# 深層学習を用いた動画像圧縮に関する研究開発

Researches on Learned Video Compression

**甲藤　二郎**（Jiro Katto, Ph. D.）
**早稲田大学　理工学術院　教授**
(Professor, Waseda University, Faculty of Science and Engineering)
電子情報通信学会　情報処理学会　映像情報処理学会　IEEE　ACM 他
受賞：電子情報通信学会・業績賞（2026 年）　電気通信普及財団・電気通信普及財団賞（2023 年）　高柳健次郎財団・高柳健次郎業績賞賞（2020 年）
映像情報メディア学会・フェロー（2020 年）　電子情報通信学会フェロー（2015 年）　他
著書：IT Text: 情報通信ネットワーク, オーム社（2015 年）　他
研究専門分野：マルチメディア信号処理　情報ネットワーク

**あらまし**

本研究では、深層学習を用いた動画像圧縮（学習型動画像符号化）に関する研究成果について報告を行う。研究目的と研究背景について説明した後に、初めに、Transformer のアテンション機構を活用し、明示的な動き検出を行わない学習型動画像符号化の報告を行う。次に、Transformer ベースの学習型動画像符号化におけるアテンション機構、特にトークンミキサーの特性解析について報告を行う。最後に、行動認識をタスクとする学習型動画像符号化の機械向け符号化拡張について報告を行う。それぞれの研究成果は、著名な国際学会、並びに査読付き論文誌に採択されている。

## １．研究の目的

本研究は、深層学習を用いた動画像符号化（以下、学習型動画像符号化、LVC: Learned Video Compression）とその要素技術の確立を目標とする。

2017 年頃から深層学習を用いた静止画像符号化（以下、学習型画像符号化、LIC: Learned Image Compression）の研究開発が開始され、JPEG 等の既存の国際標準方式の圧縮効率を凌ぐと共に、2025 年には、学習型画像符号化に基づく世界初の国際標準 JPEG-AI が規格化された。学習型動画像符号化は、2021 年頃から成果発表が急増し、MPEG の最新の国際標準方式である VVC（Versatile Video Coding）を凌ぐ圧縮効率が報告されている。

また、最近の AI 技術の発展を背景に、現在、JPEG・MPEG では、ICM（Image Coding for Machines）、並びに VCM（Video Coding for Machines）に関する議論が進められている。これは、従来の JPEG・MPEG が、人間の視聴による符号化を前提としていたのに対し、AI などの「機械」による認識・解析を前提にした符号化技術を示す（「機械向け符号化」と呼ばれる）。学習型符号化の流れと同様に、静止画像を対象とする ICM の研究開発が先行して進められ、最近は動画像を対象とする VCM の研究開発が活発化している。

以上を背景に、本研究では、学習型動画像符号化の研究開発、並びに、学習型動画像符号化を用いた機械向け符号化に関する研究開発を進めた。

## ２．研究の背景

学習型動画像符号化（LVC）では、動画像の符号化・復号処理をニューラルネットワークで行う。従来の HEVC や VVC 等の MPEG 規格では、人間の技術者が符号化の手順を細かく設計してきたのに対し、LVC では大量の動画像データを使い、深層学習モデルに「どう圧縮すれば効率が良いか」を学習させる。静止画像を対象とする学習型画像符号化（LIC）では既に既存の国際標準方式を上回る圧縮効率を示しているが、動画像では異なるフレーム間の時間方向の情報も扱う必要があり、技術的な難易度が高くなり、静止画像に比べると進化がやや緩やかになるとされている[1]。

LVC を代表する研究例として、MSRA（Microsoft Research Asia）の DCVC（Deep Contextual Video Compression）シリーズが挙げられる[2]。DCVC は、過去の動画像符号化とは異なり、明示的なフレーム差分を取ることなく、「前フレームをコンテキストと考え、これを手がかりに現フレームを予測して符号化する」（Conditional Coding）という考え方を一貫し、継続的に改良が重ねられている（図 1 参照）。DCVC の最初の発表は NeurIPS 2021 で行われ、以降、DCVC-HEM（MM 2022）で VVC の圧縮効率を凌ぎ、DCVC-DC（CVPR 2023）では時間・空間両方向のコンテキストの

