Paper

Optimization of mid-roll advertisement placement in video content based on contextual analysis of multimodal data

Tomoki Haruyama†, Syunsuke Tsukatani† and Takuya Kitade†

Abstract: Video content delivered via video streaming services often includes mid-roll ads, a type of video advertising. While mid-roll ads are highly effective, they can disrupt the viewing experience of the video content and cause viewers to feel annoyed. To address this issue, methods have been proposed to optimize the placement of mid-roll ads by utilizing viewer information acquired by sensors attached to smartphones and by focusing on the distribution and changes of objects appearing in the video. However, obtaining sensor information is difficult in practice due to privacy and device constraints. Furthermore, focusing solely on the distribution and changes of objects appearing in the video does not provide a global understanding of the video context, and these methods are unable to be applied to long videos or take audio information into account. Therefore, in this study, we propose a method to automatically determine the placement of mid-roll ads without using non-visual information such as sensor information. By utilizing only multimodal data acquired from the video, we consider the video context using Transformer , and analyze audio information, we propose a method to automatically determine the placement of mid-roll ads without disrupting the viewing experience. Finally, through subject experiments, we confirmed that the proposed method can accurately determine ad positions that do not disrupt the viewing experience.

Keywords: Video Distribution Service, Advertisment Insertion, Mid-roll Advertisement, Video Understanding, Multimodal Analysis

1. Introduction

In 2019, Japan's internet advertising market surpassed television advertising in terms of advertising expenditures and has continued to grow ever since. This growth is largely due to the widespread use of smartphones and advances in high-speed communication technology, making internet advertising an increasingly important part of consumers' daily lives. 1) Internet advertising comes in a variety of forms, including banner ads, listing ads, and video ads. The market for mid-roll ads, which are video ads inserted into video content on video streaming services, has experienced particularly remarkable growth. This expansion trend is predicted to continue, with the overall video advertising market expected to exceed 1 trillion yen by 2027. 2) Given this background, academic research has focused on mid-roll ads, with active research focused on two key themes: "video characteristics that enhance advertising effectiveness" and "distribution methods that maximize the effect on purchasing behavior." 3)-7) The former approach aims to enhance advertising effectiveness through visual and content appeal . While the latter deals with the design of distribution strategies that focus on how viewers perceive mid-roll ads and how they affect their subsequent behavior, this study focuses on the latter and aims to determine the insertion position of mid-roll ads that is effective for advertisers without disrupting the viewing experience. Reference [4] points out that mid-roll ads disrupt the viewing experience of video content and cause viewers to feel aversion. Furthermore, it suggests that the insertion position of mid-roll ads, which is less likely to cause viewers to feel aversion, is where the context of the video is interrupted. Reference [5] also reports that when viewers feel aversion to mid-roll ads, their desire to purchase the introduced product or service tends to decrease. Therefore, inserting mid-roll ads at appropriate positions that do not interrupt the viewing of video content is effective in increasing viewers' desire to purchase. From the above, it can be considered that inserting mid-roll ads in a position that does not impair the viewer's viewing experience is beneficial for both viewers and advertisers.

On the other hand, in recent years, a huge amount of video content is being published on the Internet every day, so it is no longer possible to manually determine the insertion positions of mid-roll ads for all of these videos.

2025 年 1 月 14 日受付，2025 年 3 月 12 日再受付，2025 年 5 月 9 日採録
†株式会社 NTT ドコモ R&D イノベーション本部 クロステック開発部
（〒 100–6150 東京都千代田区永田町 2 丁目 11 番 1 号，TEL03-5156-1111）

マルチモーダルデータを活用した映像の文脈考慮に基づく動画コンテンツにおけるミッドロール広告の挿入位置最適化

Optimization of mid-roll advertisement placement in video content based on contextual analysis of multimodal data

正会員　春山知生†,　塚谷俊介†,　北出卓也†

Tomoki Haruyama†, Syunsuke Tsukatani† and Takuya Kitade†

あらまし　映像配信サービスにおいて配信される動画コンテンツには，動画広告の1種であるミッドロール広告が挿入されることがある．ミッドロール広告は，広告効果が高い一方で動画コンテンツの視聴体験を妨げる要因となり，視聴者に嫌悪感を抱かせることが問題とされている．そこで，スマートフォンに付帯するセンサにより取得される視聴者に関する情報を活用したり，動画中に映る物体の分布やその変化に着目することで，ミッドロール広告の挿入位置を最適化する手法が提案されている．しかしながら，センサ情報を取得することは，実際にはプライバシーやデバイス制約の観点で困難であることや，動画中に映る物体の分布やその変化への着目だけでは，映像の文脈を大域的に捉えることはできず，長尺映像への適用や音声情報の考慮ができていないことが課題として挙げられた．そこで，本研究では，センサ情報といった映像以外の情報は用いずに，映像から取得されるマルチモーダルデータのみを活用し，映像の文脈の考慮を Transformer を用いて行うことに加えて，音声情報を解析することで，視聴者の視聴体験を妨げないミッドロール広告の挿入位置を自動で決定する手法を提案する．最後に我々は，被験者実験を通じて，提案手法が視聴体験を妨げない広告位置の決定を高精度に行えることを確認した．

キーワード：映像配信サービス，広告挿入，ミッドロール広告，映像理解，マルチモーダル解析

1. ま え が き

日本のインターネット広告市場は，2019年にその広告費がテレビ広告を上回り，以降も成長を続けている．この成長の背景には，スマートフォンの普及と高速通信技術の進展が大きく寄与しており，インターネット広告は消費者の日常生活においてますます重要な存在となっている[1)]．インターネット広告には，バナー広告，リスティング広告，動画広告など様々な形態の広告が存在するが，特に，映像配信サービスにおける動画コンテンツに挿入される動画広告であるミッドロール広告の市場は目覚ましい成長を遂げている．今後もその拡大傾向は続くと予測され，2027年には動画広告市場全体の規模が1兆円を超える見通しである[2)]．

