計測自動制御学会東北支部 第 360 回研究集会 (2026.9.16)
資料番号 360-1

双方向状態空間モデルを用いた音響・視覚統合認識における入力表現の検討

Investigation of Input Representations for Audio-Visual Recognition Using a Bidirectional State Space Model

○西村悠汰，田中大介，張山昌論

○ Yuta Nishimura, Daisuke Tanaka, Masanori Hariyama

東北大学

Tohoku University

キーワード： 状態空間モデル (state space model)，音響・視覚統合認識 (audio-visual recognition)，メルスペクトログラム (mel spectrogram)，転移学習 (transfer learning)，シーン識別 (scene classification)

連絡先： 〒 980-8579　仙台市青葉区荒巻字青葉 6-3-09 東北大学 大学院情報科学研究科
田中大介，Tel.: (022)795-7155, E-mail: daisuke.tanaka.b7@tohoku.ac.jp

1. はじめに

上空から撮影した画像に基づく土地被覆のシーン識別は，地表面の状態を広域に把握する手段として利用されている．しかし画像のみによる識別は安定しない．季節や植生の変化で同じ地点の見え方が変わり，日照条件，天候，遮蔽の影響も受けるためである [1)]．

これに対し，同じ地点で収録された環境音は画像とは独立した情報を与える．水面の音，鳥の声，車両や人の音は，上空からの情報が似ていても土地の使われ方によって異なる．ただし音のみによる識別も容易ではなく，実環境では複数の音源と背景雑音が重なるため，本稿が対象とするベンチマークにおける音響単独の識別性能は画像単独の半分に満たない [1)]．重要なのは二つのモダリティが劣化する条件が互いに一致しない点であり，両者を統合することによってそれぞれの弱点を補い合うことが期待できる．

一方で統合には計算資源上の代償がある．各モダリティをトークン列として連結すると入力系列長が増加し，Vision Transformer (ViT)[2)] に代表される自己注意機構では，トークン数 L に対して計算量とメモリが $O(L^2)$ で増大してしまう．この制約に対する枠組みが状態空間モデル (state space model, SSM) であり，状態を逐次更新する構造から系列長に対して線形の計算量で処理できる．Vision Mamba (Vim)[3)] はこれを画像に適用した符号化器であり，ImageNet-1K の画像分類において同規模の Transformer 系モデルと比較しうる精度が報告されている．

Vim は画像を対象として設計された符号化器であるが，その入力に求められるのは 2 次元の配列である．音響信号もスペクトログラムに変換すれば時間・周波数の 2 次元表現となるため，単一の符号化器で両モダリティを扱い，大規模画像データによる事前学習の恩恵を音響側にも

及ぼせる可能性がある．

ただし転移にあたっては，音響をどのような形式で符号化器に入力するかという問題が生じる．スペクトログラムを時間軸方向の系列として扱うか，画像と同一の 2 次元データとして扱うかによって，事前学習で得られた知識の利用のされ方が変わるためである．著者らの従来研究 [4] は前者を採り，各時刻の周波数ベクトルを一つのトークンとしていた．本稿では Vim を基盤とするこの音響・視覚統合の枠組みを ViAuM (Vision-Audio Mamba) と呼び，従来の入力表現による構成を 1D-ViAuM と表記する．

本稿では，対数メルスペクトログラムを画像とまったく同一の 2 次元パッチに分割する 2D-ViAuM を提案し，その効果を同一のデータ分割・同一の学習条件のもとで従来手法と比較して定量化する．

