修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01 / 26 / 2026 (MM/DD/YYYY)

専攻名（専門分野） Department	基幹理工・ 情報通信専攻	氏 名 Name	速見 泰雅	指 導 教 員 Advisor	渡辺 裕 印 Seal
研究指導名 Research guidance	オーディオビジュアル 情報処理研究	学籍番号 Student ID number	CD 5124F088-1		
研究題目 Title	暗黙的ニューラル表現における効率的な動画埋込 Efficient Video Embedding in Implicit Neural Representation				

1. まえがき

信号をニューラルネットワークによって連続関数として表現する Implicit Neural Representation (INR) は，複雑な信号を容易にモデル化できる点から注目を集めている．特に，動画信号をネットワークの重みや潜在埋め込みとして保持し，順伝播によってフレームを復号する Neural Video Representation は，従来の映像符号化方式とは異なる新しい動画圧縮の枠組みとして，近年活発に研究が進められている．既存研究[1][2][3]では，多様なアーキテクチャ設計や復号戦略が提案されているものの，フレームの高周波成分の再構成は依然として困難な課題である．加えて，フレーム間で類似する情報をどのように扱うかという観点では，時間冗長性の制御と時間的一貫性の両立も重要な課題である．

本稿では，これらの課題に対して，周波数分解に基づく二流路復号と空間構造を保持するパッチベース復号という二つの方針を提案する．前者は高周波・低周波の役割分担によって学習を安定化し，後者は空間構造の保持と局所復号を両立させることで global-to-local な復号を実現する．実験では，提案手法が既存手法と比較して再構成品質の向上に寄与することを確認する．

2. 関連研究

INR は一般に，信号を入力座標 $\boldsymbol{x}$ と出力信号 $\boldsymbol{y}$ の関係として，重み θ を持つニューラルネットワーク Φ によって $\boldsymbol{y} = \Phi_\theta(\boldsymbol{x})$ として表現する枠組みである．NeRV[1]は動画のフレームインデックスからフレーム全体をマッピングするインデックス構造を導入することで，高速な復号を実現する．しかし，急激なアップサンプリングやニューラルネットワークのスペクトルバイアスという，信号の高周波成分よりも低周波成分を優先的に学習する性質により，しばしばフレームの高周波成分の表現に苦労する．

既存の研究では，この課題を軽減するためのアーキテクチャ設計や復号戦略を提案している．HNeRV [2] ではフレームインデックスの代わりにフレームから抽出される潜在埋込を用いるハイブリッド構造を導入し，表現能力の向上を示している．Boosting NeRV[3]は既存のデコーダフレームワークにコンディションを導入することで，従来手法の性能を向上させる．また，フレーム全体の代わりに空間的に分割したパッチ単位で復号する手法[4]も提案されている．しかし，パッチ境界で不連続性が生じやすい課題が残っている．

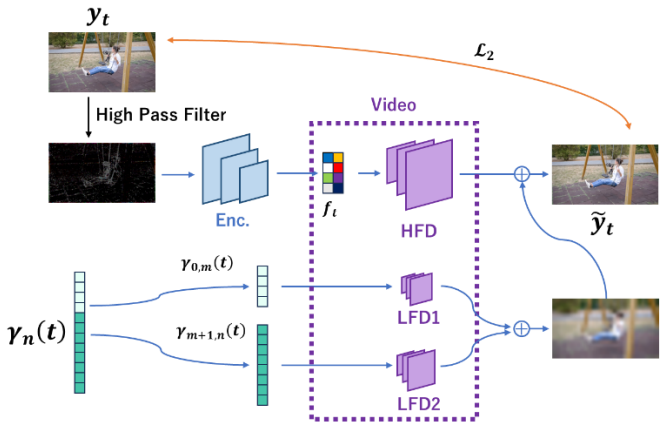

図 1 二流路復号アーキテクチャの概要

3. 周波数分解に基づく二流路復号

時間冗長性の抑制と時間的一貫性の維持を目的として，高周波成分を担う HF-stream と低周波成分を担う LF-stream からなる二流路構成の INR ベース動画表現手法を提案する．図 1 に示すように，全体はハイブリッド構造を基盤とし，HF-stream はフレーム固有入力（潜在特徴）に基づいて高周波成分を復号する．LF-stream はフレーム時刻に基づいて時間的に滑らかな低周波成分を生成する．最終再構成は両ストリームの出力を加算統合して得る．この役割分担により，低周波側は時間的に安定した構造・色調を担い，高周波側は局所的な細部を補完する形となる．

HF-stream では，隣接フレーム間で類似しやすいフレーム全体特徴の冗長化を抑えるため，入力フレームにハイパスフィルタを適用して高周波成分 I_t^{HF} を抽出し，これをエンコーダに入力して低次元特徴へ変換する．高周波成分は時間方向の変動が相対的に大きく，そこから得られる特徴は冗長性が低下し，細部表現の強化に寄与する．エンコーダに ConvNeXt ブロックを用い，デコーダ (HFD) は HNeRV ブロック[2]を積層して構成する．

LF-stream はインデックス構造とし，正規化した時刻 t に対する Positional Encoding を入力として低周波側を復号する．MLP のパラメータ削減のため，Positional Encoding を周波数帯域で二分し，二つの部分ネットワ

表 1 DAVIS データセットにおける再構成品質

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV(All)	28.91	0.8865	0.3663
HNeRV(All)	30.94	0.9160	0.2955
Ours(All)	**31.81**	**0.9284**	**0.2780**
HNeRV(Suc)	31.19	0.9261	0.2873
Ours(Suc)	**31.91**	**0.9361**	**0.2674**

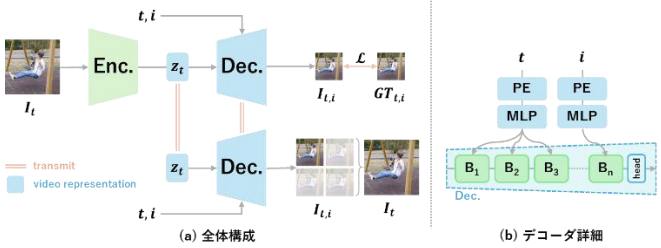

図 2 SPP に基づく復号の概要.

ーク(LFD1, LFD2)へ入力する設計を導入する. これにより, 表現能力の低下を抑えつつ, 入出力次元に依存して増大しやすいパラメータを抑制する. 損失は再構成フレームと元フレームの L2 損失で最適化する.

実験ではDAVIS データセット[5](640 × 1280 クロップ)を用いて既存手法[1][2]と比較し, 総パラメータ数約 1.5M の条件で評価した. 表 1 に定量評価の結果を示す通り, 提案手法は複数の評価指標において一貫して良好な傾向を示した.

4. 空間構造を保持するパッチベースの復号

空間構造の保持と局所復号の両立を狙い, Structure-Preserving Patch(SPP)に基づくパッチ単位復号を提案する. 本手法は, 図 2 に示すように, フレームを単純なタイル分割で扱うのではなく, PixelUnshuffle に類似した決定論的画素再配置により, 元フレームの空間レイアウトを保持した複数のパッチ画像集合へ変換する. これにより, パッチ間で一貫した空間配置が維持され, パッチ間冗長性を活用しつつ大域構造を捉えやすくなる.

デコーダは global-to-local 戦略に基づき, 前段層を時間インデックス t のみにより条件付けしてフレーム共通の大域構造をモデル化し, 後段層でパッチインデックス i を導入して局所的細部を精緻化する. この設計により, まず全体として整合的な構造を形成し, 次にパッチ固有の情報で局所ディテールを補う.

学習では, パッチごとの複合損失を用いる. 画素 L1 損失$\mathcal{L}_1$と MS-SSIM 損失$\mathcal{L}_{\text{MS-SSIM}}$に加え, FFT 差分に基づく周波数領域損失$\mathcal{L}_{\text{freq}}$を導入し, 高周波成分の整合を促す.

$$\mathcal{L}_i = w_i\left(\alpha\mathcal{L}_1(x_i, \hat{x}_i) + \beta\mathcal{L}_{\text{MS-SSIM}}(x_i, \hat{x}_i)\right) + \mathcal{L}_{\text{freq}}(x_i, \hat{x}_i) \quad (1)$$

$$\mathcal{L}_{\text{freq}}(x_i, \hat{x}_i) = \mathcal{L}_1(\text{FFT}(x_i), \text{FFT}(\hat{x}_i)) \quad (2)$$

さらに, パッチ間差分に基づく適応重み w_i を導入し, 隣接パッチと大きく異なる難易度の高い領域へ学習を重点配分することで, 構造的一貫性と局所忠実度の両立を図る.

$$w_i = \frac{\sum_{j\neq i}||x_i - x_j||_1}{\sum_k\sum_{j\neq k}||x_k - x_j||_1 + \epsilon} \quad (3)$$

ここで, ϵ はゼロ除算を回避する微小定数である.

表 2 各データセットにおける再構成品質 (PSNR/MS-SSIM)

Dataset	DAVIS	MCL-JCV	UVG
NeRV	28.60/0.8811	31.64/0.9217	31.22/0.8937
HNeRV	30.69/0.9146	33.47/0.9417	29.44/0.8470
Boost	33.53/0.9604	35.60/0.9638	33.45/0.9311
Ours	**34.23/0.9643**	**35.94/0.9654**	**33.70/0.9312**

実験では DAVIS[5], MCL-JCV[6], UVG[7]データセットを用いて既存手法[1][2][3]と比較した. 表 2 は各データセットにおける PNSR と MS-SSIM の評価平均を表す. いずれの評価指標においても提案手法が優れた再構成品質を達成した.

5. むすび

本研究では, INR に基づく動画表現における課題(時間冗長性, 高周波復元, 空間・時間的一貫性)に対し, 周波数分解二流路(HF-stream/LF-stream)による冗長性低減と時間的一貫性の強化と, SPP に基づく global-to-local パッチ復号による構造保持と局所精緻化, という二つの復号戦略を提案した. 実験においては, 標準ベンチマークを用いた評価により, 再構成品質の改善が確認された.

参考文献

[1] H. Chen, B. He, H. Wang, Y. Ren, S. Lim, and A. Shrivastava, "NeRV: Neural Representations for Videos," Advances in Neural Information Processing Systems (NeurIPS), pp. 21557-21568, Dec. 2021.

[2] H. Chen, M. Gwilliam, S. Lim, and A. Shrivastava, "HNeRV: A Hybrid Neural Representation for Videos," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10270-10279, Jun. 2023.

[3] X. Zhang, R. Yang, D. He, X. Ge, T. Xu, Y. Wang, H. Qin, and J. Zhang, "Boosting Neural Representations for Videos with a Conditional Decoder," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2556-2566, Jun. 2024.

[4] Y. Bai, C. Dong, and C. Wang, "PS-NeRV: Patch-wise Stylized Neural Representations for Videos," IEEE International Conference on Image Processing (ICIP), pp. 41-45, Sep. 2023.

[5] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 724-732, Jun. 2016.

[6] H. Wang *et al.*, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," IEEE International Conference on Image Processing (ICIP), 2016, pp. 1509-1513, Aug. 2016.

[7] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development," ACM Multimedia Systems Conference, pp. 297-302, May 2020.

2025 年度

早稲田大学大学院　基幹理工学研究科 情報理工・情報通信専攻　修士論文

暗黙的ニューラル表現における効率的な動画埋込

Efficient Video Embedding in Implicit Neural Representation

速見 泰雅
(5124F088-1)

提出日：2026 年 1 月 26 日

指導教員：渡辺 裕 教授

研究指導名：オーディオビジュアル情報処理研究

論文要旨

動画は配信・通信・記録の主要なメディアとして広く利用されており，限られた帯域・記憶容量の下で高品質な動画を扱うためには，高効率な圧縮が不可欠である．近年，動画信号をニューラルネットワークのパラメータへ埋め込み，暗黙的関数として表現する暗黙的ニューラル表現（Implicit Neural Representation：INR）が，動画表現・圧縮の新たな枠組みとして注目されている．INR に基づく動画表現は，復号を比較的単純なニューラルネットワークの順伝播に帰着できる一方で，(i) フレーム由来の入力特徴が時間方向に冗長化しやすいこと，(ii) スペクトルバイアスやアップサンプリングに起因して高周波成分の再構成が難しいこと，(iii) 局所的な復号に伴う空間的不連続や時間的一貫性の低下，といった課題が残る．

本論文では，これらの課題に対して二つのアプローチを提案する．第一に，高周波成分と低周波成分を分離した二流路（HF-stream ／ LF-stream）構成を導入する．HF-stream ではフレームの高周波成分から特徴を抽出することで入力冗長性を抑制し，細部表現の強化を図る．一方，LF-stream では時間インデックスに基づく復号により時間的に滑らかな成分を安定に生成し，時間的一貫性の維持を図る．両流路の出力を統合することで，高周波の忠実度と時間的一貫性を両立する設計指針を与える．第二に，構造保持パッチ（Structure-Preserving Patch：SPP）に基づくパッチ単位復号を提案する．本手法は PixelUnshuffle に類似した画素再配置により，各フレームを空間構造を保持したパッチ画像群へ分解し，パッチ境界の不連続を抑えつつ局所詳細を復号する．さらに，デコーダを global-to-local な方針で条件付けすることで，フレームの大域構造を捉えた後に局所的細部を精緻化する復号過程を実現する．

標準的なベンチマーク動画に対する評価の結果，提案する二流路設計および SPP に基づく復号戦略は，既存の INR に基づく動画表現手法と比較して，再構成品質および圧縮性能の観点から有効であることを確認した．

キーワード

暗黙的ニューラル表現, 動画表現, 動画圧縮

Abstract

Implicit Neural Representations (INRs) embed video signals into neural network parameters and represent them as implicit functions, providing a promising framework for video representation and compression. Although INR-based methods can reduce decoding to a simple forward pass of a compact neural network, they still face several challenges: (i) temporal redundancy in frame-derived input features, (ii) difficulty in reconstructing high-frequency details due to spectral bias and upsampling, and (iii) degraded spatial continuity and temporal consistency caused by local decoding strategies.