以上の背景から，学術的な研究においても，ミッドロール広告に関連する研究テーマは注目を集めており，特に「広告効果を高める動画の特徴」と「購入行動への効果を最大化する配信方法」という2つの主要なテーマに基づいて，活発な研究が行われている[3)~7)]．前者では，視覚的・内容的な魅力を通じて広告効果を高めることを目指したアプローチを扱う一方で，後者は，視聴者がミッドロール広告をどのように受け止め，その後の行動にどのような影響を与えるかに焦点を当てた配信戦略の設計に関わる研究である．本研究では，後者に注目し，視聴者が視聴体験を損なうことなく，広告主にとっても効果的なミッドロール広告の挿入位置を決定することを目的とする．

ここで，文献 [4] では，ミッドロール広告は動画コンテンツの視聴体験を妨げる要因となり，視聴者に嫌悪感を抱かせることを指摘している．さらに，視聴者が嫌悪感を感じにくいミッドロール広告の挿入位置として，映像の文脈の流れが途切れる箇所が適していることを示唆した．また，文献 [5] では，視聴者がミッドロール広告に対して嫌悪感を感じる場合，紹介された商品やサービスに対する購買意欲が低下する傾向があるため，動画コンテンツの視聴を中断させない適切な位置にミッドロール広告を挿入することが，視聴者の購買意欲を向上させるために有効的であることが報告された．以上より，視聴者の視聴体験を損なわないような位置にミッドロール広告を挿入することは，視聴者と広告主の双方にとってメリットであると考えられる．

一方で，近年では日々膨大な量の動画コンテンツがインターネット上に公開されるため，そのすべての動画に対して手作業でミッドロール広告の挿入位置を決定することは

2025年1月14日受付，2025年3月12日再受付，2025年5月9日採録
†株式会社 NTT ドコモ R&D イノベーション本部 クロステック開発部
（〒100–6150 東京都千代田区永田町2丁目11番1号，TEL03-5156-1111）

Therefore, in many cases, a method is adopted in which the insertion position is determined at regular intervals according to the length of the video content. 4) However, with such a method, mid-roll advertisements are inserted in scenes where the context of the video is not interrupted, such as during tense scenes or conversations, which often leads to complaints from viewers.

Therefore, when inserting mid-roll advertisements into video content, a method is required to automatically determine the position of advertisements so as not to impair the viewer's viewing experience.

Methods that utilize sensor information [6] and methods that focus on the distribution and changes of objects in a video [7] have been proposed. In [6], an accelerometer is used to measure changes in the viewer's posture and the optimal ad placement for the viewer is determined based on the measurement results. In [7], object detection is applied to every frame of video content to estimate not only the objects and their distribution but also their time-series changes. Based on the estimation results, the optimal ad placement is determined to be a point where the number of objects and object classes do not change rapidly. However, acquiring sensor information as in [6] is difficult in practice due to privacy and device constraints, making it unrealistic for practical applications. In addition, while [7] focuses on the distribution and changes of objects in a video, it is unable to grasp the global context of the video, and it is not possible to apply it to long videos such as movies and dramas, or to consider audio information.

Therefore, this study proposes a method to automatically determine mid-roll ad insertion positions that do not disrupt the viewer's experience. This is achieved by utilizing only multimodal data extracted from video—without employing non-visual information such as sensor data—while considering the video's context and analyzing audio information. Specifically, we assume positions where the video context naturally breaks as scene boundaries. By analyzing the video based on a Transformer[10]), we extract scene boundaries that consider the video context. Here, the Transformer addresses a limitation of Recurrent Neural Networks (RNNs)[8] and Long Short-Term Memory (LSTMs)[9]—representative techniques in natural language processing—where information degradation often occurs when analyzing long-range dependencies between tokens due to their recursive structure. By utilizing a self-attention mechanism, the Transformer enables comprehensive and efficient capture of dependencies between tokens. This characteristic enables analysis with minimal degradation even for long token sequences. Furthermore, the Transformer's design supports parallel computation, making it highly efficient at processing lengthy token sequences. In recent years, due to these properties, the Transformer has seen increasing application beyond natural language processing into fields such as image analysis and video analysis. In natural language processing, words within sentences were treated as tokens. However, by treating each patch resulting from image patch segmentation or the frames themselves as tokens, it becomes possible to consider the context of moving images. Therefore, utilizing Transformers is expected to enable video analysis that captures the global context of long-duration videos.

Furthermore, inserting advertisements where conversation or narration is taking place may cause viewers to feel annoyed. Therefore, in this study, we extract and analyze audio data from videos to extract sections containing conversation or narration (hereafter referred to as conversation sections). Ultimately, we identify candidate points for inserting mid-roll advertisements at scene boundaries that do not correspond to conversation sections. This allows us to determine the insertion positions of mid-roll advertisements so as not to disrupt the viewer's video viewing experience. In our experiments, we evaluate the effectiveness of the proposed method through subject experiments and analyze the impact of the proposed method on the viewing experience.

In Section 2, we explain the proposed method. We will now explain the experiments, and finally conclude in Section 4.

2. Proposed method

In this study, we propose a method to automatically determine the insertion position of mid-roll advertisements that does not disrupt the viewer's viewing experience by utilizing only multimodal data obtained from the video, without using information other than the video, such as sensor information, and by analyzing audio information in addition to considering the video context. An overview of the proposed method is shown in Figure 1. In the proposed method, scene boundaries are defined as positions where the video context naturally stops, and scene boundaries are extracted using TranS4mer11), a video analysis method based on Transformer10). Furthermore, conversation segments are extracted using inSpeechSegmenter18), a speech recognition method. Finally, points that are scene boundaries but do not correspond to conversation segments are set as candidate points for mid-roll advertisement insertion positions.

In Section 2.1, we explain how to extract scene boundaries, and in Section 2.2, we explain how to extract dialogue. Finally, in Section 2.3, we explain how to determine candidate points for inserting mid-roll advertisements.