# 深層学習を用いた動画像圧縮に関する研究開発

Researches on Learned Video Compression

統合により次世代標準規格 ECM をも上回り、DCVC-FM (CVPR 2024)では誤差累積を回避するための特徴量の周期的リフレッシュ機構を導入した。DCVC-RT (CVPR 2025)では、明示的な動き検出を行わない（暗黙的動きモデルによる）設計によって、100fps 超という LVC の実時間実装を実現し、さらに DCVC-UF (CVPR 2026)では、複数フレームをまとめて 1 つの潜在表現に符号化・並列復号するなど、さらなる高速化を図っている。ただし、DCVC シリーズは一貫して CNN (Convolutional Neural Networks) のみを使用し、Transformer 等の新しい深層学習モデルは使用していない。

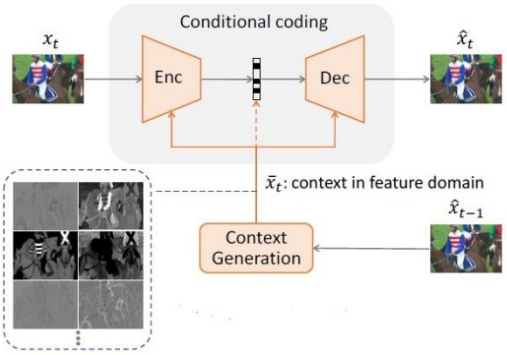


図 1　残差符号化を行わない DCVC シリーズ [2]

また、2022 年には、Google から、Transformer ベースの LVC である VCT (Video Compression Transformer) が発表されている。初期の DCVC 等の CNN ベース手法は、動き推定等の事前構造に依存していたが、VCT はこれらを排除し、各フレームを潜在表現へマッピングした上で、過去フレームの潜在表現を用いたセルフアテンションおよびクロスアテンションにより、現フレームの潜在表現の確率分布を予測させる（図 2 参照）。これは、動き情報を明示的に扱わなくても、アテンション機構によって、複雑な動きパターンをデータから学習できることを示している。また、Disney Research らによる ST-XCT では、チャネル方向の共分散に基づくアテンション(cross-covariance attention)を計算し、トークンを細かく分割せずにグローバルな時空間相関を低計算量で捉える設計となっている。最近はまた、計算量の軽いスライディングウィンドウ型のアテンションを組み込んだり、CNN と組み合わせる手法なども提案されている。

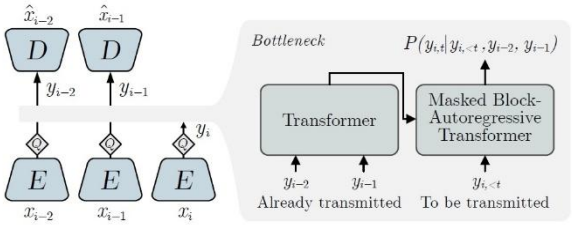


図 2　暗黙的動きコンテキストによる VCT [3]

このように、LVC 研究は大きく CNN ベースの DCVC 系と Transformer・暗黙モデリングベースの VCT 系に分類され、両者の長所を組み合わせるハイブリッド設計も、重要な研究の方向性となっている。

また、機械向け符号化 (ICM/VCM) は、ETH による学習型画像符号化の「レート・識別率特性」の最適化[5]をトリガにして、JPEG、MPEG 共に 2019 年に検討が開始され、これまでの「人間が見てきれいに見える」ことを目的とした（人間向け）符号化とは異なり、「AI が物体検出や認識などのタスクを正確にこなせる」ことを目的とした符号化を目指すものである。監視カメラ映像の解析や自動運転など、人間が直接見るよりも、AI による処理映像の増加を想定している技術である。図 3 は、JPEG-AI で紹介されている ICM の構成例を示す[6]。受信側で、人間向けの復号に加え、各種の画像処理とコンピュータビジョンタスクを実行する構成を示している。

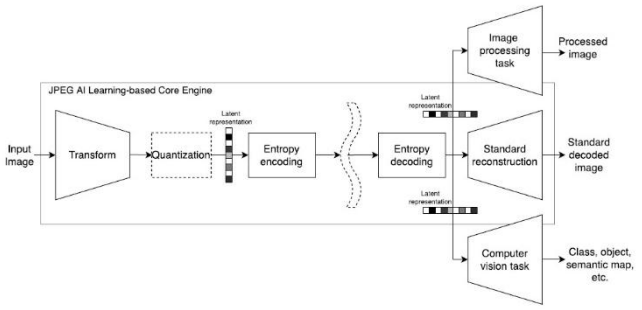


図 3　ICM/VCM の基本構成 [6]

