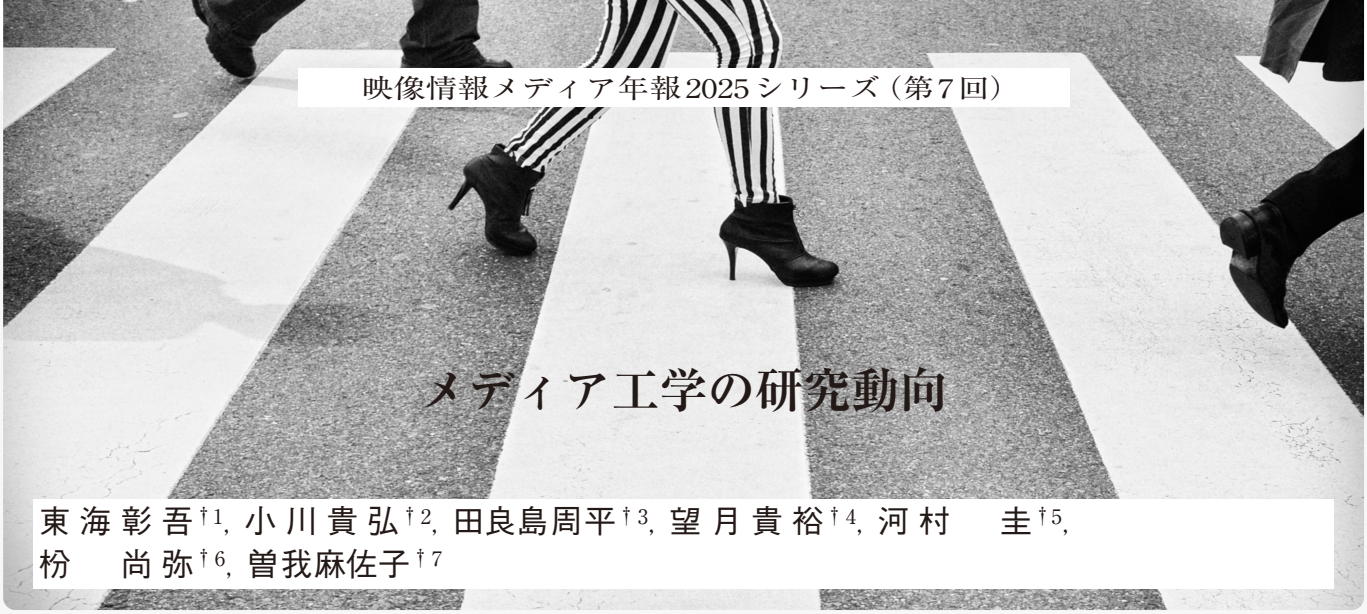

映像情報メディア年報2025シリーズ（第7回）

メディア工学の研究動向

東海彰吾[†1]，小川貴弘[†2]，田良島周平[†3]，望月貴裕[†4]，河村　圭[†5]，
枌　尚弥[†6]，曽我麻佐子[†7]

1．まえがき

メディア工学は，映像メディアの生成，処理，検索等を主な対象とする分野である．他の技術分野と相互に関連する映像メディアの研究領域であり，非常に広いアスペクトを持った研究分野である．近年は，深層学習（Deep Learning: DL）や大規模言語モデル（Large Language Model: LLM）などのAI技術を利用した映像メディア処理の研究が非常に多く行われていて，映像コンテンツ生成技術，映像コンテンツの中身に踏み込んだ要約技術や検索技術，学習のためのデータセット構築に関する技術，さらに利用者と空間的にインタラクションする3次元的メディア技術等に展開している．

そこで，本稿ではメディア工学研究委員会の委員を中心とした著者らにより，メディア工学分野の最新動向について紹介する．第2章では，映像をはじめとする大規模データセットの情報圧縮に関わるデータセット蒸留の動向について述べる（小川）．第3章では，マルチモーダルなタスクに対応するためのAIエージェント技術の動向について述べる（田良島）．第4章では，日々提供される膨大な映像メディア情報をコンパクトな表現メディアに変換して利用しやすくする，映像自動要約技術について紹介する（望月）．第5章では，デバイスの開発も進む3次元動画像の取り扱いのための動的メッシュデータ技術について紹介する（河村）．第6章では，言語情報を利用した画像ハンドリング技術として注目される，視覚言語モデルを用いた画像検索について紹介する（枌）．さらに，第7章では，3次元的インタラクションを含む3Dコンテンツの構成や提示技術の動向を紹介する（曽我）．最後に，第8章でメディア工学の今後の展開について述べる（東海）．

2．データセット蒸留

データセット蒸留（Dataset Distillation: DD）とは，大規模データセットに内在する情報を，ごく少量の合成サンプルに圧縮し，特定の学習アルゴリズムまたはモデルに対して，元のデータセットで学習した場合と同等の汎化性能を実現することを目的とする技術である[1)2)]．この手法は，学習効率の向上およびデータの保存や配布に伴う計算資源の削減を目指すものであり，これまでの研究は，最適化ベース手法から生成モデルベース手法，さらに分布マッチングベース手法へと段階的に発展してきた．実運用においては，生成データの再利用性，異なるネットワークアーキテクチャへの適用可能性，高解像度画像および多クラス分類への拡張性が特に重要であり，これらの要件が各手法の設計方針に直接的な影響を与えている．

初期の研究では，合成サンプルを直接最適化する「最適化ベース」手法が主流であり，これらは大きく「勾配マッチング」と「パラメータマッチング」に分類される．前者は，実データによる学習で得られる勾配と，合成データによる学習で得られる勾配を一致させることで，蒸留過程を最適化問題として定式化する[3)]．後者は，合成データで学習した際のモデルパラメータが，実データで学習した場合のパラメータに近づくように更新を行うものであり，あらかじめ取得された専門家モデルの学習軌跡（Expert Trajectory）を利用することで性能を向上させる[4)]．これらの手法は小規模データセットでは有効であったが，二段階最適化に伴う計算負荷の増大，初期値およびモデル構造への高い依存性，さらに高解像度データや多クラス環境へのスケーラビリティの制約など，実用上の課題が指摘されている[2)]．加えて，クラスあたりの画像数や入力解像度の増加に伴い，合成データ更新の計算コストが急激に増大し，再学習の際に同等の計算資源を必要とするため，再現性および再利用性の確保が困難となる．これらの制約を克服する試みとし

†1 福井大学 学術研究院 工学系部門
†2 北海道大学 大学院情報科学研究院
†3 NTTドコモビジネス株式会社/東京都立大学
†4 NHK 放送技術研究所
†5 株式会社KDDI総合研究所
†6 NEC ビジュアルインテリジェンス研究所
†7 龍谷大学 先端理工学部

"Recent Researches on Media Engineering" by Shogo Tokai (University of Fukui, Fukui), Takahiro Ogawa (Faculty of Information Science and Technology, Hokkaido University, Sapporo), Shuhei Tarashima (NTT DOCOMO Business Inc., Tokyo/Tokyo Metropolitan University, Tokyo), Takahiro Mochizuki (Science & Technology Research Laboratories, NHK, Tokyo), Kei Kawamura (KDDI Research, Inc., Saitama), Naoya Sogi (NEC Corporation, Kawasaki), and Asako Soga (Ryukoku University, Otsu)

て，生成モデルを活用した手法が提案されている．

生成モデルベースのDDは，合成サンプルそのものを最適化して情報を圧縮する代わりに，事前学習済み生成モデルの潜在空間や階層的表現を最適化領域として利用するものである．代表的な例であるGLADは，生成器の潜在空間内で中間表現を少数最適化することにより，ピクセル空間での直接最適化に比べて，異なるモデル間での汎化性能およびロバスト性を向上させた[5]．その後の研究では，代表的な生成モデルである敵対的生成ネットワーク（Generative Adversarial Network: GAN）の条件付き拡張である条件付きGANを導入して大域的構造と局所的詳細の両立を図る手法[6]，ロジットマッチングと自己蒸留を統合して分布マッチング性を高める枠組み[7]，および生成器層ごとの階層的パラメータ化により最適化領域を拡張する設計[8]などが提案されている．これらは初期手法で顕在化した計算負荷や依存性の問題に対する有効な解決策である．しかしながら，生成モデルの事前分布とタスク分布との不一致や，高解像度条件下における多様性の維持といった課題が依然として残されており，これらに対して拡散モデルの導入が注目されている．