2. 背景と準備

2.1 状態空間モデル

連続時間の線形状態空間モデルは，状態 $h(t)$，入力 $x(t)$，出力 $y(t)$ を用いて式 (1)，(2) で表される．

$$\frac{dh(t)}{dt} = Ah(t) + Bx(t) \tag{1}$$

$$y(t) = Ch(t) \tag{2}$$

ここで $h(t)$ は n 次元の状態ベクトル，A，B，C は系の係数行列である．系列モデルとして用いる際には，離散化ステップ Δ による離散化を施し，式 (3)，(4) の形で用いる．

$$h_k = \bar{A}h_{k-1} + \bar{B}x_k \tag{3}$$

$$y_k = Ch_k \tag{4}$$

$\bar{A}$，$\bar{B}$ は零次ホールド等により A，B，Δ から定まる離散系の係数行列である．式 (3) が示すように各時刻の計算は直前の状態のみに依存するため，系列長 L に対する計算量は $O(L)$ であり，保持すべき状態の次元 n は系列長によらず一定である．

2.2 選択機構

文献 [5] の S4 では，A，B，C，Δ が入力によらない定数であり，どの入力に対しても同一の応答しか返せない．Mamba[6] はこの制約に対し，B，C，Δ を入力 x_k の関数とする選択機構を導入した．これにより，入力の内容に応じて状態への取り込み方と忘却の速さを変化させられる．ただし係数が時刻ごとに変化するため，係数が一定であることを前提とした並列化は適用できない．Mamba では並列スキャンにより逐次更新を並列化することで，$O(L)$ の計算量を保ったまま選択機構を実現している．

2.3 Vision Mamba

Vim[3] は，画像を 16×16 の非重複パッチに分割し，線形射影によりトークン列へ変換して SSM ブロックで処理する．画像のトークン列には時系列のような因果的順序が存在しないため，順方向と逆方向の両方向に SSM を走らせる双方向ブロックを用いる点と，空間情報を保持するために学習可能な位置埋め込みを加える点が特徴である．本稿では ImageNet-1K で事前学習された Vim-tiny を基盤とし，双方向構成はそのまま用いる．

2.4 音響・視覚統合による航空シーン認識

Hu ら [1] は，航空画像と地理タグ付き音響の対 5,075 組を 13 カテゴリに分類した ADVANCE データセットを構築し，音響・視覚統合による航空シーン認識を最初に定式化した．同研究では，航空画像データセット AID[7] と大規模音響イベントデータセット AudioSet[8] でそれぞれ事前学習した ResNet の表現を単純に連結しただけでは画像単独の性能を下回り，クロスタスク転移を加えて初めて F1 スコアが 74.58 へ改善したことが報告されている．これより統合による利得は，音響表現をどう扱うかという設計から生じているといえる．

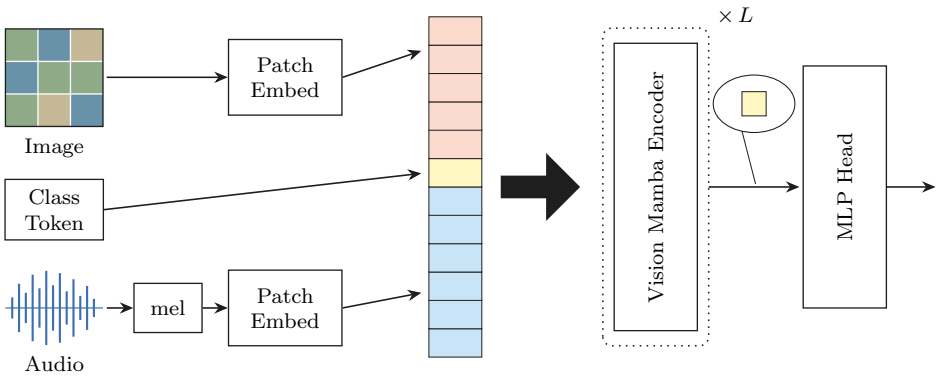


Fig. 1　提案手法 2D-ViAuM の全体構成

2.5　画像用モデルの音響への転移

画像向けに事前学習されたモデルを音響へ転移する試みには，対数メルスペクトログラムをパッチ単位で ViT に入力する AST[9]，DeiT-B/384 を音響側に転移する TFAVCNet[10]，双方向 SSM ブロックを音響に特化させた Audio Mamba (AuM)[11] がある．これらに共通するのは，音響側の表現をまず固定し，事前学習済みのモデル側をそれに合わせて調整するという順序である．その結果，入力寸法の変更やパッチ埋め込み層の再学習が必要となり，事前学習で得られた入力側の構造は部分的に失われる．