VCM では、LVC のような学習型の技術が望ましいとする考え方がある。従来の MPEG 規格をそのまま使う方式は、元々、人間の視覚向けに設計された符号

# 深層学習を用いた動画像圧縮に関する研究開発

Researches on Learned Video Compression

器を流用しているため、特に高圧縮時に AI のタスク精度が大幅に劣化する。これに対して、LVC ベースの手法であれば、「AI タスクの精度最適化」を損失関数に組み込み、符号化モデル全体をエンドエンドに最適化できる。実際、MPEG-VCM のフィーチャー符号化トラックの評価基準に対し、画像から抽出した特徴量(AI が認識に使う中間特徴量)をニューラルネットワークでエンドツーエンドに圧縮する学習型手法が提案されている[7]。

ただし、現状の LVC は概して計算量が大きく、リアルタイム処理や省電力ハードウェアへの実装が難しいという指摘がある。また、GPU 間の演算精度(TF32/FP16 等)の非決定性に起因して、エンコーダとデコーダの計算結果が微妙にずれ、算術符号化が同期しなくなるという「クロスプラットフォーム問題」も知られており、異なる世代・異なるベンダーの GPU 間で一貫した復号を保証することも実用化に向けた検討課題となっている。

## ３．研究の方法

### (1) 特徴空間アテンションによる学習型動画像符号化[8]

LVC は、圧縮処理を特徴量領域へ移すことで大きく性能を伸ばしてきた。多くの手法は動き推定・動き補償と CNN ベースのコンテキスト抽出を組み合わせるが、動きへの依存と CNN ベースのコンテキストモデルへの依存が、汎化性能とグローバルな知覚能力を制限してきた課題がある。

そこで本提案では、これらの課題に対処するため、Transformer ベースのフレームワークでフレーム全体を明示的に知覚し、限定的な動き情報に依存しない特徴空間アテンション（FLA: Feature-level Attention）を提案している。FLA の核心的な仕組みは、高レベルのローカルパッチ埋め込みを 1 次元のバッチ単位ベクトルに変換し、従来のアテンション重みをグローバルなコンテキスト行列に置き換え、グローバルな知覚を実現している点にある。さらに、埋め込み投影前に局所特徴を保持するための Dense overlapping Patcher (DP) の導入も行っている。

複数フレームの入力に対し、無圧縮フレームを特徴量空間にエンコード・デコードする画像エンコーダ/デコーダを用意し、エントロピー符号化には過去 2 フレームから学習したコンテキストを利用する。DP は特徴マップをベクトル系列へ変換しエントロピーモデルに渡す役割を担う。

図 4 に示す提案モデル(FLAVC: Feature Level Attention Video Compression）では、UVG データセット上で HM-16.26 をベースラインとした BD-rate 評価により、計算コストと性能のトレードオフを満足できることを示した。また、FLA は、CNN ベース手法の知覚限界を超え、かつ、VCT で用いられる標準的なアテンションよりも高い冗長性削減効果を実現できることを示した。

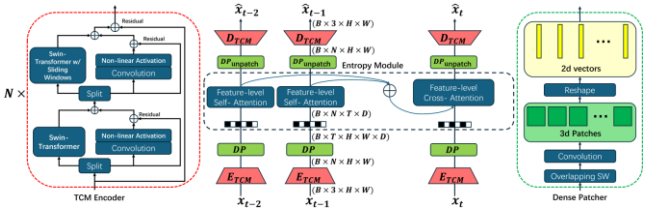


図 4　FLAVC の構成（潜在変数領域における
アテンションの活用）

### (2) Transformer ベースの学習型動画像符号化のためのトークンミキサーの特性解析［9］

LVC はコンテキストモデリングの高度化により性能を伸ばしてきたが、その代償としてフレームワークの設計・計算コストが複雑化し、どの要素が性能向上に寄与しているかが見えにくくなっている。特に前述の VCT [3] は、アテンション機構でフレーム間依存性を暗黙的に捉える一方で、アテンション計算の演算量の増加がボトルネックになっている。

そこで本提案では、VCT をテストベッド基盤として使用し、エントロピーモデルの核心部分である「トークンミキサー」(トークン間で特徴を伝播させる機構)の寄与を定量評価するために、モジュール化された ATEM (Abstracted Transformer Entropy Model) の提案を行った（図 5 参照)。そして、空間コンテキストと時間コンテキストのモデリングを担う部分に、Pooling-Mixer、MLP-Mixer、CNN-Mixer、Attn-Conv/Conv-Attn のハイブリッド（CNN とアテンションの組み合わせ)、Swin-Mixer（スライディン