現在最も成功している生成モデルである拡散モデルを用いた手法は，分布の忠実性と多様性を同時に確保しつつ，モデル非依存な再利用性を向上させることに成功している．D4Mでは，潜在拡散を利用して実画像と合成画像間の空間的一貫性を保持し，クラスプロトタイプを導入することで再利用性の高い合成データセットを生成している[9]．さらに，拡散過程にミニマックス基準を導入して，代表性と多様性を理論的に制御する手法も提案されており，限られた計算資源下でも高い性能を示している[10]．近年では，多様性低下を防ぐ自己適応メモリーの導入[11]，下流タスクの難易度分布に基づくタスク指向サンプリング[12]，およびプロトタイプ情報とコンテキスト情報を統合的に強化する情報誘導型サンプリング[13]などが報告されている．これらの研究では，クラスあたりの画像数が低い条件下でのモード崩壊の抑制，高解像度画像におけるテクスチャ多様性の維持，および異なるモデルへの転用時における再蒸留の不要化などが主要な評価指標となっており，生成モデルベース手法の実応用上の有効性を高めている．

一方，従来の最適化ベース手法や生成モデルベース手法では，合成データを更新するたびに，まず合成データのみでモデルを一度学習し，そのモデルが実データで学習された場合の挙動にどれだけ近いかを評価し，その結果に基づいて合成データを更新する，といった多段階の入れ子構造を取る必要があり，計算コストが高かった．これに対して分布マッチング（Distribution Matching: DM）に基づく手法は，実データと合成データの分布間の差異を一つの目的関数として直接最小化することで，この入れ子構造を単一段階に統合し，計算効率を高めようとするアプローチである．例えばNCFMは，特性関数に基づく一般化指標を提案し，ニューラルネットワークを介して周波数空間での分布差への感度を最適化することで，高解像度条件下でも安定した距離計量を実現している[14]．また，最適輸送理論に基づくWMDDは，事前学習済み分類器が抽出した特徴のワッサースタイン重心に合成データをマッチングさせ，クラス内統計を保持することで単純かつ高性能な手法を実現している[15]．さらに，階層的および木構造的関係を表現するために，HDDは特徴空間に双曲幾何を導入し，ローレンツモデルにおける測地距離に基づくマッチングを行うことで，分布の幾何的特性を明示的に表現しつつ，安定性と情報凝縮性を両立させている[16]．これらの研究は，学習時間の短縮やメモリー効率の向上に加え，合成データの評価・更新プロセスを単純化し，生成モデルベース手法を理論的に補完するものである．

総じて，データセット蒸留は「最適化ベース手法」，「生成モデルベース手法」，「分布マッチングベース手法」という三つの相補的方向に発展してきた．初期手法が抱える計算コストおよびスケーラビリティの課題に対し，生成モデルベース手法は強力な事前分布と誘導的サンプリングにより対応し，分布マッチング手法は理論的に整備された距離計量および幾何表現を通じて効率性と性能を両立させている．今後は，異なるモデル間での汎用性のさらなる向上，高解像度・多クラス・多モダリティ環境下における評価プロトコルの標準化，データ拡張および初期化条件の統一的定義，さらに合成データセットの配布に関するライセンスおよびプライバシー上の課題の明確化が求められる[2]．これらの課題に対する体系的な取り組みは，学術的知見と産業応用を接続する上で重要な基盤となり，データセット蒸留研究のさらなる発展に寄与することが期待される．（小川）

3．マルチモーダルAIエージェント技術の動向

複雑かつ多様なマルチモーダルタスク（VQA[29)30)50)53)71)]，グラウンディング[17)27)]，数学的推論[52)55)]等）を解決可能な技術が注目されている．大規模モデルのエンドツーエンド学習はプリミティブなタスクを解く性能を飛躍的に向上させる一方，タスク毎にデータ収集を伴うファインチューニングが必要であり，判断根拠の説明可能性も低い．そこで近年では，LLMの高いタスク分割/プランニング/推論能力[31)～33)]を活用し，画像映像認識/生成器を含む外部ツールを自動連係させてタスクを解く手法が数多く提案されている．本章ではこのようなマルチモーダルAIエージェント技術について，その登場（3.1節），発展（3.2節），展望（3.3節）を俯瞰する．

3.1　マルチモーダルAIエージェント技術の登場

2022年後半から2023年にかけて，言語のみを入力とするLLMへのプロンプティングの工夫でマルチモーダルAIエージェントを実現する技術が複数提案された．例えば

VISPROG[39]は，画像処理/認識の諸機能をPythonのクラスとして定義しておき，タスクとそれを解くためのクラス群の使用例をプロンプトとして与える（In-Context Learning: ICL）ことで，GPT-3＊1が未知のタスクを解くためのクラス呼び出し手順を出力可能であることを示した．ViperGPT[41]は，Codex＊2に画像映像処理/認識のAPIをシステムプロンプトとして与えると，タスクとそれを解くPythonコード例に基づくICLによりタスクを解くためのPythonコードが得られることを示した．MM-ReAct[36]は，LLMに「思考→ツール呼出→観察」の反復でタスクを解決させるReActプロンプト[42]をマルチモーダルデータに拡張させた．Visual ChatGPT[37]はChatGPTと視覚基盤モデルとを連携させるためのプロンプトマネージャを設計し，対話形式による複雑な画像生成/編集/変換を実現した．HuggingGPT[44]では，ChatGPTをコントローラとして使用しHugging Face＊3上のAIモデルを自動的に選択・結合しタスクを解くフレームワークが提案された．Chameleon[43]はGPT4[35]等のプランニング能力を活かすことで，画像理解に加えウェブ検索・数学演算・表理解等のより広範なモジュール群を組み合わせたタスク解決を実現した．IdealGPT[40]では，推論結果の確信度に応じて追加の質問生成とその回答を行う自己訂正型のシステムが提案された．

3.2 マルチモーダルAIエージェント技術の発展

本節では，以下六つの観点でより最近の技術発展について述べる．

(1) 効率化

VPD[47]では，LLMが外部ツールを用いて得た正解情報を視覚言語モデルへ蒸留することで，蒸留モデルの単発推論でも複雑なタスクが解決可能なことを示し，ツール連携に伴う低いスループットやシステムの脆弱性を改善した．LLMs as Programmers[56]は，LLMがタスクの解からIn-Context事例を自動生成できることを示し，従来手法における事例作成コストの削減に成功した．

(2) 視覚情報を利用したChain-of-Thought（CoT）

CCoT[49]は画像入力にも対応するLLM[34)38)]からシーングラフ[18]を生成し，それをLLMに再入力する二段推論によって画像中の物体間の関係性理解性能を向上させた．Visual Sketchpad[54]では画像への線や矩形の描画を伴うCoTを導入し，数学の図形問題等のタスクで有効性を示した．Image-of-Thought Prompting[45]では，サブタスクの推論結果と解釈を画像とテキスト双方で提示することで説明可能性を向上させた．DIEM[48]では画像中の注目領域の特定機構の導入，VAT[59]では画像の線画化やスケッチ化によって，背景等推論に寄与しない画像情報の影響抑制に成功した．

(3) チューニングの導入

CLOVA[46]は，人間のフィードバックから誤推論の原因となった箇所を特定しそのプロンプトをチューニングすることで推論精度を高めた．Olympus[62]やMMAT-1 M[65]はより多くの画像映像ツールと連携するため，LLMをファインチューニングするためのデータセットを構築した．DWIM[66]ではツール仕様と出力のずれを手掛かりとすることで，人間のフィードバックやデータセットの構築をせずにLLMをファインチューニングする方法が提案されている．CATP-LLM[70]はツールのトークン化とLLMの強化学習によって，ツールのコスト（処理時間，メモリー消費量，利用料等）を考慮しつつ，複雑な順序を伴うプランニングを実現した．VisTA[58]は，良質なツール選択を能動的に探索する枠組みを用いてLLMをファインチューニングし，学習データ分布から外れたタスクに対しても推論性能を向上させた．

(4) 動的制御

HYDRA[51]では強化学習ベースのコントローラを導入し，タスクの推論状態に応じてLLMが提示するプランを改善/選択する手法が提案された．NAVER[64]では決定的有限オートマトンと確率付き論理プログラミング言語の導入により，グラウンディングタスクで中間推論結果の自己修正が可能なシステムを実現している．

(5) ツールの「生成」

PyVision[60]では，GPT-4.1＊4やClaude-4.0-Sonnet＊5を用いることで，タスクを解くために必要なツールのPythonコードが（事前のAPI登録を必要とせず）動的に生成可能なことが示された．VADAR[61]では，3D空間認識タスクを前提として，タスクを解くAPIの設計と実装とを担う二つのエージェントが協調してツールを生成する手法がベンチマークとともに提案された．

(6) タスクの拡がりと深化

VIoTGPT[57]では監視映像解析のシナリオを想定し，ツールに加え映像中の解析対象フレームをLLMが選択し処理を効率化する方法が提案されている．TANGO[63]ではプロンプティングベースのマルチモーダルAIエージェント[39)41)]を拡張しEmbodied AIに適用，シミュレータ内でのナビゲーション等のタスクで高い性能が得られることを示した．Preacher[67]では，論文を入力としてその映像を生成するというより実用志向のタスクに対しマルチモーダルAIエージェントでアプローチする方法が提案されている．

