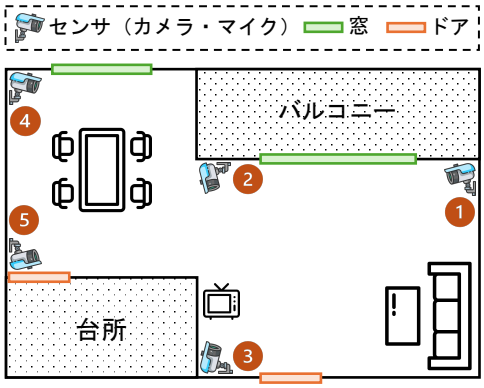


図 2: MultiSensor-Home データセットにおける多視点カメラの配置と部屋のレイアウト．

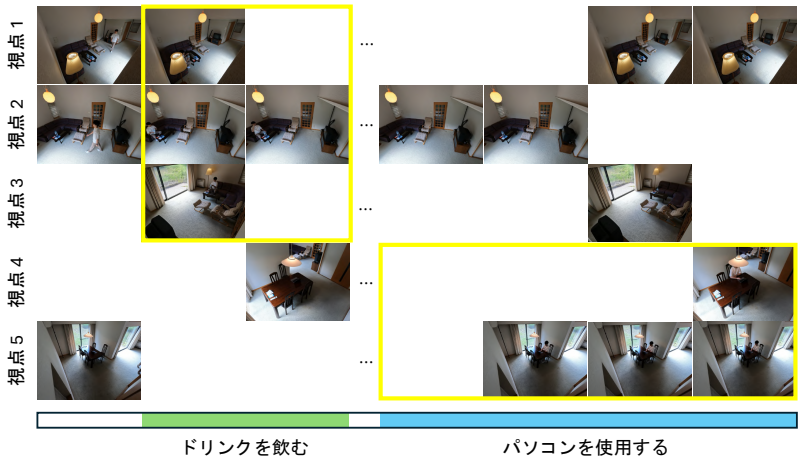


図 3: MultiSensor-Home データセット中の多視点から捉えた行動の様子の例．

トである．本研究の貢献は以下の通りである．

- 多視点マルチモーダル MultiSensor-Home データセット．提案データセットは，多視点から記載したフレーム単位の詳細なラベルを付与した複数の行動を含む映像を提供する．
- 多視点マルチモダール MultiTSF 手法．映像を構成する音声と動画像のモダリティを統合し，Transformer に基づく時間モデリングと多視点統合になる行動認識手法を提案する．さらに，人物検出モジュールを導入することで，人物活動を含むフレームに注目させる．

本論文の構成は以下の通りである． 2. で関連研究を紹介する． 3. と 4. で提案するデータセットと手法を説明する． 5. で実験結果を述べ， 6. でまとめと今後の方向性を示す．

2. 関 連 研 究

2.1 多視点マルチモーダルデータセット

行動認識技術の進展は，多視点マルチモーダルデータセットの開発に支えられている．これらのデータセットは，複数の視点とマルチモーダルデータを提供し，人間の行動を広域で理解することを可能にする．NorthWestern-UCLA Multi-view Action 3D [10] データセットは，多視点から RGB+D データを提供している．一方，NTU RGB+D [11] 及びその拡張版 NTU RGB+D 120 [12] データセットは，大規模な多視点マルチモーダルデータを提供している．しかし，これらのデータセットは，統制された狭い範囲の短時間動画像が中心で，現実世界への適用性に制限がある．一方，広域を対象としたデータセットとしては，Yasuda らがオフィス環境とコンビニエンスストア環境で撮影された RGB ＋音声録画を含む MM-Office [8] および MM-Store [9] データセットを提供している．しかし，これらのデータセットは動画像系列単位のラベルや限定的なフレーム単位のラベルしか提供せず，系列単位の教師あり学習タスクにしか適用できない．これらの課題に対応するため，本研究では MultiSensor-Home データセットを提案し，異なる時間帯における複数の行動を記録し，フレーム単位の詳細なラベルを提供する．