グウィンドウアテンション）の、計 6 種類のトークンミキサーを差し替えた系統的な特性評価を行った。また時間方向の結合方式も、系列方向に結合する VCT 方式から、チャネル方向に結合する Channel-Concatenated Temporal Mixer に変更し、アテンション行列サイズを 128×128 から 64×64 に削減する工夫も導入した。

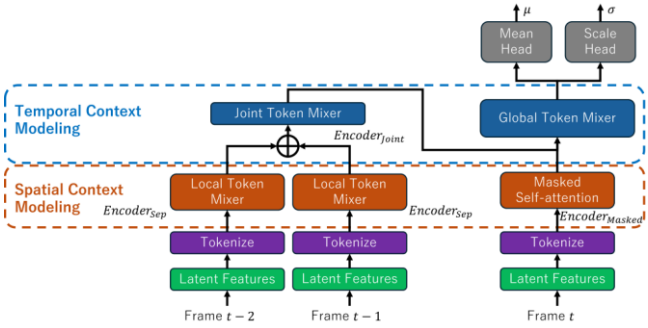


図 5　空間コンテキストと時間コンテキストのモデリングを行う ATEM の構成

複数の動画像データセットを用いた実験結果として、以下の知見を得た。

① パラメータフリーの Pooling-Mixer は下限の性能を示し、暗黙的な空間コンテキストモデリングが不可欠であることを確認した。

② 画像分類タスクで有効な MLP-Mixer は、圧縮タスクでは性能を悪化させることを確認した。

③ CNN-Mixer はバニラのアテンションと同等の性能を達成し、アテンション自体よりもマクロな Transformer 構造が性能に寄与するという MetaFormer 仮説を確認した。

④ CNN とアテンションの組み合せは最も高い圧縮性能を実現でき、特に Conv→Attn の順序が、Attn→Conv の順序よりも優れることを確認した。

⑤ Swin-Mixer は、計算量と圧縮性能のバランスを両立し、ベースラインの VCT に対して、演算量を抑えつつ、良好な圧縮性能を実現できることを確認した。

⑥ チャネル方向の時間結合は、Swin-Mixer のような局所的なトークンミキサーにとって必須な構造であり、系列方向結合に戻すと演算量も圧縮性能も劣化することを確認した。

以上の評価により、図 6 に示すように、画像分類に有効な MLP-mixer が圧縮用途では有効ではないこと、演算量と圧縮性能のバランスを考えると、Swin-Mixer がベストな選択と考えられること等、今後の LVC の符号器設計に有効な知見を明らかにした。

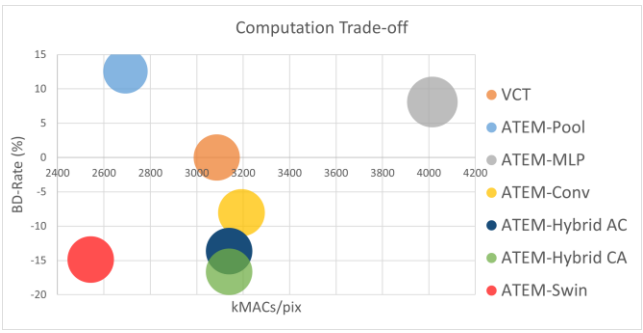


図 6　演算量と圧縮性能のトレードオフ
（左側ほど演算量が少なく、下側ほど高い圧縮性能）

### (3) 時空間潜在変数分割による機械向け動画像符号化 [10]

従来の VCM は、物体検出などの空間タスクの検討は数多く行われているが、行動認識のような時間依存タスクの検討は十分に行われていなかった。また、VCM で頻繁に行われる、復号画像に戻してからの特徴抽出（圧縮→復号→特徴抽出）は、演算の無駄を生んでいる。

そこで、本提案では、図 7 に示すような、圧縮領域で完結する（復号画像に戻さない）VCM フレームワークを提案している。LVC としては DCVC-DC を使用し、そのコンテキストデコーダを改造することで、潜在変数領域における空間コンテキストと時間コンテキストを活用する VCM 方式の提案を行っている。