3.3 マルチモーダルAIエージェント技術の展望

3.2節の各観点で，今後更なる発展が見込まれる．複数の観点を同時に改善させる方法も登場するだろう．既存研究は画像映像入力を前提とするが，今後は点群等も入力対象とな

＊1 https://github.com/openai/gpt-3
＊2 https://github.com/openai/codex
＊3 https://huggingface.co/
＊4 https://openai.com/index/gpt-4-1/
＊5 https://www.anthropic.com/news/claude-4

るのではないか．ドメインや特定のシナリオに特化した手法も，データセットと対になって今後さらに増えると予想する．Physical AIやSpatial AIとの融合にも期待したい．（田良島）

4．映像自動要約技術

放送やインターネットを通して，膨大な数の映像コンテンツにアクセス可能となり，映像内容を短い時間で把握するための技術が求められる時代となった．そのような技術のひとつが，映像から重要と思われるシーンを自動で抽出して短尺動画を生成する「映像自動要約技術」であり，さまざまな手法やアプリケーションの開発が進められている．

4.1　基礎研究の動向

これまで主流であった，視覚的な特徴のみに基づく手法の高度化に加えて，マルチモーダル特徴やLLMを利用するアプローチが数多く提案されている．

Chavesらは，フレーム画像系列のグラフ表現に基づく映像要約技術VideoSAGEを提案した[72]．グラフニューラルネットワーク（Graph Neural Network: GNN）の初期状態として，各フレーム画像をノードとし，時刻の差が閾値以下のフレーム同士をエッジで結合する．その際，無向グラフだけでなく，時刻が後のノードのみを結んだ順方向グラフと，時刻が前のノードのみを結んだ逆方向グラフも構築する．ノードを要約動画に使われるものと使われないものに分類するクラスタリング問題としてGNNを学習し，使われると判断されたノードのフレーム画像から要約動画を生成する．

GuoらがCFSumは，映像と音声，およびクエリとなるテキストを入力としたマルチモーダル自動要約技術であり，クエリの内容に適した要約動画を生成することができる[73]．まず，事前学習済オートエンコーダを用いて，各モダリティの入力データを特徴ベクトルに変換する．次に，それらの特徴ベクトルをTransformerに入力して各モダリティの情報を融合し，出力を分解してモダリティ毎の新たな特徴ベクトルを生成する．最後に，映像と音声の各特徴ベクトルにクエリの特徴ベクトルを加えるためのTransformerを通して各クリップの重要度を算出し，要約動画を生成する．

Leeらは，マルチモーダル大規模言語モデル（M-LLM）とLLMを組み合わせた映像要約手法を考案した[74]．まず，映像の各フレームを学習済M-LLMを用いてキャプションに変換する．次に，数フレーム分のキャプションを，インストラクション，キャプションに対する重要度の例，および「中央フレームの重要度を算出してください」というクエリとともにLLMに入力し，インストラクション，重要度の例，クエリ，および回答の埋め込みベクトルを取得する．最後に，それらの埋め込みベクトルをSelf-Attention BlockやMLPを通してフレーム毎の重要度に変換し，重要度の高いフレームを抽出して要約動画を生成する．

4.2　映像制作現場における実用化の動向

スポーツ映像は，試技の成功や得点といった「要約動画で使うべき重要シーン」が明確であるため，さまざまなスポーツコンテンツの制作現場において自動要約技術が実用化されている．

IBM社は，テニスの試合のハイライト動画とその解説を自動で生成する技術を開発し，2023年のウィンブルドン選手権でサービス化を実現した[75]．過去の試合映像から大量に収集した得点シーンを学習したAIを用いて，観客の歓声，選手の動き，選手の点数などを解析し，試合における重要シーンを自動抽出してハイライト動画を生成する．さらに，基本用語，得点にかかわるキーワード，スポーツ記事などを学習させた生成AIを用いて，ハイライト動画に解説を自動付与する．高度なテクニックを含む“マニアックな”シーンにうまく解説を付けられないなどの課題はあるが，選手やボールの動きをより細かく解析することにより，性能向上が期待できる．

WSC Sports社は，ユニフォーム認識技術，スタッツ情報，過去の大量のスポーツ映像の重要シーンを学習したAIなどを用いたハイライト動画自動生成技術を開発し，Jリーグ，セリエA，Bリーグ，2023年のWBCなど，多岐にわたるスポーツシーンにサービス展開している[76)～80)]．WOWOW社は，WSC Sports社の技術を用いて，UEFAチャンピオンズリーグやATPテニスなどのハイライト動画配信を迅速化し，動画視聴数の倍増を達成した[81]．

パリ五輪においても，AIによるハイライト自動生成技術が活躍した．オリンピック放送サービス（OBS）はIntel社と連携し，過去の膨大なオリンピック競技映像を利用して学習した競技毎のモデルを用いて，特定のチーム・選手・アクションなどに関連したハイライト動画を自動生成する技術を開発した[82]．この技術はオリンピックの14競技を対象に実用化され，各競技のハイライト動画を作成するコストの大幅な削減を実現した．

スポーツ以外の分野では，編集スタッフの着眼点や視聴者の嗜好の違いなどにより「重要シーン」を定めるのが難しい．しかし，メタデータの活用やジャンルに特化したAIの開発などにより，試験運用や実用に至った技術が登場している．

博報堂DYメディアパートナーズとCBCテレビは，AIを活用して生成した縦長サイズのドラマ要約動画を配信するサービスの実証実験を実施した[85]．まず，出演者の発話やテロップなどのメタデータを自然言語処理して各シーンの重要度を判定し，指定の長さの要約動画を生成する．次に，物体検出AIを用いた被写体のクロッピング処理などにより，動画を1：1にリサイズする．最後に1：1の動画の両サイドを機械的に削除して，縦長サイズ（9：16）の動画を生成する．この技術を利用して，実際に放送されたドラマの要約動画を生成し，番組の公式サイトやSNSの番組公式

アカウントで配信した．

テレビ業界特化のベンチャー企業であるNAXA社は，AIを活用した自動動画編集サービス「Short Video Generator」の提供を開始した[86]．音声解析で番組の「盛り上がりポイント」を検出し，指定した長さの要約動画を短時間で生成できる．簡単な操作でのIN/OUT点調整や主要SNSへのワンクリック投稿などの機能も搭載されており，動画の編集時間を大幅に削減できるツールとして，放送局やメディア企業で試験導入が進んでいる．

NHKは，ニュース番組を対象とした映像自動要約システムの実用化を実現した[83) 84)]．放送用ニュースの重要シーンならではの“画作り”を学習したAIと音声認識技術の統合利用により，番組制作者が手動で編集したものに近い品質の要約動画を自動で生成できる．さらに，利用者の嗜好とこだわりにもきめ細かく対応するために，自動生成された動画を簡単な操作で修正できる機能を実装した．このシステムを利用して，これまでに1,000本を超える要約動画が制作され，SNS等で配信されている．（望月）

5．動的メッシュデータ

近年，メタバース技術の進展に伴い，世界各地でボリュメトリックスタジオが開設され，実写映像から高密度の3Dデータを生成するキャプチャ技術が急速に普及している．これにより，人や物体の動きを高精度に記録・再現するニーズが高まり，3Dデータの効率的な圧縮・伝送技術が求められている．動的メッシュは時間的に変化する頂点や形状を直接扱うため，より自然でリアルな動きを再現できる．特に，リアルな体験をメタバース空間で提供するための基盤技術として，動的メッシュ符号化技術が不可欠である．

従来，3D表現には動的点群やTボーン型メッシュが用いられてきた．動的点群は高密度かつ写実的な表現が可能であり，MPEGではV-PCC（Video-based Point Cloud Compression）として標準化が完了している．詳細については本会の会誌記事を参照されたい[87]．しかしながら，GPUによるレンダリング支援が限定的であり，対象を拡大した際に点同士のすき間が生じないような高度なレンダリングが必要などの課題が顕在化している．

一方，従来のTボーン型メッシュは骨格構造に基づきキャラクタの動きを制御する方式で，主にアニメーションやゲームで使われている．特に人物や衣服の細かな変形を忠実に表現できるため，メタバースやボリュメトリック映像に適している．

動的メッシュの構成要素は，各フレームの頂点座標，ポリゴン構造を定義するトポロジー，テクスチャ座標などの頂点属性，時間軸に関するメタデータ，そして平面に展開したテクスチャ情報である．符号化方式には，トポロジーが時間的に変化しない場合に用いるインタメッシュ符号化，トポロジーが変化する場合に用いるイントラメッシュ符号化がある．前者はトポロジーを再利用して頂点座標のみを時間方向に予測符号化し，後者はトポロジーと頂点座標の両方を符号化する．