This thesis addresses these issues from two complementary perspectives. First, we introduce a two-stream representation that separates frequency bands. The high-frequency stream (HF-stream) extracts content-adaptive features from high-frequency components to suppress temporal redundancy and better preserve fine details. In contrast, the low-frequency stream (LF-stream) reconstructs temporally smooth components from time indices to improve temporal consistency. By integrating the outputs of the two streams, the proposed design aims to achieve both high-frequency fidelity and temporal consistency. Second, we propose a patch-wise decoding strategy based on Structure-Preserving Patch (SPP). By applying a deterministic pixel rearrangement analogous to PixelUnshuffle, each frame is transformed into a set of structure-preserving patch images, which helps suppress boundary discontinuities while enabling local refinement. We further design a global-to-local conditioning scheme in the decoder so that the global layout is captured first and local details are refined afterward.

Experiments on standard benchmark videos show that the proposed designs improve reconstruction quality and compression performance over existing INR-based baselines.

Keywords

Implicit neural representation, video representation, video coding.

目次

図目次

表目次

第 1 章

序論

1.1 研究背景

近年，動画配信サービスの普及，オンライン会議の常態化，モバイル端末による撮影・共有の一般化に伴い，映像コンテンツの生成・流通量は継続的に増加している．動画は時間方向の情報を含むため，静止画像や音声と比較してデータ量が大きい．そのため，限られた通信帯域や記憶容量の下で高品質な動画を安定して扱うには，効率的な動画圧縮技術が不可欠である．

従来の動画圧縮は，Moving Picture Experts Group（MPEG）を中心とする標準化活動により発展してきた．AVC/H.264 [1]，HEVC/H.265 [2]，VVC/H.266 [3] などの標準方式は，ブロック分割，動き補償予測，量子化，エントロピー符号化といった処理を人手設計の規則として精緻化することで高い圧縮性能を達成している．一方で，さらなる高圧縮化を目的としたモジュール追加や探索範囲の拡大は，計算量増大および実装の複雑化を招く．加えて，エンコーダのみならずデコーダにおいても，リアルタイム再生や省電力動作を満たす計算資源制約が一層顕在化しつつある．

近年は深層学習の発展に伴い，ニューラルネットワークを用いた学習型動画圧縮が活発に研究されている．学習型手法 [4, 5] はデータ駆動で符号化過程を最適化できるため，特定条件下で従来方式を上回る圧縮効率を示す例も報告されている．しかし，ネットワークの大規模化・複雑化により推論計算コストが増大し，復号速度や消費電力の観点で課題を抱える場合がある．

これらとは異なる枠組みとして，動画信号そのものをニューラルネットワークへ埋め込み，ネットワークパラメータを圧縮表現として扱う暗黙的ニューラル表現（Implicit Neural Representation：INR）に基づく動画表現が注目されている．INR は，時空間座標を入力し，対応する信号値を出力する連続関数をニューラルネットワークで近似する枠組みであり，複雑な信号を解像度非依存に表現できる点に特徴がある．特に，NeRV [6] に代表されるフレーム生成型の動画 INR では，フレーム番号やフレーム特徴などの低次元入力からフレーム全体を生成することで，復号をニューラルネットワークの順伝播として実行できる．この性質は，復号計算の規則性や並列性の観点で利点を持つ一方で，入力にフレーム由来の特徴を用いる場合，時間方向に冗長な情報が混入しやすく，圧縮という目的と整合しない可能性がある．また，ニューラルネットワークが低周波成分を優先的に学習する傾向（スペクトルバイアス）などに起因して，高周波成分（エッジ，テクスチャ）の再構成が難しいことも課題として挙げられる．さらに，パッチ単位などの局所的な復号を採用する場合には，パッチ

境界に起因する空間的不連続が生じ得る．加えて，動画としての知覚品質を左右する時間的一貫性の確保も重要であり，INR に基づく動画埋め込みと復号戦略の設計には，表現効率と再構成品質を両立する観点が求められる．

1.2 研究目的

本研究の目的は，INR に基づく動画表現において，(i) 入力表現の冗長性を抑制しつつ，(ii) 高周波成分の再構成と時間的一貫性を改善し，(iii) 復号計算量およびモデル容量の増大を抑えた効率的な復号戦略を確立することである．具体的には，フレーム特徴の時間冗長性に着目した表現設計と，空間構造を保持したパッチ単位復号の設計を通じて，INR 動画表現の圧縮性能と再構成品質の向上を目指す．

1.3 本論文の貢献

本論文の主な貢献は以下のとおりである．

- 高周波成分と低周波成分を分離する二流路復号（HF-stream / LF-stream）に基づき，フレーム特徴に含まれる時間冗長性の抑制と，時間的一貫性を考慮した再構成を両立する動画表現手法を提案する．
- PixelUnshuffle に類似した画素再配置によりフレームを空間構造を保持したパッチ画像群へ分解し，Structure-Preserving Patch（SPP）に基づく復号戦略を提案する．これにより，局所詳細の再構成とパッチ境界に起因する不連続の抑制を両立する．
- 標準的なベンチマーク動画を用い，PSNR，MS-SSIM，LPIPS に基づく評価を通じて，提案法の有効性を検証する．

1.4 本論文の構成

本論文の構成を以下に示す．

第 1 章では，本研究の背景，目的，および本論文の貢献を述べる．

第 2 章では，INR および動画 INR に関する関連研究を整理し，本研究の位置付けを明確化する．

第 3 章では，冗長性削減と一貫性保持のためのニューラル動画表現について述べる．

第 4 章では，効率的なニューラル動画表現のための構造保持パッチ復号（Structure-Preserving Patch：SPP）について述べる．

第 5 章では，本研究を総括し，限界と今後の展望を述べる．

第 2 章

関連研究

2.1 まえがき

本章では，本論文に関連する既存研究を整理し，提案手法の位置付けを明確にする．まず，暗黙的ニューラル表現（Implicit Neural Representation: INR）の基本概念と学習特性を概観する．次に，動画に対する INR の代表的な定式化を整理し，入力表現・埋め込み・復号戦略という設計論点を中心に Neural Video Representation について述べる．最後に，既存の動画圧縮技術と INR による動画圧縮の枠組みを整理する．

2.2 暗黙的ニューラル表現

暗黙的ニューラル表現（Implicit Neural Representation：INR）は，画像・音声・三次元形状などの信号を，座標入力に対する連続関数としてニューラルネットワークで表現する枠組みである．INR では，信号を離散サンプルの集合として保持するのではなく，座標 $\boldsymbol{x}$ を入力すると対応する信号値 $\boldsymbol{y}$ を返す関数としてモデル化する点に特徴がある．形式的には，入力座標 $\boldsymbol{x}$ と出力信号値 $\boldsymbol{y}$ の対応を，学習パラメータ θ をもつニューラルネットワーク Φ_θ により次式で表す．

$$\boldsymbol{y} = \Phi_\theta(\boldsymbol{x}) \tag{2.1}$$

ここで，$\boldsymbol{x}$ は信号の定義域上の座標（画像であれば 2 次元空間座標，動画であれば時刻を含む時空間座標など）を表し，$\boldsymbol{y}$ はその座標に対応する信号値（RGB 値，輝度値，密度など）である．

INR の学習は，観測サンプル $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ に対して，ネットワーク出力と観測値の誤差を最小化する最適化問題として定式化される．一般に，損失関数 $\mathcal{L}(\cdot)$ を用いて

$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(\Phi_\theta(\boldsymbol{x}_i),\ \boldsymbol{y}_i) \tag{2.2}$$

のように表される．代表例として Neural Radiance Fields（NeRF）[7] は，三次元空間座標と視線方向を入力とし，放射輝度や密度を出力する連続関数を MLP で近似することで，ボリュームレンダリングを介した高精度な新規視点合成を実現する．

一方で，標準的なニューラルネットワークは低周波成分を優先して学習する傾向（スペクトルバイアス）[8, 9, 10, 11] をもつため，信号の高周波成分の復元が難しくなる場合がある．この問題に対しては，入力座標

の表現（Positional Encoding や周波数特徴）やネットワーク構造（活性化関数，階層的復号）の工夫により，高周波成分の表現能力を高める研究が進められている．例えば SIREN [12] は，正弦活性化関数を用いることで，高周波を含む信号を効率的に表現できることを示している．これらの成果は，INR における入力表現とネットワーク設計が，複雑な信号の復元性能を左右することを示唆している．

2.3 ニューラル動画表現

INR を動画へ適用する際には，ネットワーク入力 $\boldsymbol{x}$ と出力 $\boldsymbol{y}$ の設計をどのように扱うかが重要となる．例えば，時空間座標を入力として対応する画素値を出力する座標ベース手法は，解像度やフレームレートに依存しない連続的な再構成が可能である一方，高解像度かつ長尺の動画に対してはサンプル数が膨大となり，学習・推論コストが増大しやすい．この課題に対し，フレーム全体を直接生成するフレームベースの枠組みとして Neural Representations for Videos（NeRV）[6] が提案されている．

2.3.1 NeRV

NeRV は，フレームインデックスとフレーム全体を直接対応付ける frame-wise な暗黙表現として動画を表現する．すなわち，動画 $\{\boldsymbol{I}_t\}_{t=1}^{T}$ を，フレーム時刻 t を入力として対応フレーム $\boldsymbol{I}_t$ を出力する関数 f_θ として定式化し，動画表現をネットワーク重み θ に埋め込む．

具体的には，フレーム時刻 $t \in \{1 \dots T\}$ を $(0, 1]$ に正規化したスカラーとして入力し，Positional Encoding $\mathrm{PE}(\cdot)$ により高次元特徴 $\boldsymbol{z}_t$ へ写像した後，デコーダ D によりフレーム $\hat{\boldsymbol{I}}_t$ を生成する．

$$\hat{\boldsymbol{I}}_t = D(\boldsymbol{z}_t \,;\, \theta) \tag{2.3}$$

$$\boldsymbol{z}_t = \mathrm{PE}(t) \tag{2.4}$$

$$\mathrm{PE}(t) = \big(\sin(b^0\pi t), \cos(b^0\pi t), \dots, \sin(b^n\pi t), \cos(b^n\pi t)\big) \tag{2.5}$$

ここで b および n は埋め込みの周波数帯域を規定するハイパーパラメータである．推論時は t を走査して順伝播を実行するだけで全フレームを生成できるため，復号は CNN ベースの規則的演算として実装でき，高い並列性と高速な復号が期待できる．

NeRV のデコーダは，(i) 低解像度特徴マップを生成する **Stem**，(ii) 解像度を段階的に拡大する **NeRV ブロック列**，(iii) RGB 空間へ写像する **Head** から構成される．概念的には次式で表される．

$$\boldsymbol{F}_0 = \mathrm{reshape}(\mathrm{FC}(\boldsymbol{z}_t)) \tag{2.6}$$

$$\boldsymbol{F}_{k+1} = \mathcal{U}_k(\boldsymbol{F}_k) \quad (k = 0, \dots, K-1) \tag{2.7}$$

$$\hat{\boldsymbol{I}}_t = \mathrm{Head}(\boldsymbol{F}_K) \tag{2.8}$$

ここで $\boldsymbol{F}_0 \in \mathbb{R}^{C_0 \times H_0 \times W_0}$ は低解像度特徴マップであり，K 段のアップサンプリングにより最終解像度 (H, W) に到達する．NeRV ブロック $\mathcal{U}_k$ は，畳み込みと PixelShuffle によるアップサンプリング，および活性化関数から構成され，次のように表せる．

$$\boldsymbol{F}_{k+1} = \mathrm{act}(\mathrm{PS}(\mathrm{Conv}(\boldsymbol{F}_k), s_k)) \tag{2.9}$$

ここで PS(·) は PixelShuffle，s_k はアップサンプリング倍率である．一方で，入力がフレーム時刻のみで情報量が小さいこと，および段階的なアップサンプリングに伴う表現制約により，細線や微細テクスチャ等の高周波成分の再構成が難しくなる場合がある．この点を改善するため，後続研究では入力設計の拡張や復号戦略の改良が提案されている．

2.3.2 HNeRV

HNeRV [13] は，暗黙表現（動画ごとに学習されるデコーダ）と明示表現（内容依存の埋め込み）を組み合わせたハイブリッド表現として NeRV を拡張する．具体的には，ConvNeXt ブロック等からなる動画固有エンコーダにより各フレームをコンパクトな埋め込みへ写像し，その埋め込みを入力として動画固有デコーダがフレームを復号する．フレーム固有埋め込みは小さく設計され，動画の大部分はデコーダ重みに暗黙的に保持されるため，暗黙表現の簡潔さを維持しつつ，内容適応な入力により再構成性能と収束性の改善が期待できる．一方で，フレームごとに埋め込みを保持する構造は，フレーム数に比例して埋め込みサイズが増加し得る．したがって，HNeRV の総表現サイズは「埋め込みサイズ＋デコーダサイズ」として捉えられ，復号性能と表現サイズとのトレードオフ設計が重要となる．

2.3.3 Boosting NeRV

Boosting NeRV [14] は，NeRV 系列の復号過程において，中間特徴と目標フレームの同一性情報（特に時間情報）の整合を明示的に強化するため，条件付きデコーダ（conditional decoder）を導入する枠組みである．従来の多くの手法では，フレーム t の同一性情報は入力側の時間埋め込み（あるいはフレーム埋め込み）に主として局在し，中間特徴が目標フレームと十分に整合しないことが再構成品質のボトルネックとなり得る．これに対し Boosting NeRV は，時間情報を用いて中間特徴を段階的に変調することで，復号全体を通じた特徴整合を促す．まず，動画フレーム $\boldsymbol{I}_t$ から内容依存の埋め込み $\boldsymbol{y}_t$ を得る埋め込み生成器 $E(\cdot)$（ハイブリッド表現ではフレームエンコーダ）と，正規化したフレーム時刻 t から時間埋め込み $\boldsymbol{z}_t$ を得る小規模 MLP $M(\cdot)$（時間埋め込み生成器）を用意する．このとき，Boosting NeRV における復号（表現）プロセスは次式で表される．

$$\boldsymbol{y}_t = E(\boldsymbol{I}_t\ ;\ \phi) \tag{2.10}$$

$$\boldsymbol{z}_t = M(\mathrm{PE}(t)\ ;\ \psi) \tag{2.11}$$

$$\hat{\boldsymbol{I}}_t = F(\boldsymbol{y}_t, \boldsymbol{z}_t\ ;\ \theta) \tag{2.12}$$