2.1 Extracting Scene Boundaries

In this study, to take into account the context of long videos, we extract scene boundaries using TranS4mer11), a method that applies Transformer10) to video analysis. Before explaining the method in detail, we first define the terms frame, shot, scene, shot boundary, and scene boundary, as shown in Figure 2. Here, a shot is defined as a series of frames, and frames within a shot are assumed to be shot uninterrupted from the same camera position. Next, a series of shots constitutes a scene, and a scene is defined as a unit of context in video content, i.e., a story. A shot boundary is defined as the last frame of a shot at the moment when the shot ends and the next shot begins.

困難である．そのため，多くの場合，動画コンテンツの長さに応じて一定の間隔で挿入位置を決定する手法が採用されている[4]．しかしながら，そのような手法では，緊迫した場面や会話の途中など，映像の文脈が途切れないシーンにミッドロール広告が挿入されてしまい，視聴者から不満の声が寄せられることが多々ある．

そこで，動画コンテンツにミッドロール広告を挿入する際，視聴者の視聴体験を損なわないような位置を自動で決定する手法が求められている．先行研究として，スマートフォンに付帯するセンサにより取得した視聴者に関する情報を活用する手法[6] や動画中に映る物体の分布やその変化に着目する手法[7] が提案されている．文献 [6] では，加速度センサにより視聴者の姿勢変化を測定し，その測定結果から視聴者にとって最適な広告位置を決定する．また，文献 [7] では，動画コンテンツのすべてのフレームに対して物体検出を適用することで，動画中に映る物体やその分布に加えて，その時系列的な変化を推定する．その推定結果より，物体数や物体のクラスの変化が急激ではない点を最適な広告位置として決定する．しかしながら，文献 [6] のようにセンサ情報を取得することは，実際にはプライバシーやデバイス制約の観点で困難であり，実応用を考えると現実的ではない．また，文献 [7] では，動画中に映る物体の分布やその変化に着目しているが，映像の文脈を大域的に捉えることはできず，映画やドラマのような長尺映像への適用や音声情報の考慮ができていないことが課題として挙げられた．

そこで本研究では，センサ情報といった映像以外の情報は用いずに，映像から取得されるマルチモーダルデータのみを活用し，映像の文脈の考慮を行うことに加えて，音声情報を解析することで，視聴者の視聴体験を妨げないミッドロール広告の挿入位置を自動で決定する手法を提案する．具体的には，映像の文脈が自然に途切れる位置をシーン境界と仮定し，Transformer[10] に基づき動画を解析することで，映像の文脈を考慮したシーン境界の抽出を行う．ここで，Transformer は，自然言語処理の分野で代表的な手法である Reccurent Neural Network（RNN）[8] や Long Short-Term Memory（LSTM）[9] の，再帰的構造によりトークン間の長距離依存関係を解析する際に情報が劣化しやすいという課題を解決している．Transformer は自己注意機構 (Self-Attention) を活用することで，トークン間の依存関係を全体的かつ効率的に捉えることが可能である．この特性により，長いトークン列に対しても劣化の少ない解析が実現されている．さらに，Transformer は並列計算が可能な設計を持つため，長いトークン列を効率的に処理できる点でも優れている．近年では，このような特性から，Transformer は自然言語処理の領域に留まらず，画像解析や動画解析などの分野への応用が進んでいる．自然言語処理においては，文章における単語をトークンとして捉えていたが，画像をパッチ分割した際の各パッチやフレーム自体をトークンとして捉えることで，動画像の文脈を考慮することが可能となる．以上より，Transformer を活用することで，長尺映像の文脈を大域的に捉えた動画解析が可能となることが期待される．

また，会話やナレーションが行われている位置に広告が挿入されると，視聴者に嫌悪感を抱かせる可能性がある．そのため，本研究では，動画から音声データを抽出して解析することで，会話やナレーションが存在する部分（以降，会話部分）を抽出する．最終的には，シーン境界でありながら会話部分には該当しない箇所をミッドロール広告の挿入位置の候補点とする．以上より，視聴者の動画視聴体験を妨げないミッドロール広告の挿入位置の決定を実現する．実験では，提案手法の有効性を被験者実験により評価し，提案手法が視聴体験に与える影響を分析する．

以降，**2** 章では提案手法について説明する．続いて，**3** 章では実験について説明し，最後に **4** 章でむすびとする．

2. 提 案 手 法

本研究では，センサ情報といった映像以外の情報は用いず，映像から取得されるマルチモーダルデータのみを活用し，映像の文脈の考慮を行うに加えて，音声情報を解析することで，視聴者の視聴体験を妨げないミッドロール広告の挿入位置を自動で決定する手法を提案する．提案手法の概要図を図 **1** に示す．提案手法では，映像の文脈が自然に途切れる位置をシーン境界と定義し，Transformer[10] に基づく動画解析手法である TranS4mer[11] を用いてシーン境界を抽出する．さらに，音声認識手法である inSpeechSegmenter[18] を用いて，会話部分を抽出する．最終的に，シーン境界でありながら会話部分には該当しない箇所をミッドロール広告の挿入位置の候補点とする．

以降，**2.1** 節でシーン境界の抽出について説明を行い，**2.2** 節で会話部分の抽出について説明を行う．最後に，**2.3** 節でミッドロール広告の挿入位置の候補点の決定について説明を行う．

2.1 シーン境界の抽出

本研究では，長尺な映像の文脈を考慮するために，Transformer[10] を動画解析へ応用した手法である TranS4mer[11] を用いてシーン境界の抽出を行う．まず，手法の具体的な説明の前に，図 **2** で示すように，フレーム（Frame），ショット（Shot），シーン（Scene），ショット境界（Shot Boundary），シーン境界（Scene Boundary）について用語の定義を行う．ここでは，連続する複数のフレームがショットを構成し，同一ショット内のフレームは，同じカメラ位置から途切れることなく撮影されるものとする．次に，連続する複数のショットがシーンを構成し，シーンは動画コンテンツにおける文脈すなわちストーリーがひと段落するような単位とする．なお，ショット境界は，あるショットが終わって次のショットが始まる瞬間において，あるショットの最後のフレームと定義する．また，シーン境界は，あるシー