2.2 多視点からのマルチモーダルな行動認識

最近の行動認識研究は，多視点およびマルチモーダルを活用して認識精度を向上させることに焦点を当てている．既存手法は，主に狭域のデータセットを対象としている．教師あり対照学習 [4] により視点変動に対する特徴の頑健性を高めたり，教師なし表現学習 [5] により視点変動に対する特徴を作成する研究が進められている．広域の行動認識では，Yasuda ら [8] が分散型センサ間の関係をモデル化する MultiTrans を提案し，Guided-MELD [9] はセンサ観測の断片化や冗長情報を補完し，行動表現を実現している．また，John ら [13] は弱い教師ありの潜在埋め込みモデルを，我々 [14] は擬似フレームレベルのラベル生成でフレームレベルラベル不足に対応する手法を提案している．これらに対し，本研究では，多視点行動認識タスクにおける精度向上を目指すため，Transformer に基づく時間モデリングと多視点統合を導入する MultiTSF 手法を提案する．

3. MultiSensor-Home データセット

本研究では，室内の家庭環境における現実的な多視点マルチモダール行動認識のためのベンチマーク MultiSensor-Home データセット(注1)を提案する．提案データセットは，広範囲の環境に戦略的に配置された 5 台の同期カメラで撮影された 未編集の映像を特徴としており，図 2 に示すように，一人の被験者が部屋内で動作を行う様子が記録される．この多視点構成により，多様な視点から行動を記録することが可能となり，環境内の空間的な動態を覆うことができる．各映像には，高解像度かつ高フレームレートで記録された音声と RGB 動画像が含まれており，高度な認識タスクのための豊富なマルチモーダルデータを提供する．提案データセットは，昼夜の異なる時間帯，さまざまな衣服の種類，そして活動環境の自然な変動を含む多様な条件下で行われた行動を捉えている．また，このデータセットは多視点からのフレーム単位の詳細なラベルを提供するため，時

(注1)：本データセットは近い将来公開予定である．詳細は https://github.com/thanhhff/MultiSensorHome-Dataset を参照のこと．