現在，MPEGでは動的メッシュ符号化の標準化が検討されている．このうち，イントラメッシュ符号化はMPEGEB（Edge Breaker）として[89]，インタメッシュ符号化と映像符号化によるテクスチャ符号化を組合せたフレームワークはV-DMC（Video-based Dynamic Mesh Coding）として検討されている．詳細は学会誌の記事を参照されたい[88]．

なお，V-DMCはデコード方式のみが標準化されており，符号化効率を高める手法は利用者に任されている．そのため，例えばボリュメトリックスタジオで取得した高密度なメッシュから，トポロジーに時間方向の一貫性を抽出し，高品質な動的メッシュに変換・生成する手法が提案されている[90) 91)]．このような領域でのメディア工学研究のさらなる発展が期待されている．（河村）

6．視覚言語モデルと画像検索

CLIP[114]などの視覚言語モデルの登場により，文入力による画像検索（以降，画像検索）が大きく進展している．本章では，画像検索の概要と最新の研究動向を概説する．

6.1　画像検索の概要

画像検索は，画像群から入力文に関連する画像を発見するタスクであり，入力文と各画像の関連度の推定によって実現される．関連度の推定法は，① 文特徴と画像特徴の類似度を用いる埋め込み型検索と，② 文と画像が関連しているか否かの2値分類スコアを用いる分類型検索の二つに大別される．埋め込み型検索では，画像特徴は事前計算が可能なため，検索を高速に実行できる．一方，分類型検索では，文と画像を相互作用させスコアを推論するため，検索精度は高い反面，対象画像数分だけ視覚言語モデルの推論が必要となるため検索速度が遅くなる．精度と速度の両立のために，埋め込み型検索で絞り込んだ画像を分類型検索で再評価する手法も多く利用されている．以下では，分類型検索の概要を述べた後に，埋め込み型検索の最新の研究動向を概説する．

6.2　分類型検索

分類型検索では，文から得たトークンと画像から得たトークンをTransformer[137]へ入力し，2値分類スコアを推論する手法が多い[93) 95) 97) 99) ～102) 145)]．これらは，画像トークンの抽出法で3カテゴリーに分類できる．① 物体領域特徴に基づく手法では，Faster R-CNN[135]などで検出した物体領域をそれぞれCNNへ入力し得た特徴を一つのトークンとする[93) 95) 97) 99) 102)]．物体間・単語間の関係性を陽に考慮できる一方で，検知できるクラスが限定的，計算コストが高い，といった課題がある．そこで，② グリッド特徴に基づくPixelBERT[101]では，画像全体をResNet[142]へ入力

し得た特徴マップをトークンとする．さらなる高速化のために，③パッチ特徴に基づくViLT[100]では，スライディングウィンドウで得た画像パッチに対し線形変換を行いトークンとする．

6.3 埋め込み型検索

埋め込み型検索では，識別向けCNNが出力する特徴量を基に計量学習を行う手法が初期に提案されていた[92)113)143)]．近年はCLIP[114]をベースに，学習方法やモデル構造等の観点で改善が続けられている．

(1) 対照学習の改善

視覚言語モデルの学習には，InfoNCE損失[136]を用いた対照学習が広く利用されている．FLIP[117]やEVA-CLIP[118]では，学習時間の短縮と精度向上のために，画像の一部をマスクし対照学習を行う．COSMOS[121]では，画像の顕著性が低い部分へも注意を向けるために，画像と文の一部をマスクし対照学習を行う．InfoNCEよりも計算効率のよいsigmoid損失[94)111)]がある．softmax正規化が必要なInfoNCEと異なり，sigmoid損失はバッチ内のすべてのデータを参照する必要がないため，分散学習においても効率的に計算できる[94]．上記は画像と文章が1対1に紐づくデータを用いた学習であるが，1対多/多対多で紐づくデータへの拡張も提案されている[107)108)129)131)]．

(2) マルチタスク学習の活用

対照学習と同時に複数のタスクを学習することで検索精度を改善する研究も多くある．CoCa[105]では，特徴量を入力とする画像キャプショニングモデルも学習する．その他にも，Masked Image/Language Modeling (MIM/MLM)[96)106)116)]，分類型検索[96)103)104)116)]，単一モダリティ内での対照学習[115]を利用する手法がある．検索精度の高いBLIP-2[104]では，対照学習と画像キャプショニング，分類型検索を同時に学習している．

(3) 大規模データの活用

埋め込みモデルの学習には，画像と文ペアの集合 $\{I_i, T_i\}_{i=1}^{N}$ が利用される．画像I_iと文T_jは，その添え字が同じ場合 $(i=j)$ は関連しており，異なる場合 $(i \neq j)$ は関連していない．学習データは低品質であっても大規模であることが重要である[116]．ここで低品質であるとは，文の曖昧性が高いものや，同じ添え字だが画像と文が関連していないデータ，異なる添え字だが画像と文が関連しているデータを含むことなどを指す．

このような低品質なデータを自動的に高品質化し学習する手法がある．BLIP[103]では，事前に用意したモデルで画像キャプショニングと関連度推定を行い，曖昧性の低い文章を生成したうえで関連度が高そうなデータのみを選択し学習に利用する．ALIP[144]は，画像－文ペアの品質を重みとしてInfoNCE損失に導入することで，関連していない可能性の高い画像-文ペアの影響を抑制する．FFF[134]は，バッチ内に含まれるデータから自動的に関連する画像－文ペアを発見し学習に利用する．このように文章の洗練とペア付けの洗練によって，検索精度の改善が続けられている．

(4) 推論方法とモデルの改善

CLIPでは，画像と文それぞれ独立に特徴抽出を行う．画像特徴抽出時に別データを参照することで，より良い特徴量を得る手法がある[112)133)]．FLAIR[133]では，画像特徴抽出の最後に入力文を参照し精度改善を実現している[133]．一般に，画像や文は一つの特徴量で表現されるが，これでは画像の多義性を充分に表現できない．そこで，確率分布[109)110)148)149)]やベクトル集合[104)129)132)146)150)]を用いた表現方法も提案されている．例えば，BLIP-2[104]では画像を32本のベクトルで表現している．

また，生成特化のLLMや視覚言語モデルを検索に活用する研究もある．LLMは，文エンコーダへの応用が検討されている[124)130)]．LLM2CLIP[124]では，文のみを用いた対照学習などをLLMに施した後にCLIPの画像エンコーダと併せて対照学習を行い，LLMを文エンコーダへ変換する．生成向け視覚言語モデルは，画像エンコーダとLLMから構成されることが多い．LLMに事後学習を施しLLMを画像と文の埋め込みモデルへ変換する手法がある[127)128)]．VladVA[128]では，CLIPと同様の対照学習をLLMに施すことで生成向け視覚言語モデルを埋め込みモデルへ変換する．

6.4 まとめ

本章では，視覚言語モデルを用いた画像検索の最新の研究動向について主に埋め込み型検索の観点から概説した．自由文による画像検索精度は高まっているものの，実用性を高めていくには長文への対応[120)122)130)132)]や合成性の欠如[138)141)]などの課題への対応が必要である．本章では触れられなかった生成型画像検索[139)140)]も研究が進んでおり，今後の発展が期待される．（枌）

7．3Dコンテンツ構成・提示技術

7.1 3Dデータの計測・提示機材の動向

3Dデータの計測技術として，アニメやVTuberなど，人体アニメーションのコンテンツ作成に使われるようになったモーションキャプチャシステムと，近年一般的に普及してきたVR/AR装置の動向を紹介する．

モーションキャプチャシステムは人体動作を計測する装置として，2000年代に反射マーカを赤外線で撮影する光学式システム（Vicon，Motion Analysisなど）が一般的となったが，価格は数千万円するものが多く，手軽に購入できるものではなかった．その後，2010年代には深度センサと骨格推定技術により，人体の姿勢がリアルタイムに推定できるkinectが登場し，スマートフォンで撮影した画像や映像からリアルタイムで関節の位置を高精度に推定できるライブラリー（OpenPose，MediaPipeなど）も普及した．しばらくはマーカやセンサを使用することで高精度に計測できるものと，精度は低いがマーカレスで手軽に計測できもの

の二極化が進んでいたが，近年ではAIで骨格推定を行うマーカレスAI方式がモーションキャプチャ装置にも組み込まれ，従来の光学式システムは，赤外線カメラに加え，各社AIカメラを搭載したマーカレスのシステムも併用できるようになっている．今後，計測後に編集を行うという作業は一般的には減っていくものと思われる．また，2024年に登場したmocopiは，小型軽量のセンサ6個で全身動作の計測が可能である．精度的にはkinectやOpenPoseと同等で大まかな動きしか計測できないが，安価で小型軽量なフルトラッキング機材として活用されていくと思われる．