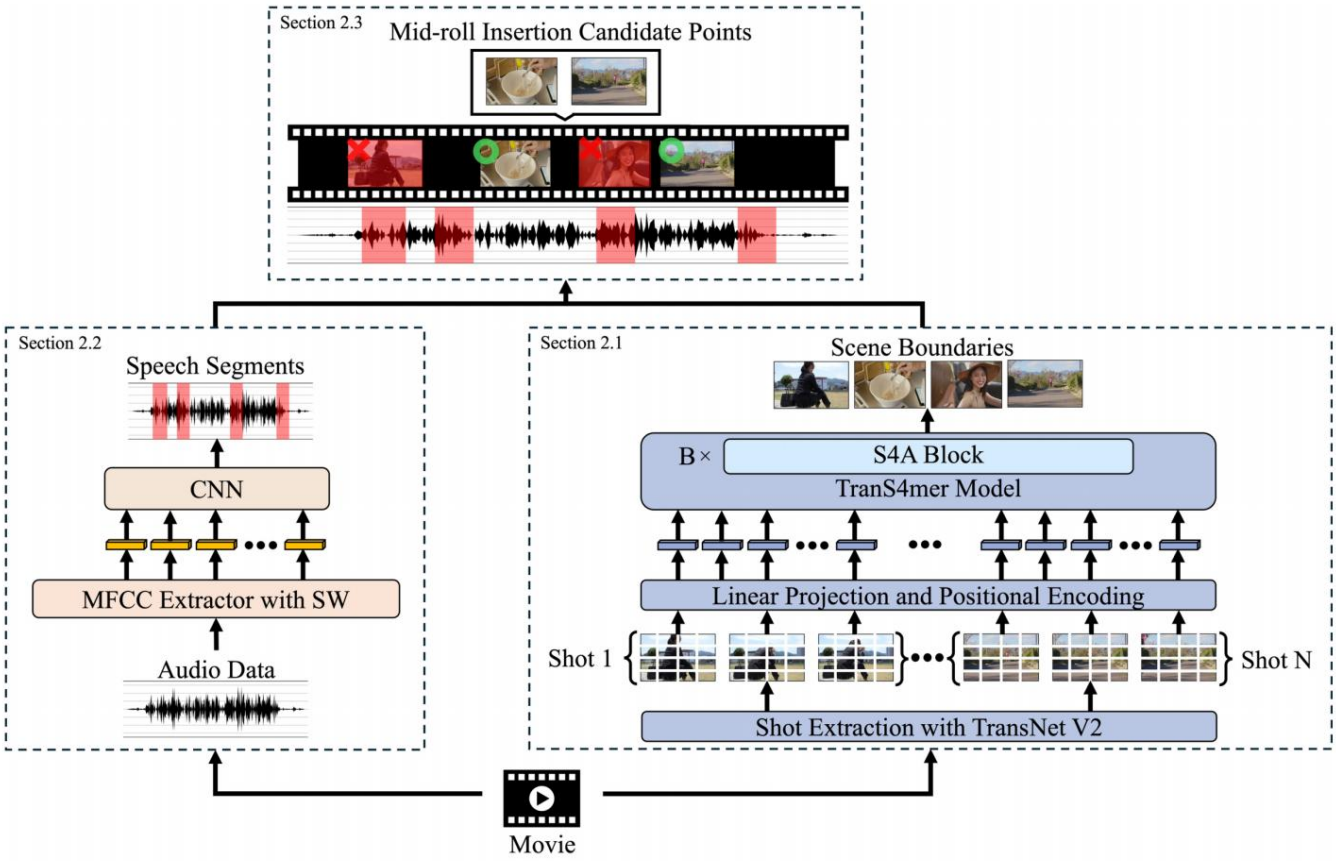


Figure 1. Overview of the proposed method

It is defined as the last frame of the last shot of a scene at the moment when the scene ends and the next scene begins.

2.1.1 Shot boundary extraction using TransNet V2

In this study, first, as preprocessing, we perform the following steps on the input video:

By applying TransNet V212, shot boundaries are extracted.

TransNet V2 is an improved version of TransNet13), which has a structure that processes data using a Deep Convolutional Neural Network14) , incorporating batch normalization to stabilize gradients and add noise during learning. It also combines skip connections and average pooling to reduce dimensions, thereby speeding up

calculations. Specifically, each frame of the input video is processed by TransNet. By inputting the data into V2 , feature values are extracted. Furthermore, the feature values of adjacent frames are compared to calculate similarity. Cosine similarity and RGB color histograms are used to calculate similarity. Furthermore, based on the change in similarity between the frame in question and the k frames before and after it, a total of 2k + 1 frames, the frame is classified as being a shot boundary or not. In doing so, the probability that the frame in question is a shot boundary is output, and if this exceeds a pre-set threshold ÿ , the frame in question is extracted as a shot boundary.

2.1.2 Scene boundary extraction using TranS4mer

In this study, we define Vi = {siÿm, …, si, … , si+m} as a set of N (= 2m + 1) consecutive shots extracted in Section 2.1.1.

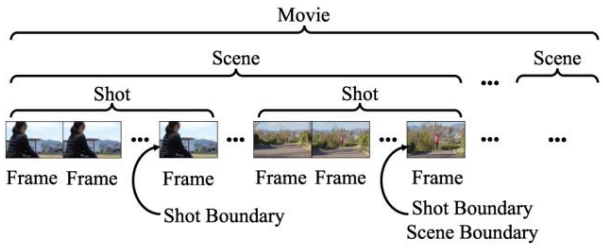


Figure 2 Definitions of Frame, Shot, Scene, Shot Boundary, and Scene Boundary in the proposed method

The input frames are input to TranS4mer, which predicts whether shot si is a scene boundary. Here, shot si being a scene boundary means that the last frame of shot si is the last frame of a scene, and a new scene begins from the next frame, as defined in Section 2.1. Also, Vi ÿ RN×K×3×H×W is composed of K frames extracted evenly from each of N consecutive shots . H and W represent the height and width of the frame. Here, the input frames are divided into P non-overlapping patches of size p×p according to the Vision Transformer15) , where P = HW/p2 . Next, each patch is input to a linear layer to convert it into a D -dimensional vector (Linear Projection in Figure 1). Position information is also added to each patch (Linear Projection in Figure 1) .