本稿はこの順序を逆にする．すなわち，事前学習側の入力寸法とパッチ分割を固定したうえで，それに整合するように音響フロントエンドのパラメータを逆算する．音響側を可変とみなすことで，パッチ埋め込み層と位置埋め込みを含む入力側の処理全体を，事前学習重みのまま転用できる．この方針が本稿の設計の出発点であり，具体的な導出は 3.3 節で述べる．

3.　提案手法：2D-ViAuM

3.1　全体構成

提案する 2D-ViAuM の全体構成を Fig. 1 に示す．航空画像はパッチに分割され，線形射影によりトークン列へ埋め込まれる．音響も対数メルスペクトログラムへ変換されたのち，画像とまったく同じパッチ埋め込み層で同様にトークン化される．系列全体の情報を集約するために付加する学習可能なトークン (クラストークン) を二つのモダリティの境界に挿入し，連結した系列を共有の ImageNet 事前学習済み Vim 符号化器で処理する．最終層におけるクラストークンの隠れ状態を正規化し，MLP ヘッドへ入力して 13 クラスのスコアを出力する．

3.2　画像

標準的な Vim-tiny の構成に従い，航空画像を 224×224 画素にリサイズして一辺 16 画素の非重複パッチ 196 個に分割し，ImageNet 事前学習済みのパッチ埋め込み層によりトークンへ射影する．

3.3　音響：スペクトログラムの画像化

二つの入力表現の対比を Fig. 2 に示す．

従来手法である 1D-ViAuM[4] では，音響を短時間フーリエ変換 (FFT 長 1024，ホップ長 816，Hann 窓) により線形周波数軸のスペクトログラムへ変換し dB 表現する．得られる 513×196 の表現に対し，時間軸上の 1 フレーム，すなわち 513 次元の周波数ベクトルを単一のトークンとして線形射影する (Fig. 2 左)．音響は 196 トークンの 1 次元系列となり，周波数軸は各トークン内に畳み込まれる．ImageNet 事前学習済み

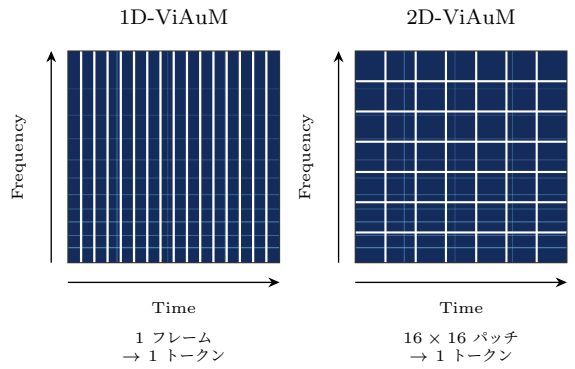


Fig. 2 音響の入力表現の対比 (模式図). 1D-ViAuM は時間フレームを，2D-ViAuM は 16×16 パッチをトークンの単位とする．いずれも系列長は 196 トークンとなる