3Dデータの提示機器として一般化してきたVR機器は，主にアウトサイドイン方式とインサイドアウト方式がある．アウトサイドイン方式はカメラなどの外部センサを利用してHMDやコントローラの位置や動きを検出する．インサイドアウト方式は，HMD（Head Mounted Display）などの計測するもの自体にセンサが内蔵されているものである．具体例としては，HTC VIVEのようにBaseStationを設置して空間を構成するものがアウトサイドイン方式，ヘッドセッドのみでハンドキャプチャができるMeta Quest 3はインサイドアウト方式である．近年の動向としては，外部センサなしで手軽に計測できるインサイドアウト方式が増えてきており，HTC社からはインサイドアウト式のトラッカも登場している．また，3DCGコンテンツが提示できるARデバイスとしては，2016年頃からHololensやMagicLeapなどが登場したが，VR機器に比べると高価で重いという印象であった．近年では，Meta Quest 3のパススルー機能でAR表示が可能になったことやXrealなどの小型軽量ウェアラブルデバイスが登場したことで，ARコンテンツも普及していくと思われる．

7.2　コンテンツ応用

CG，XR，インタラクティブ技術を用いたコンテンツ応用として，2024年11月にロンドンで行われたDigital Body Festival[151]で展示・上演された作品を紹介する．このイベントは，ディジタル化した人体に関する映像作品，パフォーマンス，VRコンテンツ，インタラクティブ展示などを集めたものである．主催者が振付家であり，特にダンスに関するコンテンツが多く，研究者ではなく一般人やアーティストがVRやインタラクティブ技術に触れることができるユニークなイベントであった．

インタラクティブデモでは，Azure Kinectセンサを用いたものがほとんどであった．Vast Body[152]はカメラの前で動くと，各姿勢に合わせてリアルタイムに類似姿勢の写真を検索して提示する．入力の動きに合わせて滑らかに動きながらも，異なる身体とアイデンティティの間を絶えず揺らめきながら投影する（図1（左））．VRギャラリーでは六つのVRデモを展示しており，そのうち最大三つの体験が可能であった．Body of Mine[153]は別の性別の身体に入り込み，トランスジェンダーの人々の物語に触れることができる．Dazzle:Solo[154]は海軍迷彩のリズムと色彩を取り入れ，ファッションと振付を融合させたインスタレーションであり，実際の衣装も一緒に展示されていた（図1（右））．Soul Paint[155]は，仮想空間のアバターの表面と内部をコントローラで3Dドローイングし，自身の動きと連動して見ることができる．他の人が作った創作物や録音したメッセージの再生も可能である．

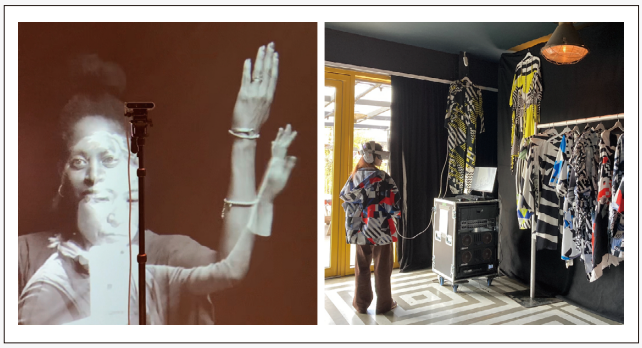

図1 Digital Body Festivalでの体験の様子
（左：VAST BODY，右：VRギャラリーよりDazzle:Solo）

AI技術を活用した作品として，Colette Sadler氏によるARK 1[156]というパフォーマンスは，未来が舞台のSFの物語であり，AIで人類の歴史を再構築していく様子がプロジェクション映像と1人のダンサーにより演じられていた．カタコトの単語とロボットのような断片的な人体動作が徐々に再構成され，流暢な発話と動きになっていく様子をダンサーが演じていた．

Superradiance[157]は，脳が他者の動きを映し出す「身体性シミュレーション」という認知現象を活用したもので，人体の動きが森や海，生物などの映像にシームレスに変化していくものである．画像処理および彩色によるものと，シミュレーションおよび生成AIを活用したものの2部構成であり，AI技術をアートに活用した作品として完成度が高かった．

全体的な考察として，技術を単純に適用するだけでなく，他の要素と融合させてそのコンテンツに合った調整が必要であると感じた．今後は，使い方次第で新しい表現手法やさまざまな社会実装を見いだすことが期待される．（曽我）

8．メディア工学技術の展開

今後のメディア工学技術については，以下の3点の展開が主に考えられる．

まず，改めて言うまでもないが，AI技術としてのDLやLLM技術の利活用が今後も進むことが予想される．LLMやVLM技術によって，従来別々に取り扱われた画像・映像と文字情報が渾然一体となって利用される形態がさらに一般的になり，メディア工学のカバー範囲が映像メディアからマルチモーダル，マルチメディアに広がることが考えられる．その展開においては，個々のアプリケーションに

特化した技術の発展も期待される．

上記にも関連するが，OpenAI社のSORAに代表されるプロンプトから画像・映像を生成する技術，NeRFに代表される実撮影像から任意視点を生成する画像・映像の生成技術，さらに，RGB-Dカメラや，全方位画像カメラ等，安価で高性能な撮影デバイスの提供が今後も続くことが予想される．メディア工学として，計測デバイスの進化を背景として動的な3次元状況の獲得，生成，処理の技術については，今後も継続的に研究が進むと考えられる．

さらに，その提示先であり利用形態として，3次元的な映像メディアを利用した，さまざまなアプリケーションを実現する映像メディア処理への期待は，今後ますます増大すると考えられる．大量のデータを適切に変換して保持したり，圧縮したりする技術の他，シーンや映像と単一ユーザの間のインタラクションだけでなく，そのようなシーンやメディアを介して複数のユーザ間の相互インタラクション等が今後も継続的に取り扱われると考えられる．（東海）

9．むすび

本稿では，メディア工学分野について，いくつかの切り口でその技術動向を紹介した．AI技術の発展の中で映像メディアは欠かせない処理対象であり，AI技術の発展と並走しながら今後のメディア工学分野の研究も進んで行くと考えられる．一方で，SNSで拡散される偽情報や誤情報は映像メディアを伴うことも多く，メディア工学として安全・安心な社会実現に向けた課題を意識しつつ，技術動向に注目したい．