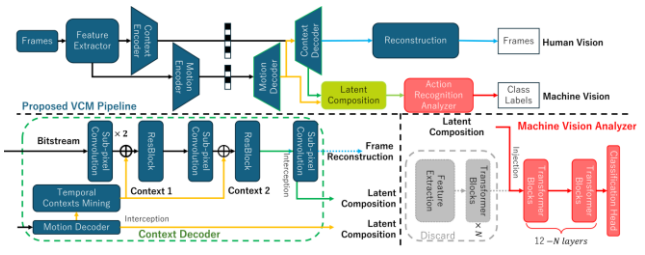


図 7　提案する VCM 方式の構成（圧縮領域における行動認識）

本提案では、これを STDLC（Spatial-Temporal Decoupled Latent Composition Module）と呼び、空

間パスは Swin-Transformer ブロック、時間パスは Temporal-Swin（T-Swin）ブロックによる処理を行い、最後に合成ブロックで統合する構成を取っている。行動認識器には既存の UniFormerV2 を採用し、その特徴抽出層はバイパスして STDLC の出力を直接 Transformer バックボーンへ注入する構成として、復号画像に戻す冗長な計算コストの削減を実現している。

その上で、2 つの行動認識データセットを用いた評価実験を行い、復号画像に戻す従来手法に対して、同じビットレートで高い行動認識率を実現できることを確認した（図 8 参照）。これは特に低ビットレート域で顕著であり、UCF101 データセットを用いた実験では、非圧縮映像を用いたベースラインからの精度低下はわずか 2%に留まっている。さらに、空間パスの処理は CNN より Swin Transformer が優れていること、時間パスを取り除くと精度が 22%低下すること、なども明らかにしている。

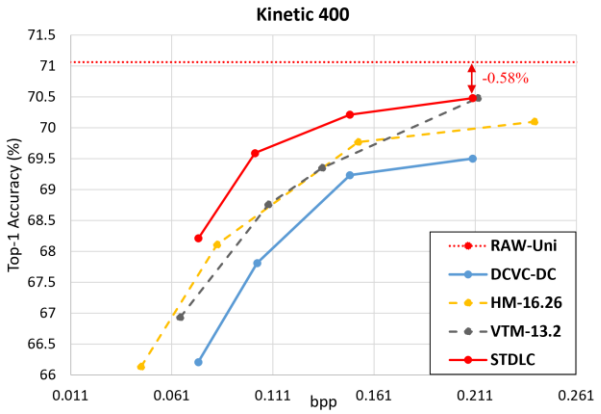


図 8　各種 VCM におけるレート･認識率の比較（RAW は無圧縮画像の認識率、STDLC は提案方式）

### (4) その他の研究成果

上記以外の研究成果として、マルチモーダル大規模言語モデル（M-LLM）を活用した学習型画像符号化[11]、状態空間モデル（Mamba）を活用した学習型画像符号化[12]、深層学習を用いた超解像モデルのネットワーク圧縮[13]に関する成果発表も行った。

## 4．将来展望

本研究では、学習型動画像符号化とその機械向け符号化拡張について検討を行った。JPEG では、2025 年に学習型画像符号化に基づく JPEG-AI の標準化が完了した。一方、MPEG では、Beyond VVC として従来の信号処理技術の延長線に位置付けられる ECM (Enhanced Compression Model) や、ニューラルネットワークを取り込んだ NNVC (Neural Network Video Coding) の検討が進められている。NNVC は深層学習を動画像符号化に導入する技術であるが、現状は従来の信号処理手法とのハイブリッド構成に重点が置かれている。MPEG とは異なる標準化団体 MPAI においても、AI ベースの動画像符号化（すなわち LVC）の標準化が模索されているものの、現状は MPEG ほどのアクティビティには至っていない。

このため、LVC の研究開発はアカデミックな国際学会を中心に進められており、とりわけ MSRA の DCVC シリーズが顕著な研究成果として知られている。興味深いことに、数年前までは処理負荷の重さが課題とされていた LVC であるが、最新の深層学習技術の知見を取り入れることで、実時間実装の観点において VVC や ECM を凌駕しつつある。従って、次世代の動画像符号化標準（仮称：H.267）は ECM をベースに策定されると推測されるが、その次の世代、あるいはそれと並行する形で、LVC が国際標準の主役に躍り出る日もそう遠くはないと考えられる。