ここで PE(·) は Positional Encoding，$F(\cdot)$ は条件付きデコーダ（フレーム再構成ネットワーク）である．

条件付き復号器 F の中核は，時間埋め込み $\boldsymbol{z}_t$ に基づいて中間特徴をアフィン変換する Temporal-aware Affine Transform（TAT）である．既存の条件付け手法で広く用いられる AdaIN [15] は，正規化と条件付きアフィン変換を結合しており，データへの過適合を前提とする INR の性質と必ずしも整合しない．この点を踏まえ，Boosting NeRV では正規化を伴わない TAT を導入し，中間特徴 $\boldsymbol{f}_t$ を次式により変調する．

$$\mathrm{TAT}(\boldsymbol{f}_t \mid \boldsymbol{\gamma}_t, \boldsymbol{\beta}_t) = \boldsymbol{\gamma}_t \odot \boldsymbol{f}_t + \boldsymbol{\beta}_t \tag{2.13}$$

ここで $\boldsymbol{\gamma}_t$ および $\boldsymbol{\beta}_t$ は $\boldsymbol{z}_t$ から生成されるチャネルごとのスケールおよびシフトであり，$\odot$ は要素積（空間方向にブロードキャスト）を表す．実装上は，NeRV-like なアップサンプリングブロックの後段に TAT 残差

ブロックを挿入し，復号の各段で同一性情報を注入する構成を採る．さらに Boosting NeRV は，NeRV-like ブロックにおける活性化関数が中間特徴の多様性に影響する点に着目し，GELU の代わりに正弦活性を用いた Sinusoidal NeRV-like（SNeRV）ブロックを導入する．

なお，NeRV，HNeRV，および Boosting NeRV（HNeRV-Boost）の全体構成の比較を図 2.1 に，各手法で用いられるブロック（NeRV block，SNeRV block，TAT block）の詳細を図 2.2 に示す．

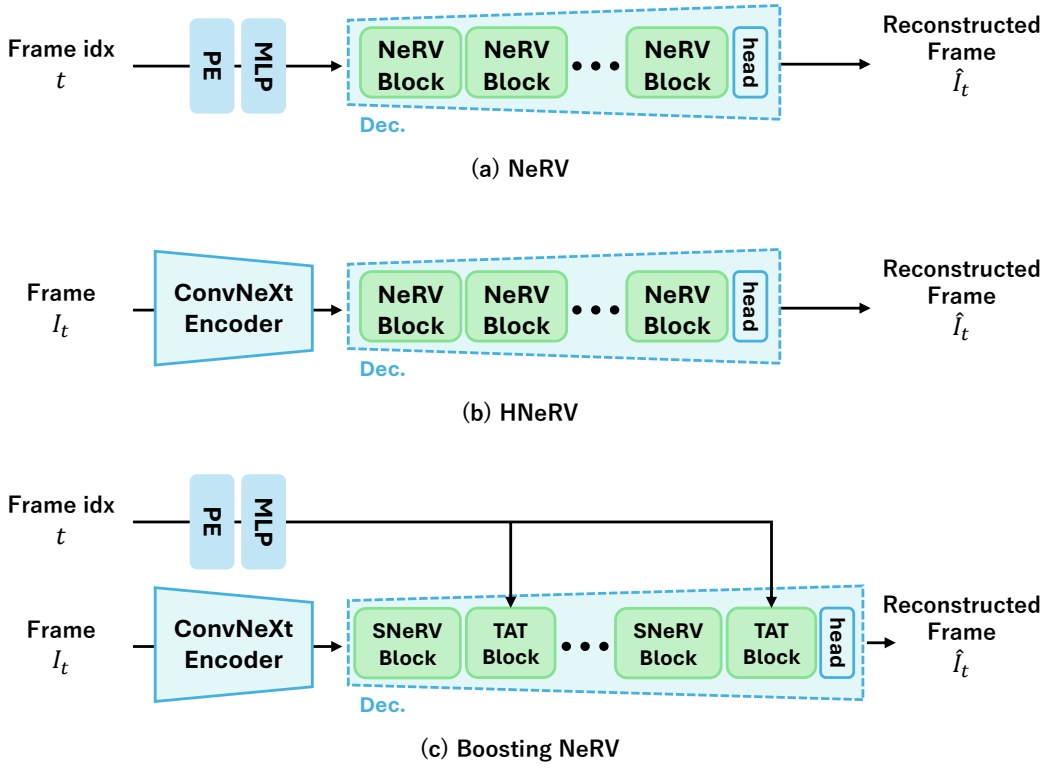

図 2.1: Neural Video Representation における代表手法のアーキテクチャ比較．（a）NeRV は時刻 t の Positional Encoding を入力としてフレームを生成する．（b）HNeRV はフレームから抽出した潜在埋め込みを入力とするハイブリッド構造を導入する．（c）Boosting NeRV は時間埋め込みに基づく条件付け（TAT）を復号途中へ挿入し，中間特徴の整合を強化する．

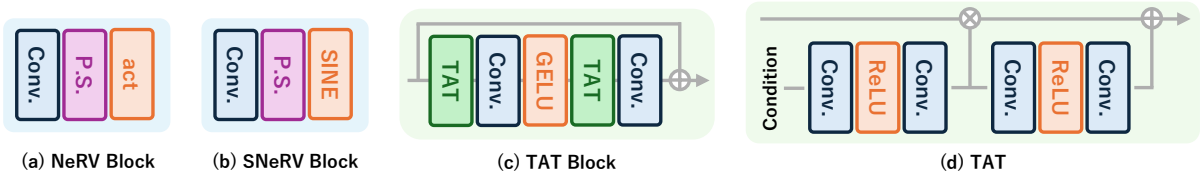

図 2.2: NeRV 系列で用いられる各ブロック（NeRV block，SNeRV block，TAT block）の構成．

2.4　動画圧縮

動画圧縮は，一般にレート（bitrate）を抑えつつ歪み（画質劣化）を最小化するレート歪み最適化（Rate–Distortion Optimization：RDO）の枠組みで捉えられる．従来の符号化方式 [1, 2, 3] では，動き補償予測，変換・量子化，ループ内フィルタ，エントロピー符号化等のツールを組み合わせ，RDO に基づいて量子化パラメータやモード等の符号化パラメータを選択することで高い圧縮性能を達成してきた．学習型圧縮では，オートエンコーダに基づく潜在表現を量子化し，確率モデルに基づくエントロピー符号化によってレートを見積もりつつ，歪みとの和を最小化する形で学習することが多い．

INR による動画圧縮では，図 2.3 に示すように，動画をデコーダ重み θ と必要に応じた入力埋め込み $\{\boldsymbol{z}_t\}$ として表現し，これらを量子化・符号化して伝送する点に特徴がある．したがって，圧縮効率は，(i) 重みおよび埋め込みの量子化方式，(ii) 符号化に用いる確率モデル，(iii) どの成分へ容量を割り当てるか（復号戦略・アーキテクチャ設計），に強く依存する．本論文では，復号戦略の工夫により再構成品質を改善しつつ，既存手法と比較してレート歪み特性を評価する．

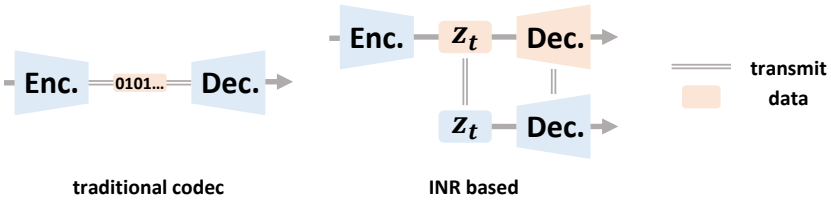

図 2.3: 従来の動画符号化と INR ベース動画圧縮の処理フロー比較.

2.5　むすび

本章では，本論文の提案手法を適切に位置付けるために，まず暗黙的ニューラル表現（INR）の定式化と学習特性を整理した．具体的には，座標入力に対する連続関数として信号を表現する枠組みと，学習において生じ得るスペクトルバイアス等の性質を概観し，高忠実度な復元に向けた入力表現・ネットワーク設計の重要性を確認した．次に，INR を動画へ適用した Neural Video Representation として，NeRV を中心に入力表現とデコーダ設計を整理し，HNeRV や Boosting NeRV がそれぞれ埋め込みの内容適応化や条件付けによって再構成品質と整合性を高めることを述べた．さらに，従来の符号化方式が予測・変換・量子化を核にレート歪み最適化を行うのに対し，INR ベース圧縮はデコーダ重みと埋め込みを量子化・符号化して伝送する点に本質的な違いがあることを整理した．

第 3 章

冗長性削減と一貫性保持のためのニューラル動画表現

3.1 はじめに

近年，インターネットの普及と動画配信サービスの拡大に伴い，高品質な動画を効率的に伝送・蓄積するための動画圧縮技術がますます重要になっている．Moving Picture Experts Group（MPEG）を中心として多くの主要な圧縮技術が標準化されており，HEVC/H.265 [2] や VVC/H.266 [3] に代表される従来の動画符号化規格は，高い圧縮効率と実用的な復号計算量を両立してきた．一方で，さらなる高圧縮化を目的としたツールの追加や探索範囲の拡大により複雑化が進むと，計算量増大および実装負荷の増加を招きやすい．この背景の下，深層学習の発展に伴い，ニューラルネットワークを用いた学習型動画圧縮が活発に研究されている．中でも，動画信号をニューラルネットワークのパラメータとして埋め込み，暗黙的関数として表現する暗黙的ニューラル表現（Implicit Neural Representation：INR）に基づく動画表現は，復号を比較的単純なネットワークの順伝播として実装できる点から注目されている．

INR は，対象信号に対してニューラルネットワークを過学習させ，ネットワーク自体を信号の表現として扱う枠組みである．したがって，ネットワークパラメータ（および付随する潜在埋め込み）を符号化・圧縮することは，信号の圧縮に対応する．また，複雑な信号を連続関数としてコンパクトに表現できる点から，画像 [12, 16, 17, 18, 19] や動画 [6, 13, 20, 21, 22] の表現・圧縮，3D シーン [7, 23, 24, 25, 26]，3D 形状再構成 [27, 28, 29] など幅広い分野に応用されている．

INR による動画圧縮は，従来の複雑なパイプラインと比較して，より単純なネットワーク構造により復号における計算コストを低減できる．代表的な枠組みとして，Neural Representations for Videos（NeRV） [6] はフレーム時刻（インデックス）を入力し，対応するフレームを出力するインデックスベース構造を導入した．この構造はフレームサイズに依存しない入力を用いるため，INR で一般的に用いられる座標ベース手法と比較して学習・推論が効率的である．一方で，動画内容に適応したフレーム固有情報を直接入力に与えないため，複雑なテクスチャや局所的なディテールの再構成が難しい場合がある．

これに対し，HNeRV [13] は各フレームから抽出した特徴（潜在埋め込み）を入力として用いるハイブリッドベース構造を提案し，動画固有のパターンを捉えることで再構成品質の向上を図る．しかし，隣接フレーム

が類似している動画では，抽出される特徴も類似しやすく，入力特徴に時間方向の冗長性が生じる．動画圧縮の観点では冗長情報の削減が不可欠であるため，入力特徴の設計を通じて冗長性を抑制する必要がある．さらに，深層学習に内在するスペクトルバイアス（ネットワークが高周波成分よりも低周波成分を優先して学習する傾向） [8, 9, 10, 11] により，フレーム間で変化しやすい高周波ディテールの再構成は難しくなり得る．また，ハイブリッドベース構造は時間情報を明示的に入力しないため，フレーム間関係のモデル化や時間的一貫性の確保という観点でも改善余地が残る．

本章では，フレーム固有情報と時間情報の双方を活用し，各フレームの高周波成分と低周波成分を分離して復号する二流路構成を提案する．提案手法は，高周波成分の再構成を担う High-Frequency stream（HF-stream）と，低周波成分の再構成を担う Low-Frequency stream（LF-stream）から構成される．HF-stream ではフレームの高周波成分のみを入力とし，時間冗長性を抑制した特徴抽出により細部表現の強化を図る．一方，時間的に相関の強い低周波成分は時間情報に基づいて生成し，時間的一貫性の維持を図る．両者を統合することで，本手法は再構成品質と圧縮効率の向上を目指す．

3.2 関連研究

3.2.1 動画のための暗黙的ニューラル表現

近年，INR は画像・動画・3D シーンなど多様な信号表現に適用されている．画像やシーン表現など多くのタスクでは，画素位置を表す座標 x および y（あるいは三次元座標）を入力し，対応する信号値を出力する座標ベースの INR [7, 12, 16, 17, 18, 19, 23, 24, 25, 26] が広く用いられる．動画へ拡張する場合は，画素位置と時刻を表す時空間座標 (x, y, t) を入力とするが，フレーム数や解像度が増加すると学習サンプル数が大幅に増え，学習および推論の計算量が増大しやすい．この制約の下で，NeRV [6] に代表されるフレームベースの動画 INR が提案されている．

フレームベースの INR は，低次元の入力からフレーム全体を一括生成する点に特徴がある．NeRV はフレーム時刻（インデックス）を入力とするインデックスベース構造を導入した．この入力形式はフレームサイズに依存しないため，座標ベースの手法と比較して学習・推論時間を抑制できる．NeRV を基盤とした後続研究では，E-NeRV [20] が復号器設計の改良により表現能力の向上を図っている．また，DS-NeRV [22] は動画の動的成分と静的成分を分離することで再構成品質をさらに高めている．一方，HNeRV [13] は各フレームから抽出した特徴量（潜在埋め込み）をネットワーク入力に用いるハイブリッド構造を提案し，インデックスベース手法と比べて動画固有のパターンをより効果的に捉える．さらに，DNeRV [21] は隣接フレーム間の差分画像を統合することでハイブリッド手法を拡張し，頑健な動画表現を実現している．

3.2.2 動画圧縮

従来の動画符号化規格である HEVC/H.265 [2] や VVC/H.266 [3] は，事前定義された規則とアルゴリズム（ブロック分割，動き補償予測，変換・量子化，エントロピー符号化等）に基づき，高い圧縮効率と実用性を実現してきた．一方，近年は深層学習に基づく学習型動画圧縮が活発に研究されている．学習型動画圧縮 [4, 5] はデータ駆動で符号化過程を最適化できる反面，モデル構造の複雑化に伴って学習コストが増大し，復号速度