（2025年11月10日受付）

〔文　献〕

1) T. Wang, J. Zhu, A. Torralba and A. Efros: "Dataset Distillation", ArXiv:1811.10959 (2018)
2) G. Li, B. Zhao and T. Wang: "Awesome Dataset Distillation", (2022), https://github.com/Guang000/AwesomeDataset-Distillation
3) B. Zhao and H. Bilen: "Dataset Condensation with Gradient Matching", ICLR (2021)
4) G. Cazenavette, T. Wang, A. Torralba, A. Efros and J. Zhu: "Dataset Distillation by Matching Training Trajectories", CVPR, pp.10718-10727 (2022)
5) G. Cazenavette, T. Wang, A. Torralba, A. Efros and J. Zhu: "Generalizing Dataset Distillation via Deep Generative Prior", CVPR, pp.3739-3748 (2023)
6) L. Li, G. Li, R. Togo, K. Maeda, T. Ogawa and M. Haseyama: "Generative Dataset Distillation: Balancing Global Structure and Local Details", CVPR Workshops, pp.7664-7671 (2024)
7) L. Li, G. Li, R. Togo, K. Maeda, T. Ogawa and M. Haseyama: "Generative Dataset Distillation Based on Self-knowledge Distillation", ICASSP, pp.1-5 (2025)
8) X. Zhong, H. Fang, B. Chen, X. Gu, T. Dai, M. Qiu and S. Xia: "Hierarchical Features Matter: A Deep Exploration of GAN Priors for Improved Dataset Distillation", CVPR (2025)
9) D. Su, J. Hou, W. Gao, Y. Tian and B. Tang: "D4M: Dataset Distillation via Disentangled Diffusion Model", CVPR. pp.5809-5818 (2024)
10) J. Gu, S. Vahidian, V. Kungurtsev, H. Wang, W. Jiang, Y. You and Y. Chen: "Efficient dataset distillation via minimax diffusion", CVPR, pp.15793-15803 (2024)
11) M. Li, G. Li, J. Mao, T. Ogawa and M. Haseyama: "Diversity-Driven Generative Dataset Distillation Based on Diffusion Model with Self-Adaptive Memory", ICIP, pp.415-420 (2020)
12) M. Li, G. Li, J. Mao, L. Ye, T. Ogawa and M. Haseyama: "Task-Specific Generative Dataset Distillation with Difficulty-Guided Sampling", ICCV Workshops (2025)
13) L. Ye, S. Hamidi, G. Li, T. Ogawa, M. Haseyama and K. Plataniotis: "Information-Guided Diffusion Sampling for Dataset Distillation", NeurIPS Workshops (2025)
14) S. Wang, Y. Yang, Z. Liu, C. Sun, X. Hu, C. He and L. Zhang: "Dataset Distillation with Neural Characteristic Function: A Minmax Perspective", CVPR (2025)
15) H. Liu, Y. Li, T. Xing, P. Wang, V. Dalal, L. Li, J. He and H. Wang: "Dataset Distillation via the Wasserstein Metric", ICCV (2025)
16) W. Li, G. Li, K. Maeda, T. Ogawa and M. Haseyama: "Hyperbolic Dataset Distillation", NeurIPS (2025)
17) S. Kazemzadeh, V. Ordonez, M. Matten and T. Berg: "ReferItGame: Referring to Objects in Photographs of Natural Scenes", EMNLP (2014)
18) J. Johnson, R. Krishna, M. Stark, L. Li, D. Shamma, M. Bernstein and L. Fei-Fei: "Image Retrieval using Scene Graphs", CVPR (2015)
19) J. Andreas, M. Rohrbach, T. Darrell and D. Klein: "Neural Module Networks", CVPR (2016)
20) R. Hu, J. Andreas, M. Rohrbach, T. Darrell and K. Saenko: "Learning to Reason: End-To-End Module Networks for Visual Question Answering", ICCV (2017)
21) J. Johnson, B. Hariharan, L. Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick and R. Girshick: "Inferring and Executing Programs for Visual Reasoning", ICCV (2017)
22) R. Hu, J. Andreas, T. Darrell and K. Saenko: "Explainable Neural Computation via Stack Neural Module Networks", ECCV (2018)
23) N. Xie, F. Lai, D. Doran and A. Kadav: "Visual Entailment: A Novel Task for Fine-grained Image Understanding", ArXiv:1901.06706 (2019)
24) D. Hudson and C. Manning: "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering", CVPR (2019)
25) K. Marino, M. Rastegari, A. Farhadi and R. Mottaghi: "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge", CVPR (2019)
26) R. Zellers, Y. Bisk, A. Farhadi and Y. Choi: "From Recognition to Cognition: Visual Commonsense Reasoning", CVPR (2019)
27) A. Akula, S. Gella, Y. Al-Onaizan, S. Zhu, Reddy, J. Chai, N. Schluter and J. Tetreault: "Words Aren't Enough, Their Order Matters: on the Robustness of Grounding Visual Referring Expressions", ACL (2020)
28) P. Lu, R. Gong, S. Jiang, L. Qiu, S. Huang, X. Liang and S. Zhu: "Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning", ACL (2021)
29) A. Masry, D. Long, J. Tan, S. Joty and E. Hoque: "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning", ACL Findings (2022)
30) D. Schwenk, A. Khandelwal, C. Clark, K. Marino and R. Mottaghi: "A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge", ECCV (2022)
31) W. Huang, P. Abbeel, D. Pathak and I. Mordatch: "Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents", ICML (2022)
32) T. Kojima, S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa: "Large Language Models are Zero-Shot Reasoners", NeurIPS (2022)
33) J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou: "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", NeurIPS (2022)
34) H. Liu, C. Li, Y. Li and Y. Lee: "Improved Baselines with Visual Instruction Tuning", ArXiv Preprint (2023)
35) OpenAI: "GPT-4 Technical Report", ArXiv:2303.08774 (2023)
36) Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng and L. Wang: "MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action", ArXiv:2303.11381 (2023)
37) C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang and N. Duan: "Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation

Models", ArXiv:2303.04671 (2023)
38) Z. Yang, L. Li, K. Lin, J. Wang, C. Lin, Z. Liu and L. Wang: "The Dawn of LMMs: Preliminary Explorations with GPT-4V (ision) ", ArXiv:2309.17421 (2023)
39) T. Gupta and A. Kembhavi: "Visual Programming: Compositional Visual Reasoning Without Training", CVPR (2023)
40) H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. Ayyubi, K. Chang and S. Chang: "IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models", EMNLP Findings (2023)
41) D. Surís, S. Menon and C. Vondrick: "ViperGPT: Visual Inference via Python Execution for Reasoning", ICCV (2023)
42) S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan and Y. Cao: "ReAct: Synergizing Reasoning and Acting in Language Models", ICLR (2023)
43) P. Lu, B. Peng, H. Cheng, M. Galley, K. Chang, Y. Wu, S. Zhu and J. Gao: "Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models", NeurIPS (2023)
44) Y. Shen, K. Song, X. Tan, D. Li, W. Lu and Y. Zhuang: "HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face", NeurIPS (2023)
45) Q. Zhou, R. Zhou, Z. Hu, P. Lu, S. Gao and Y. Zhang: "Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models", ArXiv Preprint Arxiv:2405.13872 (2024)
46) Z. Gao, Y. Du, X. Zhang, X. Ma, W. Han, S. Zhu and Q. Li: "CLOVA: A Closed-Loop Visual Assistant with Tool Usage and Update", CVPR (2024)
47) Y. Hu, O. Stretcu, C. Lu, K. Viswanathan, K. Hata, E. Luo, R. Krishna and A. Fuxman: "Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models", CVPR (2024)
48) X. Jiang, G. Wang, J. Guo, J. Li, W. Zhang, R. Lu and S. Tang: "DIEM: Decomposition-Integration Enhancing Multimodal Insights", CVPR (2024)
49) C. Mitra, B. Huang, T. Darrell and R. Herzig: "Compositional Chain of Thought Prompting for Large Multimodal Models", CVPR (2024)
50) X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su and W. Chen: "MMMU: A Massive Multidiscipline Multimodal Understanding and Reasoning Benchmark for Expert AGI", CVPR (2024)
51) F. Ke, Z. Cai, S. Jahangard, W. Wang, P. Haghighi and H. Rezatofighi: "HYDRA: A Hyper Agent for Dynamic Compositional Visual Reasoning", ECCV (2024)
52) P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K. Chang, M. Galley and J. Gao: "MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts", ICLR (2024)
53) L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin and Others: "Are We on the Right Way for Evaluating Large Vision-Language Models?", NeurIPS (2024)
54) Y. Hu, W. Shi, X. Fu, D. Roth, M. Ostendorf, L. Zettlemoyer, N. Smith and R. Krishna: "Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models", NeurIPS (2024)
55) K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan and H. Li: "Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset", NeurIPS (2024)
56) A. Stanić, S. Caelles and M. Tschannen: "Towards Truly Zero-shot Compositional Visual Reasoning with LLMs as Programmers", TMLR (2024)
57) Y. Zhong, M. Qi, R. Wang, Y. Qiu, Y. Zhang and H. Ma: "VIoTGPT: Learning to Schedule Vision Tools towards Intelligent Video Internet of Things", AAAI (2025)
58) Z. Huang, Y. Ji, A. Rajan, Z. Cai, W. Xiao, H. Wang, J. Hu and Y. Lee: "VisualToolAgent (VisTA) : A Reinforcement Learning Framework for Visual Tool Selection", ArXiv:2505.20289 (2025)
59) D. Liu, Z. Wang, M. Ruan, F. Luo, C. Chen, P. Li and Y. Liu: "Visual Abstract Thinking Empowers Multimodal Reasoning", ArXiv:2505.20164 (2025)
60) S. Zhao, H. Zhang, S. Lin, M. Li, Q. Wu, K. Zhang and C. Wei: "PyVision: Agentic Vision with Dynamic Tooling", Arxiv:2507.07998 (2025)
61) D. Marsili, R. Agrawal, Y. Yue and G. Gkioxari: "Visual Agentic AI for Spatial Reasoning with a Dynamic API", CVPR (2025)
62) Y. Lin, Y. Li, D. Chen, W. Xu, R. Clark and P. Torr: "Olympus: A Universal Task Router for Computer Vision Tasks", CVPR (2025)
63) F. Ziliotto, T. Campari, L. Serafini and L. Ballan: "TANGO: Training-free Embodied AI Agents for Open-world Tasks", CVPR (2025)
64) Z. Cai, F. Ke, S. Jahangard, M. Banda, R. Haffari, P. Stuckey and H. Rezatofighi: "NAVER: A Neuro-Symbolic Compositional Automaton for Visual Grounding with Explicit Logic Reasoning", ICCV (2025)
65) T. Gao, Y. Fu, W. Wu, H. Yue, S. Liu and G. Zhang: "MMAT1M: A Large Reasoning Dataset for Multimodal Agent Tuning", ICCV (2025)
66) F. Ke, V. G, X. Leng, Z. Cai, Z. Khan, W. Wang, P. Haghighi, H. Rezatofighi and M. Chandraker: "DWIM: Towards Tool-aware Visual Reasoning via Discrepancy-aware Workflow Generation and Instruct-Masking Tuning", ICCV (2025)
67) J. Liu, L. Yang, H. Luo, F. Wang, H. Li and M. Wang: "Preacher: Paper-to-Video Agentic System", ICCV (2025)
68) Y. Qin, L. Kang, X. Song, Z. Yin, X. Liu, X. Liu, R. Zhang and L. Bai: "RoboFactory: Exploring Embodied Agent Collaboration with Compositional Constraints", ICCV (2025)
69) T. Wei, Y. Yang, J. Xing, Y. Shi, Z. Lu and D. Ye: "GTR: Guided Thought Reinforcement Prevents Thought Collapse in RL-based VLM Agent Training", ICCV (2025)
70) D. Wu, J. Wang, Y. Meng, Y. Zhang, L. Sun and Z. Wang: "CATP-LLM: Empowering Large Language Models for Cost-Aware Tool Planning", ICCV (2025)
71) S. Yin, T. Lei and Y. Liu: "ToolVQA: A Dataset for Multi-step Reasoning VQA with External Tools", ICCV (2025)
72) J-M.R. Chaves, S. Tripathi: "VideoSAGE: Video Summarization with Graph Representation Learning", CVPR, pp.2527-2534. (2024)
73) Y. Guo, J. Xing, X. Hou, S. Xin, J. Jiang, D. Terzopoulos, C. Jiang, Y. Liu: "CFSum: A Transformer-Based Multi-Modal Video Summarization Framework with Coarse-Fine Fusion", ICASSP. (2025)
74) M-J. Lee, D. Gong, M. Cho: "Video Summarization with Large Language Models", DOI: 10.48550/arXiv.2504.11199 (Accepted to CVPR2025) (2025)
75) “テニスのウィンブルドン選手権にAI解説者登場その実力は”, https://www3.nhk.or.jp/news/special/international_news_navi/articles/feature/2023/07/13/33001.html
76) “スポーツのダイジェスト映像を簡単に生成するWSC Sports”, https://k-tai.watch.impress.co.jp/docs/event/dvd2020/1234075.html
77) “イマジカJリーグのハイライト映像，AIで迅速に”, https://www.nikkei.com/article/DGXMZO50796080Z01C19A0000000/
78) “セリエA，AIの自動化ツール導入でハイライトをよりリアルタイムに”, https://www.all-stars.jp/news/seriea-highlight-ai-wscsports/
79) “りくなグループB.LEAGUE 2024-25ニュース用ハイライト映像のご案内”, https://www.bleague.jp/files/user/media/pdf/highlight_video_guide.pdf
80) “2023ワールド・ベースボール・クラシックのハイライト動画を日本中のファンに届ける”, https://prtimes.jp/main/html/rd/p/000000002.000123916.html
81) “AI技術革新でWOWOWのスポーツハイライト動画視聴者数増加を促進”, https://prtimes.jp/main/html/rd/p/000000005.000123916.html
82) "Live from Paris 2024: How OBS is using AI to speed up the creation of highlights", https://www.svgeurope.org/blog/headlines/live-from-paris-2024-how-obs-is-using-ai-tospeed-up-the-creation-of-highlights/
83) 望月，河合，藤森，吉澤，遠藤，浅見：“ニュース要約映像作成支援システムの試作”，映情学誌，77，2, pp.262-271（2023）
84) “ニュース映像自動要約システムの実用化”，NHK放送技術研究所技研だより，2023年3月号，TOP NEWS（2023）
85) “博報堂DYメディアパートナーズとCBCテレビ，AIによるドラマコンテンツ利活用の実証実験をスタート”, https://prtimes.jp/main/html/rd/p/000000148.000038657.html
86) “NAXA株式会社がAI動画編集サービス「Short Video Generator」をリリース！ ショート動画の編集，運用を90%効率化！”, https://prtimes.jp/main/html/rd/p/000000005.000049850.html
87) 中神：“MPEGにおけるPoint Cloud圧縮の標準化”，映情学誌，74,

