検証用セット 100 枚から構成される。平均解像度は $1,978 \times 1,452$ ピクセルで、本研究において使用する評価データセットの中では最も高解像度である。図 2.16 に DIV2K Validation の画像例を示す。

第3章 Perceptual lossの再考

3.1 導入

SR では，視覚的に自然な画像の再構成を目指す手法として，Perceptual loss [43] が広く用いられている．これは，SR 画像と GT 画像を ImageNet [12] で事前学習された画像分類器の中間特徴空間上で比較することで，テクスチャや構造の視覚的整合性を向上させるアプローチである．従来は CNN ベースの VGG ネットワーク [11] が特徴抽出器として主に用いられてきたが，以下の2点に課題がある．

第一に，VGG は局所的な畳み込み演算に基づくため，受容野の範囲が限られ，大域的な依存関係を十分に捉えられない．このため，高周波成分や複雑なテクスチャの再現に限界があり，出力画像にグリッド状のアーチファクトが生じる傾向がある．第二に，VGG は画像分類を目的として学習されているため，特徴表現がクラス識別に特化しており，画像全体の意味的整合性を十分に反映できる特徴表現を得るには限界がある．そのため，より高次の意味的特徴を捉えられる表現学習を導入することで，知覚品質のさらなる向上が期待される．

本章では，これらの課題を克服するため，Vision Transformer (ViT) [13] をベースとし，CLIP [8] で学習された特徴空間を用いた新たな Perceptual loss を提案する．ViT は画像をパッチ単位に分割し，自己注意機構を用いて大域的な依存関係を学習するため，局所構造に依存せず，画像全体の意味的整合性を捉えることができる．これにより，アーチ

ファクトの抑制とともに、より自然な画像再構成が期待できる。さらに、CLIP は画像と言語のペアに基づいて学習されており、単なる見た目の一致ではなく、画像が持つ意味や構造を反映した特徴表現を獲得している。これにより、SR 画像と GT 画像の比較において、人間の直感に近い損失関数の設計が可能となる。本研究では、CLIP の ViT エンコーダから抽出される特徴ベクトルを用いた損失を定義することで、アーチファクトの抑制による歪み品質の改善と、知覚的に自然な画像再構成の両立を図ることを目的とする。

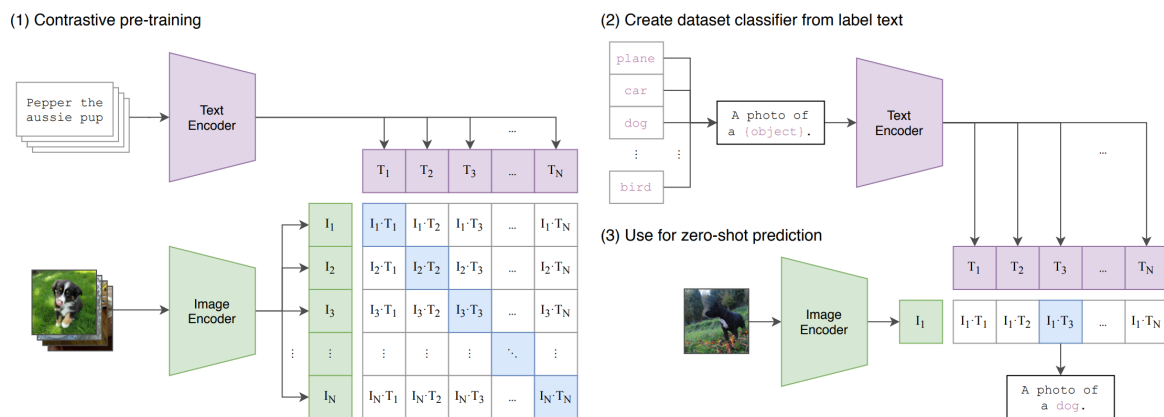


図 3.1: 画像と言語のペア学習によりゼロショット分類を実現する CLIP の構成. [8] より引用.

3.2 事前知識

3.2.1 Contrastive Language-Image Pre-training (CLIP)

CLIP [8] は、言語と画像の対応関係を対照学習によって獲得するマルチモーダル学習モデルであり、その全体構成を図 3.1 に示す。従来の ImageNet [12] に基づく画像分類モデルが、固定されたクラスラベルとの対応を学習するのに対し、CLIP は画像エンコーダとテキストエンコーダを同時に学習し、入力された画像とテキストのペアが正しい組み合わせであるかを識別するよう学習される。この学習には、Web 上から収集された約 4 億

組の画像・テキストペアが用いられており、多様な概念や文脈を横断的に理解できる表現を獲得している。

また, CLIP から得られる特徴表現は, 多様な視覚的概念を獲得しており, コンピュータビジョンの幅広いタスクにおいて有効性が確認されている. まず, 高レベルな視覚理解タスクとして, 動画認識 [65] や意味的領域分割 [66] などにおいて, CLIP の意味的特徴を活用した手法が多数提案されている. これらの研究では, CLIP の表現がカテゴリ情報や意味的整合性の保持に優れていることが示されている. 一方で, スタイル変換 [67], 3D シーン再構成 [68], 画質評価 [69] などの低レベルな画像処理タスクにおいても, CLIP の特徴は有効であることが報告されている. 特に, 色や質感といった低次視覚情報を含んだ特徴表現が, 各種の学習や評価において有用であることが確認されている.

このように, CLIP の獲得する視覚特徴は, 高次の意味理解から低次の視覚処理に至るまで, 多様なレベルのタスクにおいて汎用的に活用可能であることが, 多くの先行研究により示されている. 以上を踏まえ, 次節では提案手法について述べる.

3.3 提案手法

3.3.1 手法の流れ

本研究では, 図 3.2 に示すように, Perceptual loss の特徴抽出器として ViT を用いた SR 学習手法を提案する. まず, Bicubic 補間により高解像度画像 I_{HR} を縮小して低解像度画像 I_{LR} を生成する. 次に, I_{LR} を SR モデルに入力して, 再構成された SR 画像 I_{SR} を得る. その後, I_{SR} と I_{HR} の L1 loss $\mathcal{L}_1$ を計算する. 加えて, 特徴抽出器として ViT を用い, I_{SR} と I_{HR} をそれぞれ入力し, 出力される CLS トークンに対応する特徴ベクトルを抽出する. これらの特徴ベクトルをそれぞれ $\psi_{\text{ViT}}(I_{\text{SR}})$ および $\psi_{\text{ViT}}(I_{\text{HR}})$ とし, それらの差分に基づ

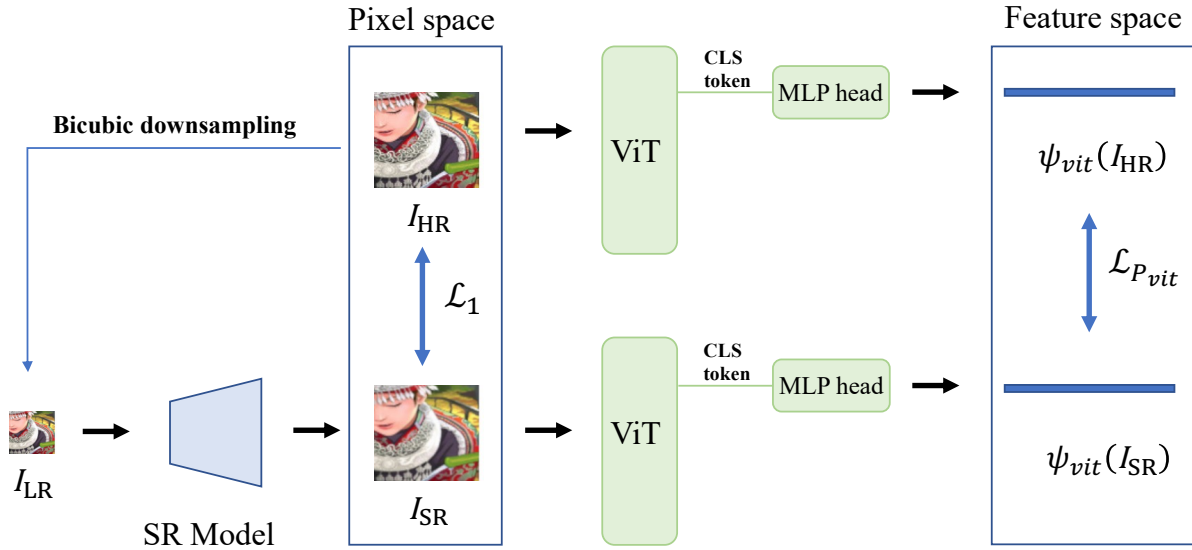


図 3.2: 提案手法の概要. Perceptual loss の算出に, VGG ネットワークではなく, ViT ベースの CLIP モデルを用いて学習を行う.

いて Perceptual loss $\mathcal{L}_{P_{ViT}}$ を算出する.

3.3.2 ViT を用いた損失関数

本研究で提案する ViT ベースの Perceptual loss は, 以下のように定義される:

$$\mathcal{L}_{P_{ViT}} = |\psi_{ViT}(I_{HR}) - \psi_{ViT}(I_{SR})|. \quad (3.1)$$

ここで, $\mathcal{L}_{P_{ViT}}$ は ViT をベースとした Perceptual loss を表し, $\psi_{ViT}(\cdot)$ は ViT から得られる特徴ベクトルを示す. ViT の事前学習には, ImageNet-1k, ImageNet-21k, および CLIP を使用する. ViT のアーキテクチャは, パッチサイズが 16×16 の ViT-B/16 と 32×32 の ViT-B/32 を採用する. ImageNet で学習した ViT では最終出力層の 1 つ手前の 768 チャンネルの特徴ベクトルを, CLIP で学習した ViT では 512 チャンネルの最終層から出力された特徴ベクトルを使用する.

学習に使用する損失関数の構成と対応するモデルは以下の通りである:

- $\mathcal{L}_1 \rightarrow \text{Baseline}$
- $\mathcal{L}_1 + \lambda \mathcal{L}_{P_{\text{VGG}}}(\text{IN1k}) \rightarrow \text{VGG Loss (IN1k)}$
- $\mathcal{L}_1 + \lambda \mathcal{L}_{P_{\text{ViT}}}(\text{IN1k}) \rightarrow \text{ViT Loss (IN1k)}$
- $\mathcal{L}_1 + \lambda \mathcal{L}_{P_{\text{ViT}}}(\text{IN21k}) \rightarrow \text{ViT Loss (IN21k)}$
- $\mathcal{L}_1 + \lambda \mathcal{L}_{P_{\text{ViT}}}(\text{CLIP}) \rightarrow \text{ViT Loss (CLIP)}$

λ は Perceptual loss の重み係数であり, モデルの歪み品質と知覚品質のバランスを制御するために使用される. なお, 括弧内は使用した事前学習の種類を示している.

3.4 実験

3.4.1 実験設定

SR モデル. 本研究では, CNN ベースのモデルである RCAN-it [70] を主に使用する. RCAN-it は, RCAN [33] で使用されている活性化関数を, ReLU から SiLU へ変更したモデルであり, より安定した学習を可能にする. また, 3.4.3 項では, 同じく CNN ベースの EDSR [29] も使用する.

データセット. 学習データセットは DIV2K [5] を用いる. 評価データセットは, SR 分野で広く使用されている Set5 [61], Set14 [62], Urban100 [63], BSD100 [57], および DIV2K Validation [5] を用いる.

評価指標. 評価には, 歪み品質を評価する PSNR と, 知覚品質を評価する LPIPS [19] を使用する. PSNR は, YCbCr 色空間に変換後の Y チャネル (輝度成分) に対して算出する. また, PSNR と LPIPS がトレードオフの関係にあることを踏まえ, 一部の実験では

Perceptual loss の重み係数 λ を変化させて学習したモデルを用い、PSNR-LPIPS 曲線により性能を比較する。このとき、曲線上で右下に位置するモデルほど、歪み品質と知覚品質のバランスに優れたモデルであると評価する。

学習設定. 超解像倍率は 4 倍に設定し、低解像度パッチサイズは 48×48 とする。データ拡張として、水平方向・垂直方向の反転および 90 度回転をランダムに適用する。最適化手法 Lamb オプティマイザ [71] を使用し、初期学習率は 3.2×10^{-3} 、バッチサイズは 64、総イテレーション数は 16 万とする。VGG ベースの Perceptual loss に関しては ImageNet-1k で事前学習された VGG19 [11] を使用し、第 4 最大値プーリング層の後にある第 4 畳み込み層（活性化後）の特徴ベクトルを用いる。

3.4.2 従来の Perceptual loss との比較

本研究では、提案する ViT ベースの Perceptual loss の有効性を検証するため、従来用いられてきた VGG ベースの Perceptual loss と比較実験を行った。図 3.3 に、各損失関数で学習したモデルによる視覚的な出力例を示す。L1 loss のみで学習した Baseline モデルでは、テクスチャの復元が不十分であり、全体的に滑らかで情報量の少ない出力となっている。一方、VGG Loss (IN1k) を適用したモデルでは、テクスチャの再現性は向上する一方で、局所的に格子状のアーチファクトが生じる傾向がある。これに対して、ViT Loss を用いたモデルでは、同等の解像感を維持しつつ、このようなアーチファクトは抑制されている。特に、ViT Loss (CLIP) によるモデルでは、細部の復元精度が最も高く、自然で破綻の少ない結果となっており、視覚的品質の観点から最も優れた結果が得られている。

図 3.4 は、各損失関数に対するモデル出力の PSNR と LPIPS のトレードオフ関係を示している。VGG Loss (IN1k) では $\lambda = \{0.005, 0.02, 0.05\}$ 、ViT Loss (IN1k/IN21k) では $\lambda = \{1, 5, 10\}$ 、ViT Loss (CLIP) では $\lambda = \{10, 25, 50\}$ を設定し、各条件下での性能を比較

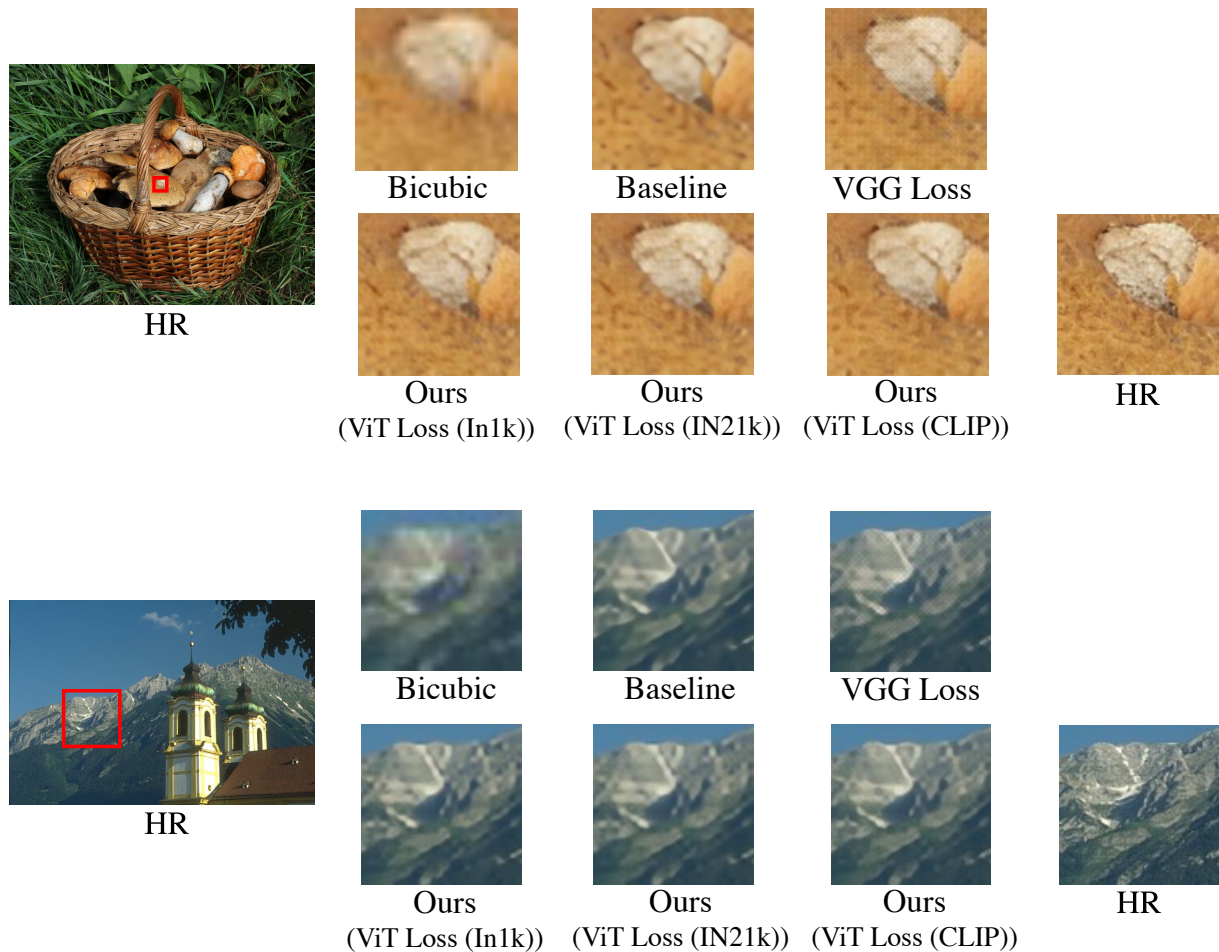


図 3.3: 学習に使用した損失関数による定性比較. HR は正解画像で, 赤枠部分の領域を比較する.

した. その結果, ViT Loss (CLIP) を適用したモデルは図中で最も右下に位置し, PSNR と LPIPS の両面において最良のバランスを示した. 一方, ViT Loss (IN21k) を用いたモデルは, ViT Loss (IN1k) の 10 倍にあたる学習画像を使用したにもかかわらず, 性能の改善は限定的であり, 学習データの内容や事前学習方法の違いが出力品質に与える影響の大きさを示唆している.

また, 表 3.1 に定量評価結果を示す. ViT Loss (CLIP) を用いたモデルは, 従来の VGG Loss を用いた場合と比較して, PSNR および LPIPS の両指標において優れた性能を示した.

以上の結果から, Perceptual loss の特徴抽出器として ViT が有効であることが示された.