が課題となる場合がある．

INR に基づく動画圧縮の多くは，動画データをニューラルネットワークのパラメータ（および付随する潜在埋め込み）へ効率的に埋め込むことに主眼を置いている．具体的には，動画を埋め込んだネットワーク重みに対して枝刈り，量子化，エントロピー符号化などを適用することで符号量を削減する．この枠組みでは，復号が比較的簡潔なネットワークの順伝播で構成されるため，高速な復号と動画ごとの最適化に適する．一方で，各動画に対して埋め込みと圧縮を個別に学習する必要があり，汎用性の観点では制約がある．また，INR ベース動画圧縮の圧縮効率をさらに高めるために，エントロピー符号化と量子化モデルを統合した枠組み [30] も提案されている．

3.3 提案手法

3.3.1 概要

動画圧縮において，限られた符号量で動画を効率的に表現するには，時間方向の冗長性を抑制しつつ，高周波ディテールの再構成と時間的一貫性の維持を両立する復号戦略が重要である．しかし，ハイブリッドベース構造においてフレーム全体から特徴を抽出すると，隣接フレーム間の類似性に起因して抽出特徴も類似しやすく，入力特徴の冗長化を招き得る．また，ニューラルネットワークのスペクトルバイアス [8, 9, 10, 11] により高周波成分の再構成は難しく，時間情報を明示的に利用しない場合には時間的一貫性の確保にも課題が残る．

これらの課題に対処するために，本章では図 3.1 に示す二流路構成を提案する．提案手法は，入力フレーム I_t から高周波成分を抽出して特徴化し，高周波側の復号を担う High-Frequency stream（HF-stream）と，フレーム時刻に基づいて低周波側を復号する Low-Frequency stream（LF-stream）を組み合わせる．最終的な再構成フレームは，残差接続 [31, 32] と同等の役割を果たす加算統合により，両ストリームの出力を合成して得る．

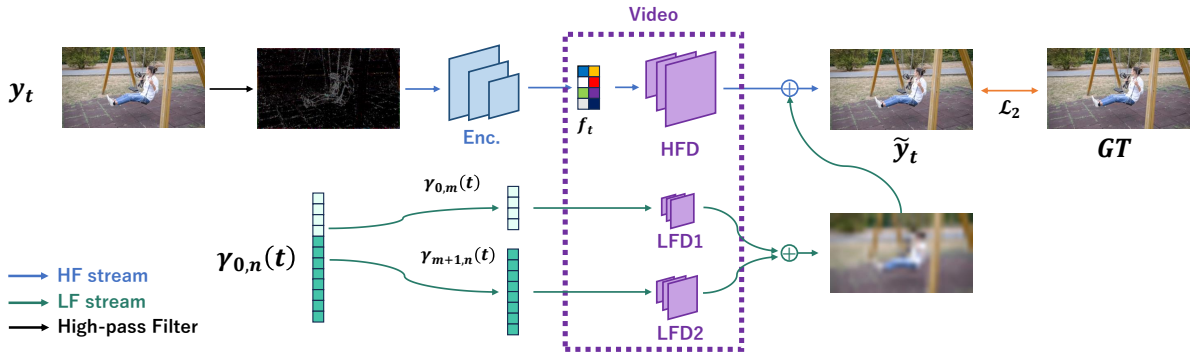

図 3.1: 提案手法のアーキテクチャ概要．HF-stream（青矢印）と LF-stream（緑線）を用い，両者の出力を加算することで再構成フレームを得る．図の画像 [33] は CC BY-NC 4.0 ライセンスの下で使用している．(文献 [7] より転載．)

3.3.2 High-Frequency stream

HF-stream は，高周波成分の再構成を主として担うネットワーク（High-Frequency Decoder：HFD）からなる．HFD はハイブリッドベース構造を採用し，フレーム固有の入力としてエンコーダが抽出するコンパクトな特徴（潜在埋め込み）を用いる．ハイブリッド構造では，特徴サイズは極めて小さい（例：$16 \times 2 \times 4$）設計のため，限られた次元で有用な情報を保持するには効率的な特徴抽出が不可欠である．ただし，HNeRV のようにフレーム全体 I_t から特徴を抽出する場合，隣接フレームの類似性が高い傾向にあることから特徴も類似しやすく，入力特徴の冗長性が生じる．図 3.2 は，連続フレームに対する $16 \times 2 \times 4$ の特徴を 8×16 へ整形して可視化した例であり，HNeRV では隣接フレーム間で類似した特徴が生成されやすいことを示している．このような類似特徴に基づく復号では，フレーム間で差異が大きい高周波成分が十分に表現されず，細部が失われやすい．

この問題に対処するため，提案手法では入力フレームにハイパスフィルタを適用して高周波成分 I_t^{HF} を抽出し，これをエンコーダに入力して特徴化する．高周波成分は一般に時間方向の変動が大きく，隣接フレーム間での類似性が比較的小さいため，そこから得られる特徴は冗長性が抑えられ，細部表現の強調に寄与すると期待される．高周波成分 I_t^{HF} はエンコーダ $E(\cdot)$ により低次元特徴 $\boldsymbol{f}_t$ へ変換され，HFD により高周波側の再構成 $\hat{I}_t^{\mathrm{HF}}$ が得られる．

$$\boldsymbol{f}_t = E\big(I_t^{\mathrm{HF}}\big) \tag{3.1}$$

$$\hat{I}_t^{\mathrm{HF}} = HFD(\boldsymbol{f}_t) \tag{3.2}$$

ここで，エンコーダには ConvNeXt ブロック [34] を用い，HFD は HNeRV ブロック [13] を複数段積層して構成する．

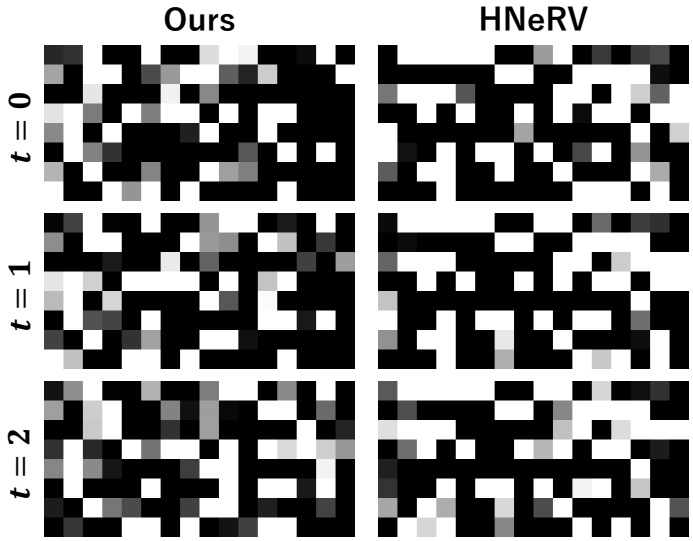

図 3.2: エンコーダから抽出された特徴 $\boldsymbol{f}_t$ の比較．提案手法では隣接フレーム間の特徴の類似が抑制され，フレームごとの高周波差分が強調される．(文献 [7] より転載．)

3.3.3 Low-Frequency stream

LF-stream は，低周波成分の再構成を担うインデックスベース構造であり，多層パーセプトロン（MLP）と複数の NeRV ブロック [6] から構成される．ここで，ネットワークパラメータ数を削減するため，LF-stream のデコーダを Low-Frequency Decoder 1（LFD1）および Low-Frequency Decoder 2（LFD2）の二つに分割する．LFD1 および LFD2 は HF-stream の HFD に比べて小規模に設計し，低周波側の再構成を安定に担わせる．低周波成分は一般に時間方向の相関が高く，フレーム固有情報を用いずともフレーム時刻に基づく生成により時間的一貫性を保ちやすい．また，インデックスベース手法はフレームから抽出された特徴を保持する必要がないため，低周波成分の再構成において効率的であり，動画表現の全体サイズを増加させない．

提案手法では，フレーム時刻 t を $(0, 1]$ の範囲に正規化し，Positional Encoding [6, 7, 35] により次元拡張した表現を入力として低周波側の復号を行う．Positional Encoding $\gamma_{0,n}(t)$ は次式で定義する．

$$\gamma_{0,n}(t) = \big(\sin(b^0\pi t), \cos(b^0\pi t), \ldots, \sin(b^n\pi t), \cos(b^n\pi t)\big) \tag{3.3}$$

ここで b および n はハイパーパラメータである．$\gamma_{0,n}(t)$ の各成分は，b^x の値に応じて異なる時間スケールに敏感に反応し，b^x が小さいほど長期的な変化，大きいほど短期的な変化を表現しやすい．一方で，n を大きくすると多様な時間変化を表現しやすくなる反面，$\gamma_{0,n}(t)$ の次元が増加するため MLP のパラメータ数が増大しやすい．

そこで本手法では，ハイパーパラメータ m（$0 < m < n$）を用いて $\gamma_{0,n}(t)$ を $\gamma_{0,m}(t)$ と $\gamma_{m+1,n}(t)$ の二つに分割し，それぞれを二つの部分ネットワーク LFD1，LFD2 へ入力する．この分割により，入力・出力次元に依存する MLP パラメータを削減できる．例えば m を n のおよそ半分に設定すると，MLP パラメータ数をおよそ半減できる．また，$\gamma_{0,n}(t)$ は周波数帯域ごとに異なる時間スケールの変化を表すため，周波数帯域を分けて処理する本分割は表現能力の低下を抑えつつパラメータ削減を実現できると考えられる．LF-stream の出力は次式で表される．

$$\hat{I}_t^{\mathrm{LF1}} = \mathrm{LFD1}(\gamma_{0,m}(t)) \tag{3.4}$$

$$\hat{I}_t^{\mathrm{LF2}} = \mathrm{LFD2}(\gamma_{m+1,n}(t)) \tag{3.5}$$

3.3.4 損失関数

提案法の最終出力フレームは，HF-stream および LF-stream の出力を加算して得る．

$$\hat{I}_t = \hat{I}_t^{\mathrm{HF}} + \hat{I}_t^{\mathrm{LF1}} + \hat{I}_t^{\mathrm{LF2}} \tag{3.6}$$

各ネットワークは，再構成フレーム $\hat{I}_t$ と元フレーム I_t の L_2 損失により最適化する．

$$\mathcal{L} = \sum_t \left\| \hat{I}_t - I_t \right\|_2^2 \tag{3.7}$$

3.4 実験

3.4.1 データセットと設定

実験では，DAVIS データセット [33] の動画を用いる．本データセットは解像度 1080×1920 の動画 50 本からなり，フレーム数は 25 から 104 である．本章では，各フレームを 640×1280 へクロップして用いた．高周波成分の抽出にはハイパスフィルタを用い，周波数帯域の 80% を除去する設定とした．学習は 300 エポックで行った．モデル規模は，特徴量およびデコーダ（HFD，LFD1，LFD2）の総パラメータ数が約 1.5M となるように調整し，HFD・LFD1・LFD2 のモデルサイズ比は 20:1:5 とした．Positional Encoding のハイパーパラメータは，$m = 10$，$n = 30$ とし，b は 1.25 とした．

動画圧縮の評価では，既存研究と同様に特徴量および各デコーダ（HFD，LFD1，LFD2）のパラメータに量子化を適用し，量子化後の重みに対して Huffman 符号化を行うことで圧縮した．量子化係数は 6 および 8 に設定する．評価指標には PSNR，MS-SSIM，LPIPS を用いる．

3.4.2 定量的評価

各動画シーケンスにおける HNeRV と提案手法の PSNR 差分（提案手法 −HNeRV）を図 3.3 に示す．縦軸が正であるほど，提案手法が HNeRV より高い PSNR を達成していることを意味する．ハイブリッドベース構造では，一部のシーケンスにおいて学習が十分に収束しない場合が報告されている [36]．図 3.3 では，HNeRV で学習が不安定であったシーケンスを赤，提案手法で同様の問題が生じたシーケンスを橙で強調表示している．例えば “m-jump” シーケンスでは，HNeRV の PSNR が 14.36 dB に留まり，学習が収束しなかった．一方，提案手法はデコーダを HFD，LFD1，LFD2 に分割し，小規模なネットワークを併用する構成であるため，学習が十分に収束しない場合でも一定の再構成品質を確保できる．これらの特定シーケンスを除けば，提案手法は概ね HNeRV を上回る性能を示す．

定量評価の集約として，再構成動画に対する各評価指標の平均を表 3.1 に示す．同表では，全シーケンス平均（All）に加え，ハイブリッドベース構造の HNeRV および提案手法については “m-jump”，“m-fly”，“p-launch” を除外し，学習に成功したシーケンスのみの平均（Suc）も併記している．

表 3.1: DAVIS データセットにおける再構成品質の平均値の比較

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV(All)	28.91	0.8865	0.3663
HNeRV(All)	30.94	0.9160	0.2955
Ours(All)	**31.81**	**0.9284**	**0.2780**
HNeRV(Suc)	31.19	0.9261	0.2873
Ours(Suc)	**31.91**	**0.9361**	**0.2674**

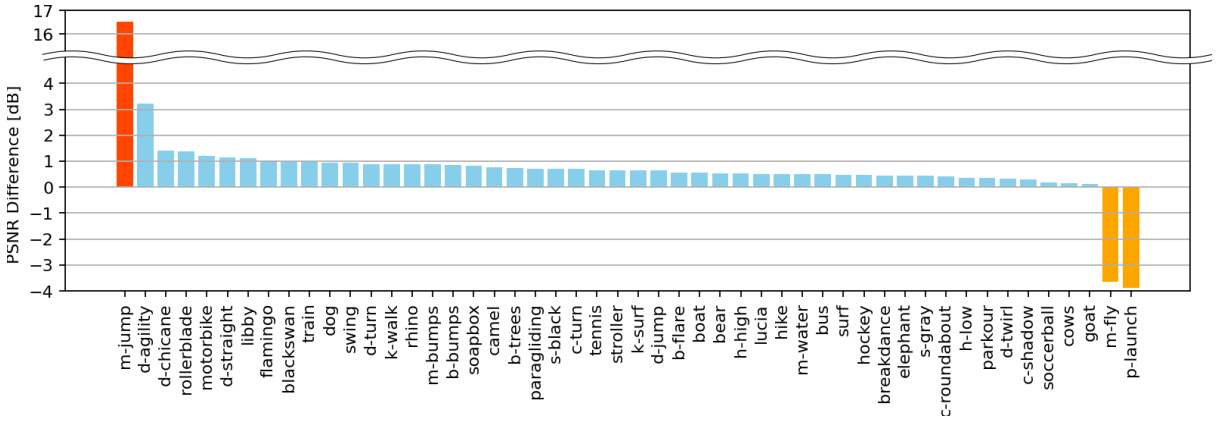

図 3.3: DAVIS データセット各動画シーケンスにおける PSNR 差分（提案手法 –HNeRV）．横軸は動画シーケンス，縦軸は PSNR 差分であり，正の値は提案手法が HNeRV を上回ることを示す．赤は HNeRV で学習が不安定であったシーケンス，橙は提案手法で学習が不安定であったシーケンスである．(文献 [7] より転載．)