2, pp.352-355（2020）
88）徐，西村，岸本：“動的メッシュ符号化の標準化”，映情学誌，78，2, pp.167-170（2024）
89）J. Zhao, S. Zhang, W. Zou and F. Yang: "MPEG Edgebreaker: An Efficient Static and Dynamic Mesh Codec in MPEG VDMC", ICIP2025 (2025)
90）Y. Liu, J. Xu, K. Kawamura and H. Watanabe: "Tracked QEM Algorithm: Adding Temporal Consistency to Dynamic Mesh Simplification Based on Mesh Registration", ITE Trans.on MTA, 12, 3, pp.175-189 (2024)
91）X. Jin, J. Xu and K. Kawamura: "Geometry Parametrization Stabilization for Dynamic Mesh Coding", ICIP2025 (2025)
92）A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato and T. Mikolov: "Devise: A deep visual-semantic embedding model", Adv. In NeurIPS, 26 (2013)
93）J. Lu, D. Batra, D. Parikh and S. Lee: "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-andlanguage tasks", Adv. In NeurIPS, 32 (2019)
94）X. Zhai, B. Mustafa, A. Kolesnikov and L. Beyer: "Sigmoid loss for language image pre-training", ICCV, pp.11975-11986 (2023)
95）Y. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng and J. Liu: "Uniter: Universal image-text representation learning", ECCV, pp.104-120 (2020)
96）A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach and D. Kiela: "Flava: A foundational language and vision alignment model", CVPR, pp.15638-15650 (2022)
97）D. Qi, L. Su, J. Song, E. Cui, T. Bharti and A. Sacheti: "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data", ArXiv:2001.07966 (2020)
98）L. Li, M. Yatskar, D. Yin, C. Hsieh and K. Chang: "What does BERT with vision look at?", Annual Meeting of the Association for Computational Linguistics, pp.5265-5275 (2020)
99）L. Li, M. Yatskar, D. Yin, C. Hsieh and K. Chang: "Visualbert: A simple and performant baseline for vision and language", ArXiv:1908.03557 (2019)
100）W. Kim, B. Son and I. Kim: "Vilt: Vision-and-language transformer without convolution or region supervision", ICML, pp.5583-5594 (2021)
101）Z. Huang, Z. Zeng, B. Liu, D. Fu and J. Fu: "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers", ArXiv:2004.00849 (2020)
102）H. Tan and M. Bansal: "LXMERT: Learning Cross-Modality Encoder Representations from Transformers", Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp.5100-5111 (2019)
103）J. Li, D. Li, C. Xiong and S. Hoi: "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation", ICML, pp.12888-12900 (2022)
104）J. Li, D. Li, S. Savarese and S. Hoi: "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", ICML, pp.19730-19742 (2023)
105）J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini and Y. Wu: "CoCa: Contrastive Captioners are Image-Text Foundation Models", Trans. on Machine Learning Research (2022)
106）Z. Ma, F. Xu, J. Liu, M. Yang and Q. Guo: "SyCoCa: Symmetrizing Contrastive Captioners with Attentive Masking for Multimodal Alignment", ICML, pp.34038-34052 (2024)
107）J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan and J. Gao: "Unified contrastive learning in image-text-label space", CVPR, pp.19163-19173 (2022)
108）L. Yuan, D. Chen, Y. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li and Others: "Florence: A new foundation model for computer vision", ArXiv:2111.11432 (2021)
109）U. Upadhyay, S. Karthik, M. Mancini and Z. Akata: "Probvlm: Probabilistic adapter for frozen vison-language models", ICCV, pp.1899-1910 (2023)
110）S. Chun, W. Kim, S. Park and S. Yun: "Probabilistic Language-Image Pre-Training", International Conference on Learning Representations (2025)
111）M. Tschannen, A. Gritsenko, X. Wang, M. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa and Others: "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization and dense features", ArXiv:2502.14786 (2025)
112）C. Xie, S. Sun, X. Xiong, Y. Zheng, D. Zhao and J. Zhou: "Ra-clip: Retrieval augmented contrastive language-image pre-training", CVPR, pp.19265-19274 (2023)
113）F. Faghri, D. Fleet, J. Kiros and S. Fidler: "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives", British Machine Vision Conference (2018)
114）A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark and Others: "Learning transferable visual models from natural language supervision", ICML, pp.8748-8763 (2021)
115）Z. Guo, T. Wang, S. Pehlivan, A. Radman and J. Laaksonen: "PiTL: cross-modal retrieval with weakly-supervised vision-language pre-training via prompting", International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.2261-2265 (2023)
116）C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. Le, Y. Sung, Z. Li and T. Duerig: "Scaling up visual and vision-language representation learning with noisy text supervision", ICML, pp.4904-4916 (2021)
117）Y. Li, H. Fan, R. Hu, C. Feichtenhofer and K. He: "Scaling language-image pre-training via masking", CVPR, pp.23390-23400 (2023)
118）Q. Sun, Y. Fang, L. Wu, X. Wang and Y. Cao: "Eva-clip: Improved training techniques for clip at scale", ArXiv:2303.15389 (2023)
119）W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. Mohammed, S. Singhal, S. Som and Others: "Image as a foreign language: Beit pretraining for vision and vision-language tasks", CVPR, pp.19175-19186 (2023)
120）W. Wu, K. Zheng, S. Ma, F. Lu, Y. Guo, Y. Zhang, W. Chen, Q. Guo, Y. Shen and Z. Zha: "Lotlip: Improving language-image pre-training for long text understanding", Adv. In NeurIPS, 37, pp.64996-65019 (2024)
121）S. Kim, R. Xiao, M. Georgescu, S. Alaniz and Z. Akata: "Cosmos: Cross-modality self-distillation for vision language pre-training", CVPR, pp.14690-14700 (2025)
122）B. Zhang, P. Zhang, X. Dong, Y. Zang and J. Wang: "Longclip: Unlocking the long-text capability of clip", ECCV, pp.310-325 (2024)
123）W. Kuo, A. Piergiovanni, D. Kim, Luo, B. Caine, W. Li, A. Ogale, L. Zhou, A. Dai, Z. Chen, C. Cui and A. Angelova: "MaMMUT: A Simple Architecture for Joint Learning for MultiModal Tasks", Trans. on Machine Learning Research (2023)
124）W. Huang, A. Wu, Y. Yang, X. Luo, Y. Yang, L. Hu, Q. Dai, C. Wang, X. Dai, D. Chen and Others: "Llm2clip: Powerful language model unlocks richer visual representation", ArXiv:2411.04997 (2024)
125）D. Schnaus, N. Araslanov and D. Cremers: "It's a (Blind) Match! Towards Vision-Language Correspondence without Parallel Data", CVPR, pp.24983-24992 (2025)
126）Y. Ji, X. Xiao, G. Chen, H. Xu, C. Ma, L. Zhu, A. Liang and J. Chen: "Cibr: Cross-modal information bottleneck regularization for robust clip generalization", International Conference on Artificial Neural Networks, pp.247-259 (2025)
127）T. Jiang, M. Song, Z. Zhang, H. Huang, W. Deng, F. Sun, Q. Zhang, D. Wang and F. Zhuang: "E5-v: Universal embeddings with multimodal large language models", ArXiv:2407.12580 (2024)
128）Y. Ouali, A. Bulat, A. Xenos, A. Zaganidis, I. Metaxas, B. Martinez and G. Tzimiropoulos: "VladVA: Discriminative Fine-tuning of LVLMs", CVPR, pp.4101-4111 (2025)
129）H. Wang, C. Ju, W. Lin, S. Xiao, M. Chen, Y. Huang, C. Liu, M. Yao, J. Lan, Y. Chen and Others: "Advancing myopia to holism: Fully contrastive language-image pre-training", CVPR, pp.29791-29802 (2025)
130）A. Cao, X. Wei and Z. Ma: "FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training", CVPR, pp.4080-4090 (2025)
131）K. Zheng, Y. Zhang, W. Wu, F. Lu, S. Ma, X. Jin, W. Chen and Y.