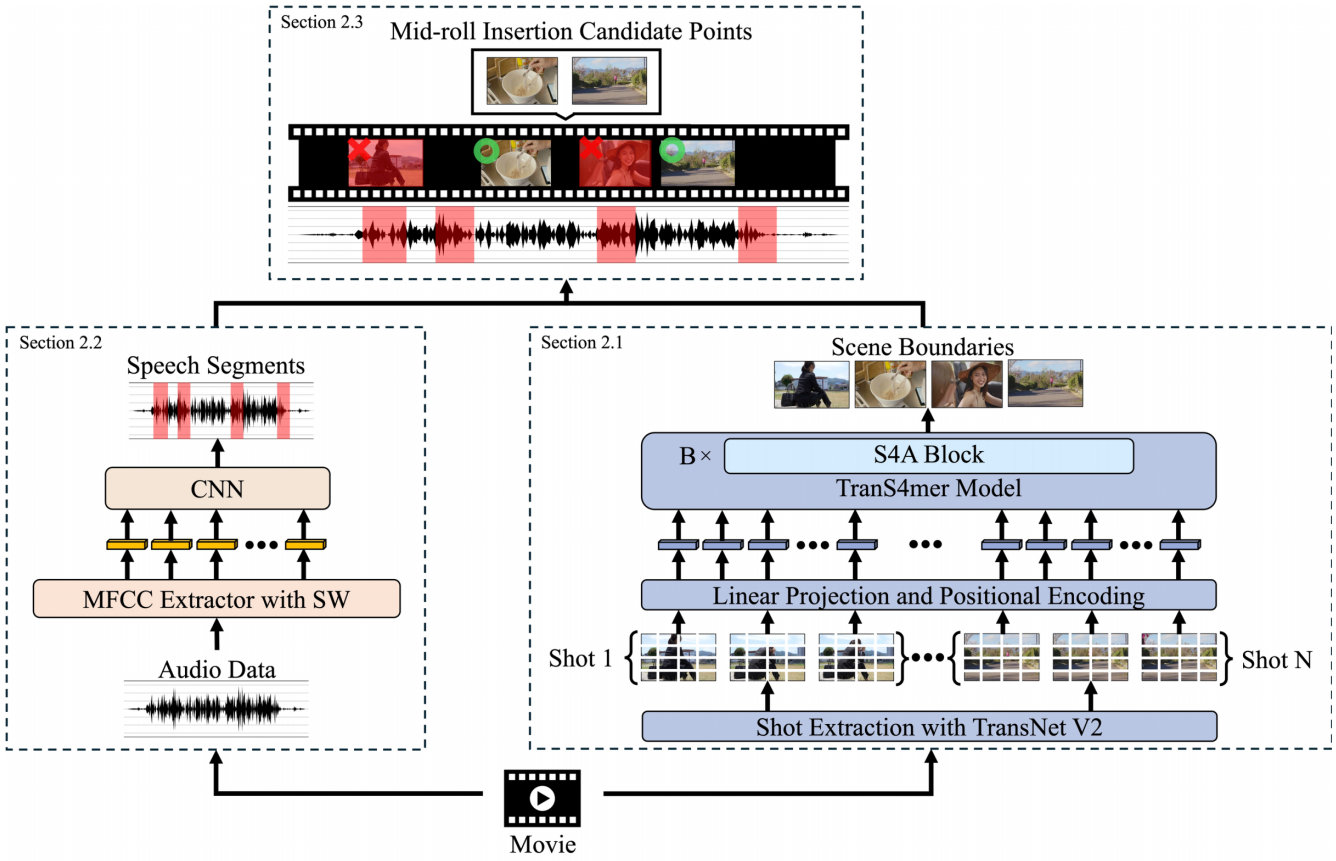


図 1　提案手法の概要図

ンが終わって次のシーンが始まる瞬間において，あるシーンの最後のショットにおける最後のフレームと定義する．

2.1.1 TransNet V2 によるショット境界の抽出

本研究では，まず，前処理として入力映像に対して TransNet V2[12] を適用することでショット境界の抽出を行う．TransNet V2 は，TransNet[13] を改良したモデルであり，Deep Convolutional Neural Network[14] で処理する構造を持ち，それらにバッチ正規化を組み込み，勾配の安定化や学習時にノイズの付与を行っている．また，スキップ接続と平均プーリングを組み合わせ，次元削減を行う事で計算の高速化を実現している．

具体的には，入力された動画の各フレームを TransNet V2 に入力することで，特徴量を抽出する．さらに，隣接するフレームの特徴量を比較し，類似度を算出する．類似度の算出には，コサイン類似度と RGB カラーヒストグラムを用いる．さらに，該当のフレームとその前後 k フレーム，合計 $2k+1$ フレームの類似度の変化に基づいて、そのフレームがショット境界か否かの分類を行う．その際，該当のフレームがショット境界である確率を出力し，あらかじめ設定した閾値 θ を超えた際に，該当のフレームをショット境界として抽出する．

2.1.2 TranS4mer によるシーン境界の抽出

本研究では，**2.1.1** 節で抽出された $N(=2m+1)$ 個の連続するショットの集合である $\mathcal{V}_i = \{s_{i-m}, \cdots, s_i, \cdots, s_{i+m}\}$

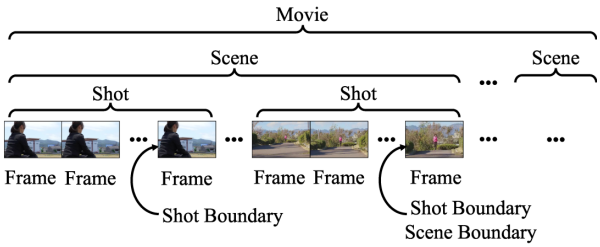


図 2　提案手法におけるフレーム（Frame），ショット（Shot），シーン（Scene），ショット境界（Shot Boundary），シーン境界（Scene Boundary）の定義