3.4.3 定性的評価

定性的評価として，“hockey” シーケンスの連続フレームに対する可視化例を図 3.4 に示す．図中の赤字は各フレーム全体の PSNR を表す．NeRV は全体的に再構成品質が低い一方，HNeRV では複数フレームにおいてスティックやボールなどの物体が欠落するなど，時間方向に一貫した再構成が困難な場合が観察される．これに対し提案手法は，動画全体にわたり物体の形状や輪郭をより安定に再構成し，高い品質を維持する傾向を示す．

また，“soapbox” および “stroller” シーケンスの可視化例を図 3.5 に示す．提案手法は，木目模様や微細な線といった高周波ディテールをより忠実に再構成できていることが確認できる．これは，高周波成分から特徴を抽出することで，フレームごとの差異を反映した多様な特徴が得られ，細部表現の補完に寄与するためと考えられる（図 3.2 参照）．

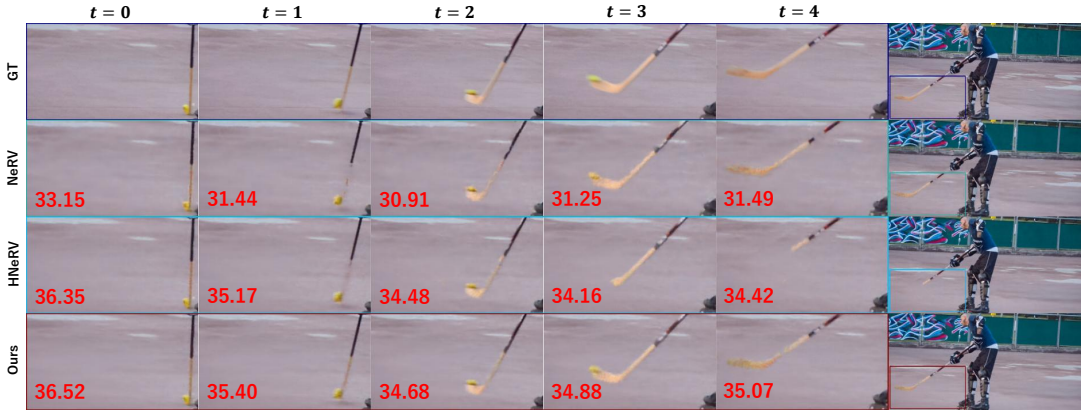

図 3.4: “hockey” 動画における連続フレームの再構成例．赤色の数値は各フレーム全体の PSNR を示す．図の画像 [33] は CC BY-NC 4.0 ライセンスの下で使用している．(文献 [7] より転載．)

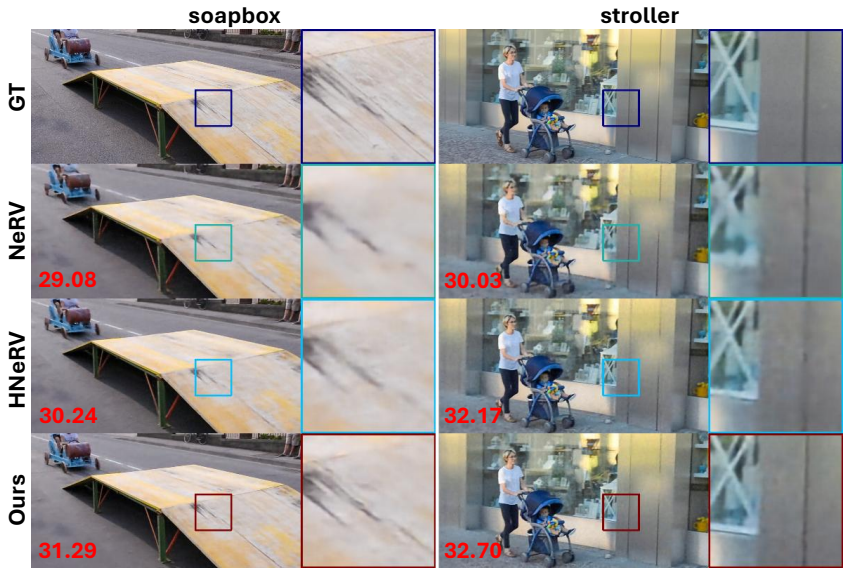

図 3.5: “soapbox” および “stroller” シーケンスにおける再構成結果の可視化例．赤色の数値は各フレームの PSNR を示す．図の画像 [33] は CC BY-NC 4.0 ライセンスの下で使用している．(文献 [7] より転載．)

3.4.4 圧縮性能の比較

DAVIS データセットにおける圧縮性能の比較結果を図 3.6 に示す．複数のモデルサイズに対して特徴量およびモデル圧縮を適用し，提案手法と既存手法を比較することで，提案手法の圧縮性能を評価する．図 3.6 より，提案手法は従来の INR ベース手法と比較して，良好な率歪み特性を示すことが分かる．

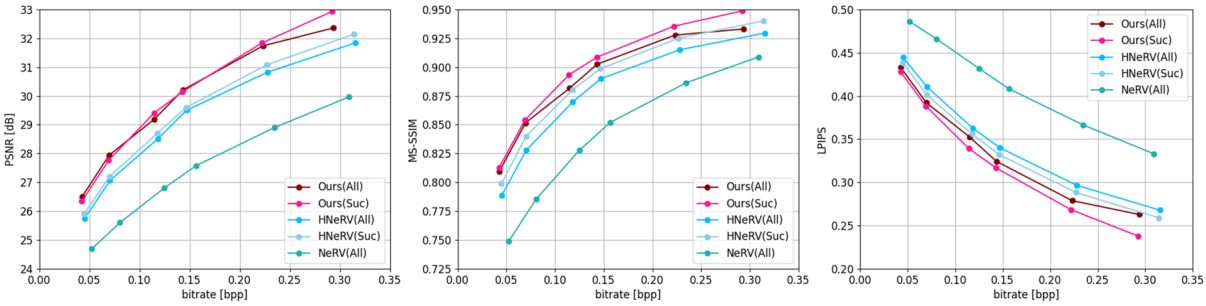

図 3.6: DAVIS データセットにおける圧縮性能の比較結果．(文献 [7] より転載．)

3.5 まとめ

本章では，フレーム固有情報と時間情報の双方を組み込む二流路構成により，高周波ディテールの再構成と時間的一貫性の両立を図る INR ベースの動画表現手法を提案した．HF-stream では，フレームの高周波成分に特化した特徴抽出を行うことで入力特徴の冗長性を抑制しつつ，細部表現を強化した．LF-stream では，フレーム時刻に基づく復号により時間的に滑らかな低周波成分を安定に生成し，時間的一貫性の維持を図った．

さらに，LF-stream のデコーダを分割することで MLP のパラメータ数を抑え，限られたモデル容量の下で再構成品質と時間的一貫性を両立する設計とした．実験では，DAVIS データセットにおいて再構成品質および圧縮性能の観点から提案手法の有効性を確認した．今後の課題として，各動画シーケンスに対するハイパーパラメータの最適化や，一部シーケンスで生じる学習不安定性の要因解析が挙げられる．

第 4 章

効率的なニューラル動画表現のための構造保持パッチ復号

4.1 はじめに

Implicit Neural Representation（INR）は，複雑な信号を明示的な表現で記述する従来手法に対し，コンパクトかつ解像度非依存の代替表現として注目を集めている．ニューラルネットワークにより空間座標および時間座標を入力とする連続関数として信号をパラメータ化することで，INR はメモリ使用量を大幅に削減しつつ，任意の解像度における滑らかな補間と再構成を可能にする．この特性により，INR は新規視点合成 [7, 23, 37] や静止画像表現 [16, 17, 18, 19] に加え，近年では動画表現 [6, 13, 20] へも応用が広がっている．

既存の INR に基づく動画表現手法は，復号の粒度に応じて主に次の 3 種類に分類できる．(1) 座標ベース手法 [12, 38]：時空間座標から各画素値を予測する．(2) フレームベース手法 [6, 13, 20]：コンパクトな入力からフレーム全体を生成する．(3) パッチベース手法 [39, 40]：空間的に分割されたパッチを生成し，再配置してフレームを再構成する．座標ベース手法は，微細な空間制御と解像度スケーラビリティに優れる一方，画素単位で独立に予測を行うためシーン全体の構造を捉えにくく，計算コストが大きい．これらの課題に対し，フレームベース手法は時間埋め込みやコンテンツ埋め込みといった低次元の入力からフレーム全体を直接生成することで，高速な推論と大域構造のモデル化を実現する．しかし，アップサンプリングに起因するアーティファクトや，ニューラルネットワークが低周波成分を優先して学習するスペクトルバイアス [8, 9, 10, 11] の影響により，高周波成分（細部テクスチャ）の再構成に課題を抱えることが多い．また，パッチベース手法は両者の中間に位置し，局所的な再現性と大域構造のモデル化を両立し得るが，一様な空間分割では独立に生成された領域間の整合性が不足し，パッチ境界に不連続が生じやすい．その結果，継ぎ目などの視覚的アーティファクトが発生し，知覚品質が低下することがある．

これらの課題を克服するため，本章ではニューラル動画表現のための構造保持パッチ（Structure-Preserving Patch：SPP）に基づくデコード手法を提案する．従来のようにフレームを空間的に分離したパッチへ分割するのではなく，本手法は PixelUnshuffle に類似した画素再配置を適用し，元の空間レイアウトを保持したパッチ画像群を生成する．この画素再配置によりパッチ間の相対的な空間関係が維持され，ネットワークはまずシーンの大域構造を捉えた上で，局所的な細部の精緻化へ段階的に焦点を移すことが可能となる．これを実現するため，デコーダ前段ではコンテンツ埋め込みと時間インデックスにより大域的構造をモデル化し，後段

でパッチインデックスを導入して局所的精緻化を行う構成とする．この global-to-local な再構成戦略により，パッチ境界に起因するアーティファクトを低減し，空間的な整合性を向上させる．実験により，本手法は既存の INR ベース動画表現手法と比較して，再構成品質および圧縮性能の両面で優れた性能を示すことを確認した．

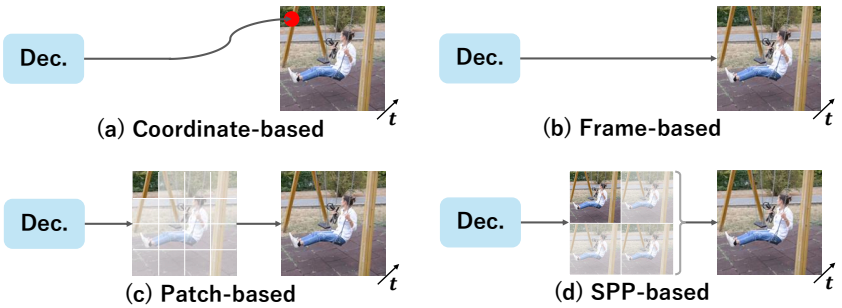

図 4.1: ニューラル動画表現における復号戦略の比較．(a) 座標ベース：時空間座標から各画素値を直接予測する．(b) フレームベース：埋め込みからフレーム全体を生成する．(c) パッチベース：一様分割したパッチを生成・復号し，再配置してフレームを再構成する．(d) SPP ベース（提案手法）：構造保持パッチ（Structure-Preserving Patch：SPP）を用い，空間整合を保ったパッチ復号と global-to-local 復元によりフレームを再構成する．図の画像 [33] は CC BY-NC 4.0 ライセンスの下で使用している．(文献 [6] より転載．)

4.2 関連研究

4.2.1 暗黙的ニューラル表現

INR の中核概念は，信号をニューラルネットワークによって実装される連続関数としてモデル化する点にある．形式的には，信号は学習可能パラメータ θ をもつネットワーク Φ_θ により次のように表される．

$$\boldsymbol{y} = \Phi_\theta(\boldsymbol{x}) \tag{4.1}$$

ここで，$\boldsymbol{x}$ は連続座標，$\boldsymbol{y}$ はその座標における信号値を表す．Φ_θ は与えられたターゲット信号に適合するよう最適化され，最適化後のネットワーク自体が信号の表現として機能する．本分野の代表的な先駆研究として，3D シーンを空間座標と視線方向により定義される放射輝度場（Radiance Fields）としてモデル化する Neural Radiance Fields（NeRF） [7] が挙げられる．NeRF は，カメラ光線に沿った色と密度を予測する全結合ネットワークを最適化することで，フォトリアリスティックな新規視点合成が可能であることを示した．計算負荷は高いものの，NeRF は明示的なメッシュやボクセル表現に依存せず，複雑な連続構造をニューラル関数として表現できる可能性を示した点で重要である．また，周期的な活性化関数を用いる SIREN [12] は，ReLU ベースのネットワークに内在するスペクトルバイアスを緩和し，微細なディテールやテクスチャの表現能力を向上させる．これらの成果を背景として，INR は空間・時間を含む時空間信号へと適用範囲を拡張し，とりわけ動画データへの応用が活発に研究されている．