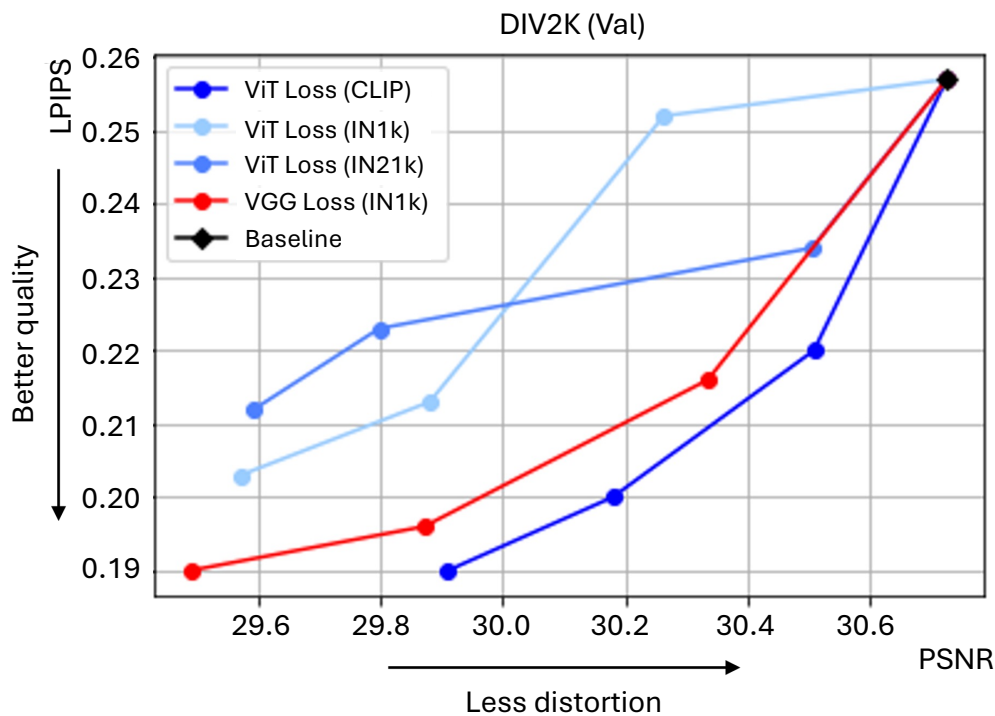


図 3.4: 学習に使用した損失関数による知覚-歪み曲線の比較

表 3.1: 異なる損失関数の組み合わせによる定量評価

Loss function	Set5		Set14		Urban100		BSD100		DIV2K (Val)	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Baseline	32.58	0.17	28.90	0.28	26.91	0.21	27.76	0.36	30.73	0.26
VGG Loss (IN1k)	31.67	0.15	28.24	0.25	25.68	0.21	27.25	0.34	29.49	0.19
ViT Loss (IN1k)	31.49	0.13	28.14	0.24	25.36	0.24	27.18	0.34	29.57	0.20
ViT Loss (IN21k)	31.42	0.14	28.11	0.25	25.28	0.25	27.18	0.35	29.59	0.21
ViT Loss (CLIP)	31.94	0.15	28.42	0.25	25.84	0.22	27.40	0.33	29.91	0.19

特に, CLIP で事前学習された ViT は, 他の事前学習設定 (IN1k, IN21k) よりも一貫して高い性能を示しており, 事前学習におけるデータ多様性とマルチモーダル学習が SR 性能に貢献している可能性が高いと考えられる. ただし, CLIP で事前学習された ViT は約 4 億の画像と言語ペアで学習されており, ImageNet-21k (約 1,419 万枚) の ViT との比較には学習規模の差がある. このため, 性能向上がマルチモーダルな事前学習に由来するのか, 単に画像枚数の増加によるものかを厳密に切り分けることは現時点では困難である.

表 3.2: 様々な SR モデルによる定量評価

SR model	Set5		Set14		Urban100		BSD100		DIV2K (Val)	
	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
RCAN-it (Baseline)	32.58	0.17	28.90	0.28	26.91	0.21	27.76	0.36	30.73	0.26
RCAN-it (VLC)	31.94	0.15	28.42	0.25	25.84	0.22	27.40	0.33	29.91	0.19
EDSR (Baseline)	32.55	0.17	28.85	0.26	26.75	0.21	27.73	0.36	30.65	0.26
EDSR (VLC)	32.10	0.13	28.58	0.24	26.25	0.20	27.50	0.31	30.11	0.17

3.4.3 SR モデルによる比較

提案する Perceptual loss の汎用性を検証するため, RCAN-it とは異なる CNN ベースの SR モデルとして EDSR を用いた実験を行った. 損失関数には Baseline (L1 loss) および ViT Loss (CLIP) を適用し, 重み係数 λ は 50 に設定した. 表 3.2 に, 各モデルにおける定量評価結果を示す.

その結果, EDSR においても RCAN-it と同様に, ViT Loss (CLIP) を用いたモデルでは PSNR が低下した一方, LPIPS が大きく改善された. これは, 提案する ViT ベースの Perceptual loss が, 学習する SR モデルのアーキテクチャに依存せず, CNN ベースの SR モデル全般に対して有効であることを示している.

3.4.4 CLIP モデルを使用した損失の重み付けによる比較

CLIP モデルを用いた Perceptual loss において, 重み係数 λ の変化が SR 性能に与える影響を検証するため, ViT ベースおよび ResNet ベースの CLIP モデルを用いて評価を行った. ViT ベースには ViT-B/32 を, ResNet ベースには RN50 を使用した.

表 3.3 および表 3.4 に, それぞれのモデルに対する定量的評価結果を示す. ViT ベースのモデルでは, λ の値が上がるにつれて PSNR は徐々に低下し, 一方で LPIPS は向上した. 一方, ResNet ベースのモデルでは, λ の増加により PSNR は同様に低下したが, LPIPS の変化は小さく, 知覚品質の改善は限定的であった.

図 3.5 は, Perceptual loss の重み係数を変化させて得られたモデル群における PSNR

表 3.3: ResNet ベースの CLIP モデルによる Perceptual loss の重み係数 λ による定量評価

λ	Set5		Set14		Urban100		BSD100		DIV2K (Val)	
	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
10	32.60	0.14	28.89	0.28	26.90	0.20	27.75	0.36	30.71	0.25
100	32.36	0.17	28.77	0.27	26.64	0.21	27.67	0.36	30.50	0.24
200	32.05	0.17	28.50	0.29	25.96	0.24	27.48	0.37	30.15	0.24
400	31.89	0.17	28.40	0.29	25.73	0.25	27.40	0.38	29.98	0.24

表 3.4: ViT ベースの CLIP モデルによる Perceptual loss の重み係数 λ による定量評価

λ	Set5		Set14		Urban100		BSD100		DIV2K (Val)	
	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
5	32.53	0.14	28.88	0.25	26.88	0.20	27.73	0.35	30.63	0.24
10	32.36	0.13	28.78	0.26	26.65	0.20	27.66	0.34	30.51	0.22
25	32.11	0.14	28.55	0.24	26.14	0.21	27.51	0.33	30.18	0.20
50	31.94	0.15	28.42	0.25	25.84	0.22	27.40	0.33	29.91	0.19

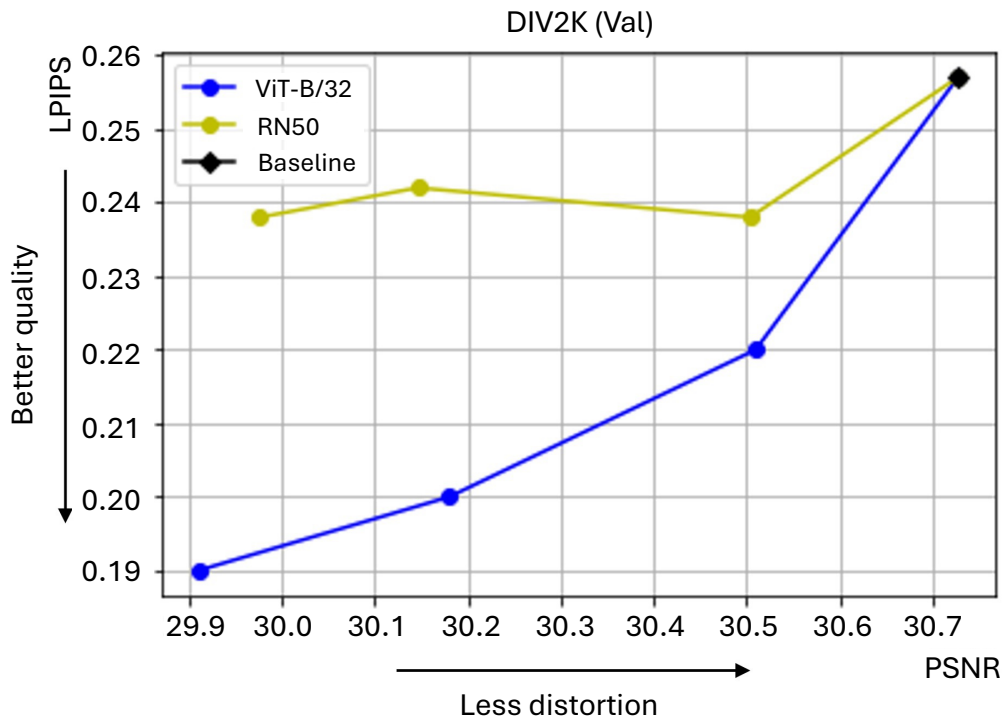


図 3.5: CLIP モデルのアーキテクチャによる知覚-歪み曲線の比較

と LPIPS のトレードオフ関係を可視化したものである。Baseline および両 CLIP モデルを比較した結果, ViT-B/32 によるモデルが最も右下に位置し, 歪み品質と知覚品質のバランスにおいて優れた結果を示した。

また, 視覚的な出力例として, 図 3.6 および図 3.7 に, 各 CLIP モデルを用いた SR モデル

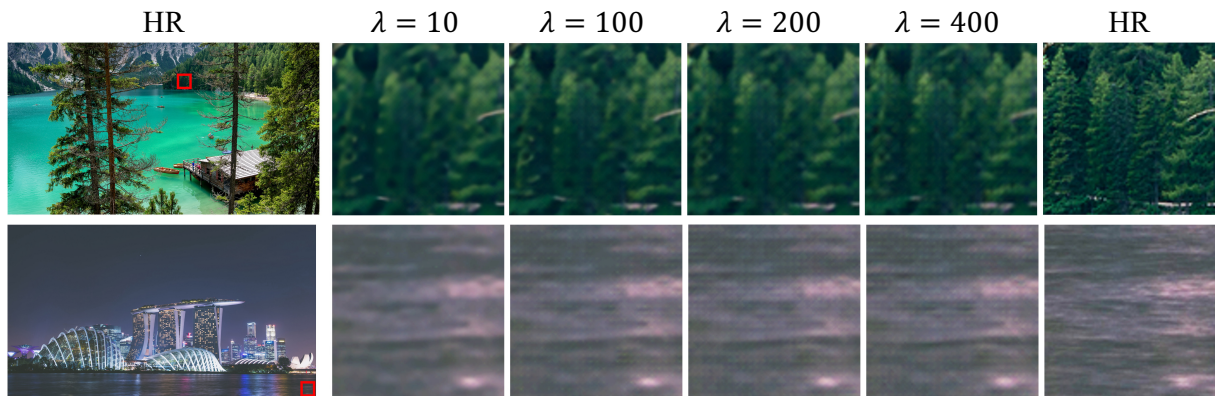


図 3.6: ResNet ベースの CLIP モデルによる Perceptual loss の重み係数の値による定性比較

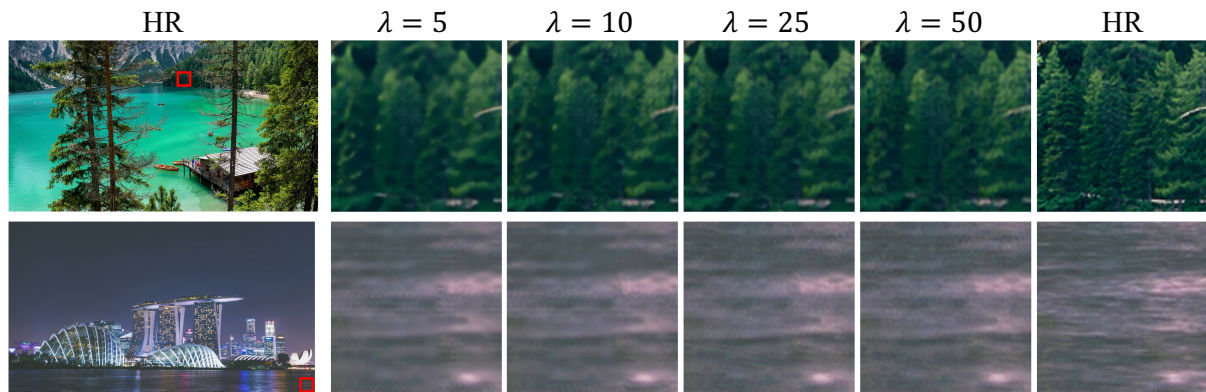


図 3.7: ViT ベースの CLIP モデルによる Perceptual loss の重み係数の値による定性比較
 の出力結果を示す. 特に, ResNet ベースの CLIP モデルにおいては, 重み係数の増加に伴い, 出力画像にグリッド状のノイズパターンが顕著に現れ, 畳み込み演算に起因するアーチファクトの可能性が示唆された.

これらの結果から, CLIP モデルを Perceptual loss として導入する際には, ViT ベースのアーキテクチャを採用することで, PSNR および LPIPS の両面で高い性能が得られることが確認された.

表 3.5: CLIP モデルのアーキテクチャによる定量評価

CLIP model	Set5		Set14		Urban100		BSD100		DIV2K (Val)	
	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
RN50	31.88	0.17	28.40	0.29	25.73	0.25	27.40	0.38	29.98	0.24
RN101	31.90	0.18	28.41	0.29	25.75	0.26	27.42	0.38	30.02	0.24
ViT-B/16	31.88	0.16	28.41	0.27	25.75	0.24	27.40	0.36	29.98	0.21
ViT-B/32	31.94	0.15	28.42	0.25	25.84	0.22	27.40	0.33	29.91	0.19

3.4.5 CLIP モデルのアーキテクチャによる比較

CLIP モデルのアーキテクチャと SR 性能の関係を調査するため, 異なるアーキテクチャを持つ 4 種類の CLIP モデルを用いて評価を行った. 使用したモデルは, ResNet ベースの RN50 および RN101, および ViT ベースの ViT-B/16, ViT-B/32 である. 損失の重み係数 λ は, RN50 および RN101 に対しては 400, ViT-B/16 および ViT-B/32 に対しては 50 に設定した.

表 3.5 に各 CLIP モデルを用いた SR 結果の定量評価を示す. PSNR が同程度である場合, ViT ベースの ViT-B/16 および ViT-B/32 は, ResNet ベースの RN50 や RN101 よりも LPIPS が一貫して低く, 知覚品質に優れていた. 特に ViT-B/32 では, ViT-B/16 よりもさらに LPIPS が改善される傾向が確認された.

これらの結果から, ViT ベースの CLIP アーキテクチャは, SR における Perceptual loss として有効であることが確認された. また, 入力パッチサイズが大きい ViT-B/32 を用いた場合, ViT-B/16 と比較して知覚品質が高い結果が得られたことから, 本実験の範囲内ではパッチサイズの増加が性能向上に寄与することが示唆された.

3.5 まとめ

本章では, 従来広く用いられてきた VGG に基づく Perceptual loss に代えて, 視覚と言語の大規模事前学習を行った基盤モデルである CLIP で学習された ViT の特徴表現を活用

し，新たな Perceptual loss を設計して SR タスクへの適用を行った．実験では，ImageNet で事前学習した VGG ベースおよび ViT ベースの損失関数をそれぞれ用いて学習を行い，出力結果を比較した．その結果，ViT ベースの損失では，従来課題とされていたグリッド状のアーチファクトの発生が抑制されることが確認された．さらに，CLIP で事前学習された ViT を特徴抽出器として用いることで，意味的整合性を保ちながら，従来の Perceptual loss と同等の解像感を維持しつつ，歪み品質と知覚品質の両立が可能であることを示した．

第4章 SR学習データ構築の再考

4.1 導入

第3章では、従来広く用いられてきたVGGネットワークに代えて、CLIPで事前学習されたVision Transformer (ViT) の特徴表現を用いた新たなPerceptual lossを設計し、歪み品質と知覚品質の両立を実現した。本章では、歪み品質の改善を目的として、SRモデルの性能を支えるもう一つの基盤である学習データセットの構築方法を再検討する。

第1章で述べたように、ネットワークアーキテクチャの改良に加え、データセットの品質と規模がモデル性能の向上においてますます重要になっている [58]。代表的な高解像度画像データセットには800枚のDIV2K [5] や2,650枚のFlickr2K [6] がある。これらのデータセットは、高解像度かつ圧縮ノイズを極力含まない高品質な画像で構成されており、SR学習を成功させる基盤を築いた。これらを組み合わせたDF2Kは、これまで多くのSRモデルの学習に利用されてきた。2023年には、DF2Kの25倍の画像枚数から構成されるLSDIR [58] が提案され、大規模なSRデータセットによる性能向上が確認されている [30, 37, 72–74]。

ここで、従来のSRデータセットは以下の観点から構築されている：

- **品質 [58]**. 画像の品質評価は人手による定性評価で実施される。評価者がJPEG圧縮による劣化 (例：アーチファクト) を確認した場合、画像はデータセットに追加されない。

- **多様性 [75].** 被写体（例：人物）や照明条件、色、テクスチャなどが含まれる。これらの要素を満遍なく含むデータセットは、幅広いドメインで優れた性能を発揮するモデルの学習に役立つとされている。
- **解像度 [59].** 指定された解像度の閾値を満たさない画像はデータセットから除外される。通常、HD, 2K, 4K の画像がデータ使用のための閾値として用いられる。

上記の観点から構築された SR データセットは、SR モデルの性能を大幅に向上させることが示されている。一方で、Web 上の画像は、軽量化のために JPEG 圧縮画像や低解像度画像が大半を占めており、上述の品質・多様性・解像度の観点から、SR データセットとして直接利用できる画像はごく一部に限られると考えられてきた。そのため、SR データセットの大規模化は難しく、SR 分野で最大級とされる LSDIR でも約 8.5 万枚に留まっており、昨今の Transformer をベースとする SR モデルの学習にはデータスケールの面で心許ないとされる。

そこで本論文は、1) 品質、2) 多様性、3) 解像度の観点から SR データセットの構築を再考した。本論文の実験を通して、我々の貢献は以下のようにまとめられる：

1) 品質の再考. 品質については、人間の定性評価により圧縮ノイズを観察した場合にその画像を除外していた。しかし、本論文では Blockiness 分布 [76] を導入することで、画像の品質を自動的に定量評価できることを明らかにした。具体的には、Blockiness measure (Blockiness) [76] のカーネル密度推定に基づいて画像品質を推定する。この方法により、データセット内の圧縮に伴うアーチファクト量の分布（Blockiness 分布）を推定する。そして、閾値以上の品質を満たしていれば、学習に使用する画像が低解像度であっても SR モデルの性能を向上できることを示した。

2) 多様性の再考. 多様性はデータセット構築時の定性評価により、多くの物体が多様な

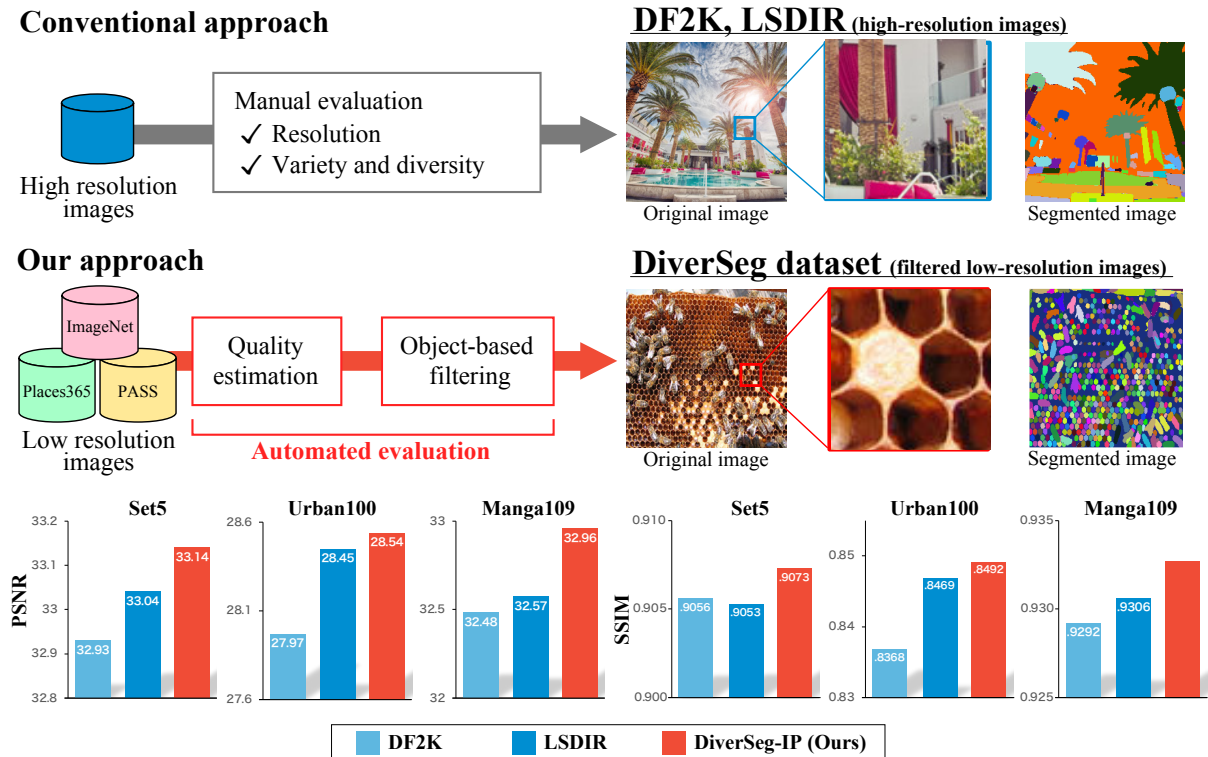


図 4.1: SR 学習データセットを構築する提案パイプライン. このパイプラインを通して得られる DiverSeg は, 低解像度だが高品質かつ多様性の高い画像で構成される. DiverSeg で学習した SR モデルは, DF2K や LSDIR などの高解像度画像データセットで学習したモデルよりも高い性能を示す.