のパッチ埋め込み層は局所的な 16×16 の空間構造を捉えるよう学習されているため，このフレーム単位の射影は当該重みを継承できず，新たに学習する必要がある．

提案する 2D-ViAuM では，2.5 節の方針に従い，事前学習側の入力寸法 224×224 を固定して音響フロントエンドのパラメータを逆算する．標本化周波数 16 kHz に対し，ホップ長を 715 標本，メルフィルタバンクのビン数を 224 と設定すればよい (FFT 長 1024，Hann 窓)．これにより，ADVANCE の 10 秒の収録から時間軸方向の補間も切り出しも伴わない 224×224 のネイティブな対数メルスペクトログラムが得られる．単チャネルのスペクトログラムは，事前学習済みパッチ埋め込み層に合わせて 3 チャネルへ複製し，画像と同一の 16×16 パッチに分割する (Fig. 2 右)．音響も同じく 196 個のパッチトークンとなるため系列長は 1D-ViAuM と等しく，パッチ埋め込み層を事前学習重みのまま再利用できる．なおこの設定は収録長 10 秒を前提としており，可変長の音響には別途の調整を要する．

3.4 モダリティの統合

統合条件では，224×224 の航空画像と 224×224 の対数メルスペクトログラムを縦方向に連結し，単一の 448×224 のテンソルとする．これを縦長の 1 枚の画像として共有パッチ埋め込み層と Vim 符号化器で処理することは，196 個の画像トークンと 196 個の音響トークンを合わせた 392 トークンの系列を扱うことに相当する．

位置埋め込みは，事前学習済みの値からクラストークン分を除いて 14×14 のグリッドへ戻し，bicubic 補間により 28×14 へ拡大したうえで，系列の中央にクラストークンを再挿入して用いる．補間は学習可能なパラメータを増やさないため，位置に関するパラメータ数は事前学習時から変化しない．音響単独条件ではグリッドが 14×14 のままであり，補間は恒等変換となる．この処理は 1D-ViAuM と 2D-ViAuM に共通である．

クラストークンを系列の中央 (視覚領域と音響領域の境界) に置くのは，中央配置が最良とする Vim[3] および AuM[11] のアブレーション結果に従う．

なお 1D-ViAuM は画像とスペクトログラムを別々のパッチ埋め込み層でトークン化するため，音響単独条件に比べて画像側の埋め込み層が追加で必要となる．これに対し 2D-ViAuM は，両モダリティが同一の形状をもつため単一の埋め込み層を共有でき，モダリティを追加しても入力側のパラメータが増えない．

4. 実験と評価

4.1 実験設定

評価には 2.4 節の ADVANCE[1] を用いる．各対は地理タグ付きの Freesound 収録音 (10 秒) と対応する座標の航空画像からなる．クラス分布は偏っており，最大の住宅地 1,056 対に対し最小の草地は 150 対である．この約 7 倍の不均衡が損失の重み付けを比較する動機となっている．前処理として航空画像は 224×224 にリサイズし，音響は 3.3 節の各設定でスペクトログラムへ変換したうえで，全特徴次元を平均 0，分散 1 に正規化する．

データセットは学習 70%，検証 10%，評価 20%に分割する．ADVANCE では少数クラスが座標をずらして水増しされ，4 枚の追加画像が 1 つの収録音を共有するため [1]，これらが異なる部分集合にまたがると情報の漏洩が生じる．そこでファイル名の先頭 5 文字が示す座標識別子で標本をグループ化したうえで分割した．評価集合は 1,067 対からなり，両手法は同一の分割を共有する．

両手法とも ImageNet-1K で事前学習された Vim-tiny[3] (24 ブロック，隠れ次元 192，SSM 次元 16，約 7.0M パラメータ) を基幹部分とし，ADVANCE 上でファインチューニングを行う．最適化には AdamW[12] を用い，学習率 1×10^{-5}，バッチサイズ 64，最大 200 エポックとし，検証損失が 50 エポック改善しない場合に早期終了する．損失はラベル平滑化 (0.1) を伴う交差エントロピーとし，データ拡張は画像に左右反転と色相・明度の摂動，音響にランダム消去 [13] を用いた．両手法で異なるのは 3.3 節で述べた音響の入力表現，すなわちトークン化の単位と周波数軸の表現の二点のみである．