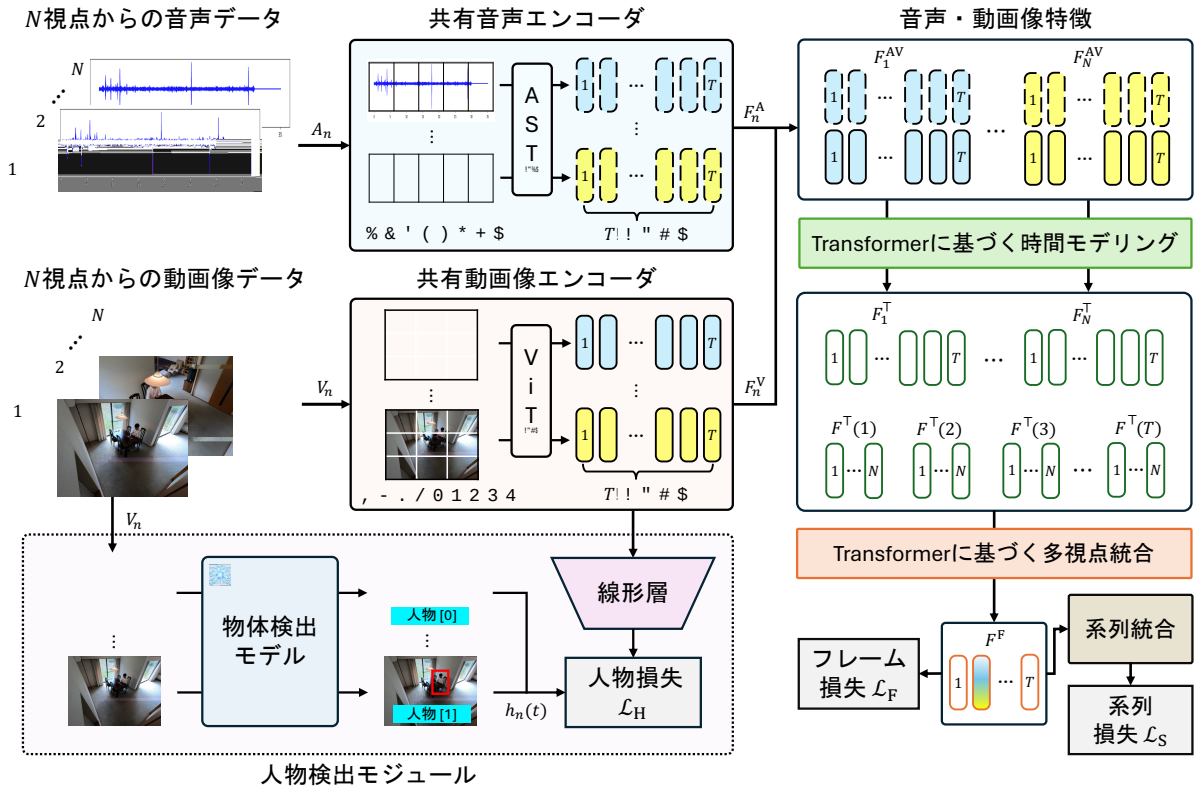


図 4: 提案する MultiTSF 手法の概要．提案手法は次の 3 つの主要な構成要素からなる：(1) 音声データと動画像データを識別的特徴を抽出するためのマルチモーダル特徴抽出モジュール，(2) 人物の存在を検出するための人物検出モジュール，(3) 時間的依存関係や多視点統合のための Transformer に基づく時間モデリングと多視点統合モジュール．

空間的な分析が可能である．図 3 に，多視点から記録された行動の例を示す．

4. 提案手法：MultiTSF

4.1 概　　要

4.1.1 問 題 設 定

MultiTSF は，取得されたマルチモーダルデータを入力として，動画像内の行動を予測することを目的とする．N 視点からの入力データは，音声データ $A = \{A_1, A_2, \ldots, A_N\}$ と対応する動画像データ $V = \{V_1, V_2, \ldots, V_N\}$ で構成される．音声データ $A_n \in \mathbb{R}^{T \times F}$ $n \in \{1, 2, \ldots, N\}$ はスペクトログラムで表され，T はフレーム数，F は周波数ビン数を示す．同様に，動画像データ $V_n \in \mathbb{R}^{T \times D \times H \times W}$ は，画像系列で表され，T がフレーム数，D がチャネル数，H と W がフレームの縦横の画素数を表す．本研究の目的は，音声・動画像及び時間的特徴を抽出し，マルチラベルの行動を予測することである．これには，で各フレーム t における行動ラベル $L_t \in \{0, 1\}^C$ を予測するタスクと，系列全体の行動ラベル $L \in \{0, 1\}^C$ を予測するタスクが含まれる．ここで，C は行動クラス数を表す．

4.1.2 MultiTSF の概要

提案手法は図 4 に示す 3 つの主要な構成要素からなる．マルチモーダル特徴抽出モジュールでは，音声と動画像から特徴を抽出する．人物検出モジュールでは，物体検出モデルを使用して人物を検出し，擬似正解ラベルを生成して人間の活動を学習する．Transformer に基づく時間モデリングと多視点統合モジュールでは，Transformer を用いて時間的依存関係を捉え，多視点の時空間特徴を統合する．最後に，フレーム，系列，人物損失関数を最適化し，高度な時空間的理解により正確な行動認識を実現する．

4.2 マルチモーダル特徴抽出

このモジュールでは，共有音声エンコーダと共有動画像エンコーダを用いて音声・動画像入力を処理する．全ての視点で同一のモデルパラメータを共有し，一貫性と効率的な特徴抽出を実現する．

4.2.1 共有音声エンコーダ

各視点 $n \in \{1, 2, \ldots, N\}$ の生音声信号 A_n は，ログメルスペクトログラムに変換され，時間的および周波数的特徴を捉える．変換後のログメルスペクトログラムは，共有音声エンコーダにより，音声特徴を抽出する．ここでは，音声エンコーダとして，Audio Spectrogram Transformer（AST）モデル [15] を利用する．AST モデルは，ログメルスペクトログラムをパッチに分割し，埋め込みを生成して Transformer 層で処理する．抽出された音声特徴は次式で表される：

$$F_n^{\mathrm{A}} = f_a(A_n), \quad F_n^{\mathrm{A}} \in \mathbb{R}^{T \times D_{\mathrm{A}}}. \tag{1}$$