背景下で撮影されているかで評価されてきた. そこで本論文では, 画像中に含まれるセグメント数が被写体の多様性と密接に関係するという仮説から SR モデルの性能向上に繋がった. セグメント数が 100 を超える画像を選定することで, 低解像度画像でも SR 学習を成功させることを示した.

3) 解像度の再考. 上記 1) 品質・2) 多様性の考慮, つまり Blockiness 分布及び画像セグメント数を計測して 2 段階フィルタリングを施すことにより, SR 学習のために高解像度画像は必ずしも必要ではなく, 大部分を低解像度画像が占める Web 画像でも十分であることが明らかにした. この知見により, ImageNet [12] や PASS [77] など, 本来は SR を目的として構築されていない既存の大規模画像データセットに対しても, 2 段階フィルタリングを施すことで, SR 学習に適したデータとして活用できることが示された.

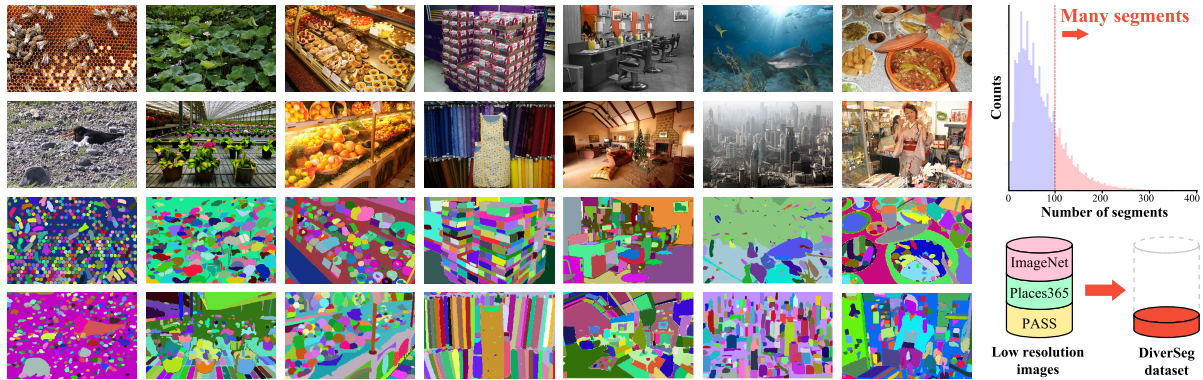


図 4.2: DiverSeg の画像とそのセグメンテーションマスクを示す. DiverSeg は, 大規模な低解像度画像群に対して自動画像評価パイプラインを適用することで得られる.

4) データセットの構築. 上記 3 つの観点を考慮することで, 低解像度画像が中心の Web 画像でありながら, 高性能な SR モデルを実現する Diverse Segmentation (DiverSeg) データセットを提案する. 図 4.1 に示すように, DiverSeg は, ImageNet-1k や PASS など, 多様性を考慮して設計されたデータセットから得られる大量の低解像度画像にフィルタリングを適用して構築される. 実験では, DiverSeg で学習した SR モデルが, DF2K や LSDIR などの高解像度 SR データセットで学習した SR モデルよりも優れていることを示す.

4.2 DiverSeg データセット

本章では, 多様なセグメント化されたオブジェクト領域を持つ低解像度であるが高品質な画像データセットである DiverSeg を紹介する. 我々は ImageNet や PASS の各画像に対して, 1) 品質を Blockiness, 2) 多様性をセグメント数にて推定し, その結果から低解像度であるが学習に使用可能な画像を選択する. DiverSeg を構築するためのパイプラインは, ソース選択とオブジェクトベースのフィルタリングという 2 つのステップで構成されており, それぞれが我々の解像度と多様性を再考する観点から設計されている. 我々のアプローチは, 高解像度画像の収集および品質評価のコストを軽減する.

4.2.1 ソース選定

$\mathcal{X}$ を低解像度の画像データセットの集合と定義する. このステップでは, 各データセット $X \in \mathcal{X}$ に対して品質 $\hat{q}_X \in [0, 1]$ を推定し, $\hat{q}_X < 0.9$ のものを除外することで, 低品質なデータセットを排除する. これは, 低品質の画像を多く含むデータセットが SR モデルの学習に有害であるという仮定の下で行われる. Blockiness 分布間の Kullback-Leibler (KL) divergence に基づく品質推定方法は以下の通りである.

画像データセット. Web から収集された低解像度データセット $\mathcal{X} = \{\text{ImageNet-1k}, \text{Places365}, \text{PASS}\}$ を使用する.

品質の定義. Y を JPEG 画像のデータセットとする. 品質 q_Y を, 平均 JPEG 品質, つまり,

$$q_Y \triangleq \frac{1}{|Y|} \sum_{y \in Y} Q(y), \quad (4.1)$$

によって定義する. ここで, $0 \leq Q(y) \leq 1$ は画像 y の JPEG 品質である (1 の場合は JPEG 非圧縮とする). 品質推定の目的は, データセット X が与えられたときに q_X を推定することである. なお, 推定段階では, データセットに JPEG 以外のフォーマットの画像が含まれることがあり, 真の品質 q_X は観測不可能である.

Blockiness 分布. 品質推定には, Blockiness measure [76] を用いる. 具体的には, 各データセット $X \in \mathcal{X}$ に対して, JPEG 品質 q を指定して圧縮した画像に対する Blockiness 値の分布 $p_{X,q}(b)$ を, 以下のカーネル密度推定で求める:

$$p_{X,q}(b) = \frac{1}{h|X|} \sum_{x \in X} K \left(\frac{b - B(\phi_q(x))}{h} \right). \quad (4.2)$$

ここで, x は画像, ϕ_q は JPEG 圧縮関数, K はガウスカーネル, h は Scott の方法で決定さ

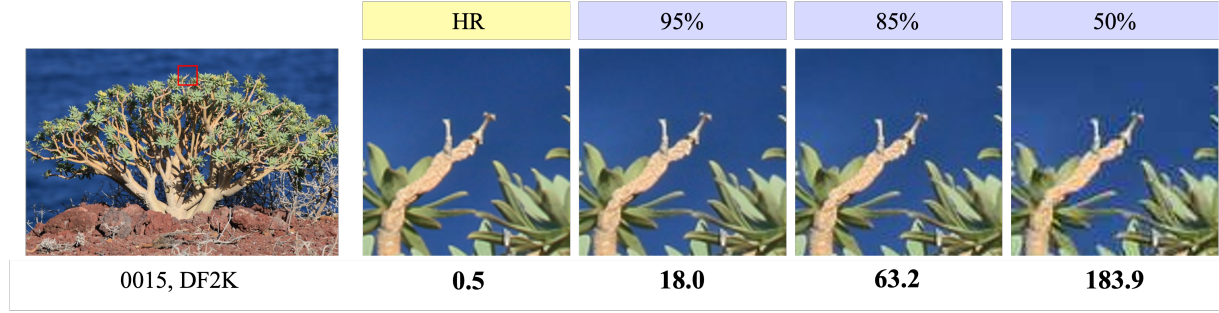


図 4.3: JPEG 品質による画像劣化の比較 (青). 各画像から算出された Blockiness 値を併記している. JPEG 品質が低下し, アーティファクトが増加するにつれて, Blockiness 値もそれに応じて上昇する傾向が見られる.

れるバンド幅である.

Blockiness 指標 B は, サブバンド DCT 係数を計算することによって JPEG 圧縮によるブロック状アーチファクトの強度を測定する. $B(x)$ は以下で定義される:

$$B(x) = \sum_{i=1}^P \sum_{j=1}^P \left| \frac{\bar{V}_{\text{crop}}(i, j) - \bar{V}(i, j)}{\bar{V}(i, j)} \right|, \quad \bar{V}(i, j) = \frac{1}{WH} \sum_{h=1}^H \sum_{w=1}^W V_{w,h}(i, j). \quad (4.3)$$

ここで $V_{w,h}(i, j)$ は各パッチとその隣接パッチ内の DCT 係数の分散, $V_{\text{crop}}(i, j)$ は先頭 4 列 4 行を除いた画像において同様に算出した分散を表す. 本研究では $P = 8$ を採用する. 非圧縮画像では $B(x)$ が低く, 圧縮画像では高くなる傾向がある (図 4.3 参照).

品質推定. $p_{X,1.0}$ と $\{p_{Z,q}\}_{q \in S}$ を比較することにより, 品質を推定する. ここで, $p_{Z,q}$ は基底分布であり, 固定品質 q の画像の Blockiness 分布である. より具体的には, 推定された品質は次のように与えられる:

$$\hat{q}_X = \sum_{q \in S} q \frac{\exp(-D_{\text{KL}}(p_{X,1.0} \| p_{Z,q}))}{\sum_{q' \in S} \exp(-D_{\text{KL}}(p_{X,1.0} \| p_{Z,q'}))}. \quad (4.4)$$

ここで, Z は非圧縮画像のみを含むデータセットである. 本論文では, DF2K を使用する.

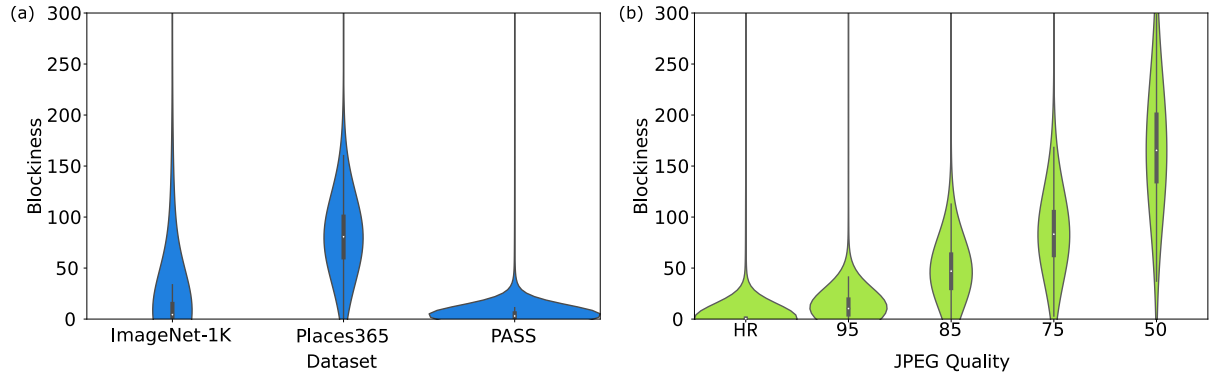


図 4.4: (a) ImageNet-1k, Places365 および PASS の Blockiness 分布. (b) DF2K の基底分布 ($q = 0.5, 0.75, 0.85, 0.95, 1.0$). 各データセットの Blockiness 分布と基底分布の KL Divergence からデータセットの品質を推定する.

D_{KL} は KL divergence であり, $S = \{1.0, 0.95, 0.85, 0.75, 0.5\}$ は離散的にサンプリングされた品質値である. 図 4.4 は, 3 つの低解像度画像データセットの Blockiness 分布 $p_{X,1.0}$ と基底分布 $p_{Z,q}$ を示している.

選択結果. ImageNet-1k, Places365, PASS の推定品質は, それぞれ 95.5%, 75.0%, 99.8% となった. この結果から, Places365 は除外される.

4.2.2 オブジェクトベースのフィルタリング

ソース選択後, SR データセットとしての効果を向上させるため, フィルタリングを適用する. これは, 均一または単調な領域を持つ画像よりも, 多様なオブジェクト領域を持つ画像の方が効果的であるという仮定の下で行われる. フィルタリング適用後のデータセット $\tilde{X}$ は

$$\tilde{X} = \{x \in X : R(x) \geq \theta\}, \quad (4.5)$$

によって与えられる. ここで, $R(x)$ はオブジェクト領域の数であり, θ は閾値である.

本研究では, 小さな領域や詳細を含む細粒度の検出から, より大まかなオブジェクト認

表 4.1: データセットの統計情報. HR: 高解像度画像, LR: 低解像度画像, #Images: 画像数, #Pixels: 平均ピクセル数, Blockiness: Blockiness 値の中央値, #Segments: 平均セグメント数. 提案データセットは DiverSeg- $\{I, P, IP\}$.

Dataset	Task	HR	LR	#Images	#Pixels	Blockiness	#Segments
DIV2K [5]	Super-resolution	✓		800	2.8M	0.47	104
DF2K [5, 6]	Super-resolution	✓		3,450	2.8M	0.47	103
LSDIR [58]	Super-resolution	✓		84,991	1.1M	0.82	92
ImageNet [12]	Image recognition		✓	1,281,167	237k	4.39	71
Places365 [80]	Image recognition		✓	1,803,460	366k	80.71	100
PASS [77]	Image recognition		✓	1,439,589	178k	3.03	74
DiverSeg-I (Ours)	Super-resolution		✓	259,448	233k	2.83	146
DiverSeg-P (Ours)	Super-resolution		✓	267,055	179k	4.39	146
DiverSeg-IP (Ours)	Super-resolution		✓	526,503	206k	3.61	146

識まで, 異なる粒度の検出が SR 性能に与える影響を評価するため, 2 種類のオブジェクトベースフィルタリング手法を導入する.

セグメンテーションベースのフィルタリング. ViT-H をバックボーンとする SAM [78] を使用し, セグメントマスク数を R と定義する. 典型的なセグメンテーションモデルがクラスラベルに基づいて意味的領域分割を実行するのに対し, SAM は意味的制約を課さない領域分割が実行できるため, 多様なオブジェクトのより細かい領域分割を可能にする. $\theta = 100$ として ImageNet-1k に適用した結果, 約 26 万枚の画像が残った.

検出ベースのフィルタリング. ViT-B をバックボーンとする Detic [79] を使用し, 検出されたオブジェクト数を R と定義する. $\theta = 18$ を設定し, 約 26 万枚の画像が残った.

4.2.3 データセット統計

表 5.1 は, データセットの統計を示す. 我々は, ImageNet-1k, PASS, およびそれらを統合して構築した DiverSeg-ImageNet (DiverSeg-I), DiverSeg-PASS (DiverSeg-P), およびそれらの統合である DiverSeg-IP という 3 つの DiverSeg データセットを作成した. DF2K

や LSDIR といった SR データセットと比較して, DiverSeg は画像数が多いものの, 全て低解像度という特徴がある. また, Blockiness 値の中央値は ImageNet-1k では減少し, PASS ではフィルタリング後に増加している. これはオブジェクトベースのフィルタリングが Blockiness 値と相関しておらず, SR データ構築において 2 つのステップが必要であることを示している.

4.3 実験

本項では, DiverSeg データセットを用いて各種 SR モデルを学習させた実験結果を示す. 高解像度画像を使用せずに SR モデルを学習可能であることを示すとともに, SR 性能向上のためにデータセットの観点から重要となる要因を詳細に分析する.

4.3.1 実験設定

SR モデル. CNN ベースのモデルとして MSRResNet [30], EDSR [29], RCAN [33], Transformer ベースのモデルとして SwinIR [37] および HAT [41] を使用する.

学習データセット. 2 つの高解像度データセット: DF2K [5,6] および LSDIR [58] と DiverSeg データセットを比較する. DF2K は DIV2K [5] と Flickr2K [6] を統合したデータセットである.

評価データセット. Set5 [61], Set14 [62], BSD100 [57], Urban100 [63], Manga109 [64] を用いる.

評価指標. YCbCr 色空間の Y チャンネル（輝度成分）における PSNR および SSIM を用いる.

学習設定. 全てのモデルに対して初期学習率は 2×10^{-4} に設定し, 最適化には Adam オプティマイザ [81] を使用する. パラメータは $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ とする.

学習スケジュールは以下の通りである. MSRResNet は 100 万イテレーションで学習し, CosineAnnealingRestartLR スケジューラを用いる. EDSR および RCAN は 30 万イテレーションで学習し, 20 万イテレーションごとに学習率を半減させる. SwinIR は 50 万イテレーションで学習し, 25 万, 40 万, 45 万, 47.5 万で学習率を半減させる. HAT は 80 万イテレーションで学習し, 30 万, 50 万, 65 万, 70 万, 75 万で学習率を半減させる. EDSR, RCAN, SwinIR, HAT には MultiStepLR を使用する.