損失は，クラス頻度の逆数に比例する重みを付した交差エントロピー (以下，クラス重み付き損失．表中は「重み付き」) と通常の交差エントロピー (以下，一様損失) の 2 条件を評価する．評価指標はクラスの標本数で重み付けした平均の F1 スコア，再現率，適合率とする．なお本稿の結果はいずれも単一分割・単一試行であり，試行間のばらつきは評価していない．

4.2　音響単独条件

音響単独条件と統合条件の結果を Table 1 にまとめて示す．

入力表現の変更は音響単独条件で大きな効果をもたらした．F1 スコアは一様損失で 15.92%から 24.83%へ 8.91 ポイント，クラス重み付き損失で 13.35%から 23.76%へ 10.41 ポイント改善した．両手法は同一の基幹部分を共有し，系列長もともに 196 トークンである．

改善の内容は混同行列に表れる．一様損失における 1D-ViAuM (Fig. 3(a)) では，全予測の 68%にあたる 722 標本が住宅地という単一の列に集中し，13 クラス中 8 クラスで正解が得られていない．事前学習重みを利用できない 1D-ViAuM は，限られた学習データのもとで多数クラスへ退化する傾向をもつ．これに対し 2D-ViAuM (Fig. 3(b)) では，最頻の予測クラスへの集中が 41%まで緩和され，正解の得られないクラスは 3 つに減少した．森林が 69 から 134 へ，低木地が 0 から 18 へ増加する一方，過剰に予測されていた住宅地は 156 から 107 へ減少しており，予測が多数クラスから他のクラスへ再配分されたといえる．

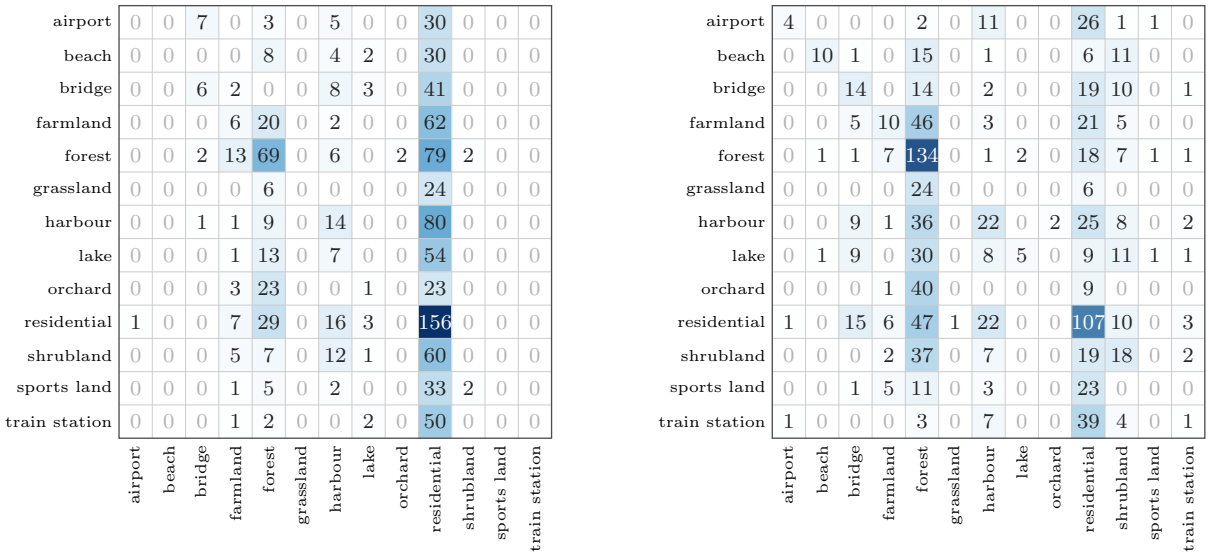
クラス重み付き損失の混同行列は紙面の都合により省略するが，同様の傾向が観察される．1D-ViAuM では正解総数が 164 にとどまり 2 クラスがまったく正解を得られないのに対し，2D-ViAuM では総数が 256 へ増加し，正解の得られないクラスは存在しない．クラス重み付けが予測を有効に再配分するためには，音響表現そのものがあらかじめ一定の分離性をもっている必要があるといえる．