を TranS4mer へ入力し，ショット s_i がシーン境界か否かの予測を行う．ここで，ショット s_i がシーン境界であるということは，2.1 節で定義したように，ショット s_i における最後のフレームが，あるシーンの最後のフレームであり，その次のフレームから新しいシーンが始まることを意味する．また，$\mathcal{V}_i \in \mathbb{R}^{N\times K\times 3\times H\times W}$ は，N 個の連続するショットにおいて，それぞれのショットから均等に K 枚ずつ抽出したフレームから構成される．また，H および W はフレームの高さおよび幅を表す．ここで，入力されるフレームは，Vision Transformer[15] に従って，各フレームをサイズ $p\times p$ の P 個の重複しないパッチに分割する．なお，$P = HW/p^2$ である．次に，各パッチは線形層に入力され各パッチを D 次元のベクトルに変換する（図 **1** における Linear Projection）．また，各パッチに位置情報を追加する（図 **1** におけ

Finally , a tensor Vi ÿ RN ×K×P ×D is obtained, which is input to the TranS4mer model.

The TranS4mer model consists of B layers of S4A Blocks . As shown in Figure 3 , the S4A Blocks are divided into Intra-Shot Modules , which can express short-distance dependencies between shots , and Inter-Shot Modules , which can express long-distance temporal dependencies between shots. By stacking S4A Blocks with the same structure in Layer B and processing the output of the previous block as the input for the next block in a stepwise manner, it is possible to obtain features with high expressive power that capture short- and long-distance temporal dependencies. As described above, by expressing short- and long-distance temporal dependencies, scene boundaries can be extracted from long video with high accuracy.

The Intra-Shot Module takes as input Vi ÿ RN×K×P ×D , a collection of N shots . Here, each shot is represented by xL , where x is a frame, perceptron and xi ÿ RD, Si ÿ x1, ..., L = K × P. Next, a multi-layer (MLP) and residual connections are applied. After that, Multihead-Self-Attention (MSA) is applied. Note that a normalization layer, Layer Normalization (LN), is applied immediately before the MLP and MSA . The above can be formulated as follows:

$$x = \mathrm{MLP}(\mathrm{LN}(x_{in})) + x_{in} \quad x_{out} = \mathrm{MSA}(\mathrm{LN}(x)) + x \tag{1}$$

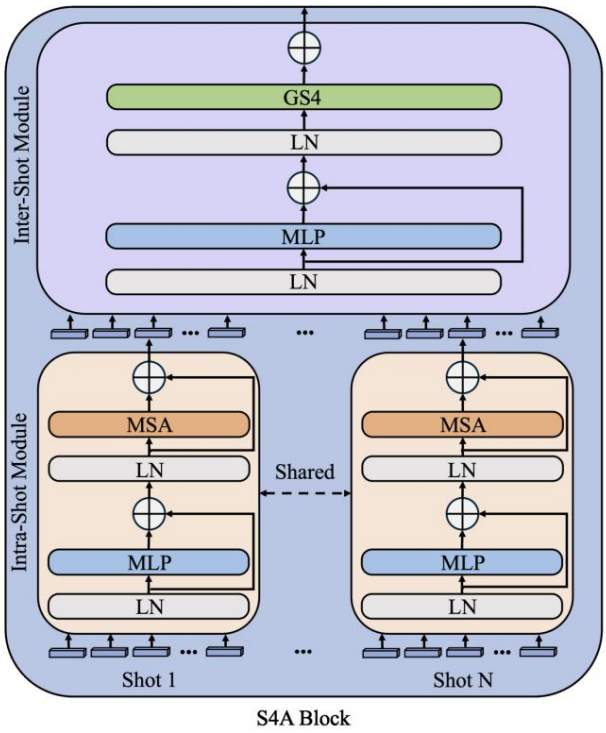
As a result, the Intra-Shot Module makes it possible to express short-distance dependencies.

Next, the output V of the Intra-Shot Module¡ ÿ RN×K×P ×D is input to the Inter-Shot Module . The Inter-Shot Module uses the Intra-Shot Allows long-distance dependencies to be expressed for modules Gated S4 (GS4)16) will be introduced. Specifically, first, Intra-Tensor V input from the Shot Module ¡ is expanded to z1, ..., zL, where zi ÿ RD and L = N × K × P. Next, a multilayer perceptron (MLP) and residual connection are applied, and then GS4 is applied. Note that LN is applied immediately before the MLP and GS4 . The above can be formulated as follows:

$$z = \mathrm{MLP}(\mathrm{LN}(z_{in})) + z_{in} \quad z_{out} = \mathrm{GS4}(\mathrm{LN}(z)) + z \tag{2}$$

Finally, zout is input to the next S4A . Compared to the Intra-Shot Module , whose sequence length is L = K × P , the Inter-Shot Module , whose sequence length is L = N × K × P , is able to express long-distance dependencies. 2.1.3 Training method and loss function

This section explains the training method and loss function of TranS4mer. Training of TranS4mer is performed in two stages: pre-training and incremental training, following the reference [11] .



S4A Block

Figure **3** S4A Block structure

Pre-training: The purpose of pre-training in TranS4mer is to learn the temporal features of the entire video. Self-supervised learning using pseudo-labels indicating whether a shot is a scene boundary is performed using two loss functions: Shot-Scene Contrastive Loss (SSCL) and Pseudo-Boundary Prediction Loss (PBPL) . Here, a shot being a scene boundary means that the last frame of a shot is the last frame of a scene, and a new scene begins from the next frame, as defined in Section 2.1. Below, we will explain SSCL and PBPL in detail. SSCL is based on the idea that shots in the same scene should have similar features, and shots in different scenes should have dissimilar features, and is expressed as follows:

$$\mathrm{IC}(r, \bar{r}) = -\log \frac{S(r, \bar{r})}{S(r, \bar{r}) + \sum_{rn} S(rn, r) + \sum_{rn} S(r, rn)} \tag{3}$$