ここで，$f_a(\cdot)$ は AST モデル，D_{A} は音声特徴の次元を示す．

4.2.2 共有動画像エンコーダ

共有動画像エンコーダを用いて，各視点 n の動画像 V_n から空間的特徴を抽出する．ここでは，動画像エンコーダとして

Vision Transformer（ViT）モデル [16] を利用する．ViT モデルは，各フレームを $P \times P$［画素］の非重複パッチに分割し，これをベクトル化して埋め込みに射影する．埋め込みは Transformer 層で処理され，空間的依存関係を学習する．抽出された視覚的特徴は次式で表される：

$$F_n^{\mathsf{V}} = f_v(V_n), \quad F_n^{\mathsf{V}} \in \mathbb{R}^{T \times D_{\mathsf{V}}}. \tag{2}$$

ここで，$f_v(\cdot)$ は ViT モデル，D_{V} は視覚的特徴の次元を表す．

4.2.3 音声・動画像特徴

音声特徴 F_n^{A} と動画像特徴 F_n^{V} は，各フレーム $t \in \{1, 2, \ldots, T\}$ ごとに結合され，次式のように統合された音声・動画像特徴を形成する：

$$F_n^{\mathsf{AV}}(t) = [F_n^{\mathsf{A}}(t); F_n^{\mathsf{V}}(t)]. \tag{3}$$

ここで，$[\cdot;\cdot]$ は特徴の結合を表す．音声・動画像特徴の統合系列は $F_n^{\mathsf{AV}} \in \mathbb{R}^{T \times D_{\mathsf{AV}}}$ で表され，$D_{\mathsf{AV}} = D_{\mathsf{A}} + D_{\mathsf{V}}$ である．統合特徴は時間モデリングおよび多視点統合段階でさらに処理される．

4.3 人物検出モジュール

空間的特徴学習を強化するため，人物損失関数 $\mathcal{L}_{\mathrm{H}}$ の擬似正解ラベルを生成する人物検出モジュールを導入する．各視点 $n \in \{1, 2, \ldots, N\}$ の動画像 V_n は，物体検出 You Only Look Once（YOLO）v10 モデル [17] に入力され，フレーム $t \in \{1, 2, \ldots, T\}$ ごとに以下の値を出力する：

$$h_n(t) = \begin{cases} 1, & \text{フレーム } t \text{ に人物が検出された場合,} \\ 0, & \text{それ以外の場合.} \end{cases} \tag{4}$$

全フレームの $h_n(t)$ は擬似正解ラベルとして，モデルが人物を含むフレームに集中し，無関係なフレームを無視するよう誘導する．これらのラベルは，訓練時に人物損失関数 $\mathcal{L}_{\mathrm{H}}$ で使用される．

4.4 Transformer に基づく時間モデリングと多視点統合

4.4.1 Transformer に基づく時間モデリング

音声・動画像特徴の時間的依存関係を捉えるため，共有時間エンコーダを使用する．エンコーダへの入力は，各視点 $n \in \{1, 2, \ldots, N\}$ の統合特徴 $F_n^{\mathsf{AV}}(t)$（$t \in \{1, 2, \ldots, T\}$）である．共有時間エンコーダは自己注意機構 [18] を用いて，視点内のフレーム間の時間的関係をモデル化する．出力は，次式のように各視点 n における時間的に強化された特徴の集合である：

$$F_n^{\mathsf{T}} = \left[F_n^{\mathsf{T}}(1), F_n^{\mathsf{T}}(2), \ldots, F_n^{\mathsf{T}}(T) \right] \in \mathbb{R}^{T \times D_{\mathsf{T}}}. \tag{5}$$

ここで，D_{T} は時間的特徴の次元数を表す．

4.4.2 Transformer に基づくの多視点統合

N 視点から抽出された時間的特徴 F_n^{T} を基に，Transformer を用いて複数視点の情報を統合し，統一表現を生成する．各フレーム t において，N 視点の時間的特徴は次式のように表される：

$$F^{\mathsf{T}}(t) = \left[F_1^{\mathsf{T}}(t), F_2^{\mathsf{T}}(t), \ldots, F_N^{\mathsf{T}}(t) \right] \in \mathbb{R}^{N \times D_{\mathsf{T}}}. \tag{6}$$