4.3　音響・視覚統合条件

視覚情報を加えることでいずれの構成でも性能が大きく向上し，2D-ViAuM の F1 スコアは音響単独条件の 24.83%から 80.71%へ上昇した．入力表現による改善はこの条件でも同じ向きに現れ，一様損失で 1.65 ポイント，クラス重み付き損失で 3.48 ポイント上昇している．改善幅が小さいのは，シーン識別に必要な情報の大半が視覚モダリティに存在し，音響表現の質が寄与できる余地が限られているためとみられる．ただし単一試行であるため，一様損失における 1.65 ポイントの差は傾向を示すにとどまる．

Table 1　音響単独条件および音響・視覚統合条件の結果

条件	手法	損失	F1 [%]	再現率 [%]	適合率 [%]
音響単独	1D-ViAuM (従来)	重み付き	13.35	15.37	19.01
	1D-ViAuM (従来)	一様	15.92	23.52	15.25
	2D-ViAuM (提案)	重み付き	23.76	23.99	29.85
	2D-ViAuM (提案)	一様	24.83	30.46	31.43
統合	1D-ViAuM (従来)	重み付き	77.37	77.79	78.04
	1D-ViAuM (従来)	一様	79.06	79.48	79.98
	2D-ViAuM (提案)	重み付き	80.85	81.26	81.49
	2D-ViAuM (提案)	一様	80.71	80.97	81.26



(a) 1D-ViAuM　　(b) 2D-ViAuM

Fig. 3　音響単独・一様損失における混同行列．縦軸は正解ラベル，横軸は予測ラベルであり，両図は共通の濃度尺度で描いている

統合条件における各クラスの正解数を Fig. 4 に示す．改善は特定のクラス群に集中している．一様損失では果樹園 (27 から 40)，低木地 (53 から 63)，農地 (57 から 65)，湖 (56 から 62) の正解数が増加した．いずれも上空からの画像だけでは区別しにくい植生地または水域であり，鳥の声や水の音といった音響的な手がかりが寄与しうるクラスである．一方で港湾は 90 から 75 へ，森林は 164 から 155 へ減少しており，改善は再配分を伴うが，正解総数は 848 から 864 へ増加している．

クラス重み付き損失についても同様の傾向が確認される．果樹園が 29 から 39 へ，低木地が 59 から 68 へ，港湾が 77 から 85 へ増加しており，このうち果樹園と低木地は一様損失でも増加したクラスである．損失の重み付けによらず，音響表現の改善が寄与する対象が同じ植生地に定まっているといえる．

ただし，以上の改善が入力表現のどの変更点に由来するかは，音響単独条件と統合条件のいずれにおいても分離できていない．事前学習済みパッチ埋め込み層の利用，2 次元の局所構造という帰納バイアスの導入，周波数軸のメル尺度への変更が同時に生じているためであり，切り分けには追加実験を要する．

5.　おわりに

本研究では，状態空間モデルである Vision Mamba を基盤とする音響・視覚統合認識にお

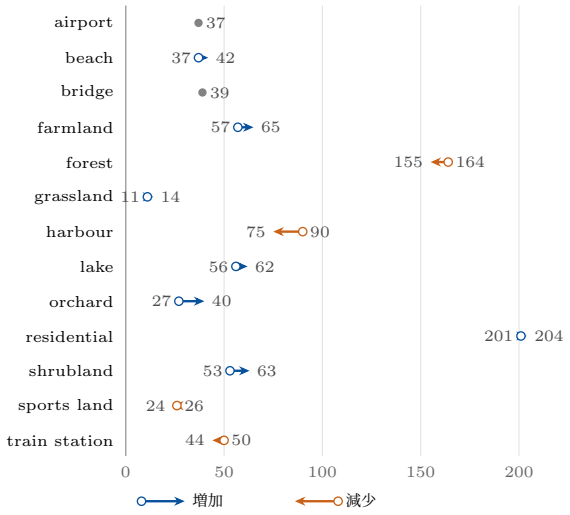


Fig. 4　音響・視覚統合・一様損失における各クラスの正解数．白丸が 1D-ViAuM，矢頭が 2D-ViAuM を表す．正解総数は 848 から 864 へ増加している