4.2.2 ニューラル動画表現

近年，INR は動画表現および動画圧縮に対する新たな枠組みとして研究が進んでいる．代表例である NeRV [6] は，フレームインデックスを入力として対応する RGB フレームを出力するネットワークを学習し，そのネットワーク重み（および付随する設計）により動画全体を符号化する．座標マッピングに基づく従来の INR とは異なり，NeRV は動画を「フレームインデックスからフレームへの写像」として扱い，動画コンテンツをネットワークパラメータへ効率的に埋め込む．これにより，復号は指定されたフレームインデックスに対する順伝播のみで実現される．また，重みの枝刈りや量子化などによりモデルを圧縮することが，そのまま動画圧縮に対応する．この系譜として，E-NeRV [20] は空間的要素と時間的要素を分離することで NeRV を改良している．具体的には，領域ごとに異なる埋め込みを学習し，復号時に Adaptive Instance Normalization（AdaIN） [15] を介して結合することで，表現効率と圧縮性能の向上を図る．HNeRV [13] はさらに，フレームインデックスから導出されるコンテンツ非依存の埋め込みを，フレーム内容に適応した埋め込みへ置き換えるハイブリッド構造を導入した．これにより，複雑なフレームや高周波成分を多く含むフレームに対して，より多くのモデリング能力を割り当てることが可能となる．さらに Boosting NeRV [14] は条件付きデコードフレームワークを導入し，既存の多様な INR ベース動画表現モデルを強化可能とした．フレームインデックスに基づく時間アフィン変換により中間特徴を調整することで，異なるアーキテクチャに対して一貫した再構成性能の改善を実現している．

一方で，再構成品質やスケーラビリティの向上を目的として，デコード戦略そのものを工夫する研究も行われている．例えば PS-NeRV [39] は，フレーム全体を一度に復号するのではなく，パッチ単位で動画フレームを再構成する．この設計は高解像度コンテンツに対して柔軟性を与え，並列計算にも適する一方，空間的連続性が損なわれることで，パッチ境界に沿った可視アーティファクトが生じやすい．他のいくつかの研究では，階層的なモデル構造を導入する手法 [22, 36, 41]，動画固有のダイナミクスをモデル化する手法 [21]，明示的な残差接続を組み込む手法 [32, 42]，あるいは周波数領域特性の保持を重視する手法 [43, 44, 45] などにより，時間的ちらつきや細部復元の不十分さといった課題の軽減が試みられている．以上を踏まえ，本研究はパッチベース表現に焦点を当てる．従来のパッチベース手法とは異なり，境界アーティファクトを軽減し知覚的一貫性を向上させるため，復号過程において大域的な空間構造を明示的に保持することを目的とする．

4.3 提案手法

4.3.1 概要

本研究では，図 4.2 に示すように，構造保持パッチ（Structure-Preserving Patch：SPP）に基づくニューラル動画表現手法を提案する．本手法は，各動画フレームを空間的に整列した複数のパッチ画像の集合として扱い，それらを復号することでフレーム全体を再構成する．元フレームの空間レイアウトを維持するため，従来のパッチベース手法のようにフレームを独立なタイルへ単純に分割するのではなく，PixelUnshuffle に類似した決定論的な画素再配置によりパッチを生成する．この操作により，パッチ間で一貫した空間配置が保たれ，ネットワークはパッチ画像間に存在する冗長性を活用しながらシーン全体の大域構造をモデル化しやすくなる．さらに，再構成品質を向上させるため，デコーダは global-to-local なフィッティング戦略に従うよう

設計する．具体的には，デコーダ前段ではコンテンツ埋め込みと時間インデックスに基づいてフレームの大域的構造をモデル化し，後段でパッチインデックスを導入して各パッチ内の局所的細部を精緻化する．この段階的なデコード過程により，大域的なレイアウトの整合性を保ちつつ，高周波の視覚的ディテールを含む精密な再構成を可能にする．

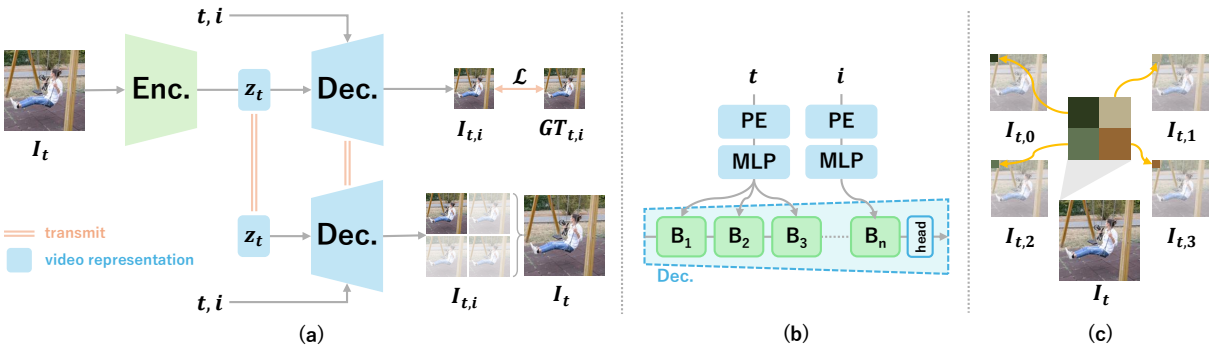

図 4.2: 提案する SPP に基づくニューラル動画表現フレームワークの概要．(a) 全体構成．入力フレーム I_t をエンコーダでフレーム埋め込み z_t へ変換し，時間インデックス t とパッチインデックス i を条件としてデコーダが各パッチ $\hat{I}_{t,i}$ を復号する．復号したパッチを逆再配置することでフレーム全体 $\hat{I}_t$ を再構成する．(b) デコーダ詳細．t および i は Positional Encoding と MLP により埋め込みへ変換する．デコーダ前段は t で条件付けして大域的構造をモデル化し，後段は i を用いて局所的なパッチ詳細を精緻化する（global-to-local 復号）．(c) SPP によるパッチ構成．各フレームを空間整合を保った複数のパッチ画像（例：$\hat{I}_{t,0}$〜$\hat{I}_{t,3}$）として表現し，パッチ間で一貫した空間配置を維持する．図の画像 [33] は CC BY-NC 4.0 ライセンスの下で使用している．(文献 [6] より転載．)

4.3.2 動機

はじめに，global-to-local 戦略の有効性を明確にするため，図 4.3 に示す単純な一次元信号フィッティング実験を行う．本実験では，以下の 4 手法を比較する．(1) *Point-wise fitting*：各座標を独立に扱い，点単位で信号をフィッティングする手法．(2) *Segment-wise fitting*：信号を固定長の区間に分割し，各区間を独立にフィッティングする手法．(3) *Global fitting*：信号全体を単一の入力からフィッティングする手法．(4) *SPP-style fitting*：SPP 機構を模倣し，固定位置サンプルに基づいて並べ替えた信号をフィッティングする手法（提案手法に対応）．ここで，フィッティングに用いるネットワークは，活性化関数を持つ 5 層の全結合ネットワークで構成される．実験結果より，*Global fitting* および *SPP-style fitting* は，信号の大域構造を保持しながら近似を行うことで，より高い適合精度を達成することが確認できる．一方，*Point-wise fitting* および *Segment-wise fitting* は局所的な適合を優先するため，大域的な一貫性を維持しにくい．以上の観察は，正確な信号再構成のためには，大域的構造のモデリングを先行させた上で局所的な精緻化を行う段階的な復号が有効であることを示唆する．本知見は，動画表現において，提案する global-to-local デコードフレームワークの設計動機を与える．

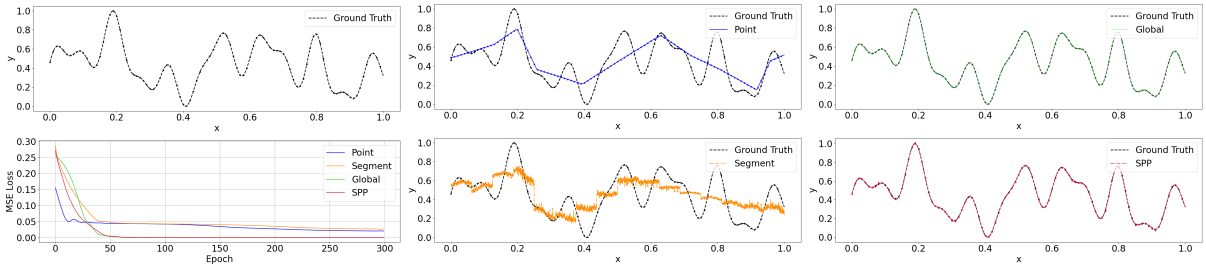

図 4.3: 1 次元信号のフィッティング性能の比較．上段（左から順に）は，正解信号，*Point-wise fitting*，および *Global fitting* の結果である．下段（左から順に）は，学習損失の推移，*Segment-wise fitting*，および *SPP-style fitting*（提案）の結果である．(文献 [6] より転載．)

4.3.3 構造保持パッチ復号

提案フレームワークは，図 4.2 に示すように，global-to-local なデコード戦略に基づくニューラルネットワークを用いて，各動画フレームをパッチ単位で予測する．時刻 t における入力フレーム I_t に対し，概念的に空間構造を保持した P 個のパッチへ分解して扱う．これを実装するため，PixelUnshuffle に類似した画素再配置操作を適用し，フレーム I_t を P 個のパッチ画像 $I_{t,0}, I_{t,1}, \ldots, I_{t,P-1}$ へ変換する．エンコーダは ConvNeXt ブロック [13, 34] から構成され，まずフルフレーム I_t を入力としてコンパクトなフレームレベル埋め込み $\boldsymbol{z}_t$ を生成する．得られた $\boldsymbol{z}_t$ はデコーダへ入力され，時間インデックス t およびパッチインデックス i の 2 つのインデックスに条件付けして，各パッチ画像 $\hat{I}_{t,i}$ を生成する．デコーダでは，前段層を t のみにより変調し，フレーム内の全パッチに共通する大域的構造をモデルが捉えられるようにする．一方，後段層は i により条件付けられ，各パッチ固有の局所的詳細を精緻化する役割を担う．全パッチ出力 $\{\hat{I}_{t,i}\}_{i=0}^{P-1}$ が得られた後，それらを空間的に再配置することでフル解像度フレーム $\hat{I}_t$ を再構成する．この階層的な条件付け機構により，デコーダはまず一貫性のある大域構造を構築し，その後に局所的内容へ特化することで，再構成品質の向上を図る．さらに，パッチ生成における構造保持特性により，単純なパッチベース復号で生じやすい境界不連続性を抑制できる．デコーダは複数のブロック B_n から構成され，各ブロックは Boosting NeRV [14] で導入された SNeRV Block と TAT Residual Block を統合した構造を有する．

4.3.4 損失関数

フレームの各パッチ i に対して，空間的忠実度，知覚品質，および周波数領域の一貫性を考慮する複合損失 $\mathcal{L}_i$ を定義する．

$$\mathcal{L}_i = w_i \left(\alpha \mathcal{L}_1(x_i, \hat{x}_i) + \beta \mathcal{L}_{\text{MS-SSIM}}(x_i, \hat{x}_i)\right) + \mathcal{L}_{\text{freq}}(x_i, \hat{x}_i) \tag{4.2}$$

$$\mathcal{L}_{\text{freq}}(x, \hat{x}) = \mathcal{L}_1(\text{FFT}(x), \text{FFT}(\hat{x})) \tag{4.3}$$

ここで，x_i および $\hat{x}_i$ はそれぞれ正解パッチ画像と予測パッチ画像を表す．$\mathcal{L}_1$ は標準的な L1 画素損失であり，$\mathcal{L}_{\text{MS-SSIM}}$ は知覚品質の向上を促す MS-SSIM 損失である．係数 α と β は各損失項の相対的な寄与を調整するハイパーパラメータである．また，周波数領域項 $\mathcal{L}_{\text{freq}}$ は，両画像に高速フーリエ変換（FFT）を適用し，その差分を L1 ノルムで評価することで算出される．

提案する SPP ベースの復号スキームでは，パッチ固有の変動はデコーダ後段層のみで調整される．この局所的な変調のみでは，隣接パッチと大きく異なる領域の学習が難しくなり，結果としてパッチの再構成精度が低下する可能性がある．そこでこの問題を解決するため，適応的なパッチ重み係数 w_i を導入し，学習を難易度の高い領域へ配分する．w_i は以下のように定義される．

$$w_i = \frac{\sum_{j \neq i} \|x_i - x_j\|_1}{\sum_k \sum_{j \neq k} \|x_k - x_j\|_1 + \epsilon} \tag{4.4}$$

ここで，ϵ はゼロ除算を避けるための微小定数である．この定式化により，全パッチにわたって重みが正規化され，モデルは一貫した構造の学習を維持しつつ，差異の大きい（すなわち難易度の高い）パッチに対してより重点的に最適化を行うようになる．

4.4 実験

4.4.1 データセット

提案する SPP ベースの動画表現手法の有効性を，Bunny [46]，DAVIS [33]，MCL-JCV [47]，UVG [48] の 4 つのベンチマークデータセットを用いて評価する．Bunny データセットは，解像度 720 × 1280，132 フレームからなる単一シーケンスで構成される．学習の安定性を確保するため，各フレームは 640 × 1280 にクロップして用いる．この調整は，高解像度入力に対して HNeRV の学習が収束しない場合が観察されたために行うものである（例：表 4.3 の “Bosphorus”）．DAVIS データセットは，50 本の自然動画シーケンスから構成され，元の解像度は 1080 × 1920 であるが，本実験では 640 × 1280 にクロップして用いる．各シーケンスの長さは比較的短く，25〜104 フレームである．同様に，MCL-JCV は 30 本のシーケンスを含み，元の解像度は 1080 × 1920 であるが，本実験では 640 × 1280 にクロップする．シーケンス長は 120〜150 フレームの範囲である．UVG データセットはフル HD 解像度（1080 × 1920）の 7 本の長尺シーケンスからなり，各シーケンスは 300 または 600 フレームを含む．

4.4.2 設定

本手法は，既存の INR ベースラインである NeRV，HNeRV，Boosting-HNeRV（Boost）と比較する．また，動画圧縮タスクにおいては，HEVC の参照ソフトウェアである HM との比較も追加で実施する．

本手法では，デコーダのストライド構成を入力解像度に合わせて調整する．具体的には，640 × 1280 ではストライド列を $[5, 4, 2, 2, 2]$，1080 × 1920 では $[5, 3, 2, 2]$ に設定する．パッチ数は $P = 4$ に固定し，損失関数の重み係数は経験的に $\alpha = 42$，$\beta = 18$ とする．モデルサイズは既定で約 1.5M パラメータとし，ただし UVG（1080 × 1920）では 3.0M パラメータのモデルを用いる．定量的評価には PSNR，MS-SSIM，LPIPS を用いる．