ミニバッチサイズは, MSRResNet, EDSR, RCAN では 16, SwinIR および HAT では 32 とする. 低解像度パッチサイズは, MSRResNet で 32×32 , EDSR および RCAN で 48×48 , SwinIR および HAT で 64×64 を使用する. EDSR については, 超解像倍率 2 倍で事前学習された重みを用いて学習を行う. 特に言及のない限り, 超解像倍率は 4 倍に設定する. 高解像度と低解像度の画像ペアは Bicubic 補間によるダウンサンプリングで作成し, 損失関数には L1 loss を用いる. データ拡張には, ランダムな 90, 180, 270 度の回転および水平方向の反転を用いる.

4.3.2 主な結果

提案手法の有効性を検証するため, DiverSeg-I および DiverSeg-P を用いて 5 つの SR モデルを学習した. 結果を表 4.2 に示す. DiverSeg-I/P で学習したモデルは, DF2K や LSDIR で学習したモデルよりも優れた性能を達成した. この結果は, 提案するデータセットの有効性を示しており, 高解像度画像を用いずに SR モデルを学習した初の成功例と考えられる.

表 4.2: 既存の高解像度 SR データセット (DF2K および LSDIR) と DiverSeg の比較. DiverSeg-I および DiverSeg-P で学習されたモデルは, 高解像度画像を一切使用せずに学習されたにもかかわらず, 優れた性能を示した.

Model (Params)	Dataset	HR LR		Set5		Set14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MSRResNet [30] (1.5M)	DF2K	✓		32.23	0.8955	28.67	0.7831	27.62	0.7374	26.23	0.7897	30.64	0.9108
	LSDIR	✓		32.15	0.8948	28.66	0.7836	27.62	0.7374	26.31	0.7918	30.57	0.9105
	DiverSeg-I		✓	32.27	0.8963	28.64	0.7837	27.64	0.7378	26.31	0.7918	30.53	0.9115
	DiverSeg-P		✓	32.09	0.8943	28.61	0.7832	27.60	0.7371	26.28	0.7918	30.36	0.9101
RCAN [33] (15.5M)	DF2K	✓		32.50	0.8990	28.87	0.7885	27.75	0.7421	26.73	0.8058	31.17	0.9165
	LSDIR	✓		32.53	0.8992	28.89	0.7894	27.75	0.7425	26.91	0.8090	31.33	0.9180
	DiverSeg-I		✓	32.70	0.9012	28.98	0.7908	27.81	0.7443	27.03	0.8116	31.58	0.9210
	DiverSeg-P		✓	32.63	0.9000	28.95	0.7898	27.77	0.7435	26.99	0.8134	31.19	0.9190
EDSR [29] (43.0M)	DF2K	✓		32.61	0.8998	28.91	0.7893	27.79	0.7434	26.84	0.8089	31.38	0.9176
	LSDIR	✓		32.57	0.8992	28.97	0.7908	27.80	0.7438	27.05	0.8131	31.47	0.9192
	DiverSeg-I		✓	32.71	0.9017	28.98	0.7913	27.85	0.7453	27.10	0.8142	31.72	0.9216
	DiverSeg-P		✓	32.57	0.9002	29.06	0.7915	27.80	0.7447	27.10	0.8163	31.33	0.9191
SwinIR [37] (11.9M)	DF2K	✓		32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
	LSDIR	✓		32.86	0.9036	29.16	0.7963	27.92	0.7492	27.79	0.8331	31.98	0.9262
	DiverSeg-I		✓	32.97	0.9053	29.23	0.7970	27.98	0.7508	27.83	0.8336	32.34	0.9283
	DiverSeg-P		✓	32.85	0.9040	29.24	0.7961	27.96	0.7502	27.85	0.8349	32.28	0.9278
HAT [41] (20.7M)	DF2K	✓		33.03	0.9056	29.16	0.7964	27.99	0.7514	27.93	0.8365	32.44	0.9292
	LSDIR	✓		32.93	0.9053	29.29	0.7988	28.01	0.7525	28.45	0.8469	32.57	0.9306
	DiverSeg-I		✓	33.15	0.9071	29.46	0.8004	28.07	0.7542	28.51	0.8477	32.90	0.9325
	DiverSeg-P		✓	33.12	0.9068	29.50	0.8002	28.04	0.7536	28.53	0.8492	32.83	0.9320

4.3.3 SOTA 手法との比較

表 4.3 は, DiverSeg を用いて学習したモデルが, 既存手法と比較してより高い性能を達成し, SOTA を更新したことを示している. 特に, DiverSeg で学習した HAT および HAT-L は, ImageNet-1k 全体で事前学習し, DF2K でファインチューニングを行ったモデルよりも, 5 つのベンチマークのうち 4 つにおいて PSNR および SSIM で優れた結果を達成している. 注目すべきは, DiverSeg-I が ImageNet-1k の 77.5 この結果は, 提案するフィルタリング手法の有効性と実用的な効率性を裏付けるものである.

表 4.3: 5つのベンチマークデータセットにおける最先端手法との性能比較. DiverSeg データセットは, Transformer ベースの2つのモデルに適用した. HR および LR のチェックマークは, 高解像度データセットおよび低解像度データセットの使用をそれぞれ示す.

Method	Training Data	HR	LR	Set5		Set14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SAN [34]	DIV2K	✓		32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IGNN [82]	DIV2K	✓		32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
HAN [83]	DIV2K	✓		32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSN [84]	DIV2K	✓		32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
RRDB [30]	DF2K	✓		32.73	0.9011	28.99	0.7917	27.85	0.7455	27.03	0.8153	31.66	0.9196
RCAN-it [70]	DF2K	✓		32.69	0.9007	28.99	0.7922	27.87	0.7459	27.16	0.8168	31.78	0.9217
EDT [60]	DF2K	✓		32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
HAT-S [41]	DF2K	✓		32.92	0.9047	29.15	0.7958	27.97	0.7505	27.87	0.8346	32.35	0.9283
IPT [36]	ImageNet		✓	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [37]	DF2K	✓		32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
SwinIR [37]	DiverSeg-I (Ours)		✓	32.97	0.9053	29.23	0.7970	27.98	0.7508	27.83	0.8336	32.34	0.9283
HAT [41]	DF2K	✓		33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
HAT [41]	ImageNet→DF2K	✓	✓	33.18	0.9073	29.38	0.8001	28.05	0.7534	28.37	0.8447	32.87	0.9319
HAT [41]	DiverSeg-I (Ours)		✓	33.15	0.9071	29.46	0.8004	28.06	0.7542	28.51	0.8477	32.90	0.9325
HAT [41]	DiverSeg-IP (Ours)		✓	33.14	0.9073	29.51	0.8007	28.07	0.7542	28.54	0.8492	32.96	0.9327
HAT-L [41]	ImageNet→DF2K	✓	✓	33.30	0.9083	29.47	0.8015	28.09	0.7551	28.60	0.8498	33.09	0.9335
HAT-L [41]	DiverSeg-I (Ours)		✓	33.28	0.9083	29.54	0.8022	28.10	0.7556	28.75	0.8529	33.14	0.9340
HAT-L [41]	DiverSeg-IP (Ours)		✓	33.20	0.9080	29.59	0.8019	28.11	0.7556	28.81	0.8547	33.19	0.9342

表 4.4: Blockiness によるフィルタリング (HAT model, ImageNet-1k) . θ' : Blockiness 値に対する閾値.

θ'	#Images	Set5		Set14		BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
10	800k	33.07	0.9068	29.34	0.8000	28.04	0.7538	28.41	0.8465	32.56	0.9310
30	939k	33.13	0.9072	29.39	0.8001	28.06	0.7538	28.43	0.8463	32.70	0.9316
100	1.08M	33.11	0.9071	29.42	0.8001	28.05	0.7537	28.41	0.8457	32.84	0.9321
-	1.15M	33.08	0.9071	29.40	0.7999	28.05	0.7535	28.41	0.8457	32.88	0.9323

4.3.4 分析 1: フィルタリングの効果

Blockiness によるフィルタリング. Blockiness が SR 性能に対して重要な要素である場合, Blockiness 値に基づく単純な閾値フィルタリング

$$\tilde{X}_B = \{x \in X : B(x) \leq \theta'\}$$

は有効なアプローチに思える. しかし, 表 5.5 に示すように, 閾値 θ' を 30 から 10 に下げても PSNR や SSIM の改善は見られず, Urban100 における SSIM のみが例外的に向上した. この結果は, 本手法がオブジェクト領域の多様性を考慮せず, 一律に画像を除外していることに起因していると考えられる.

周波数ベースのフィルタリング. 画像に含まれる高周波成分はエッジやテクスチャを含み, SR 学習に有効とされてきた. こうした背景から, 高周波成分の含有率に基づいて画像をあらかじめ選別することは, SR 性能向上に向けた有効なアプローチと考えられる. そこで, 高周波成分の含有率を定量的に評価するため, 図 4.5 に示すように周波数領域を設定し, パワースペクトルに基づく高周波含有率 H_α を以下のように定義する:

$$H_\alpha(x) = \frac{\sum_{(u,v):r>r'} P_x(u,v)}{\sum_{(u,v)} P_x(u,v)} \quad (4.6)$$

$$r = \sqrt{(u - u_0)^2 + (v - v_0)^2} \quad (4.7)$$

$$r' = \alpha R \quad (4.8)$$

ここで, $P_x(u, v)$ は画像 x に対して得られるパワースペクトル, r はスペクトル中心 (u_0, v_0) からの距離, r' はスケーリング係数 $\alpha \in (0, 1)$ に基づく半径, R は画像サイズに基づく最大半径である. なお, 図表およびフィルタリング処理における高周波含有率 H_α は, 定義式に基づく値に 100 を乗じた%単位で表記している. 実験的に得られた $\alpha = 0.3$ を設定し, 高周波含有率に基づく単純な閾値フィルタリング

$$\tilde{X}_H = \{x \in X : H_\alpha(x) \leq \theta''\}$$

を行った.

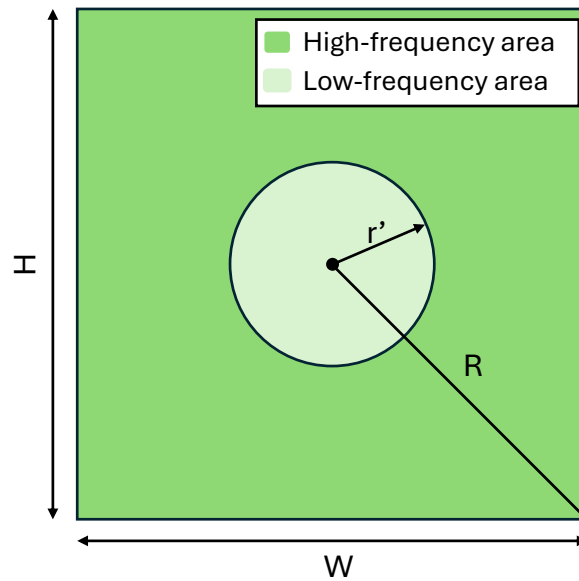


図 4.5: パワースペクトル画像における周波数領域. 図中の黒点はスペクトル中心を示し, 中心からの距離が r' を超える領域が高周波領域, $r' = \alpha R$ はスケーリング係数 α により定まる半径である. $R = \sqrt{(W/2)^2 + (H/2)^2}$ は画像サイズ (W, H) に基づく最大半径を表す.

表 4.5: 周波数成分によるフィルタリング (HAT model, ImageNet-1k). θ'' : 高周波含有率に対する閾値.

θ''	#Images	Set5		Set14		BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0.9	495k	33.02	0.9063	29.36	0.7999	28.03	0.7538	28.39	0.8466	32.57	0.9312
0.4	795k	33.14	0.9072	29.36	0.8001	28.05	0.7539	28.42	0.8465	32.73	0.9319
0.1	1.08M	32.99	0.9066	29.41	0.7999	28.05	0.7538	28.40	0.8457	32.86	0.9323
0	1.15M	33.08	0.9071	29.40	0.7999	28.05	0.7535	28.41	0.8457	32.88	0.9323

しかし, 表 4.5 に示すように, 閾値 θ'' を段階的に変化させたところ, PSNR や SSIM の顕著な改善は見られなかった. Urban100 においては, $\theta'' = 0.4$ の場合にわずかな向上が見られたものの, Blockiness によるフィルタリングと同様に他のデータセットでは性能がほぼ横ばいであり, 効果は限定的であった.

フィルタリング手法の比較. 表 4.6 は, 周波数ベース, Blockiness, 検出ベース, およびセグメンテーションベースの 4 つのフィルタリングによる比較結果を示している. いずれも ImageNet-1k を対象データセットとし, 約 26 万枚の画像に統一して学習を行っている. 結

表 4.6: フィルタリング手法の比較. 性能評価には, 約 26 万枚のフィルタリング済み画像を用いた (HAT model, ImageNet-1k) .

Filtering method	Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM
Freq.-based	28.30	0.8461	32.41	0.9302
Blockiness	28.39	0.8467	32.47	0.9304
Detection-based	28.44	0.8462	32.87	0.9322
Seg.-based	28.51	0.8477	32.90	0.9325

表 4.7: 異なる閾値 θ における性能比較 (HAT model, ImageNet-1k) .

θ	#Images	Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM
0	1.2M	28.41	0.8457	32.88	0.9323
50	663k	28.46	0.8472	32.90	0.9325
100	259k	28.51	0.8477	32.90	0.9325
150	86k	28.36	0.8452	32.83	0.9320

表 4.8: JPEG 品質の効果の分析 (HAT model).

Dataset	Quality (%)	Blockiness	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DF2K	50	165.34	31.13	0.8855	27.54	0.7533	26.32	0.7103	25.04	0.7511	30.20	0.9024
	75	83.20	32.63	0.9008	27.70	0.7457	27.70	0.7457	27.39	0.8238	31.88	0.9238
	85	46.98	32.94	0.9043	29.07	0.7941	27.93	0.7499	27.72	0.8313	32.14	0.9266
	95	10.33	32.98	0.9048	29.11	0.7951	27.99	0.7521	27.81	0.8338	32.21	0.9278
	HR	0.47	33.03	0.9056	29.16	0.7964	27.99	0.7514	27.93	0.8365	32.44	0.9292
LSDIR	50	146.43	28.50	0.8487	26.66	0.7421	24.72	0.6674	24.91	0.7657	29.39	0.8956
	75	57.14	31.87	0.8853	28.53	0.7805	27.49	0.7417	27.30	0.8231	31.83	0.9223
	85	24.14	32.84	0.9044	29.17	0.7955	27.98	0.7513	28.22	0.8412	32.38	0.9291
	95	3.71	32.97	0.9043	29.25	0.7979	28.05	0.7539	28.37	0.8452	32.42	0.9299
	HR	0.82	32.93	0.9053	29.29	0.7988	28.01	0.7525	28.45	0.8469	32.57	0.9306

果から, セグメンテーションベースの手法が最も優れた性能を示しており, 多様なオブジェクト領域を含む画像が SR 学習に有効であることが分かる. 特に, より細粒度な領域分割を行うことで, モデルの性能向上に寄与していることが示唆される.

フィルタリングの閾値. 表 4.7 は, セグメント数の閾値 θ を変更した場合の性能変化を示している. 結果として, $\theta = 100$ の設定が最も高い性能を達成した.

4.3.5 分析 2: 画像品質の影響

4.2 節では, 低品質な画像が SR モデルの学習に悪影響を及ぼすという仮定を置いた. 本節ではこの仮定を実証的に検証し, 画像品質が SR 性能に与える影響を分析する.

圧縮画像による学習. 圧縮アーチファクトを含まない高品質なデータセットである DF2K および LSDIR に対し, それぞれに JPEG 圧縮を適用してモデルを学習させた. 表 4.8 に示すように, 圧縮品質が低下するにつれて PSNR および SSIM も一貫して低下しており, 品

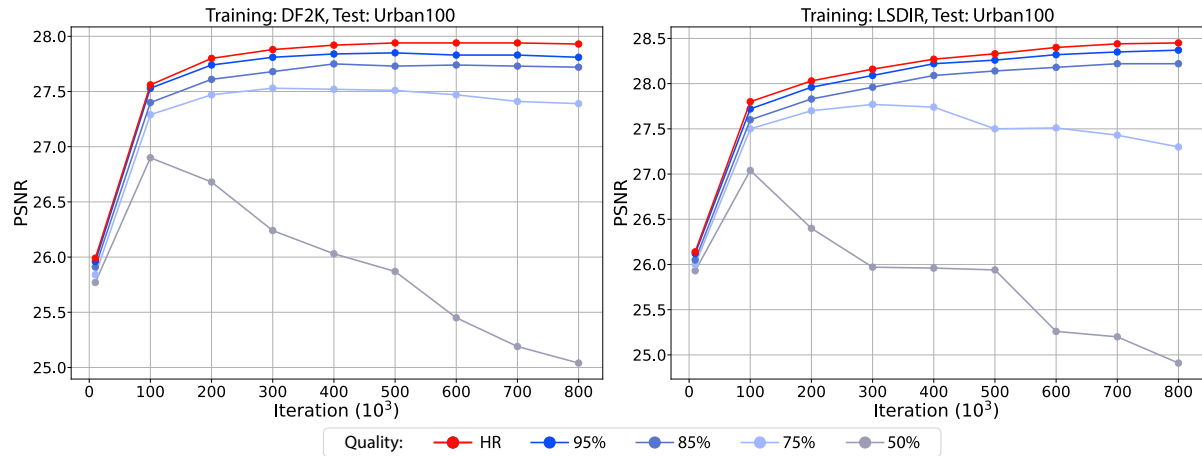


図 4.6: JPEG 圧縮による学習プロセスの比較 (HAT を使用).

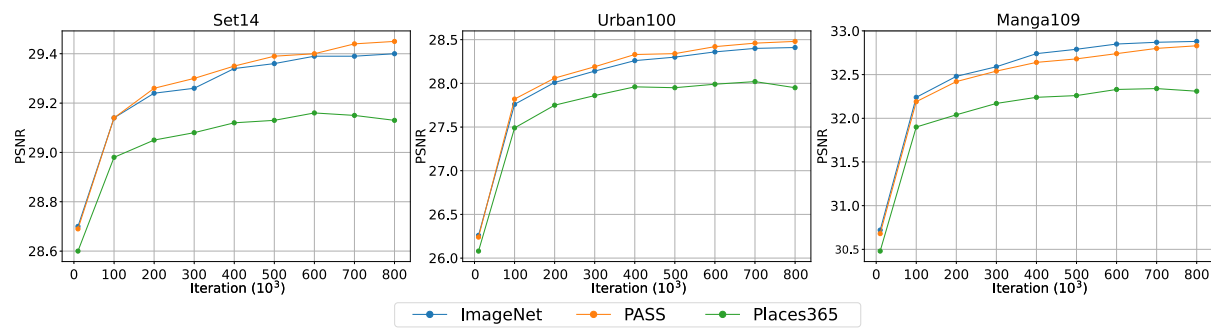


図 4.7: ImageNet-1k, PASS および Places365 における学習プロセスの比較 (HAT を使用).