これらの特徴は自己注意機構に入力され，同じフレーム t における視点間の関係性をモデル化し，各視点の行動認識への貢献度に基づき動的に重要度を割り当てる．出力として，フレーム t に対応する統合された特徴 $F^{\mathsf{F}}(t) \in \mathbb{R}^{D_{\mathsf{F}}}$ が得られる．この処理を全フレーム T に対して繰り返すことで，統合された特徴の最終系列が次式のように得られる：

$$F^{\mathsf{F}} = \left[F^{\mathsf{F}}(1), F^{\mathsf{F}}(2), \ldots, F^{\mathsf{F}}(T) \right] \in \mathbb{R}^{T \times D_{\mathsf{F}}}. \tag{7}$$

4.5 目的関数

時空間及び視点間の特徴を最適化し，効果的な行動認識を実現するために学習目的を設計する．人物損失 $\mathcal{L}_{\mathrm{H}}$ は人間を含むフレームに注目し，フレーム損失 $\mathcal{L}_{\mathrm{F}}$ と系列損失 $\mathcal{L}_{\mathrm{S}}$ はそれぞれクラス不均衡に対応する．総損失関数は $\mathcal{L} = \beta_1 \mathcal{L}_{\mathrm{H}} + \beta_2 \mathcal{L}_{\mathrm{F}} + \beta_3 \mathcal{L}_{\mathrm{S}}$ で計算される．β_1，β_2，β_3 は各損失項の重要度を制御するハイパパラメータである．

4.5.1 人物損失関数 $\mathcal{L}_{\mathrm{H}}$

この損失関数は，モデルが多視点のフレームに人間が含まれているか効果的に識別することを目的とする．各視点 $n \in \{1, \ldots, N\}$ において，共有動画像エンコーダがフレーム単位の特徴を抽出する．この特徴は線形層を通じて処理され，フレーム $t \in \{1, 2, \ldots, T\}$ における人間の存在確率 $\hat{h}_n(t)$ を予測する．次式のように，Binary Cross Entropy 損失を使用して，予測確率 $\hat{h}_n(t)$ と擬似正解ラベル $h_n(t)$ の間の差を算出する：

$$\mathcal{L}_{\mathrm{H}} = -\frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left[h_n(t) \log \hat{h}_n(t) + \left(1 - h_n(t)\right) \log \left(1 - \hat{h}_n(t)\right) \right]. \tag{8}$$

4.5.2 フレーム損失関数 $\mathcal{L}_{\mathrm{F}}$

この損失関数は，$\mathcal{L}_{\mathrm{F}} = \mathcal{L}_{\mathrm{F}}^{\mathsf{S}} + \alpha_1 \mathcal{L}_{\mathrm{F}}^{\mathsf{C}}$（$\alpha_1$ はバランスパラメータ）として表される Two-way 損失関数 [19] である．ここでフレーム単位の行動クラス予測を最適化するため，サンプル単位損失 $\mathcal{L}_{\mathrm{F}}^{\mathsf{S}}$ とクラス単位損失 $\mathcal{L}_{\mathrm{F}}^{\mathsf{C}}$ を組み合わせる．サンプル単位損失 $\mathcal{L}_{\mathrm{F}}^{\mathsf{S}}$ はフレームごとの正例と負例を識別し，不均衡を緩和する．一方，クラス単位損失 $\mathcal{L}_{\mathrm{F}}^{\mathsf{C}}$ はクラス内変動を考慮して損失を調整する．各の損失は次式のように定義される：

$$\mathcal{L}_{\mathrm{F}}^{\mathsf{S}} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{softplus}\left(-\log \sum_{n|y_t=0} e^{x_{\mathsf{S}n}} + \gamma_s \log \sum_{p|y_t=1} e^{-\frac{x_{\mathsf{S}p}}{\gamma_s}} \right). \tag{9}$$

$$\mathcal{L}_{\mathrm{F}}^{\mathsf{C}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{softplus}\left(-\log \sum_{n|y_c=0} e^{x_{\mathsf{C}n}} + \gamma_c \log \sum_{p|y_c=1} e^{-\frac{x_{\mathsf{C}p}}{\gamma_c}} \right), \tag{10}$$