圧縮性能の評価では，解像度 1080 × 1920 の UVG データセットを使用し，各シーケンスを 60 または 120 フレームにサブサンプリングする．各動画について，ネットワーク幅を調整することでモデルサイズの異なる複数のモデルを学習し，それぞれに対して量子化およびエントロピー符号化を適用することで，率歪み（Rate–Distortion：RD）解析を行う．

4.4.3 実験結果

DAVIS および MCL-JCV における定量的結果を，それぞれ表 4.1 および表 4.2 に示す．これらの表は，各データセットにおける平均再構成品質を示す．また，表 4.3 は UVG データセットの各シーケンスに対する PSNR および MS-SSIM を示す．提案手法は多くのシーケンスで既存手法を上回る指標値を達成し，一貫して高い再構成品質を示している．さらに表 4.4 は Bunny データセットにおける学習エポックごとの PSNR を示しており，既存手法と比較して再構成性能の向上に加えて，収束の速さも確認できる．

表 4.1: DAVIS データセット（1.5M，640 × 1280）における再構成品質の比較

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV	28.60	0.8811	0.4150
HNeRV	30.69	0.9146	0.3476
Boost	<u>33.53</u>	<u>0.9604</u>	<u>0.2674</u>
Ours	**34.23**	**0.9643**	**0.2539**

表 4.2: MCL-JCV データセット（1.5M，640 × 1280）における再構成品質の比較

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV	31.64	0.9217	0.4126
HNeRV	33.47	0.9417	0.3571
Boost	<u>35.60</u>	<u>0.9638</u>	<u>0.3039</u>
Ours	**35.94**	**0.9654**	**0.2936**

表 4.3: UVG データセット（3M，1080 × 1920）における再構成品質の比較（PSNR / MS-SSIM）

Method	Beauty	Bosphorus	HoneyBee	Jockey	ReadySetGo	ShakeNDry	YachtRide	Average
NeRV	32.93 / 0.8843	33.09 / 0.9288	37.08 / 0.9780	31.06 / 0.8767	24.66 / 0.8205	32.67 / 0.9264	27.75 / 0.8573	31.22 / 0.8937
HNeRV	33.18 / 0.8876	17.57 / 0.5885	38.97 / 0.9838	31.75 / 0.8884	25.09 / 0.8358	33.73 / 0.9337	27.95 / 0.8546	29.44 / 0.8470
Boost	<u>33.70</u> / <u>0.8980</u>	<u>35.81</u> / <u>0.9631</u>	<u>39.52</u> / <u>0.9852</u>	<u>33.84</u> / **0.9253**	<u>27.72</u> / **0.9070**	<u>35.55</u> / <u>0.9532</u>	<u>29.03</u> / **0.8967**	<u>33.45</u> / <u>0.9311</u>
Ours	**33.82** / **0.8992**	**36.08** / **0.9644**	**39.60** / **0.9854**	**34.21** / <u>0.9249</u>	**28.03** / <u>0.9054</u>	**35.96** / **0.9584**	**29.33** / <u>0.8945</u>	**33.70** / **0.9312**

定性的評価として，図 4.4 に DAVIS および MCL-JCV から抽出した代表フレームの再構成例を示す．提案手法は，フレーム間でシャープなエッジと一貫した空間構造を維持し，多様なシーンにおいて高い再構成品質を示す．

最後に，UVG における RD 曲線を図 4.5 に示す．なお，HNeRV は学習の収束失敗が繰り返し発生したため，平均曲線の算出から除外した．提案手法は INR ベース手法と比較して優れた圧縮性能を示す一方で，従来の動画圧縮技術である HEVC [2]（HM 18.0）にはなお及ばず，モデル圧縮の観点から改善の余地が残る．

表 4.4: Bunny 動画（1.5M，640 × 1280）における学習エポックに対する PSNR の推移

Epoch	50	100	150	200	250	300
NeRV	26.95	29.40	30.43	30.96	31.14	31.33
HNeRV	29.93	33.32	34.86	35.61	36.06	36.35
Boost	34.90	37.35	38.12	38.62	38.70	39.03
Ours	**37.80**	**38.62**	**39.00**	**39.18**	**39.31**	**39.40**

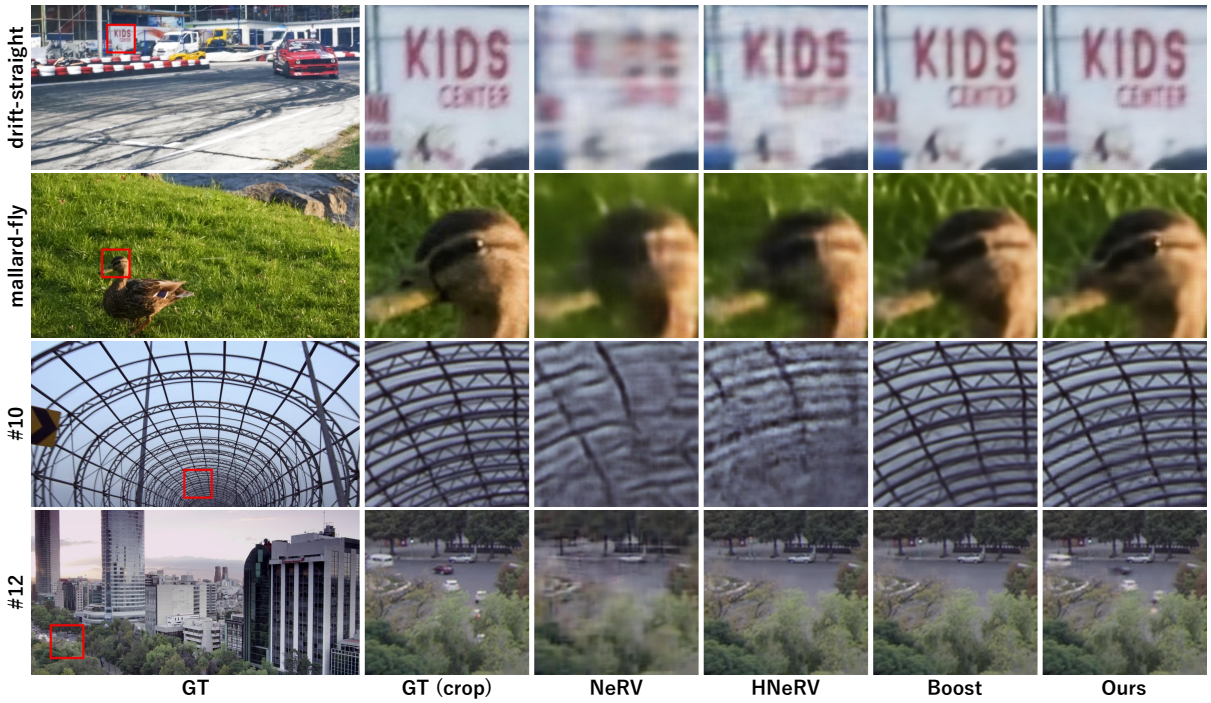

図 4.4: 複数シーケンスにおける再構成フレームの可視化比較．上から順に，DAVIS データセットの “drift-straight” および “mallard-fly”，MCL-JCV データセットのシーケンス#10 および#12 を示す．提案手法（SPP に基づく手法）は，エッジのシャープネスと空間構造の一貫性を維持し，視覚品質を改善する．図中の DAVIS dataset の画像 [33] は CC BY-NC 4.0 ライセンスの下で使用し，MCL-JCV dataset の画像 [47] は配布元の許諾条件（Copyright Notice）に基づき使用している．(文献 [6] より転載．)

4.5 まとめ

本章では，SPP に基づくニューラル動画表現手法を提案した．提案手法では，各動画フレームに対して PixelUnshuffle に類似した画素再配置を用い，元の空間構造を保持した空間的に一貫性のあるパッチ画像群へ変換して復号を行う．この構造保持の戦略により，従来手法で一般的に見られる劣化が軽減され，大域構造への適合を促進する．また，滑らかな再構成を促進するため，global-to-local なデコーダアーキテクチャを設計した．デコーダ前段では時間インデックスに条件付けてフレーム全体の大域的構造をモデル化し，後段でパッ

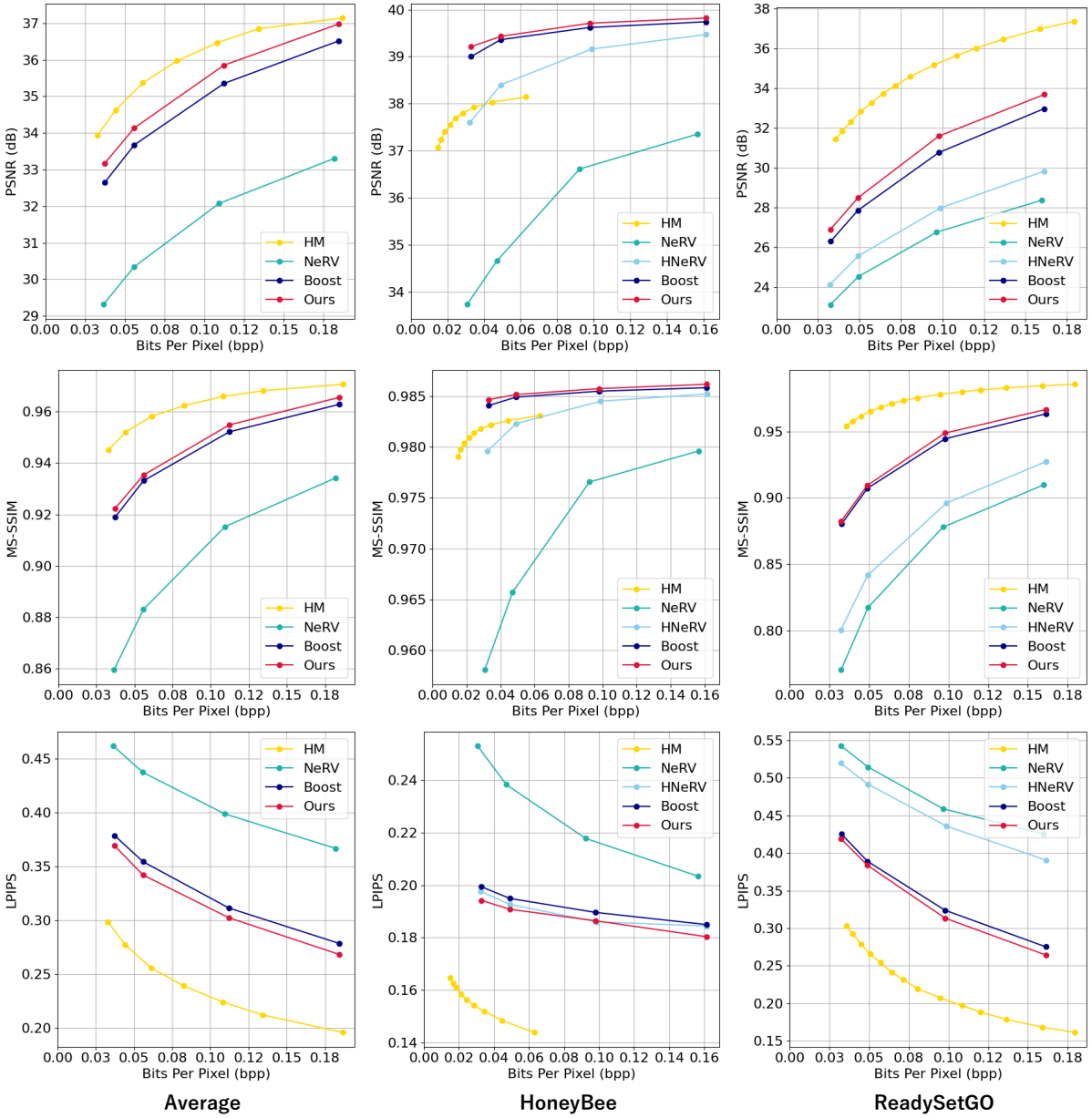

図 4.5: UVG データセットにおける圧縮性能の比較．(文献 [6] より転載．)

チ固有情報を導入して局所的詳細を精緻化する．この段階的な分離により，構造的整合性と細粒度の忠実度を効果的に両立できる．実験結果から，提案手法は既存の NeRV 系ベースラインと比較して高い再構成品質を達成することを確認した．以上の知見は，空間構造を明示的に保持するパッチ設計が，ニューラル動画表現の忠実度と効率性の向上に有効である可能性を示す．今後の課題として，シーン内容や複雑度に応じてパッチ配置や容量配分を動的に調整する適応的・内容依存型レイアウトを検討し，空間的一貫性および圧縮性能のさらなる改善を目指す．

第 5 章

結論と今後の展望

5.1 結論

本修士論文では，暗黙的ニューラル表現（Implicit Neural Representation：INR）に基づく動画表現に着目し，ニューラルネットワークへの動画埋め込みとデコード戦略の設計を通じて，再構成品質と表現効率の両立を目指した．INR に基づく動画表現は，復号をニューラルネットワークの順伝播として実現できる一方で，フレーム由来の入力特徴に含まれる時間冗長性，高周波成分の再構成の難しさ，ならびに空間・時間的一貫性の維持といった課題が残る．

第 3 章では，時間冗長性の抑制と時間的一貫性の維持を目的として，高周波成分の再構成を主に担う HF-stream と，低周波成分の再構成を主に担う LF-stream からなる二流路構成の INR ベース動画表現手法を提案した．HF-stream では，フレームの高周波成分から特徴を抽出することで，隣接フレーム間で冗長化しやすい情報の混入を抑制しつつ，細部表現の強化を図った．LF-stream では，フレーム時刻に基づく復号により時間方向に滑らかな低周波成分を安定に生成し，動画としての時間的一貫性の維持を狙った．両者の出力を加算して統合することで，高周波ディテールの忠実度と時間的一貫性を両立する設計指針を示した．

第 4 章では，Structure-Preserving Patch（SPP）に基づくデコード戦略を提案した．PixelUnshuffle に類似した画素再配置により，フレームを空間構造を保持したパッチ画像群へ分解し，global-to-local な条件付けに基づいてパッチを復号することで，フレーム全体の大域構造を捉えた後に局所的な細部を精緻化する復号過程を実現した．また，空間・周波数特性を考慮した損失関数と適応的パッチ重み付けにより，構造的一貫性と局所忠実度の両立を促進した．