質劣化が SR 性能に直接的な悪影響を与えることが確認された。

学習過程の解析. 低品質画像が性能低下を引き起こす要因を分析するため、図 4.6 に学習曲線を示す。JPEG 品質が 50% および 75% のデータセットでは、学習後半 (およそ 10 万～30 万イテレーション以降) において PSNR が減少する傾向が見られた。特に 50% の場合、最終イテレーション (80 万時点) の性能が初期 (1 万イテレーション時点) よりも低下していた。この結果は、低品質な画像の除外が SR 性能の向上において重要であることを示している。これは、低品質画像を含めた画像数の増加が性能向上に寄与することが多い画像認識など他のタスクとは対照的である。

Places365 での学習. ソース選定の段階において、Places365 は Blockiness 分布に基づき画像品質が低いと推定されたため除外された。この判断の妥当性を検証するため、図 4.7

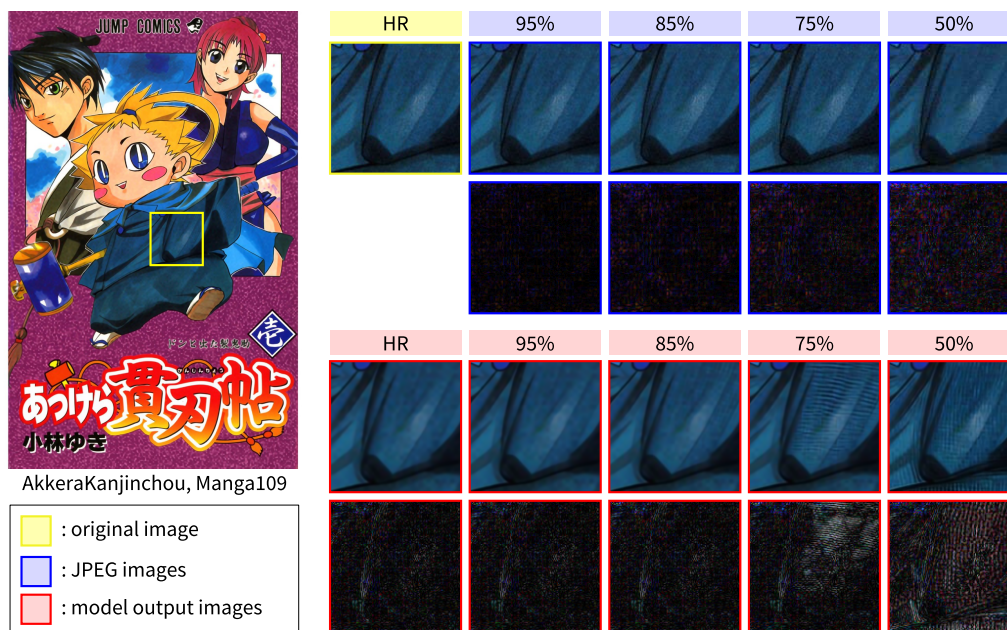


図 4.8: JPEG 圧縮によって生じたアーチファクト (青) と, 異なる JPEG 品質値で学習された SR モデル (赤, HAT モデル, DF2K) によるアーチファクトの比較. 数字は JPEG 画像の品質を示す. 1 段および 3 段: クロップ画像. 2 段および第 4 段: 差分画像.

に 3 つのデータセットで学習した場合の性能推移を示す. 結果として, Places365 で学習したモデルは ImageNet-1k および PASS を用いたモデルと比較して性能が劣り, 学習後半において PSNR の低下も確認された. この結果は, Blockiness 分布に基づくソース選定が有効であることを示している.

4.3.6 分析 3: 視覚的比較

アーチファクトの比較. 図 4.8 は, JPEG 圧縮によるアーチファクトと, 異なる品質の画像で学習された SR モデルが生成するアーチファクトの比較を示している. 50% および 75% 品質のデータで学習したモデルは, 周期的な縞や格子模様のようなアーチファクトが顕著に現れ, 元画像に JPEG 圧縮を直接適用した場合よりも強く確認された. これらのアーチファクトは, ネットワークの構造的帰納バイアス, 特に畳み込みフィルタの形状に起因している可能性があり, 今後は低品質画像を考慮したネットワーク設計が課題となる.

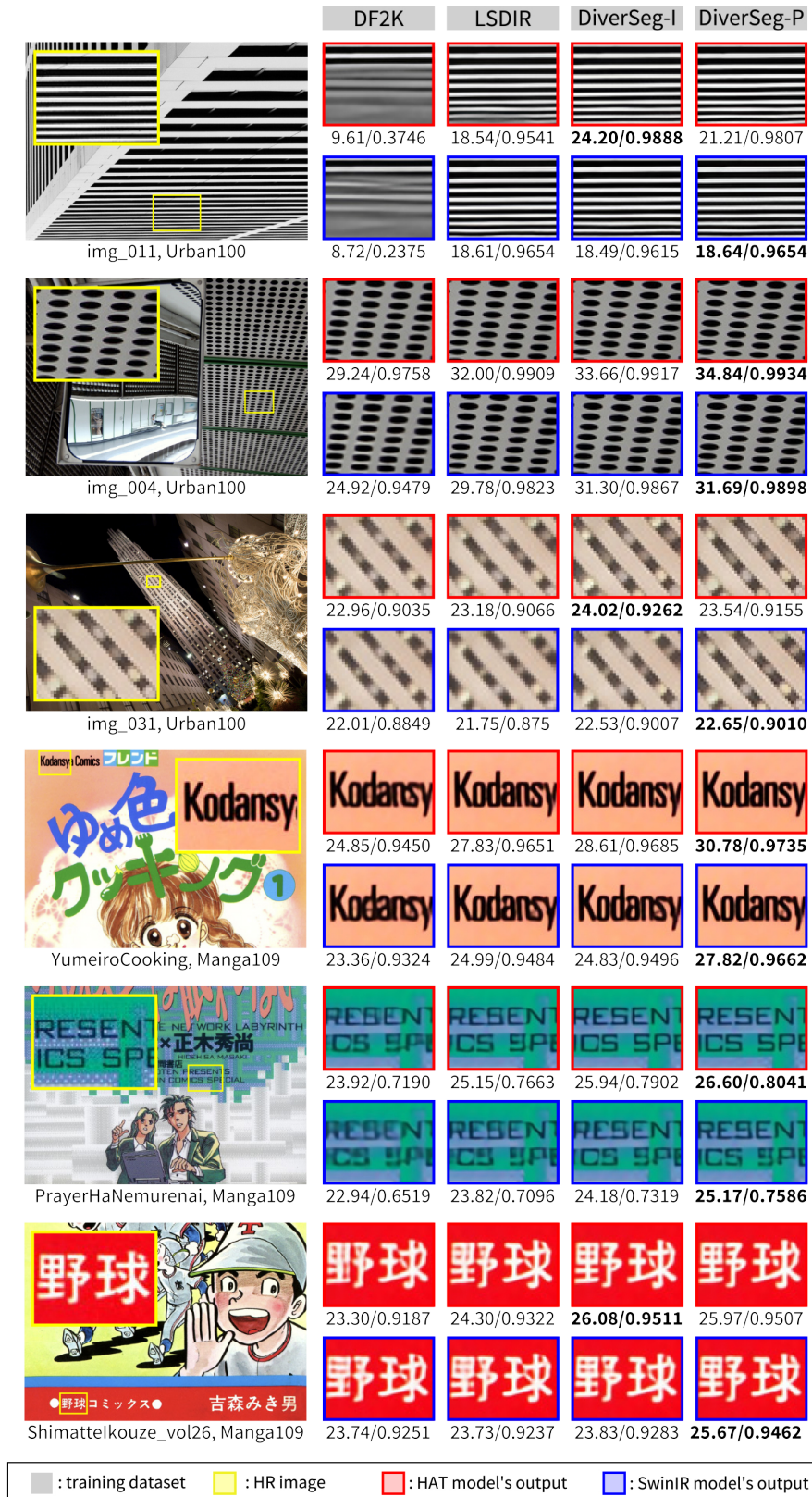


図 4.9: DF2K, LSDIR, および DiverSeg で学習した SR モデルの定性比較. 性能差をより明確に反映させるため, PSNR/SSIM はクロップしたパッチごとに個別に計算している.

定性的な比較例. 図 4.9 に, DF2K, LSDIR, DiverSeg-I, DiverSeg-P で学習した SR モデルの出力例を示す. Urban100 の “img_011” においては, DF2K で学習したモデルは繰り返しのパターンを正確に再構成できなかったのに対し, 残る 3 モデルはそれを明瞭に再現している. また, Manga109 の 3 つの画像においては, DiverSeg-I および DiverSeg-P で学習したモデルが, 文字領域において DF2K および LSDIR よりも優れた SR 性能を示している.

4.4 議論

4.4.1 高解像度データセットへの適用

本研究では, 自動画像評価パイプラインを大規模な低解像度データセットに適用し, 高解像度画像を用いずとも SR モデルの学習が可能であることを示した. この知見は, 今後の学習データセット構築やニューラルネットワークのアーキテクチャ設計に新たな視点を提供するものである. ただし, 本研究の成果は高解像度画像データセットの有用性を否定するものではない. むしろ, 提案するフィルタリング手法を高解像度データセットにも適用することで, 画像の多様性が向上し, SR 性能のさらなる改善が期待できる.

この点を検証するため, 表 4.9 に示すように LSDIR データセットに対してオブジェクト領域に基づくフィルタリングを実施した. その結果, セグメント数が 100 未満の画像を除外することで, HAT モデルの PSNR および SSIM がともに向上することが確認された.

今後の研究としては, 低解像度・高解像度の双方の多様な画像を統合的に活用する, ハイブリッド型のデータセット構築が有望であると考えられる.

表 4.9: $\theta = 100$ に設定した HAT を用いて, LSDIR に対してオブジェクトベースのフィルタリングを適用した時の性能比較.

Dataset	#Images	Set5		Set14		BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LSDIR	89,991	32.93	0.9053	29.29	0.7988	28.01	0.7525	28.45	0.8469	32.57	0.9306
w/ filtering	31,561	32.95	0.9056	29.33	0.7991	28.02	0.7526	28.53	0.8487	32.60	0.9308

4.4.2 セグメント数と周波数成分の関連性

本章で多様性の指標として導入したセグメント数は, エッジや領域の変化といった視覚的な構造を捉える特徴を持ち, 高周波成分との一定の関連があると考えられる. この関係性を定量的に検証した結果, セグメント数と高周波含有率 H_α の間には相関係数 0.23 の弱い正の相関が確認された. しかしながら, 図 4.10 に示すように, セグメント数が多いことが必ずしも高周波成分の豊富さを意味するわけではない. また, 表 4.5 に示されるように, H_α に基づく単純なフィルタリングによって SR 性能が向上するとは限らない. これは, 画像中の高周波成分を量的に捉えるだけでは, 意味的な構造や内容の多様性を十分に反映できないことを示唆している. 一方で, セグメント数に基づくフィルタリングは, 他のフィルタリング手法よりも SR 性能を向上させており (表 4.6 参照), 視覚的な複雑性や対象の多様性を間接的に捉える有効な指標であることが示された. これらの結果は, 単なる周波数情報では捉えきれない構造的な多様性を, セグメンテーションを通じて定量化することで, 学習データの質の向上に貢献することを示している. さらに, セグメント数の多い画像は, より多様な LR-HR 間の対応関係を含むとともに, 低テクスチャ領域や情報量の少ない背景が少ないため, ランダムクロップを前提としている SR 学習においても情報量の多いパッチが得られやすく, モデルの汎用性および学習効率の向上に寄与する可能性がある.

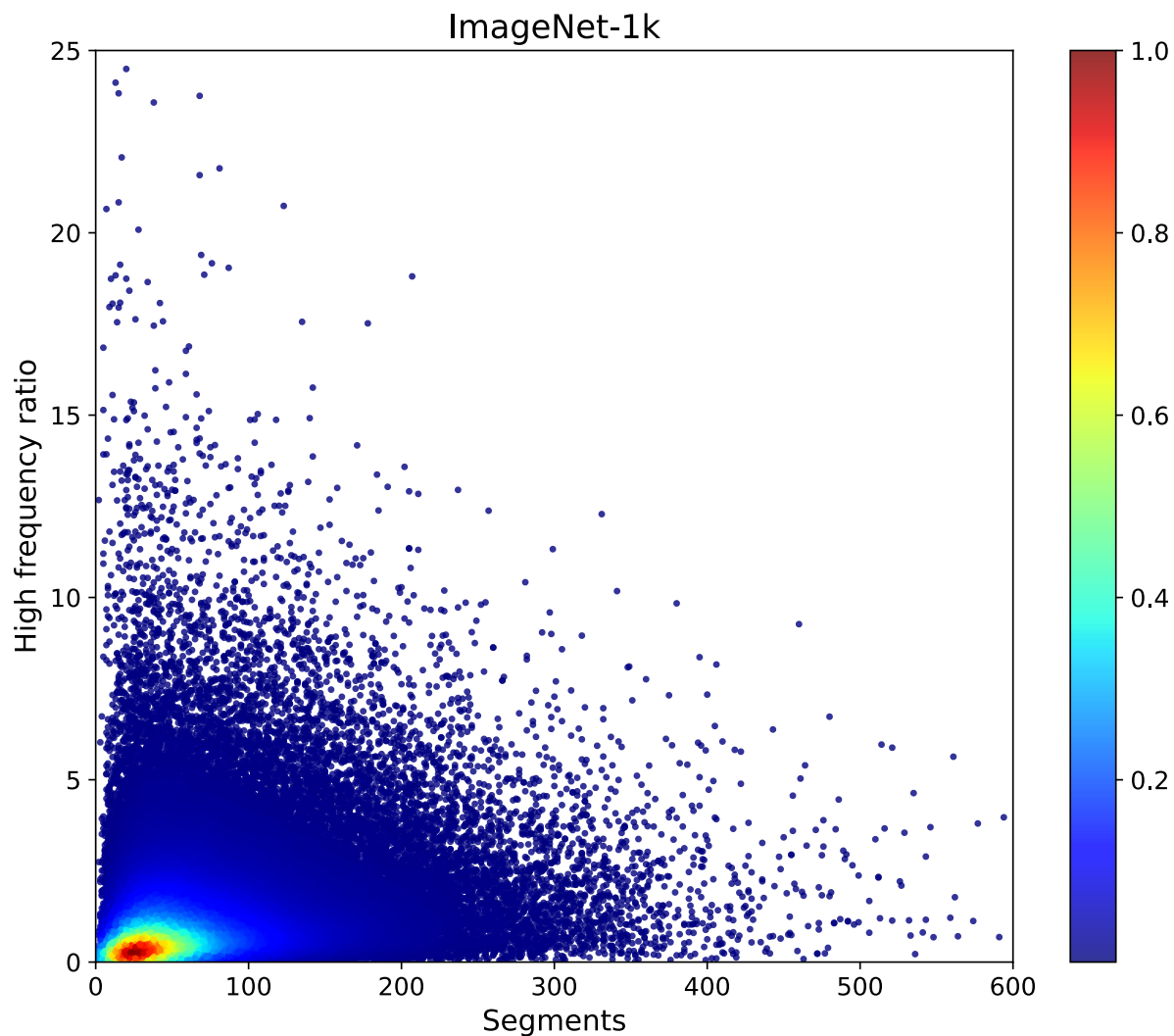


図 4.10: セグメント数と高周波含有率の分布 (ImageNet-1k). 各点は画像を表し, カラーマップはカーネル密度推定に基づく密度分布で, 最大値で正規化している. セグメント数と高周波成分の対応関係は限定的であることがわかる.

4.5 まとめ

本章では, SR 学習データの効果を調査し, 大規模な低解像度画像データセットに提案したパイプラインを適用することで, SR モデルが低解像度画像のみで学習可能であることを示した. また, 画像品質と多様性が SR 性能に与える影響を分析し, 提案したパイプラインの有効性を示した.

第5章 JPEG 圧縮画像の有効活用

5.1 導入

第4章では、高解像度画像を使用せずに、大規模な低解像度画像群から高品質かつ多様な画像を選別する自動データセット構築手法を提案した。本手法では、品質と多様性という2つの観点に基づくフィルタリング処理を導入しており、特に品質面においては、JPEG 圧縮アーチファクトの強度分布に基づき、アーチファクトの割合が高い画像を除外することで、その影響を抑えたデータセットの構築を可能にした。その結果、高解像度画像を用いることなく、SR 性能の向上が可能であることを確認した。

しかし、画像単位での除外処理には限界がある。JPEG 圧縮アーチファクトを一部に含みながらも有用な視覚情報を持つ画像まで除外される可能性があり、結果としてデータ数の減少や、多様性の損失につながるおそれがある。

本章では、この課題に対処するため、アーチファクトの強度が高いと判定された画像に対して補正処理を施し、学習に再利用可能な品質へと変換する新たなパイプラインを提案する。

図 5.1 に示すように、従来手法では圧縮アーチファクトを多く含む画像を除外することで品質を担保していたが、本手法ではそれらを補正・活用することで、品質と多様性の両立を図る。

本研究では、JPEG 圧縮画像の再利用を可能にするための学習前処理パイプライン **Re-**

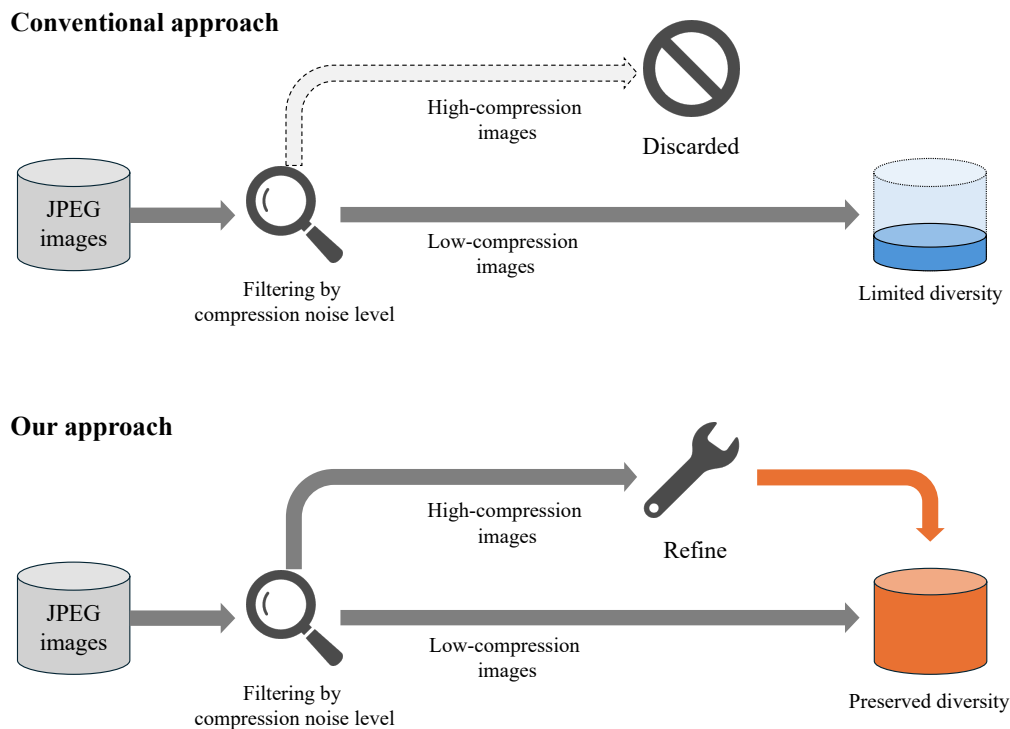


図 5.1: 従来のフィルタリング手法と、本研究で提案するデータセット構築手法との比較. 従来手法では、データセットの品質を維持するために高圧縮な JPEG 画像を除外するが、これにより学習サンプル数が大幅に減少し、多様な構造やテクスチャパターンを含む画像が失われることで、データセットの多様性が低下するという問題がある. 一方, RECA はこれらの画像を精製・再利用することで、データセットの規模と多様性の両方を維持しつつ, SR モデルの学習効果を向上させることが可能である.

finement of Compressed Artifacts (RECA) を設計する. RECA は、データセット全体に対して処理するのではなく、圧縮アーチファクトの強度に基づいて画像を選択して補正を行う. この補正処理は、JPEG 圧縮アーチファクトの除去を目的としており、元画像の構造を保持しつつ、視覚的に自然な形でアーチファクトを軽減することを目指している.