いて，音響の入力表現が認識性能に与える影響を検討した．音響を時間フレーム単位でトークン化する従来の 1D-ViAuM に対し，提案する 2D-ViAuM は対数メルスペクトログラムを画像として扱い，構造を改変することなく事前学習済み Vim 符号化器へ入力することで，パッチ埋め込み層の事前学習重みをそのまま再利用した．

ADVANCE を用いた評価により，入力表現の変更のみで F1 スコアが音響単独条件で 8.91 ポイント，統合条件で 1.65 ポイント改善することを確認した．

一方，本稿は従来手法との精度比較に焦点を当てたため，提案手法の位置づけを十分に示すには至っていない．今後の課題として，CNN を基幹部分とする既存手法や Transformer に基づく音響・視覚統合手法との精度比較が挙げられる．また，モダリティごとに独立した基幹部分を積み重ねる構成に対し，単一の符号化器の系列長を延ばすだけで統合を行う本構成がもつ利点を定量的に示すため，スループットおよびメモリ使用量を測定し，モダリティ追加に伴う計算コストの増え方を既存手法と比較する．あわせて，複数分割・複数試行による評価，周波数尺度とトークン化の寄与を分離する追加実験，および航空画像ドメインでの事前学習にも取り組む．

謝辞

本研究は JSPS 科研費 JP25K15170 の助成を受けたものである．

参考文献

1) Hu ら: Cross-task transfer for geotagged audiovisual aerial scene recognition, Proc. of the European Conference on Computer Vision (ECCV), 68/84 (2020)

2) Dosovitskiy ら: An image is worth 16x16 words: Transformers for image recognition at scale, Proc. of the International Conference on Learning Representations (ICLR) (2021)

3) Zhu ら: Vision Mamba: Efficient visual representation learning with bidirectional state space model, Proc. of the International Conference on Machine Learning (ICML), 62429/62442 (2024)

4) Nishimura, Tanaka and Hariyama: Robust audio-visual object recognition in noisy environments using state space models, Proc. of the IEEE 14th International Conference on Consumer Electronics - Taiwan (ICCE-TW) (2026)

5) Gu, Goel and Ré: Efficiently modeling long sequences with structured state spaces, Proc. of the ICLR (2022)

6) Gu and Dao: Mamba: Linear-time sequence modeling with selective state spaces, Proc. of the 1st Conference on Language Modeling (COLM) (2024)

7) Xia ら: AID: A benchmark data set for performance evaluation of aerial scene classification, IEEE Trans. Geoscience and Remote Sensing, **55**-7, 3965/3981 (2017)

8) Gemmeke ら: Audio Set: An ontology and human-labeled dataset for audio events, Proc. of the IEEE ICASSP, 776/780 (2017)

9) Gong, Chung and Glass: AST: Audio spectrogram transformer, Proc. of Interspeech, 571/575 (2021)

10) Wang ら: Two-stage fusion-based audiovisual remote sensing scene classification, Applied Sciences, **13**-21, 11890 (2023)

11) Erol ら: Audio Mamba: Bidirectional state space model for audio representation learning, IEEE Signal Processing Letters, **31**, 2975/2979 (2024)

12) Loshchilov and Hutter: Decoupled weight decay regularization, Proc. of the ICLR (2019)

13) Zhong ら: Random erasing data augmentation, Proc. of the AAAI Conference on Artificial Intelligence, **34**-7, 13001/13008 (2020)