以上の検討により，標準的なベンチマーク動画を用いた評価を通じて，提案した二流路設計および SPP に基づくデコード戦略が，既存の INR ベース手法と比較して，再構成品質と圧縮性能の観点で有効であることを確認した．本修士論文は，INR に基づく動画表現における動画埋め込みとデコード戦略の設計に対し，時間冗長性の抑制と空間構造保持の両面から改善可能性を示し，実験的検証に基づく設計指針を与えた．

5.2　今後の展望

今後の展望として，第一にレート歪み最適化（Rate–Distortion：RD 最適化）を明示的に取り込む枠組みの検討が挙げられる．本論文では量子化とエントロピー符号化により圧縮性能を評価したが，学習段階からレート項を導入し，量子化誤差や符号長を考慮した最適化へ拡張することで，より直接的に圧縮性能を改善できる可能性がある．また，モデルパラメータや埋め込みに対するエントロピーモデルの導入は，圧縮効率改善の観点から重要である．

第二に，学習の安定化と汎化性能の向上が挙げられる．INR に基づく動画表現では，解像度や系列内容に応じて学習が不安定化する場合がある．最適化の初期化，学習率スケジューリング，正則化，周波数特性の制御など，安定化手法の体系的整理が求められる．また，各動画シーケンスごとの最適化に依存する設計はエンコードに時間を要するため，事前学習やメタ学習などに基づく高速適応の検討も有望である．

第三に，コンテンツ適応型のパッチ設計が挙げられる．本論文では固定パッチ数で構造保持を実現したが，領域ごとの複雑度や運動量に応じてパッチ配置や容量配分を動的に変化させることにより，限られたモデル容量をより効果的に利用できる可能性がある．具体的には，高周波成分が卓越する領域への適応的割当てや，領域分割そのものを学習的に最適化する枠組みが考えられる．

最後に，デコードの高速化と実装面の最適化が挙げられる．INR の利点を実運用へ接続するには，GPU 並列性を活かした推論最適化，低ビット量子化，限られた計算資源下での推論効率など，システム実装上の検討が不可欠である．デコード速度，消費電力，メモリ帯域といった実行時制約を含めた評価を行うことで，応用可能性をより明確化できる．

謝辞

本研究の遂行にあたり，多くの方々より多大なご指導とご支援を賜った．ここに記して深謝する．

まず，指導教員である渡辺裕教授には，研究の方向性に関する助言から論文執筆に至るまで，終始丁寧かつ的確なご指導をいただいた．研究の進め方，課題設定の重要性，結果の解釈と議論の組み立て方など，学術研究に不可欠な姿勢を学ぶ機会を数多く与えていただいたことに，心より感謝する．

次に，渡辺研究室の皆様には，日々の議論やゼミでのコメントを通じて多くの示唆をいただいた．実験設計や実装上の課題に関する相談に乗っていただいたことに加え，研究生活を支える雰囲気づくりにも助けられた．互いに刺激を与え合える環境の中で研究に取り組めたことは，大きな財産である．ここに感謝する．

最後に，これまで常に見守り，学業と研究を継続できるよう支えてくれた両親に深く感謝する．日々の生活を支えてくれたことはもちろん，挑戦を後押ししてくれたことが，本研究を完成させる大きな力となった．

以上，支えてくださったすべての方々に，心より御礼申し上げる．

参考文献

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, pp. 560–576, Jul. 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, pp. 1649–1668, Sep. 2012.

[3] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 3736–3764, Aug. 2021.

[4] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11006–11015, Jun. 2019.

[5] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18114–18125, Dec. 2021.

[6] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "Nerv: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21557–21568, Dec. 2021.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, pp. 99–106, Dec. 2021.

[8] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International conference on machine learning*, pp. 5301–5310, Jun. 2019.

[9] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu, "Towards understanding the spectral bias of deep learning," *arXiv preprint arXiv:1912.01198*, 2019.

[10] Z. Cai, H. Zhu, Q. Shen, X. Wang, and X. Cao, "Batch normalization alleviates the spectral bias in coordinate networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25160–25171, Jun. 2024.

[11] Z. Liu, H. Zhu, Q. Zhang, J. Fu, W. Deng, Z. Ma, Y. Guo, and X. Cao, "Finer: Flexible spectral-bias tuning in implicit neural representation by variable-periodic activation functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2713–2722, Jun.

2024.

[12] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, Dec. 2020.

[13] H. Chen, M. Gwilliam, S.-N. Lim, and A. Shrivastava, "Hnerv: A hybrid neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10270–10279, Jun. 2023.

[14] X. Zhang, R. Yang, D. He, X. Ge, T. Xu, Y. Wang, H. Qin, and J. Zhang, "Boosting neural representations for videos with a conditional decoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2556–2566, Jun. 2024.

[15] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, Oct. 2017.

[16] E. Dupont, A. Goliński, M. Alizadeh, Y. W. Teh, and A. Doucet, "Coin: Compression with implicit neural representations," *arXiv preprint arXiv:2103.03123*, 2021.

[17] E. Dupont, H. Loya, M. Alizadeh, A. Goliński, Y. W. Teh, and A. Doucet, "Coin++: Neural compression across modalities," *arXiv preprint arXiv:2201.12904*, 2022.

[18] T. Ladune, P. Philippe, F. Henry, G. Clare, and T. Leguay, "Cool-chic: Coordinate-based low complexity hierarchical image codec," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13515–13522, Oct. 2023.

[19] H. Kim, M. Bauer, L. Theis, J. R. Schwarz, and E. Dupont, "C3: High-performance and low-complexity neural compression from a single image or video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9347–9358, Jun. 2024.

[20] Z. Li, M. Wang, H. Pi, K. Xu, J. Mei, and Y. Liu, "E-nerv: Expedite neural video representation with disentangled spatial-temporal context," in *European Conference on Computer Vision*, pp. 267–284, Nov. 2022.

[21] Q. Zhao, M. S. Asif, and Z. Ma, "Dnerv: Modeling inherent dynamics via difference neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2031–2040, Jun. 2023.

[22] H. Yan, Z. Ke, X. Zhou, T. Qiu, X. Shi, and D. Jiang, "Ds-nerv: Implicit neural video representation with decomposed static and dynamic codes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23019–23029, Jun. 2024.

[23] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5855–5864, Oct. 2021.

[24] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, "Efficientnerf efficient neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12902–12911, Jun. 2022.

[25] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, and N. Radwan, "Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5480–5490, Jun. 2022.

[26] X. Huang, W. Li, J. Hu, H. Chen, and Y. Wang, "Refsr-nerf: Towards high fidelity and super resolution view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8244–8253, Jun. 2023.

[27] K. Genova, F. Cole, D. Vlasic, A. Sarna, W. T. Freeman, and T. Funkhouser, "Learning shape templates with structured implicit functions," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7154–7164, Oct. 2019.

[28] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5939–5948, Jun. 2019.

[29] L. De Luigi, A. Cardace, R. Spezialetti, P. Z. Ramirez, S. Salti, and L. Di Stefano, "Deep learning on implicit neural representations of shapes," *arXiv preprint arXiv:2302.05438*, 2023.

[30] H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "Nvrc: Neural video representation compression," *Advances in Neural Information Processing Systems*, vol. 37, pp. 132440–132462, Dec. 2024.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Jun. 2016.

[32] T. Hayami and H. Watanabe, "Implicit neural representation for videos based on residual connection," in *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pp. 317–318, Nov. 2024.

[33] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, Jun. 2016.

[34] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, Jun. 2022.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, Dec. 2017.

[36] Q. Zhao, M. S. Asif, and Z. Ma, "Pnerv: Enhancing spatial consistency via pyramidal neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19103–19112, Jun. 2024.

[37] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10318–10327, Jun. 2021.

[38] A. Kayabasi, A. K. Vadathya, G. Balakrishnan, and V. Saragadam, "Bias for action: Video implicit neural representations with bias modulation," in *Proceedings of the Computer Vision and Pattern*

Recognition Conference, pp. 27999–28008, Jun. 2025.

[39] Y. Bai, C. Dong, C. Wang, and C. Yuan, "Ps-nerv: Patch-wise stylized neural representations for videos," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 41–45, Sep. 2023.

[40] S. R. Maiya, S. Girish, M. Ehrlich, H. Wang, K. S. Lee, P. Poirson, P. Wu, C. Wang, and A. Shrivastava, "Nirvana: Neural implicit representations of videos with adaptive networks and autoregressive patch-wise modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14378–14387, Jun. 2023.

[41] H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "Hinerv: Video compression with hierarchical encoding-based neural representation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72692–72704, Dec. 2023.

[42] M. Tarchouli, T. Guionnet, M. Riviere, W. Hamidouche, M. Outtas, and O. Deforges, "Res-nerv: Residual blocks for a practical implicit neural video decoder," in *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 3751–3757, Sep. 2024.

[43] T. Hayami, T. Shindo, S. Akamatsu, and H. Watanabe, "Neural video representation for redundancy reduction and consistency preservation," in *2025 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, Mar. 2025.

[44] J. Kim, J. Lee, and J.-W. Kang, "Snerv: Spectra-preserving neural representation for video," in *European Conference on Computer Vision*, pp. 332–348, Nov. 2024.

[45] T. Hayami, K. Koizumi, and H. Watanabe, "Sr-nerv: Improving embedding efficiency of neural video representation via super-resolution," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, pp. 881–884, Dec. 2025.

[46] T. Roosendaal, "Big buck bunny," in *ACM SIGGRAPH ASIA 2008 computer animation festival*, pp. 62–62, Dec. 2008.

[47] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset," in *2016 IEEE international conference on image processing (ICIP)*, pp. 1509–1513, Aug. 2016.

[48] A. Mercat, M. Viitanen, and J. Vanne, "Uvg dataset: 50/120fps 4k sequences for video codec analysis and development," in *Proceedings of the 11th ACM multimedia systems conference*, pp. 297–302, May 2020.

発表文献（国際学会）

[1] E. Ogawa, **T. Hayami**, and H. Watanabe, "Accurate and efficient surface reconstruction from point clouds via geometry-aware local adaptation," in *IIEEJ International Conference on Image Electronics and Visual Computing (IEVC)*, 2026. (to appear).

[2] **T. Hayami**, K. Koizumi, and H. Watanabe, "Sr-nerv: Improving embedding efficiency of neural video representation via super-resolution," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, pp. 881–884, Dec. 2025. DOI: 10.1109/GCCE65946.2025.11275406.

[3] K. Koizumi, **T. Hayami**, and H. Watanabe, "Semantic reconstruction for unified detection of local and logical anomalies," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, pp. 1081–1084, Dec. 2025. DOI: 10.1109/GCCE65946.2025.11275329.

[4] T. Kunitomi, **T. Hayami**, and H. Watanabe, "Clinically prioritized attention-based fusion of multi-plane knee mri for robust injury detection," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, pp. 485–488, Dec. 2025. DOI: 10.1109/GCCE65946.2025.11274947.

[5] S. Saigo, **T. Hayami**, and H. Watanabe, "Enhancing continuous emotion recognition via visually diverse frame selection," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, pp. 1275–1278, Dec. 2025. DOI: 10.1109/GCCE65946.2025.11274966.

[6] **T. Hayami**, K. Koizumi, and H. Watanabe, "Structure-preserving patch decoding for efficient neural video representation," in *2025 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 216–221, Sep. 2025. DOI: 10.1109/MMSP64401.2025.11324200.

[7] **T. Hayami**, T. Shindo, S. Akamatsu, and H. Watanabe, "Neural video representation for redundancy reduction and consistency preservation," in *2025 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, Mar. 2025. DOI: 10.1109/ICCE63647.2025.10929874.

[8] **T. Hayami** and H. Watanabe, "Implicit neural representation for videos based on residual connection," in *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pp. 317–318, Nov. 2024. DOI: 10.1109/GCCE62371.2024.10760573.

[9] H. Watanabe, L. Jin, **T. Hayami**, T. Chujoh, Y. Yasugi, S. Hong, Z. Fan, and T. Ikai, "The effect of edge information in stable diffusion applied to image coding," in *IIEEJ International Conference on Image Electronics and Visual Computing (IEVC)*, LBP–15, Mar. 2024.

発表文献（国内学会）

[1] **速見泰雅**, 渡辺裕, “構造保持パッチデコードの時間方向拡張による効率的映像表現”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P3–05, Nov. 2025.

[2] 小川英人, **速見泰雅**, 渡辺裕, “局所サイズを適応調整する点群メッシュ化手法”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P3–10, Nov. 2025.

[3] 中村悠大, 高部雅矢, **速見泰雅**, 渡辺裕, “鏡面反射を考慮した 3d convex splatting による高リアリティレンダリング”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P4–03, Nov. 2025.

[4] **速見泰雅**, 進藤嵩紘, 渡辺裕, “空間と時間的一貫性のある動画表現の一検討”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P2–15, Nov. 2024.

[5] 進藤嵩紘, **速見泰雅**, 田中頌子, 渡辺裕, “画素値の動的変化に基づく動画像表現”, 映像情報メディア学会年次大会, 13A–3, Aug. 2024.

[6] **速見泰雅**, 金洛旭, 渡辺裕, “NeRF に基づくフレーム補間手法の品質改善”, 電子情報通信学会総合大会, D–11A–26, Mar. 2024.

[7] 渡辺裕, 金洛旭, **速見泰雅**, 中條健, 八杉将伸, 洪秀俊, 范哲銘, 猪飼知宏, “エッジ・色情報を反映したプロンプトベースの画像符号化”, 電子情報通信学会総合大会, D–11A–27, Mar. 2024.

[8] **速見泰雅**, 金洛旭, 渡辺裕, 中條健, 青野友子, 八杉将伸, 洪秀俊, 范哲銘, 猪飼知宏, “NeRF および特徴マップに基づくフレーム補間手法の特性評価”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P2–05, Nov. 2023.

[9] H. Watanabe, L. Jin, **T. Hayami**, T. Chujoh, T. Aono, Y. Yasugi, S. Hong, Z. Fan, T. Ikai, “Prompt-based image coding with edge information”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P1–12, Nov. 2023.

[10] L. Jin, **T. Hayami**, H. Watanabe, T. Chujoh, T. Aono, Y. Yasugi, S. Hong, Z. Fan, T. Ikai, “Post-processing based image coding via stable diffusion”, 画像符号化シンポジウム (PCSJ) 映像メディア処理シンポジウム (IMPS), P3–08, Nov. 2023.