さらに、RECA の有効性を検証するため、補正処理を適用したデータセットと、従来のフィルタリング済みデータセットを用いて SR モデルを学習し、歪み品質の観点から比較評価を行う. 実験では、各種 SR モデルにおいて RECA による補正が性能向上に寄与することを定量的に示す. また、圧縮アーチファクトと学習効果の関係についても分析を行う.

本研究の主な貢献は以下の通りである:

- JPEG 圧縮アーチファクトを軽減し, 従来は除外されていた低品質画像を SR 学習に再利用可能とする新たな前処理パイプライン RECA を提案した.
- 未知の劣化に対応した SR モデルと Bicubic ダウンサンプリングを組み合わせることで, 不自然なテクスチャの発生を抑え, より自然な画質を実現した.
- RECA を適用したデータセットで学習した SR モデルが, 従来のフィルタリング手法を用いたモデルよりも, 一貫して優れた歪み品質を達成することを実証した.

5.2 前提知識

5.2.1 JPEG 圧縮アルゴリズム

JPEG 圧縮は, 画像のファイルサイズを大幅に削減できる非可逆圧縮技術であり, 帯域幅やストレージ容量に制約のある環境で広く利用されている. 一方で, 高い圧縮率ではブロック状のノイズや高周波成分の損失といった視認可能なアーチファクトが生じ, 画像品質の劣化を引き起こす. JPEG 圧縮は複数のステップから構成され, 復号時にはこれらを逆順に処理することで画像が再構成される.

クロマサブサンプリング. 人間の視覚は輝度成分 (Y) に比べて色差成分 (Cb, Cr) に対する感度が低いため, 視覚的品質を保ちながらデータ量を削減する目的でクロマサブサンプリングが行われる. 一般的には 4:2:0 方式が用いられ, 水平・垂直の両方向で輝度の半分の解像度で色差成分がサンプリングされる.

この処理の前に, 画像は RGB 色空間から YCbCr 色空間へと変換される.

$$\begin{pmatrix} Y \\ Cb \\ Cr \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.168736 & -0.331264 & 0.5 \\ 0.5 & -0.418688 & -0.081312 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} 0 \\ 128 \\ 128 \end{pmatrix}. \quad (5.1)$$

離散コサイン変換 (DCT) . 画像を 8×8 のブロックに分割し, 各ブロックに対して DCT (Discrete Cosine Transform) を適用することで, 空間領域の画素値を周波数領域に変換する. 低周波成分は緩やかな明暗変化を表現し, 高周波成分はエッジや細部を表現する.

周波数座標 (u, v) における DCT 係数 $C(u, v)$ は以下で定義される:

$$C(u, v) = \frac{1}{4} \alpha(u) \alpha(v) \sum_{x=0}^7 \sum_{y=0}^7 B(x, y) \beta(u, x) \gamma(v, y). \quad (5.2)$$

ここで $B(x, y)$ は空間座標 (x, y) における画素値を表す. DCT の基底関数 $\beta(u, x), \gamma(v, y)$ は以下の式で定義される:

$$\beta(u, x) = \cos \left(\frac{(2x+1)u\pi}{16} \right), \quad (5.3)$$

$$\gamma(v, y) = \cos \left(\frac{(2y+1)v\pi}{16} \right). \quad (5.4)$$

スケーリング係数 $\alpha(k)$ 次のように与えられる:

$$\alpha(k) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0 \\ 1 & \text{if } k \neq 0. \end{cases} \quad (5.5)$$

量子化. DCT で得られた周波数成分は, データ量をさらに削減するために量子化される.

人間の視覚は高周波成分に対して感度が低いため, 高周波領域には粗い量子化が適用され, 効率的な圧縮が可能となる.

量子化後の係数 $Q_{\text{quantized}}(u, v)$ は、以下の式によって得られる:

$$Q_{\text{quantized}}(u, v) = \text{round} \left(\frac{C(u, v)}{Q_{\text{table}}(u, v)} \right). \quad (5.6)$$

ここで $Q_{\text{table}}(u, v)$ は人間の視覚特性に基づいて設計された量子化テーブルの対応要素を示す。低周波成分には小さな値が割り当てられ、重要な視覚情報が保持される一方、高周波成分はより大きな値により圧縮が強くなる。

この量子化テーブルはユーザ指定の品質係数 QF に応じてスケーリングされ、圧縮率と画質のトレードオフを調整可能とする:

$$Q'_{\text{table}}(u, v) = \frac{Q_{\text{table}}(u, v) \times S}{50}, \quad (5.7)$$

$$S = \begin{cases} QF & \text{if } QF > 50 \\ \frac{5000}{QF} & \text{if } QF \leq 50. \end{cases} \quad (5.8)$$

高い品質係数では量子化が細かく（すなわち $Q'_{\text{table}}(u, v)$ が小さくなり）、画質が良好に保たれる。一方、低い品質係数では量子化が粗くなり、圧縮効率が高まる。したがって、量子化テーブルと品質係数は、視覚的忠実度と圧縮効率のバランスを左右する重要な要素である。

エンコーディング. 量子化された DCT 係数に対してハフマン符号化が適用され、頻出する値には短い符号語が割り当てられることで、データサイズの削減が実現される。

デコーディング. 復号処理では、ハフマン復号、逆量子化、逆 DCT、色空間変換が順に行われ、最終的に画像が再構成される。ただし、この過程で生じた情報の損失により、圧縮アーチファクトは不可避免的に画像内に残存する。

以上の JPEG 圧縮の構造と特性を踏まえ、次節では本研究で提案する手法について述べる。

5.3 提案手法

JPEG 圧縮アーチファクトに効果的に対処しつつ、有用な学習データを最大限に活用するため、本研究では新たな前処理パイプライン **RE**finement of **C**ompressed **A**rtifacts (**RECA**) を提案する。JPEG 圧縮は一般に高周波成分の損失を引き起こし、ブロックノイズやリングングといった視認可能なアーチファクトを生じさせる。これにより、SR 学習に不可欠なテクスチャ情報が著しく劣化し、性能の低下を招く。

図 5.2 に示すように、提案する RECA パイプラインは以下の 3 つのモジュールから構成され、各段階で圧縮アーチファクトに対処する：

- **Compression Noise Estimation (CNE) モジュール**：Blockiness measure [76] を用いて、強い JPEG アーチファクトを有する画像を検出し、学習に不適な低品質画像を特定する。
- **Quality Enhancement (QE) モジュール**
CNE で検出された低品質画像に対し、未知劣化に対応した SR モデルを適用することで、JPEG アーチファクトの除去と視覚品質の回復を行う。
- **Artifact Suppression (AS) モジュール**
QE モジュールで復元された画像に残存する微細なアーチファクトを、Bicubic ダウンサンプリングにより抑制し、知覚的に自然な画像を出力する。

これらのモジュールを統合することで、従来は圧縮アーチファクトのために除外されて

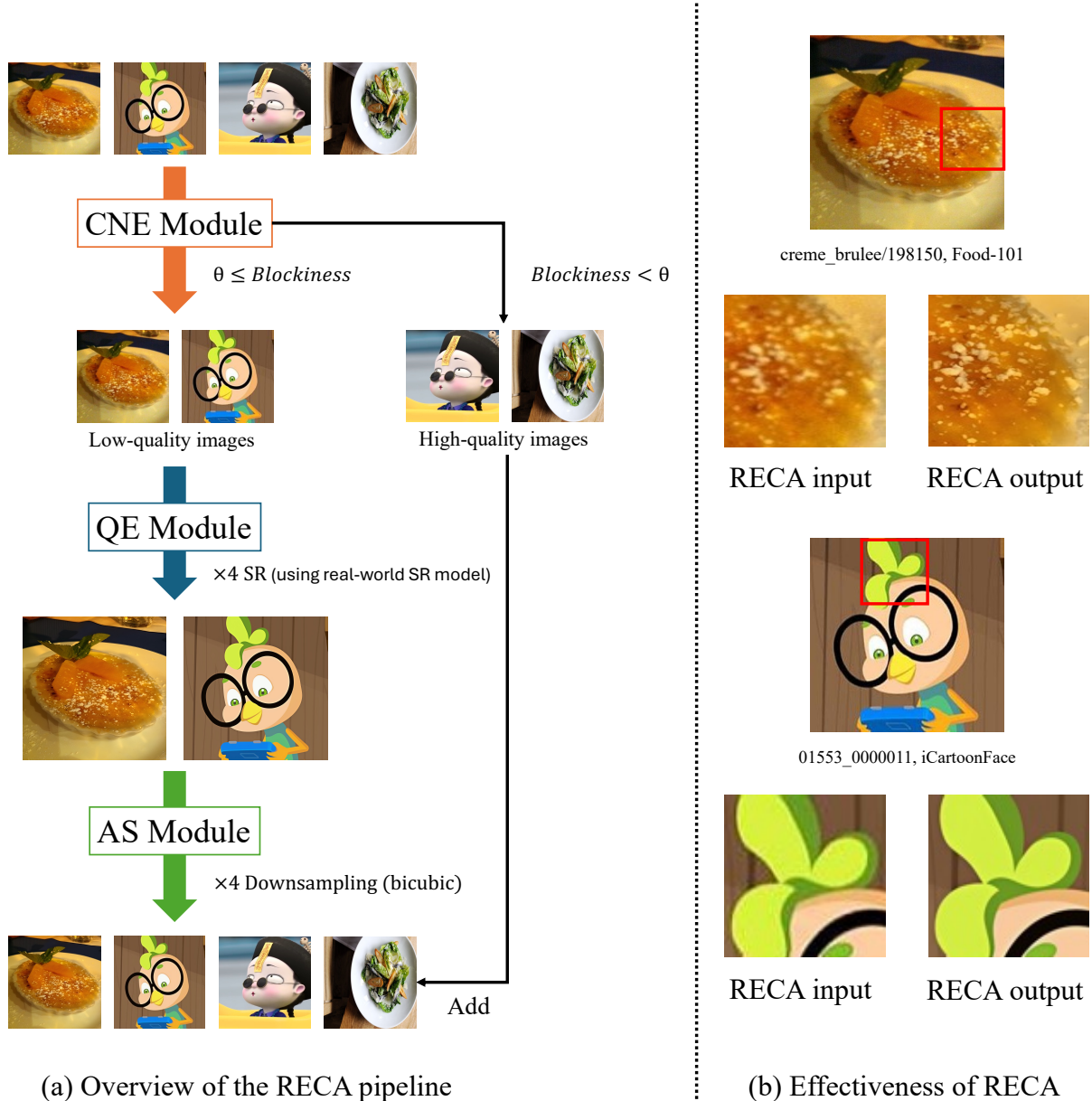


図 5.2: 提案手法 RECA パイプラインの概要とその有効性を図示する. (a) RECA は 3 つのモジュールから構成される. CNE モジュールは Blockiness 指標に基づいて低品質な画像を選別し, QE モジュールがこれらを未知劣化に対応した SR モデルにより高品質化する. ただし, この処理により新たな高周波アーチファクトが生じる可能性があるため, AS モジュールで Bicubic ダウンサンプリングを用いてそれらを抑制する. 最終的に, 補正後の画像はもともと高品質と判定された画像と統合され, 学習用データセットが構築される. (b) RECA の適用により, 入力画像と比較してノイズが軽減され, より自然なテクスチャや細部が復元されていることが視覚的に確認できる.

いた画像を, 高品質な学習データとして再活用可能にする. このパイプラインにより, データセットの規模縮小の課題を克服し, SR モデルの性能とロバスト性を大幅に向上させる

ことが可能となる.

5.3.1 Compression Noise Estimation (CNE) モジュール

CNE モジュールでは, データセット X に含まれる各 JPEG 圧縮画像 x に対して, Blockiness 値 $B(x)$ を用いて圧縮アーチファクトの強度を測定する. 所定の閾値 θ に基づき, 学習に悪影響を及ぼすと判断される画像を低品質画像として分類し, 次のようにデータセットを 2 つに分割する:

$$X_{\text{LQ}} = \{x \in X : B(x) \geq \theta\}, \quad (5.9)$$

$$X_{\text{HQ}} = \{x \in X : B(x) < \theta\}. \quad (5.10)$$

圧縮の影響を定量的に評価する指標としては, bits per pixel (bpp) も広く用いられている [59, 85]. しかしながら, bpp は画像全体の平均ビット数に基づいて情報量の損失を推定するが, JPEG アーチファクトを直接的に捉えるものではない. 特に, JPEG 画像を非圧縮形式 (PNG や BMP など) で保存し直した場合, ブロックノイズは視覚的に残存する一方でファイルサイズが大きくなるため, bpp は元の圧縮の影響を正しく反映できないという問題がある. これに対して, CNE モジュールで用いる Blockiness 値は, JPEG に特有な構造的アーチファクトに着目しており, 保存形式にかかわらず圧縮劣化の影響を安定的に捉えることができる. 5.4.4 節の実験的観察に基づき, 本研究では最適なバランスを実現する閾値として $\theta = 50$ を採用する.

5.3.2 Quality Enhancement (QE) モジュール

QE モジュールでは, CNE モジュールによって低品質と判定された画像 $x_{LQ} \in X_{LQ}$ に対し, 未知劣化に対応した SR モデルを適用することで補正を行う:

$$X_{QE} = \{RS_{\theta_{SR}}(x_{LQ}) : x_{LQ} \in X_{LQ}\}. \quad (5.11)$$

ここで, $RS_{\theta_{SR}}$ はモデルのパラメータ θ_{SR} である.

本研究では, Diffusion ベースの SR モデルである SeeSR [55] を使用し, 超解像倍率を 4 倍に設定して, アーチファクト除去と高解像度化を同時に行っている. SeeSR を選択した理由の詳細は 5.4.6 項で述べる.

Diffusion ベースのデノイジングモデルの使用も検討されたが, これらの手法はノイズ除去には優れる一方で, 不自然なテクスチャや新たなアーチファクトを生成する傾向がある. さらに, 出力解像度が入力と同一であるため, 後段の AS モジュールにおける縮小処理を適用できないという制約もある. 以上の理由から, アーチファクト除去と解像度拡張を同時に実現可能な SeeSR を本研究では採用した.

5.3.3 Artifact Suppression (AS) モジュール

QE モジュールで生成された画像は, 元の解像度の 4 倍に拡大されている. AS モジュールでは, これらの画像を Bicubic ダウンサンプリングによって元の解像度に縮小することで, 不自然な高周波成分を抑制し, SR 学習における汎化性能の向上を図る.

具体的には, 各画像 $x_{QE} \in X_{QE}$ に対して, 倍率 s に基づき以下の処理を行う:

$$x_{AS} = (x_{QE}) \downarrow_s. \quad (5.12)$$

ここで, x_{AS} はアーチファクトが抑制された画像を表す. 本研究では $s = 4$ に設定している.

この処理により得られるサブセット X_{AS} は次のように定義される:

$$X_{AS} = \{x_{AS} : x_{QE} \in X_{QE}\}. \quad (5.13)$$

最終的な学習用データセット $\tilde{X}$ は, AS モジュールによる出力と, CNE モジュールにより選別された高品質サブセット X_{HQ} を統合することで構成される:

$$\tilde{X} = X_{AS} + X_{HQ}. \quad (5.14)$$

5.3.4 データセット統計

表 5.1 に, 本研究で使用した各データセットの統計を示す. 本研究では, JPEG 圧縮アーチファクトを多く含む既存のデータセット (iCartoonFace [86], Food-101 [87]) に対して RECA を適用し, 画像数を維持したまま品質の改善を図った.

RECA 適用後, iCartoonFace の Blockiness 値は 69.44 から 1.84 に, Food-101 では 96.69 から 3.22 へと大幅に低下した. この結果は, RECA が JPEG アーチファクトを効果的に抑制しつつ, データセットの規模を損なうことなく視覚品質を大きく向上させることを示している.

表 5.1: データセットの統計情報. #Images: 画像数, #Pixels: 平均ピクセル数, Blockiness: Blockiness 値の中央値. チェックマークは, RECA が Blockiness 閾値 $\theta = 50$ で適用されたことを示す.

Dataset	RECA	#Images	#Pixels	Blockiness
DIV2K [6]	-	800	2.8M	0.47
iCartoonFace [86]	-	352,983 ⁵	139k	96.69
Food-101 [87]	-	101,000	235k	69.44
iCartoonFace (RECA) [86]	✓	352,983	139k	3.22
Food-101 (RECA) [87]	✓	101,000	235k	1.84

5.4 実験

5.4.1 実験設定

SR モデル. 3 種類のアーキテクチャに基づく 7 つの SR モデルを学習・評価する. 具体的には, CNN ベースの MSRRResNet [30], EDSR [29], RCAN [33], Transformer ベースの FreqFormer [88], HAT [41], および Mamba ベースの MambaIR-light, MambaIR [42] を使用する. MSRRResNet, EDSR, RCAN は BasicSR ライブラリ [89] を用いて実装し, その他のモデルはそれぞれの公式 GitHub 実装を用いる.

さらに, RECA の QE モジュールとして, 事前学習済みの拡散モデル SeeSR [55] も併用する. SeeSR の推論は公式実装とデフォルト設定を用い, 追加のファインチューニングは行っていない.

学習データセット. 高圧縮画像を多く含む iCartoonFace [86] および Food-101 [87] を使用する.

⁵本章では計算資源の制約により, 縦または横のいずれかの解像度が 800 ピクセル以上の画像を除外し, 389,678 枚中 352,983 枚の画像を使用した.

評価データセット. Set5 [61], Set14 [62], BSD100 [57], Urban100 [63], Manga109 [64] を使用する.

評価指標. 評価には, YCbCr 色空間の Y チャンネル (輝度成分) における PSNR および SSIM を用いる.

学習設定. 初期学習率は, MSRResNet, FreqFormer, HAT, MambaIR-light に対しては 2×10^{-4} , EDSR, RCAN, MambaIR に対しては 1×10^{-4} に設定する. 全てのモデルは Adam オプティマイザ [81] で最適化され, そのパラメータは $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ に設定する.