Here, the set of input shots Vi = {siÿm, ..., si, ..., si+m} is divided into two pseudo scenes VL = {siÿm, ..., siÿ } and VR = {siÿ+1, ..., si+m} based on Dynamic Time Warping17) . A pseudo label of shot boundary is assigned to siÿ according to the definition in Section 2.1, and other shots are assigned pseudo labels that are not shot boundaries. In addition , in equation (3), r is a representation vector obtained by applying a linear transformation to the token placed at the beginning of a sequence of a certain shot (hereafter referred to as CLS token). Note that r is

る Positional Encoding）．最終的に，$\mathbf{V}_i \in \mathbb{R}^{N\times K\times P\times D}$ のテンソルが得られ，これが TranS4mer モデルに入力される．

TranS4mer モデルは B 個の S4A Block の層から構成される．S4A Block は図 **3** に示す通り，ショット同士の短距離的な依存関係を表現可能な Intra-Shot Module と，ショット同士の長距離的な時間的依存関係を表現可能な Inter-Shot Module を持つ．同一構造を持つ S4A Block を B 層重ねて，前のブロックの出力を次のブロックの入力として段階的に処理することで，短距離的および長距離的な時間的依存関係を捉えた表現能力の高い特徴量を得ることができる．以上より，短距離的および長距離的な時間的依存関係を表現することで，長尺映像からのシーン境界の抽出を高精度に実現する．

Intra-Shot Module は，N 個のショットの集合体である $\mathbf{V}_i \in \mathbb{R}^{N\times K\times P\times D}$ を入力とする．ここで，各ショットは $S_i \in x_1, ..., x_L$ で表される．x はフレームであり，$x_i \in \mathbb{R}^D$，$L = K \times P$ である．続いて，多層パーセプトロン（MLP）と残差接続を適用する．その後，Multihead-Self-Attention（MSA）を適用する．なお，正規化層である Layer Normalization（LN）は，MLP と MSA の直前に適用する．以上を定式化すると，下記のように表される．

$$\begin{aligned}\mathbf{x}' &= \mathrm{MLP}(\mathrm{LN}(\mathbf{x}_{in})) + \mathbf{x}_{in}\\ \mathbf{x}_{out} &= \mathrm{MSA}(\mathrm{LN}(\mathbf{x}')) + \mathbf{x}'\end{aligned} \tag{1}$$

以上より，Intra-Shot Module では，短距離的な依存関係を表現することが可能となる．

続いて，Intra-Shot Module の出力 $\mathbf{V}'_i \in \mathbb{R}^{N\times K\times P\times D}$ は，Inter-Shot Module に入力される．Inter-Shot Module では，短距離的な依存関係を表現可能としていた Intra-Shot Module に対して，長距離的な依存関係を表現可能とする Gated S4（GS4）[16] を導入する．具体的には，まず，Intra-Shot Module より入力されたテンソル $\mathbf{V}'_i$ を $z_1, ..., z_{L'}$ へ展開する．ここで，$z_i \in \mathbb{R}^D$ であり，$L' = N \times K \times P$ と表される．次に，多層パーセプトロン（MLP）と残差接続を適用し，さらに，GS4 を適用する．なお，LN は，MLP と GS4 の直前に適用する．以上を定式化すると，下記のように表される．

$$\begin{aligned}\mathbf{z}' &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z}_{in})) + \mathbf{z}_{in}\\ \mathbf{z}_{out} &= \mathrm{GS4}(\mathrm{LN}(\mathbf{z}')) + \mathbf{z}'\end{aligned} \tag{2}$$

最終的に，$\mathbf{z}_{out}$ は次の S4A へ入力される．シーケンス長が $L = K \times P$ である Intra-Shot Module と比較して，シーケンス長が $L' = N \times K \times P$ である Inter-Shot Module は，長距離的な依存関係の表現を可能となる．

2.1.3 学習方法および損失関数

本節では，TranS4mer の学習方法および損失関数について説明する．TranS4mer の学習は，文献 [11] に従って，事前学習と追加学習の 2 段階で行う．

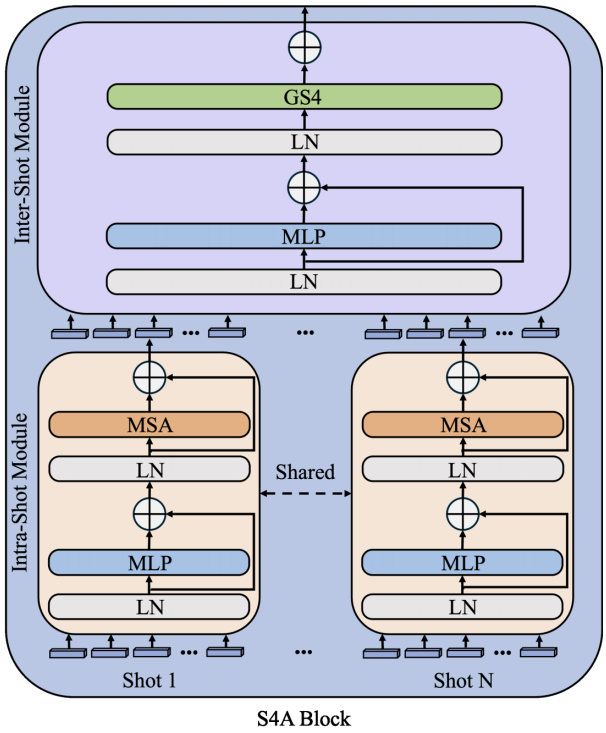


図 3　S4A Block の構造

事前学習

TranS4mer における事前学習では，映像全体の時間的特徴を学習することを目的とし，あるショットがシーン境界であるか否かの擬似ラベルを用いた自己教師あり学習を Shot-Scene Contrastive Loss（SSCL）と Pseudo-Boundary Prediction Loss（PBPL）の 2 つの損失関数を用いて行う．ここで，あるショットがシーン境界であるということは，2.1 節で定義したように，あるショットにおける最後のフレームが，あるシーンの最後のフレームであり，その次のフレームから新しいシーンが始まることを意味する．以降では，SSCL および PBPL について具体的に説明する．