Shen: "Dreamlip: Language-image pre-training with long captions", ECCV, pp.73-90 (2024)
132) M. Asokan, K. Wu and F. Albreiki: "FineLIP: Extending CLIP's Reach via Fine-Grained Alignment with Longer Text Inputs", CVPR, pp.14495-14504 (2025)
133) R. Xiao, S. Kim, M. Georgescu, Z. Akata and S. Alaniz: "Flair: Vlm with fine-grained language-informed image representations", CVPR, pp.24884-24894 (2025)
134) A. Bulat, Y. Ouali and G. Tzimiropoulos: "FFF: Fixing Flawed Foundations in contrastive pre-training results in very strong Vision-Language models", CVPR, pp.14172-14182 (2024)
135) S. Ren, K. He, R. Girshick and J. Sun: "Faster r-cnn: Towards real-time object detection with region proposal networks", Adv. In NeurIPS, 28 (2015)
136) A. Oord, Y. Li and O. Vinyals: "Representation learning with contrastive predictive coding", ArXiv:1807.03748 (2018)
137) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, LL. Kaiser and I. Polosukhin: "Attention is all you need", Adv. In NeurIPS, 30 (2017)
138) T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela and C. Ross: "Winoground: Probing vision and language models for visio-linguistic compositionality", CVPR, pp.5238-5248 (2022)
139) Y. Li, W. Wang, L. Qu, L. Nie, W. Li and T. Chua: "Generative Cross-Modal Retrieval: Memorizing Images in Multimodal Language Models for Retrieval and Beyond", Annual Meeting of the Association for Computational Linguistics, pp.11851-11861 (2024)
140) S. Kim, X. Zhu, X. Lin, M. Bastan, D. Gray and S. Kwak: "GENIUS: A generative framework for universal multimodal search", CVPR, pp.19659-19669 (2025)
141) M. Bexte, A. Horbach and T. Zesch: "EVil-Probe-a Composite Benchmark for Extensive Visio-Linguistic Probing", Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pp.6682-6700 (2024)
142) K. He, X. Zhang, S. Ren and J. Sun: "Deep residual learning for image recognition", CVPR. pp.770-778 (2016)
143) 長谷山，河村，田良島，新井：“メディア工学の研究動向”，映情学誌” (2018)
144) K. Yang, J. Deng, X. An, J. Li, Z. Feng, J. Guo, J. Yang and T. Liu: "Alip: Adaptive language-image pre-training with synthetic caption", ICCV, pp.2922-2931 (2023)
145) Y. Zeng, X. Zhang and H. Li: "Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts", International Conference on Machine Learning, pp.25994-26009 (2022)
146) D. Kim, N. Kim and S. Kwak: "Improving cross-modal retrieval with set of diverse embeddings", CVPR, pp.23422-23431 (2023)
147) H. Wu, M. Wang, W. Zhou, Z. Lu and H. Li: "Asymmetric feature fusion for image retrieval", CVPR, pp.11082-11092 (2023)
148) A. Neculai, Y. Chen and Z. Akata: "Probabilistic compositional embeddings for multimodal image retrieval", CVPR, pp.4547-4557 (2022)
149) S. Chun, S. Oh, R. De Rezende, Y. Kalantidis and D. Larlus: "Probabilistic embeddings for cross-modal retrieval", CVPR, pp.8415-8424 (2021)
150) Y. Song and M. Soleymani: "Polysemous visual-semantic embedding for cross-modal retrieval", CVPR, pp.1979-1988 (2019)
151) Digital Body Festival, https://digital-body.com/
152) Vast Body, https://www.aatoaa.com/vastbody
153) Body of Mine, https://www.bodyofminevr.com/
154) Dazzle: Solo, https://dazzle1919.com/
155) Soul Paint, https://www.soulpaint.co/
156) ARK 1, https://www.tanzforumberlin.de/en/production/ark-1/
157) Superradiance, https://superradiance.art/

東海　彰吾（とうかい　しょうご） 1996年，名古屋大学大学院工学研究科博士課程修了．京都大学助手，福井大学講師，准教授を経て，現在，福井大学工学系部門教授．コンピュータグラフィックス，映像メディア処理の研究に従事．博士（工学）．正会員．

小川　貴弘（おがわ　たかひろ） 2003年，北海道大学工学部卒業．2007年，同大学大学院情報科学研究科博士後期課程修了．同大学研究員，助教，准教授を経て，現在，同大学大学院情報科学研究院教授．マルチメディアAIに関する研究に従事．博士（情報科学）．正会員．

田良島周平（たらしま しゅうへい） 2009年，東京大学工学部卒業．2011年，同大学大学院新領域創成科学研究科修士課程修了．同年，NTT（株）入社．現在，NTTドコモビジネス（株）イノベーションセンター担当課長．2023年～2024年，東北大学非常勤講師．2023年～現在，東京都立大学客員研究員．コンピュータビジョンに関する研究開発に従事．正会員．

望月　貴裕（もちづき　たかひろ） 1994年，東京工業大学工学部情報工学科卒業，1996年，同大学大学院総合理工学部物理情報工学専攻修士課程修了．同年，NHK入局．同放送技術局を経て，1998年より，放送技術研究所にて，画像認識の研究に従事．現在，言語画像解析グループのマネジメント業務に従事．博士（工学）．正会員．

河村　圭（かわむら　けい） 2004年，早稲田大学理工学部電子・情報通信学科卒業．2005年，同大学大学院国際情報通信研究科修士課程修了．2010年，同大学大学院国際情報通信研究科博士課程修了．同年，KDDI（株）入社．現在，（株）KDDI総合研究所XR部門長．主に，動画像符号化方式の研究・開発および国際標準化に従事．博士（国際情報通信学）．正会員．

紛　尚弥（そぎ　なおや） 2017年，筑波大学情報学群卒業．2019年，同大学大学院情報科学研究科博士前期課程修了．2022年，同大学大学院情報科学研究科博士後期課程修了．同年，NEC（株）入社．現在，同社ビジュアルインテリジェンス研究所主任．画像認識に関する研究開発に従事．博士（工学）．

曽我麻佐子（そが あさこ） 名古屋大学大学院人間情報学研究科博士後期課程修了．龍谷大学理工学部助手，助教，講師を経て，2017年より，同大学准教授．おもに，人体アニメーションに関する研究に従事．博士（学術）．正会員．