学習スケジュールは以下の通りである. MSRResNet は 100 万イテレーションで学習し, CosineAnnealingRestartLR を用いる. EDSR および RCAN は 30 万イテレーションで学習し, 20 万イテレーションごとに学習率を半減させる. FreqFormer, HAT, MambaIR-light, MambaIR は 50 万イテレーションで学習し, 25 万, 40 万, 45 万, 47.5 万イテレーションで学習率を半減させる. これらのモデルには MultiStepLR スケジューラを用いる.

ミニバッチサイズは, MSRResNet, EDSR, RCAN では 16, FreqFormer, HAT, MambaIR-light, MambaIR では 32 に設定する. 低解像度パッチサイズは, MSRResNet では 32×32 , EDSR, RCAN, MambaIR-light, MambaIR では 48×48 , FreqFormer および HAT では 64×64 を用いる. 超解像倍率は 2 倍および 4 倍に設定する. EDSR, HAT, MambaIR の 4 倍超解像モデルの学習には, 対応する 2 倍超解像モデルで事前学習された重みを初期化に用いる. 高解像度と低解像度の画像ペアは Bicubic ダウンサンプリングにより生成し, 損失関数には L_1 損失を用いる. データ拡張には, 90 度, 180 度, 270 度のランダム回転および水平方向の反転を用いる.

表 5.2: ベースライン, Blockiness フィルタリング, 提案手法 RECA の性能比較 (RCAN model, iCartoonFace, $\times 4$ SR).

Method	#Images	Set5		Set14		BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	352,983	32.41	0.8977	28.70	0.7836	27.63	0.7364	26.36	0.7924	31.55	0.9181
B. filtering	107,249	32.42	0.8977	28.75	0.7844	27.64	0.7377	26.49	0.7961	31.61	0.9196
RECA (Ours)	352,983	32.49	0.8989	28.82	0.7862	27.70	0.7390	26.55	0.7971	31.73	0.9209

5.4.2 フィルタリング手法との比較

提案手法 RECA の有効性を検証するため, iCartoonFace データセットを用いて RCAN モデルによる実験を行った. 比較手法として, Blockiness 値が閾値 $\theta = 50$ を超える画像を除外する単純なフィルタリング手法を採用する.

表 5.2 に示す通り, Blockiness フィルタリングでは画像数が 352,983 枚から 107,249 枚へと大幅に削減された. その結果, 多くのベンチマークにおいて PSNR および SSIM の向上が確認されたが, これは学習データの大幅な削減を伴っている. 一方, RECA は全ての画像を保持したまま圧縮アーチファクトを効果的に除去し, 全ベンチマークにおいて最高性能を達成した.

図 5.3 は, ベースライン, Blockiness フィルタリング, RECA の出力結果を視覚的に比較したものである. RECA は他の手法と比べてエッジの鮮明さを効果的に改善しており, 特に建築物の繰り返しパターンやイラストの線画において, より自然で高精細な復元を実現している. これらの視覚的改善は, PSNR および SSIM における定量的な向上とも一致しており, RECA が構造情報を損なうことなく JPEG 圧縮アーチファクトを除去できることを示している.

総じて, Blockiness フィルタリングは画質向上に寄与するものの, データ量の大幅な削減により大規模学習の利点を損なう. 一方で RECA は, データの品質と多様性の両立を図りつつ, モデルの汎化性能および出力品質の向上に寄与する有望な手法であることが示唆

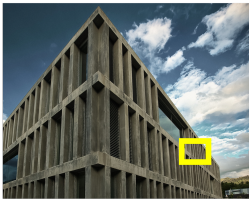
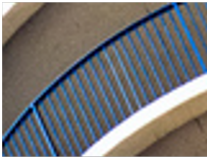
	HR	Baseline	B. filtering	RECA (Ours)
	 pepper, Set14	 26.60/0.7469	 26.89/0.7556	 27.29/0.7778
	 img027, Urban100	 21.18/0.5011	 21.44/0.5496	 21.96/0.6102
	 img100, Urban100	 28.13/0.6948	 28.47/0.7009	 28.87/0.7056
	 AosugiruHaru, Manga109	 37.46/0.9517	 38.05/0.9611	 38.21/0.9614
	 UchiNoNyansDiary_000, Manga109	 20.81/0.8071	 21.01/0.8331	 22.07/0.8781

図 5.3: ベースライン, Blockiness フィルタリング, および提案手法 RECA の視覚的比較 (RCAN model, iCartoonFace, $\times 4$ SR). 性能差をより明確に反映させるため, PSNR/SSIM はクロップしたパッチごとに個別に計算している.

された.

5.4.3 RECA の有効性

提案手法 RECA の有効性を検証するため, JPEG 圧縮アーチファクトを多く含む iCartoonFace および Food-101 データセットを用いて, 異なるアーキテクチャの SR モデルを

学習させた。その結果を、表 5.3 および表 5.4 に示す。

表 5.3: 異なるアーキテクチャを持つ SR モデルの定量的比較 (iCartoonFace)。チェックマークは RECA を使用したことを示す。

SR model (Params)	Scale	RECA	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MSRResNet [30] (1.4M)	×2	✓	37.73	0.9593	33.52	0.9148	31.98	0.8968	31.62	0.9221	38.65	0.9770
	×2	-	37.57	0.9591	33.28	0.9131	31.90	-0.8959	31.35	0.9195	38.54	0.9772
EDSR [29] (40.7M)	×2	✓	38.10	0.9610	33.83	0.9174	32.21	0.9000	32.32	0.9286	39.33	0.9789
	×2	-	37.94	0.9609	33.67	0.9165	32.14	0.8991	32.05	0.9260	39.25	0.9788
RCAN [33] (15.4M)	×2	✓	38.10	0.9610	33.86	0.9174	32.22	0.8999	32.37	0.9287	39.59	0.9793
	×2	-	37.98	0.9610	33.66	0.9161	32.14	0.8989	32.07	0.9261	39.44	0.9789
FreqFormer [88] (870k)	×2	✓	38.12	0.9610	33.82	0.9172	32.21	0.8997	32.27	0.9275	39.54	0.9789
	×2	-	37.96	0.9606	33.60	0.9157	32.13	0.8986	31.95	0.9246	39.35	0.9788
HAT [41] (20.6M)	×2	✓	38.45	0.9624	34.31	0.9217	32.49	0.9035	33.36	0.9371	40.38	0.9811
	×2	-	38.28	0.9624	34.11	0.9192	32.39	0.9023	32.88	0.9331	40.27	0.9811
MambaIR-light [42] (905k)	×2	✓	37.94	0.9603	33.68	0.9164	32.10	0.8979	31.97	0.9250	39.24	0.9781
	×2	-	37.77	0.9599	33.48	0.9145	32.02	0.8971	31.60	0.9215	39.05	0.9780
MambaIR [42] (20.4M)	×2	✓	38.34	0.9618	34.07	0.9190	32.44	0.9027	32.97	0.9334	40.33	0.9809
	×2	-	38.22	0.9619	33.89	0.9182	32.34	0.9013	32.58	0.9304	40.16	0.9807
MSRResNet [30] (1.5M)	×4	✓	32.08	0.8939	28.53	0.7800	27.51	0.7332	25.98	0.7808	30.88	0.9121
	×4	-	31.92	0.8922	28.48	0.7783	27.47	0.7316	25.89	0.7796	30.62	0.9096
EDSR [29] (43.0M)	×4	✓	32.46	0.8982	28.87	0.7881	27.71	0.7399	26.54	0.7981	31.80	0.9213
	×4	-	32.32	0.8969	28.71	0.7841	27.63	0.7368	26.39	0.7935	31.51	0.9181
RCAN [33] (15.6M)	×4	✓	32.49	0.8989	28.82	0.7862	27.70	0.7390	26.55	0.7971	31.73	0.9209
	×4	-	32.41	0.8977	28.70	0.7836	27.63	0.7364	26.36	0.7924	31.55	0.9181
FreqFormer [88] (890k)	×4	✓	32.45	0.8979	28.78	0.7843	27.66	0.7384	26.33	0.7924	31.52	0.9185
	×4	-	32.37	0.8972	28.67	0.7819	27.61	0.7360	26.29	0.7902	31.41	0.9174
HAT [41] (20.8M)	×4	✓	32.87	0.9037	29.11	0.7937	27.89	0.7468	27.32	0.8178	32.56	0.9294
	×4	-	32.66	0.9012	28.95	0.7891	27.77	0.7423	26.91	0.8080	32.23	0.9258
MambaIR-light [42] (924k)	×4	✓	32.32	0.8964	28.62	0.7812	27.53	0.7341	26.06	0.7830	31.02	0.9140
	×4	-	32.23	0.8955	28.55	0.7791	27.53	0.7327	26.01	0.7810	31.05	0.9130
MambaIR [42] (20.6M)	×4	✓	32.75	0.9017	29.03	0.7906	27.85	0.7440	26.93	0.8077	32.34	0.9266
	×4	-	32.47	0.8988	28.83	0.7857	27.71	0.7388	26.65	0.7992	31.83	0.9217

RECA を適用したモデルは、大半の構成において非適用時よりも高い性能を示した。特に iCartoonFace データセットにおいては、PSNR および SSIM の改善幅が大きく、RECA の効果がより顕著に現れた。この改善は、MSRResNet, EDSR, RCAN といった CNN ベースのモデルに加え、Transformer ベースの FreqFormer や Mamba アーキテクチャに基づく MambaIR-light といった、異なる構造のモデルにおいても一貫して確認された。これらの結果は、RECA がアーキテクチャに依存せず、幅広い SR モデルに対して有効に機能することを示しており、高圧縮データセットに対する汎用的な学習データ構築手法になり得ることを示唆している。

表 5.4: 異なるアーキテクチャを持つ SR モデルの定量的比較 (Food-101) . チェックマークは RECA を使用したことを示す.

SR model (Params)	Scale	RECA	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MSRResNet [30] (1.4M)	$\times 2$	✓	37.91	0.9603	33.52	0.9165	32.11	0.8986	31.62	0.9225	38.63	0.9778
	$\times 2$	-	37.77	0.9597	33.39	0.9151	32.05	0.8976	31.46	0.9207	38.76	0.9782
EDSR [29] (40.7M)	$\times 2$	✓	38.09	0.9611	33.95	0.9188	32.28	0.9010	32.38	0.9303	39.37	0.9795
	$\times 2$	-	37.92	0.9602	33.69	0.9170	32.17	0.8986	32.19	0.9278	39.48	0.9797
RCAN [33] (15.4M)	$\times 2$	✓	38.07	0.9608	33.94	0.9183	32.27	0.9005	32.48	0.9308	39.56	0.9799
	$\times 2$	-	38.05	0.9606	33.71	0.9169	32.16	0.8982	32.22	0.9284	39.42	0.9793
FreqFormer [88] (870k)	$\times 2$	✓	38.20	0.9612	33.92	0.9192	32.32	0.9013	32.51	0.9308	39.55	0.9796
	$\times 2$	-	38.11	0.9609	33.64	0.9177	32.24	0.8999	32.28	0.9286	39.54	0.9797
HAT [41] (20.6M)	$\times 2$	✓	38.41	0.9622	34.37	0.9227	32.47	0.9031	33.65	0.9410	40.32	0.9812
	$\times 2$	-	38.23	0.9612	34.03	0.9202	32.42	0.9026	33.03	0.9359	39.86	0.9803
MambaIR-light [42] (905k)	$\times 2$	✓	38.09	0.9611	33.78	0.9187	32.26	0.9006	32.20	0.9284	39.18	0.9788
	$\times 2$	-	37.98	0.9602	33.50	0.9157	32.12	0.8982	31.82	0.9239	39.15	0.9787
MambaIR [42] (20.4M)	$\times 2$	✓	38.27	0.9618	34.36	0.9234	32.37	0.9012	33.27	0.9381	40.05	0.9807
	$\times 2$	-	38.14	0.9610	33.94	0.9193	32.38	0.9019	32.73	0.9333	39.39	0.9790
MSRResNet [30] (1.5M)	$\times 4$	✓	32.23	0.8959	28.61	0.7829	27.59	0.7358	25.91	0.7787	30.56	0.9097
	$\times 4$	-	32.29	0.8969	28.58	0.7819	27.58	0.7357	25.94	0.7795	30.54	0.9094
EDSR [29] (43.0M)	$\times 4$	✓	32.66	0.9007	28.90	0.7898	27.77	0.7425	26.59	0.7999	31.42	0.9197
	$\times 4$	-	32.62	0.9003	28.87	0.7879	27.74	0.7414	26.51	0.7975	31.37	0.9182
RCAN [33] (15.6M)	$\times 4$	✓	32.72	0.9014	28.90	0.7893	27.77	0.7422	26.57	0.7988	31.34	0.9192
	$\times 4$	-	32.67	0.9009	28.87	0.7879	27.74	0.7415	26.55	0.7982	31.35	0.9184
FreqFormer [88] (890k)	$\times 4$	✓	32.60	0.8998	28.81	0.7869	27.73	0.7421	26.39	0.7954	31.20	0.9165
	$\times 4$	-	32.61	0.8997	28.84	0.7867	27.74	0.7419	26.48	0.7976	31.25	0.9172
HAT [41] (20.8M)	$\times 4$	✓	32.96	0.9050	29.17	0.7958	27.95	0.7497	27.60	0.8262	32.22	0.9280
	$\times 4$	-	32.91	0.9044	29.08	0.7934	27.90	0.7477	27.33	0.8199	32.07	0.9261
MambaIR-light [42] (924k)	$\times 4$	✓	32.43	0.8983	28.72	0.7842	27.66	0.7387	26.16	0.7855	30.87	0.9132
	$\times 4$	-	32.48	0.8983	28.71	0.7837	27.67	0.7387	26.14	0.7857	30.84	0.9130
MambaIR [42] (20.6M)	$\times 4$	✓	32.98	0.9035	29.15	0.7943	27.92	0.7476	27.22	0.8158	31.99	0.9253
	$\times 4$	-	32.79	0.9023	29.02	0.7909	27.84	0.7451	26.92	0.8081	31.80	0.9225

また、従来のフィルタリング手法は、圧縮アーチファクトを回避する目的で画像を除去していたのに対し、RECA は画像を一切削除することなく、アーチファクトを軽減する点に特徴がある。我々の知る限り、圧縮特性に基づくフィルタリングによって、データ量を維持しながら SR モデルの性能向上を実現した例は、本研究が初めてである。

5.4.4 CNE モジュールの閾値

表 5.5 に、CNE モジュールにおける異なる Blockiness 閾値 θ に対する性能比較を示す。 $\theta = 50$ の設定が最も良好な性能を示し、一方で $\theta = 0$ の場合は、不必要な高品質画像への処理により性能が低下した。

表 5.5: 異なる閾値 θ における性能比較 (RCAN model, iCartoonFace, $\times 4$ SR) .

θ	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0	32.28	0.8949	28.77	0.7852	26.43	0.7924	27.59	0.7352	31.26	0.9155
50	32.49	0.8989	28.82	0.7862	26.55	0.7971	27.70	0.7390	31.73	0.9209
100	32.48	0.8986	28.82	0.7861	26.56	0.7972	27.68	0.7385	31.77	0.9207

表 5.6: AS モジュールの効果. (RCAN model, Food-101, $\times 4$ SR).

AS module	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/	32.72	0.9014	28.90	0.7893	27.77	0.7422	26.57	0.7988	31.34	0.9192
w/o	32.33	0.8972	28.60	0.7812	27.57	0.7358	26.05	0.7830	30.61	0.9106

AS モジュールでは、画像を縮小することで圧縮アーチファクトを抑制するが、同時に細部の情報も失われる。 $\theta = 0$ の設定では、わずかなアーチファクトしか含まない画像まで処理対象となるため、高周波成分が過剰に抑制され、結果として重要なディテールが損なわれる。一方、 $\theta = 50$ では、強い圧縮を受けた画像のみを選択的に補正対象とすることで、高品質画像の細部を保持しつつ最適なトレードオフを実現している。

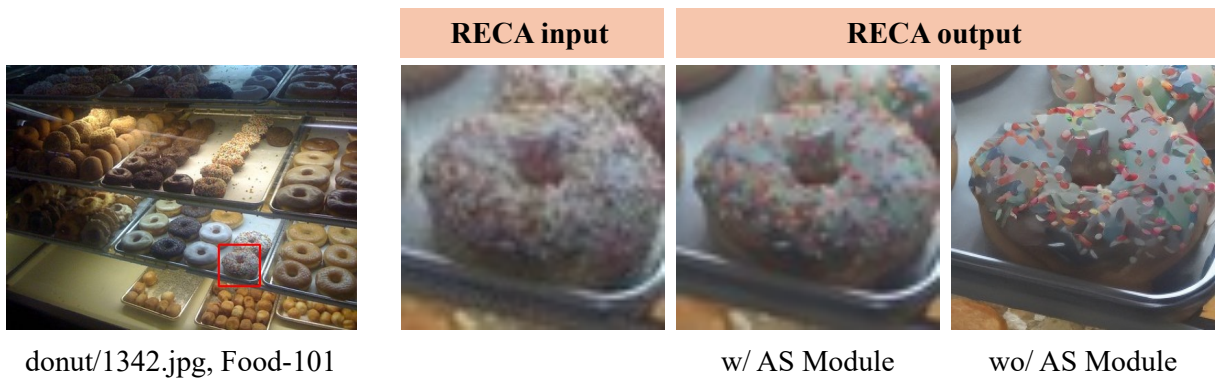


図 5.4: AS モジュールの効果を示す視覚的な例. AS モジュールを適用していない画像では、高周波アーティファクトにより過度にシャープで不自然なテクスチャが生じている。一方、AS モジュールを適用することでこれらのアーティファクトが抑制され、より滑らかで自然な外観が得られる。

表 5.7: QE モジュールにおける実世界 SR モデルの性能比較 (RCAN model, Food-101, $\times 4$ SR).

Real-world SR model	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SeeSR [55]	32.72	0.9014	28.90	0.7893	27.77	0.7422	26.57	0.7988	31.34	0.9192
Real-ESRGAN [51]	32.63	0.9005	28.87	0.7881	27.74	0.7411	26.57	0.7985	31.30	0.9188

5.4.5 AS モジュールの有効性

QE モジュールによる JPEG アーチファクト除去後に, AS モジュールが必要かどうかを検証するため, AS モジュールの有無による学習結果を比較した. 表 5.6 に示すように, AS モジュールを追加することで, 全ての評価データセットにおいて PSNR および SSIM が向上し, 再構成品質の改善に有効であることが確認された.

図 5.4 は, AS モジュールの有無による出力画像の視覚的比較を示している. AS モジュールを用いない場合, 画像は過度にシャープで人工的に見え, 不自然なテクスチャやエッジが生じている. これは, QE モジュールに使用された Diffusion ベースの SR モデルが, JPEG ノイズを除去する過程で高周波のアーチファクトを新たに生成してしまうことに起因する.

一方で, AS モジュールを適用することで, これらのアーチファクトが効果的に抑制され, より自然なエッジやテクスチャが再現されている. たとえば, ドーナツの装飾シュガー部分では, AS なしではエッジ部分が過剰に強調され, 人工的に見えるのに対し, AS ありでは現実的な外観が保たれている. これらの観察結果は, AS モジュールが視覚的自然さを維持しつつ, Diffusion モデル由来のアーチファクトを除去するうえで有効であることを示している.