ここで，y は正解ラベルを表す，$x_{\mathsf{S}}, x_{\mathsf{C}}$ は各予測値を表し，γ は調整パラメータである．$\mathrm{softplus}(\cdot) = \log(1 + \exp(\cdot))$ は Rectified Linear Unit（ReLU）関数の滑らかな近似を表す．

4.5.3 系列損失関数 $\mathcal{L}_{\mathrm{S}}$

この損失関数は，T フレームにわたる時間的特徴を集約して系列単位の予測を最適化する．統合された系列特徴は分類ヘッドを通じて処理され，損失はフレーム損失関数と同様に，$\mathcal{L}_{\mathrm{S}} = \mathcal{L}_{\mathrm{S}}^{\mathsf{S}} + \alpha_2 \mathcal{L}_{\mathrm{S}}^{\mathsf{C}}$（$\alpha_2$ はバランスパラメータである）として Two-way 損失関数 [19] を用いて計算される．

5. 実　　験

5.1 実験条件

5.1.1 処理手順

提案する MultiSensor-Home データセットおよび既存の MM-Office データセット [8] において提案手法を評価する．データ分割には，Iterative Stratification [23]

表 1: MultiSensor-Home および MM-Office データセットにおける性能比較．最良値と次点をそれぞれ太字と下線で示す．

(a) MultiSensor-Home データセットにおけるシーケンスレベルとフレームレベルの行動認識結果．

手法	系列単位		フレーム単位	
	mAP_C	mAP_S	mAP_C	mAP_S
マルチモーダル（音声 + 動画像）				
MultiTrans [8]	59.65	77.60	61.40	78.07
MultiASL [14]	58.58	77.43	73.81	85.38
MultiTSF（提案手法）	**64.48**	**87.91**	**76.12**	**91.45**
ユニモーダル（動画像のみ）				
TimeSformer [20]	50.37	71.02	–	–
ViViT [21]	43.14	67.37	–	–
X-CLIP [22]	42.57	68.83	–	–
MultiTrans [8]	57.59	76.09	60.77	75.78
MultiASL [14]	55.91	77.25	63.24	80.64
MultiTSF（提案手法）	**61.17**	**84.22**	**75.07**	**87.31**

(b) MM-Office データセット [8] における系列単位の行動認識結果．

手法	ユニモーダル		マルチモーダル	
	mAP_C	mAP_S	mAP_C	mAP_S
TimeSformer [20]	69.68	79.26	–	–
ViViT [21]	73.25	83.05	–	–
X-CLIP [22]	65.38	78.54	–	–
MultiTrans [8]	73.85	85.24	75.35	85.35
MultiASL [14]	81.13	89.52	**86.23**	92.97
MultiTSF（提案手法）	**81.71**	**91.23**	85.65	**93.03**

良く含む学習データと評価データに分割する．MM-Office データセットは著者らの分割戦略 [14] を使用し，MultiSensor-Home データセットは 70:30 の比率で分割する．実験では，固定数 T の同期フレームを抽出し，動画像フレームを 2.5 FPS で均等に抽出する．対応する音声区間はフレームの時刻情報に基づいて調整する．訓練時には，無作為な摂動を含む均等サンプリングを用いて，動画像全体から固定長 T フレームを抽出する．この方法はデータ拡張としてモデルの頑健性を向上させる．テスト時には，摂動がない均等サンプリングを適用し，結果の一貫性を確保する．

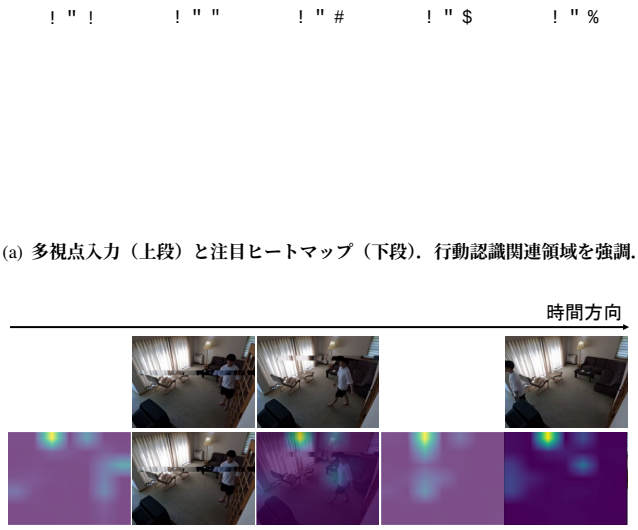
5.1.2 評 価 指 標

既存手法 [14], [19] に従い，mAP_C（マクロ平均）と mAP_S（マイクロ平均）の 2 つの mean Average Precision 指標で性能を評価する．mAP_C は各クラスの平均適合率を算出して全クラスで平均化するマルチラベル分類の主要指標であり，mAP_S は全サンプルを対象に平均適合率を計算するシングルラベル分類の標準的な指標である．