## おわりに

本研究では、深層学習を用いた動画像圧縮（学習型動画像符号化）に関する研究開発を行い、明示的な動き検出を行わない学習型動画像符号化の独自方式の提案、Transformer ベースの学習型動画像符号化におけるアテンション機構の特性解析、学習型動画像符号化の機械向け符号化拡張の独自方式の提案、について報告を行った。申請者の当初の想定以上に研究開発の進展速度が速く、今後も独自方式の提案を継続的に行うと共に、実用化や国際標準化に向けた活動にも積極的に貢献していく予定である。最後に、本助成は、学習型画像符号化に関する研究助成申請がなかなか採択されない時期に採択して頂き、深く感謝致します。

## 用語解説

*1 JPEG（Joint Photographic Experts Group）： ディ

ジタル静止画像の高能率符号化方式を定めるISO/IECの国際標準化団体

*2 MPEG（Motion Picture Experts Group）：ディジタル動画像の高能率符号化方式を定めるISO/IECの国際標準化団体

*3 LIC/LVC（Learned Image/Video Compression）：符号器全体をエンドエンドの深層学習によって構成する次世代の画像・動画像圧縮技術

*4 ICM/VCM（Image/Video Coding for Machines）：自動運転や監視カメラなどのAI処理を前提に、機械の認識精度を最適化する画像・動画像圧縮技術

*5 CNN（Convolutional Neural Network）：画像の局所的な特徴（エッジや模様など）を効率的に抽出できるニューラルネットワーク

*6 Transformer：自然言語処理から発展した、データの長距離の依存関係を高精度で捉えられるニューラルネットワーク

*7 アテンション（Attention）：処理中のデータに対して、どの情報に重点的に注目すべきかを動的に重み付けるアルゴリズム

*8 トークン（Token）：AIがテキストや画像を処理する際に、細かく分割した処理の最小単位（単語の断片や画素のブロック等）。

*9 ECM（Enhanced Compression Model）：Beyond VVCとして、基本的には従来の信号処理ベースの延長線で圧縮効率を高める次世代圧縮技術

*10 NNVC（Neural Network Video Coding）：深層学習を動画像符号化に本格導入することで圧縮効率を高める次世代圧縮技術

*11 MPAI（Moving Picture, Audio and Data Coding by Artificial Intelligence）：2020年に設立された、AIベースの符号化標準を策定する標準化団体

**参考文献**

[1] X.Sheng et al.: “VNVC: A Versatile Neural Video Coding Framework for Efficient Human-Machine Vision, IEEE TPAMI. Vol.46, No.7, July 2024.

[2] DCVC Family: https://github.com/microsoft/dcvc.

[3] F. Mentzer et al.: "VCT: A Video Compression Transformer," NeurIPS 2022, Nov. 2022.

[4] Z. Chen et al.: "Neural Video Compression with Spatio-Temporal Cross-Covariance Transformers," ACM MM 2023, Oct. 2023.

[5] E. Agustsson et al.: “Generative Adversarial Networks for Extreme Learned Image Compression,” IEEE ICCV 2019, Oct.2019.

[6] J.Ascenso et al.: “The JPEG AI Standard: Providing Efficient Human and Machine Visual Data Consumption,” IEEE Multimedia, May 2023.

[7] Y. Kim et al.: "End-to-End Learnable Multi-Scale Feature Compression for VCM," IEEE TCSVT, Vol.34, No.5, May 2024.

**発表文献**

[8] Chun Zhang, Heming Sun, Jiro Katto: “FLAVC: Learned Video Compression with Feature Level Attention,” IEEE/CVF CVPR 2025, June 2025.

[9] Chun Zhang, Heming Sun, Jiro Katto: “Benchmarking Token Mixers for Efficient Transformer-Based Learned Video Compression,” APSIPA Trans. on Signal and Information Processing (accepted).

[10] Chun Zhang, Heming Sun, Jiro Katto: “STDLC: Video Coding for Machines with Spatial-Temporal Decoupled Latent Composition,” IEEE ICASSP 2026, May 2026.

[11] Shimon Murai, Heming Sun, Jiro Katto: “LMM-driven Semantic Image-Text Coding for Ultra Low-bitrate Learned Image Compression,” IEEE VCIP 2024, Dec.2024.

[12] 村井史門・孫　鶴鳴・甲藤二郎：“状態空間モデルを用いた深層学習画像圧縮，” 電子情報通信学会・画像工学研究会，Feb.2025.

[13] 遠藤祐宇音・甲藤二郎：“Deep Compressionによる超解像モデルの圧縮と評価，” 電子情報通信学会・画像工学研究会，Mar.2025.

**この研究は、令和５年度ＳＣＡＴ研究助成の対象として採用され、令和６年度～令和７年度に実施されたものです。**