以上より, AS モジュールは RECA の重要な構成要素であり, 自然なエッジ・テクスチャの再構成とモデル性能の向上を両立させるうえで不可欠な役割を果たしている.

表 5.8: QE モジュールと AS モジュールの組み合わせと Denoising モデルの性能比較 (RCAN model, iCartoonFace, $\times 4$ SR).

Module type	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
QE+AS module	32.49	0.8989	28.82	0.7862	27.70	0.7390	26.55	0.7971	31.73	0.9209
QGCN [90]	32.52	0.8988	28.79	0.7854	27.67	0.7383	26.49	0.7961	31.64	0.9192
FBCNN [91]	32.45	0.8981	28.77	0.7848	27.65	0.7380	26.47	0.7955	31.63	0.9196
DiffBIR [92]	32.40	0.8976	28.81	0.7877	27.67	0.7393	26.61	0.8000	31.74	0.9208

5.4.6 QE モジュール内における未知劣化 SR モデルの比較

QE モジュールにおいて使用する SR モデルが性能に与える影響を検証するため, Diffusion ベースの SeeSR と, GAN ベースの Real-ESRGAN [51] を比較した. 表 5.7 に示すように, SeeSR は全てのデータセットにおいて, わずかながら一貫して Real-ESRGAN を上回る性能を示した. この差は, Diffusion モデルにおけるノイズ除去および高周波成分の再構成における安定性の高さに起因すると考えられる.

5.4.7 デノイジングモデルとの比較

QE モジュールと AS モジュールを統合した構成 (QE+AS 構成) と, JPEG 圧縮アーチファクト除去を目的とした既存のデノイジングモデルとの性能を比較し, 提案手法におけるこの構成の有用性を検証した. QE モジュールはノイズ除去と同時に解像度の向上を行い, AS モジュールは QE 出力を元の解像度に縮小することで, 過剰な高周波成分やアーチファクトを抑制する. 一方, デノイジングモデルはノイズ除去のみを行い, 出力解像度が入力画像と同一である点は QE+AS 構成と共通である.

表 5.8 および 5.9 に示されている結果から, QE+AS 構成は, Food-101 および iCartoon-Face といった異なる特性をもつデータセットにおいて, 一部の指標では他手法が上回る場合もあるものの, 総じて高い性能を達成している. 特に, CNN ベースの JPEG 除去モ

表 5.9: QE モジュールと AS モジュールの組み合わせと Denoising モデルの性能比較 (RCAN model, Food-101, $\times 4$ SR).

Module type	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
QE+AS module	32.72	0.9014	28.90	0.7893	27.77	0.7422	26.57	0.7988	31.34	0.9192
QGCN [90]	32.69	0.9009	28.87	0.7882	27.76	0.7420	26.56	0.7985	31.41	0.9188
FBCNN [91]	32.67	0.9007	28.89	0.7888	27.75	0.7420	26.59	0.7991	31.37	0.9189
DiffBIR [92]	32.62	0.9004	28.87	0.7880	27.71	0.7408	26.50	0.7975	31.25	0.9181

表 5.10: DIV2K と iCartoonFace で学習された SR モデルの性能比較 (RCAN model, $\times 4$ SR).

Training dataset	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DIV2K [6]	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
iCartoonFace [86]	32.41	0.8977	28.70	0.7836	27.63	0.7364	26.36	0.7924	31.55	0.9181

デル (FBCNN, QGCN) は iCartoonFace において全体的に性能が低く, Diffusion ベースの DiffBIR は Food-101 において性能が劣っていた. これに対し, QE+AS 構成はいずれのデータセットにおいても安定した性能を示しており, データ特性に依存しない汎用性の高さが示された.

以上の結果は, QE および AS モジュールの統合が, 高品質なデータセットの構築において重要な構成要素であり, デノイジングモデル単体よりも優れた性能を発揮することを示している.

5.5 議論

5.5.1 学習データと評価データの特徴

本研究では, 学習データとして iCartoonFace を使用した. 標準的な SR 学習データセットである DIV2K を用いた学習結果と比較したところ, 評価データセットである Manga109 に対しては, iCartoonFace を用いた方がより高い性能が得られた (表 5.10). この理由とし

て、アニメーション画像から構成される iCartoonFace と、イラスト画像が中心の Manga109 との視覚的特性の類似性が挙げられる。

具体的には、両データセットはともに明瞭なエッジ、均一な色領域、高いコントラストといった特徴を持ち、自然画像とは異なる視覚的構造を有している。自然画像は滑らかな勾配や複雑なテクスチャを含む傾向があるのに対し、アニメやイラスト画像は色の遷移が急峻で、塗りも均質であることが多い。このような視覚的特性の一致が、Manga109 における SR 性能の向上に寄与したと考えられる。

これらの結果は、学習データと評価データの特徴を適切に一致させることで、タスク固有の SR 性能を向上させる可能性を示唆している。一方で、学習データと評価データの特徴が過度に近い場合、評価結果が過大に見積もられるリスクも存在する。この問題は、Set5, Set14, BSD100, Urban100, Manga109 といった広く用いられているベンチマークにおいて特に顕著であり、意図しない過学習を引き起こす可能性がある。

このリスクを回避するためには、評価データとの特性のバランスを考慮した学習データの設計が求められるとともに、学習および評価データ間の関係性についての詳細な分析が必要である。

5.5.2 画像復元タスクへの応用

本研究では、JPEG 圧縮ノイズを効果的に軽減し、SR モデルの性能を向上させるための RECA パイプラインを提案した。JPEG 圧縮ノイズは、超解像のみならず、デノイズングや雨除去など、他の画像復元タスクにも悪影響を及ぼすことが知られている。このことから、提案する RECA を他の復元タスクへ適用することで、JPEG 圧縮アーチファクトへの汎用的な対処法としての有効性を、より広い画像復元分野において検証できる可能性がある。

5.5.3 未知劣化への拡張

本研究では主に、劣化条件が既知である設定における性能比較に焦点を当てた。しかし、より実用的な設定である未知劣化下において RECA の有効性を検証することも、今後の重要な課題である。特に、複雑かつ非定常な実世界の劣化条件下での一般化性能を評価することで、RECA の実用性をさらに明確にすることが期待される。

5.5.4 学習可能な設計の欠如

RECA は既存のモジュールを統合することで実用的な効果を示しているが、現時点では学習可能な構成要素を欠いている。そのため、適応的なアーチファクト抑制や軽量の画像復元ネットワークの導入、あるいは全体を微分可能なパイプラインとして設計することが、有望な改善の方向性として挙げられる。

5.5.5 計算コストの課題

RECA では、高性能な Diffusion ベースの SR モデルである SeeSR を用いることで、JPEG 圧縮アーチファクトの除去を実現している。しかし、SeeSR は計算負荷が高く、実運用における大きな課題となる。今後の展開においては、モデルの軽量化や高速化など、計算効率の改善に向けた最適化が不可欠である。

5.5.6 感度分析の制約

CNE モジュール内での JPEG 圧縮画像判別に用いる Blockiness の閾値 $\theta = 50$ は、iCartoonFace データセットと RCAN モデルを用いた代表的な設定に基づき実験的に決定されたが、計算資源の制約のため、他のデータセットや SR モデルを含む包括的な感度分析は実

施していない．今後は，この閾値が多様な条件下において SR 性能に与える影響を詳細に分析することが課題である．

5.6 まとめ

本章では，JPEG 圧縮アーチファクトの影響を受けた画像を学習データとして再活用するための新たなパイプライン，RECA を提案した．

RECA は，圧縮アーチファクトを含む画像を高品質な学習データとして活用することを目的とした 3 段構成の処理パイプラインである．まず CNE モジュールが各画像におけるアーチファクトの強度を定量的に評価し，続いて QE モジュールが品質の低い画像に対して解像度の向上を伴う補正を行う．最後に AS モジュールが，処理過程で新たに生じた高周波アーチファクトを抑制し，出力画像を自然な質感に整える．なお，この過程において解像度は入力画像と同じスケールに調整される．

実験結果からは，JPEG 圧縮の強い画像を単純に除外する従来のフィルタリング手法と比較して，RECA がより高い性能を達成することが示された．本手法は，これまで学習に利用されてこなかった圧縮画像を有効に再活用することで，データの量と質の両面から SR モデルの性能向上に寄与する有効なアプローチである．

第6章 結論

本研究では, SR 学習における損失関数と学習データセットの構築をデータという観点から再検討し, SR モデルの性能と実用性の両立を目指す新たなアプローチを提案した. 具体的には, (1) Perceptual loss の再考, (2) SR 学習データセット構築の再考, (3) JPEG 圧縮画像の有効活用という, 三つの観点から取り組みを行った. 以下に, 各取り組みの概要と得られた成果を示す.

Perceptual loss の再考. 従来一般的に使用されてきた VGG ベースの Perceptual loss に代えて, 視覚と言語を用いた大規模事前学習モデル CLIP の ViT エンコーダを特徴抽出器とする損失関数を設計し, その SR への適用を検証した. 実験の結果, VGG ベースの損失で頻繁に見られたグリッド状のアーチファクトが, ViT ベースの CLIP 損失を用いることで大幅に抑制され, 歪み品質の低下を最小限に抑えつつ, 知覚品質の向上を実現した. これにより, Perceptual loss の特徴抽出器を再設計することで, 歪み品質と知覚品質の両立が可能であることを示した.

SR 学習データセット構築の再考. 本研究では, SR 性能に大きく影響する学習データの質と多様性に着目し, 高解像度画像を用いず, 大規模な低解像度画像群から有効な画像を自動的に選別する手法を提案した. JPEG 圧縮アーチファクトの強度分布や画像内のセグメント構造に基づく定量的評価を用いて, 高品質かつ冗長性の低いデータセット DiverSeg を構築した. 実験により, 提案手法で構築した低解像度データセットのみを用いても, 既存の高解像度画像データセットを使用したモデルを上回る SR 性能を達成できることを明ら

かにした.

JPEG 圧縮画像の有効活用. JPEG 圧縮アーチファクトにより劣化した画像を SR 学習データとして再活用するための前処理パイプライン RECA を提案した. RECA は, 圧縮アーチファクトの定量評価 (CNE モジュール), 画像品質向上 (QE モジュール), 高周波アーチファクト抑制 (AS モジュール) という三つの段階で構成される. 実験では, RECA によって補正した画像を含むデータセットを用いて学習した SR モデルが, 従来のフィルタリング手法を一貫して上回る性能を示した. この結果から, これまで利用困難だった圧縮画像も, 適切な前処理を施すことで SR 学習に有効に活用できることを示した.

今後の展望. 本研究で提案した手法群は, 既存の SR モデルに汎用的に適用可能であり, 特に高解像度画像の取得が困難なドメインにおいて有効である. 今後は, 画像劣化の対象を JPEG 圧縮に限定せず, ブラーやカメラ由来のノイズといった多様な劣化に拡張することで, 実世界における SR モデルの汎化性能をさらに向上させる必要がある. また, 本アプローチをデノイジングや雨除去など他の画像復元タスクや画像認識タスクと統合したマルチタスク学習への展開も有望な研究方向である.

謝辞

本研究は、慶應義塾大学大学院理工学研究科後期博士課程在学中に実施したものです。本研究の執筆にあたり、多くの方々にお世話になりました。この場をお借りして、心より感謝申し上げます。

まず初めに、本研究の主査であり、指導教員である青木義満教授に、心より深く感謝申し上げます。学部4年生の研究配属から博士課程に至るまで、青木教授には一貫して多くのご指導を賜りました。また、研究に取り組む姿勢や、研究者としての在り方についても多くを学ばせていただきました。併せて、青木研究室で共に研究活動を行ってきたメンバーの皆様にも感謝申し上げます。日々の議論を通じて、多くの刺激と支えをいただきました。

また、本論文の副査をお引き受けいただいた池原雅章教授、斎藤英雄教授、五十川麻理子准教授に厚く御礼申し上げます。学位審査を通じて、研究の深度をさらに高めるための示唆を多数いただきましたこと、深く感謝申し上げます。

さらに、本研究に関連する共同研究の機会を提供してくださり、多大なご助力を賜りました産業技術総合研究所の皆様に、心より感謝申し上げます。特に、博士課程の期間を通じてご指導くださった片岡裕雄博士には、数多くの建設的な議論やご助言を通して研究の方向性を定める手助けをいただきました。この場を借りて、深く御礼申し上げます。

最後に、常に温かく見守り、支えてくれた家族にも感謝いたします。

参考文献

- [1] Hujun Yang, Zhongyang Wang, Xinyao Liu, Chuangang Li, Junchang Xin, and Zhiqiong Wang. Deep learning in medical image super resolution: a review. Applied Intelligence, Vol. 53, No. 18, pp. 20891–20916, 2023.
- [2] Daniele Berardini, Lucia Migliorelli, Alessandro Galdelli, and Manuel J Marín-Jiménez. Edge artificial intelligence and super-resolution for enhanced weapon detection in video surveillance. Engineering Applications of Artificial Intelligence, Vol. 140, p. 109684, 2025.
- [3] Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. Earth-Science Reviews, Vol. 232, p. 104110, 2022.
- [4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 184–199, 2014.
- [5] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 126–135, 2017.

-
- [6] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 114–125, 2017.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems (NeurIPS), Vol. 30, , 2017.
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763, 2021.
- [9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.
- [10] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.

-
- [11] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), 2015.
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.
- [14] 大谷豪, 片岡裕雄, 青木義満. 超解像のための画像及び言語の統合特徴を利用した perceptual loss の改善. 精密工学会誌, Vol. 90, No. 2, pp. 217–223, 2024.
- [15] Go Ohtani, Ryu Tadokoro, Ryosuke Yamada, Yuki M. Asano, Iro Laina, Christian Rupprech, Nakamasa Inoue, Rio Yokota, Hirokatsu Kataoka, and Yoshimitsu Aoki. Rethinking image super-resolution from training data perspectives. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 19–36, 2024.
- [16] Go Ohtani, Hirokatsu Kataoka, and Yoshimitsu Aoki. Reca: A pipeline for refinement of compressed artifacts in image super-resolution training. IEEE Access, Vol. 13, pp. 168361–168373, 2025.

-
- [17] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 596–612, 2021.
- [18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, Vol. 13, No. 4, pp. 600–612, 2004.
- [19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595, 2018.
- [20] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6228–6237, 2018.
- [21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690, 2017.
- [22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems (NeurIPS), 2014.

-
- [23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Proceedings of the Neural Information Processing Systems (NeurIPS), 2017.
- [24] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. IEEE Signal processing letters, Vol. 20, No. 3, pp. 209–212, 2012.
- [25] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp. 5148–5157, 2021.
- [26] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 391–407, 2016.
- [27] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1874–1883, 2016.
- [28] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654, 2016.

-
- [29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 136–144, 2017.
- [30] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), pp. 63–79, 2018.
- [31] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4799–4807, 2017.
- [32] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2472–2481, 2018.
- [33] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301, 2018.
- [34] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11065–11074, 2019.

-
- [35] Y Zhang, K Li, B Zhong, and Y Fu. Residual non-local attention networks for image restoration. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.
- [36] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12299–12310, 2021.
- [37] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1833–1844, 2021.
- [38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.
- [39] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 649–667, 2022.
- [40] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. arXiv preprint arXiv:2208.11247, 2022.

-
- [41] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22367–22377, 2023.
- [42] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 222–241, 2024.
- [43] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 694–711, 2016.
- [44] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14245–14254, 2021.
- [45] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4401–4410, 2019.
- [46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning (ICML), pp. 2256–2265, 2015.

-
- [47] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. pp. 6840–6851, 2020.
- [48] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 4, pp. 4713–4726, 2022.
- [49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, Vol. 25, , 2012.
- [50] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4791–4800, 2021.
- [51] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1905–1914, 2021.
- [52] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In ACM International Conference on Multimedia (ACMMM), pp. 1329–1338, 2022.

-
- [53] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10585–10594, 2023.
- [54] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 574–591, 2022.
- [55] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25456–25467, 2024.
- [56] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8, 2008.
- [57] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 416–423, 2001.
- [58] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1775–1787, 2023.

-
- [59] Qinghong Yang, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Lu Yuan, Gang Hua, and Nenghai Yu. Hq-50k: A large-scale, high-quality dataset for image restoration. arXiv preprint arXiv:2306.05390, 2023.
- [60] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. arXiv preprint arXiv:2112.10175, 2021.
- [61] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the British Machine Vision Conference (BMVC), 2012.
- [62] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Curves and Surfaces, p. 711–730, 2010.
- [63] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206, 2015.
- [64] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, Vol. 76, pp. 21811–21838, 2017.
- [65] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472, 2021.

-
- [66] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7086–7096, 2022.
- [67] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 10, 2023.
- [68] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5885–5894, 2021.
- [69] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI conference on artificial intelligence (AAAI), pp. 2555–2563, 2023.
- [70] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. Revisiting rcan: Improved training for image super-resolution. arXiv preprint arXiv:2201.11279, 2022.
- [71] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.
- [72] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In Proceedings

-
- of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12780–12791, 2023.
- [73] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1922–1960, 2023.
- [74] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, Junpei Zhang, Kexin Zhang, Rui Peng, Yanbiao Ma, Licheng Jia, et al. Ntire 2023 challenge on image super-resolution (x4): Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1865–1884, 2023.
- [75] Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12514–12524, 2023.
- [76] Dinesh Bhardwaj and Vinod Pankajakshan. A jpeg blocking artifact detector for image forensics. Signal Processing: Image Communication, Vol. 68, pp. 155–161, 2018.
- [77] Yuki M. Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. In NeurIPS Track on Datasets and Benchmarks, 2021.

-
- [78] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026, 2023.
- [79] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 350–368, 2022.
- [80] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Proceedings of the Neural Information Processing Systems (NeurIPS), Vol. 27, 2014.
- [81] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [82] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In Proceedings of the Neural Information Processing Systems (NeurIPS), pp. 3499–3509, 2020.
- [83] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 191–207, 2020.

-
- [84] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3517–3526, 2021.
- [85] Zheng Chen, Zongwei Wu, Eduard Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, Hongyuan Yu, Cheng Wan, Yuxin Hong, et al. Ntire 2024 challenge on image super-resolution (x4): Methods and results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 6108–6132, 2024.
- [86] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In Proceedings of the ACM International Conference on Multimedia, pp. 2264–2272, 2020.
- [87] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 446–461, 2014.
- [88] Tao Dai, Jianping Wang, Hang Guo, Jinmin Li, Jinbao Wang, and Zexuan Zhu. Freqformer: Frequency-aware transformer for lightweight image super-resolution. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 731–739, 2024.
- [89] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox. <https://github.com/XPixelGroup/BasicSR>, 2022.

-
- [90] Jianwei Li, Yongtao Wang, Haihua Xie, and Kai-Kuang Ma. Learning a single model with a wide range of quality factors for jpeg image artifacts removal. IEEE Transactions on Image Processing, Vol. 29, pp. 8842–8854, 2020.
- [91] Jiayi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4997–5006, 2021.
- [92] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 430–448, 2024.