5.1.3 比 較 手 法

提案手法 MultiTSF(注2)を，多視点マルチモーダルおよび動画像行動認識の最先端手法と比較する．多視点マルチモーダル手法としては MultiTrans [8] と MultiASL [14] と比較する．動画像行動認識手法としては，時空間関係を捉える Transformer に基づく TimeSformer [20] と ViViT [21]，さらに動画像をモデル化する X-CLIP [22] と比較する．

(注2)：モデルの詳細は https://github.com/thanhhff/MultiTSF を参照されたい．



(a) 多視点入力（上段）と注目ヒートマップ（下段）．行動認識関連領域を強調．

(b) 動画像フレーム系列（上段）と注目ヒートマップ（下段）．時間経過による行動関連領域の変化を示す．

図 5: MultiSensor-Home データセットにおける Shared Visual Encoder の注意ヒートマップの可視化．

5.2 結　　果

5.2.1 定量的評価

表 1(a) および表 1(b) は，それぞれ MultiSensor-Home データセットと MM-Office データセットにおける MultiTSF と他手法の比較結果を示す．

まず，MultiSensor-Home データセット（表 1(a)）では，MultiTSF が全ての指標と設定で最良の性能を達成した．特に，系列単位の設定では mAP_C で 64.48%，mAP_S で 87.91%，フレーム単位の設定では mAP_C で 76.12%，mAP_S で 91.45% を記録し，MultiTrans および MultiASL を大きく上回った．ユニモーダル設定でも，TimeSformer，ViViT，X-CLIP を上回り，音声特徴がない場合でも高能を示した．

一方，MM-Office データセット（表 1(b)）でも，MultiTSF は系列単位の設定で優れた性能を維持した．ユニモーダル（動画像のみ）設定では，mAP_C で 81.71%，mAP_S で 91.23% を達成し，すべての比較手法を上回った．マルチモーダル設定では，mAP_S で 93.03% を記録し，MultiASL を超える結果を示した．

5.2.2 定性的評価

図 5 は，MultiSensor-Home データセットにおける共有動画像エンコーダの注目ヒートマップを示し，空間および時間方向の行動関連領域を特定する能力を示している．図 5(a) では，多視点入力と注目ヒートマップを並べて表示し，電灯やカーテン，机，人の存在など，行動認識に重要な領域に効果的に注目していることが分かる．図 5(b) は時間的注目ヒートマップを可視化し，時間経過に伴って環境領域の変化（人の動きや物との相互

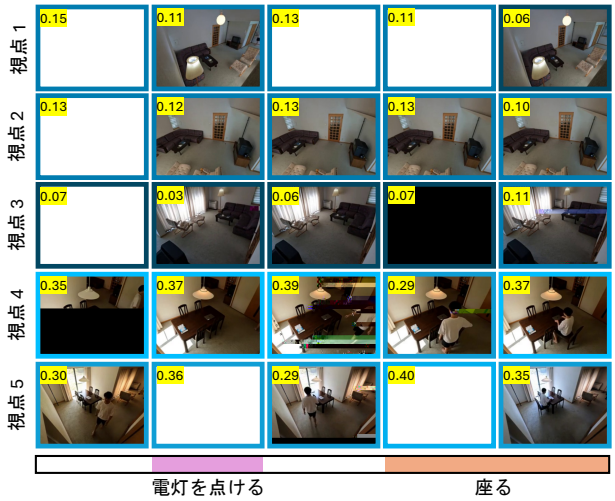


図 6: MultiSensor-Home データセットにおける Transformer に基づく多視点統合における注目スコアの可視化.

作用）を追跡する様子が分かる.

さらに，図 6 は Transformer に基づく多視点統合の注目スコアを可視化し，特定の行動（「電灯を点ける」や「座る」）に対して視点 4 と視点 5 が特に重要と判断されていることを示している．これにより，提案手法の多視点統合機構の有効性を確認できる.

6. むすび

本研究では，多視点マルチモーダル行動認識を対応するため，Transformer に基づく多視点統合 MultiTSF 手法と MultiSensor-Home データセットを提案した．提案データセットは，複数の行動の多視点マルチモーダルとフレーム単位の詳細なアノテーションを含む新しいベンチマークを提供する．提案手法は，Transformer により多視点の情報を効果的に統合し，フレームごとに人物検出を活用することで，提案の MultiSensor-Home および既存の MM-Office [8] データセットにおいて最先端の性能を実証した．今後の課題としては，効率的な展開のための軽量モデルの開発，データセットをさらに多様な家庭環境に拡張すること，ノイズが多い入力に対する頑健性の向上が挙げられる.

謝辞 本研究の一部は JSPS 科研費 JP21H03519 と JP24H00733 の助成を受けたものである.

文献

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, “Human action recognition from various data modalities: A review,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.45, no.3, pp.3200–3225, 2022.

[2] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” International Journal of Computer Vision, vol.130, no.5, pp.1366–1401, 2022.

[3] A.S. Olagoke, H. Ibrahim, and S.S. Teoh, “Literature survey on multi-camera system and its application,” IEEE Access, vol.8, pp.172892–172922, 2020.

[4] K. Shah, A. Shah, C.P. Lau, C.M. deMelo, and R. Chellappa, “Multi-view action recognition using contrastive learning,” Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, pp.3381–3391, 2023.

[5] S. Vyas, Y.S. Rawat, and M. Shah, “Multi-view action recognition using cross-view video prediction,” Proceedings of the 16th European Conference on Computer Vision, vol.27, pp.427–444, 2020.

[6] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, “Attention bottlenecks for multimodal fusion,” Advances in Neural Information Processing Systems, vol.34, pp.14200–14213, 2021.

[7] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, “Listen to look: Action recognition by previewing audio,” Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10457–10467, 2020.

[8] M. Yasuda, Y. Ohishi, S. Saito, and N. Harado, “Multi-view and multi-modal event detection utilizing transformer-based multi-sensor fusion,” Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4638–4642, 2022.

[9] M. Yasuda, N. Harada, Y. Ohishi, S. Saito, A. Nakayama, and N. Ono, “Guided masked self-distillation modeling for distributed multimedia sensor event analysis,” Computing Research Repository arXiv Preprints, arXiv:2404.08264, pp.1–13, 2024.

[10] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp.2649–2656, 2014.

[11] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp.1010–1019, 2016.

[12] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A.C. Kot, “NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.42, no.10, pp.2684–2701, 2019.

[13] V. John and Y. Kawanishi, “Frame-level latent embedding using weak labels for multi-view action recognition,” Proceedings of the 7th IEEE International Conference on Multimedia Information Processing and Retrieval, pp.235–238, 2024.

[14] T.T. Nguyen, Y. Kawanishi, T. Komamizu, and I. Ide, “Action selection learning for multilabel multiview action recognition,” Proceedings of the 2024 ACM Multimedia Asia Conference, pp.1–7, 2024.

[15] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” Computing Research Repository arXiv Preprints, arXiv:2104.01778, pp.1–5, 2021.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” Computing Research Repository arXiv Preprints, arXiv:2010.11929, pp.1–22, 2020.

[17] J. Redmon, “You only look once: Unified, real-time object detection,” Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp.1–10, 2016.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, vol.30, pp.6000–6010, 2017.

[19] T. Kobayashi, “Two-way multi-label loss,” Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7476–7485, 2023.

[20] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?,” Proceedings of the 38th International Conference on Machine Learning, pp.813–824, 2021.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “ViViT: A Video Vision Transformer,” Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, pp.6836–6846, 2021.

[22] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, “X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval,” Proceedings of the 30th ACM International Conference on Multimedia, pp.638–647, 2022.

[23] K. Sechidis, G. Tsoumakas, and I. Vlahavas, “On the stratification of multi-label data,” Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases, vol.3, pp.145–158, 2011.