SSCL は，同じシーンに含まれるショットは互いに類似する特徴を持つべきであり，異なるシーンに含まれるショットは類似しない特徴を持つべきであるという考えに基づき，下記の式で表される．

$$l_C(\mathbf{r}, \bar{\mathbf{r}}) = -\log \frac{\mathcal{S}(\mathbf{r}, \bar{\mathbf{r}})}{\mathcal{S}(\mathbf{r}, \bar{\mathbf{r}}) + \Sigma_{\mathbf{r}_n}\mathcal{S}(\mathbf{r}_n, \bar{\mathbf{r}}) + \Sigma_{\bar{\mathbf{r}}_n}\mathcal{S}(\mathbf{r}, \bar{\mathbf{r}}_n)} \tag{3}$$

ここで，入力されるショットの集合 $\mathcal{V}_i = \{s_{i-m}, ..., s_i, ..., s_{i+m}\}$ を Dynamic Time Warping[17] に基づき，擬似的に $\mathcal{V}_L = \{s_{i-m}, ..., s_{i^*}\}$ と $\mathcal{V}_R = \{s_{i^*+1}, ..., s_{i+m}\}$ の 2 つのシーンに分割する．s_{i^*} には，2.1 節の定義に従いショット境界の擬似ラベルが付与され，それ以外のショットにはショット境界ではない擬似ラベルが付与される．また，式（3）において，$\mathbf{r}$ はあるショットのシーケンスの先頭に配置されるトークン（以降，CLS トークン）に線形変換を適用して得られた表現ベクトルである．なお，$\bar{\mathbf{r}}$ は

It is the representation vector of the scene (VL or VR) that the shot belongs to , and the vector of all shots included in that scene.

It is calculated by averaging the CLS tokens. Also, rn is the representation vector of a shot other than the shot, and rn is the representation vector of a scene other than the scene, meaning that they are negative samples. Finally, the similarity between shots is calculated as S(x, y) = exp((xy)/(xy)) , and SSCL is summarized by the formula LC = IC (riÿm, rL) + IC (ri+m, rR) .

Next, PBPL is a loss function for predicting pseudo-labels and is expressed as follows using cross-entropy error:

$$LB = ÿlog(ÿb(riÿ)) ÿ log(1 ÿ ÿb(rb)) \quad (4)$$

Here, ri* is the CLS token of shot si* that has been pseudo-labeled as a scene boundary , and rb is the CLS token of shot sb that has been randomly selected and pseudo-labeled as a non-scene boundary . ÿb is a function that outputs the probability that the shot is a scene boundary. Finally, the loss function LP in pre-training is a function that outputs the probability that the shot is a scene boundary. Adding up the PBPLs gives the following:

$$LP = LC + LB \quad (5)$$

Additional learning

After pre-training in a self-supervised learning framework, The model is trained using ground truth data with shot labels that represent scene boundaries. The loss function uses the cross-entropy error, as in equation (4) described in pre-training.

2.2 Extraction of

Conversational Portions In this study, we use inaSpeechSegmenter18) to extract conversational portions. Specifically, audio data from video content is extracted using a sliding window method with a window width Q [ms] and a shift size T [ms] . Next, acoustic features are obtained by calculating the Mel-Frequency Cepstral Coefficient (MFCC) . These acoustic features are then input into a Convolutional Neural Network (CNN)19) to extract silence, conversation, and music portions. 2.3 Determining

Candidate Points for Mid-Roll Ad Insertion Positions The purpose of this study is to optimize the insertion position of mid-roll advertisements so as not to disrupt the viewer's viewing experience in video streaming services. Therefore, we select as candidate points for mid-roll advertisement insertion positions those that are scene boundaries extracted in Section 2.1 but do not correspond to the conversational portions extracted in Section 2.2.

3. Experiment

In this section, we explain the experiments to verify the effectiveness of the proposed method. In these experiments, we conduct subject experiments using actual drama footage. In the following, in Section 3.1 , we will explain the data used in these experiments. In Section 3.2 , we explain the experimental setup. Finally, in Section 3.3 , we present and discuss the experimental results.

3.1 Dataset

In this section , we explain the dataset used in the experiment .

We use MovieNet20) . MovieNet is a large-scale dataset containing 1,100 dramas and movies, each of which is labeled with scene boundary points. Following Section 2, we pre-train TranS4mer using the 1,100 videos and pseudo-labels in a self-supervised learning framework. We also perform additional training using 318 videos and the labels assigned to the dataset in a supervised learning framework.

3.2 Experimental

Settings In this section, we explain the hyperparameters, experimental method, and comparison method as experimental settings. First, we explain the hyperparameters .

For the extraction of the image, k = 50 is used, referencing 101 frames before and after each frame , and ÿ = 0.5. The number of elements in the set of shots input to TranS4mer is N = 25 , following [11]. Furthermore, the number of frames extracted from each shot is K = 3 , extracting three frames: the first, middle, and last frame of the shot . The video frames are resized to H = 224 and W = 224. Patch division is performed with a size of p = 32 , and the linear layer converts to D = 384 dimensions, with B = 12 S4A blocks . The batch size for pre-training and additional training is 256. The model is trained with a learning rate of 0.3 for 10 epochs during pre-training and a learning rate of 10 epochs during additional training . The optimization algorithm uses momentum of ^{6} 20 epochs as 0.9 and weight decay of 10 epochs . Finally, in inaSpeechSegmenter, processing was performed with a window ^{6} Adam21) width of Q = 25 [ms] and a shift size of T = 10 [ms] . Training and inference in this experiment were performed using a single Amazon Web Services (AWS) EC2 g4dn.4xlarge instance . Next, we explain the comparison methods used to verify the effectiveness of the proposed method.

Specifically, we use the following three comparison methods (CM1–3) . CM1: A method in which the video is divided equally into three parts and the division points are used as ad insertion positions. CM2: A method in which the video editor manually determines ad insertion positions. CM3: The proposed method does not consider audio information. The following explains the experimental method. In this experiment, we conducted a subject experiment to verify the effectiveness of the proposed method. First, we determined ad insertion positions for eight dramas (videos A–H) actually distributed on a video streaming service using the proposed method (PM) and CM1–3 . Here, one video is about 50 minutes long, equivalent to one episode of a drama .

そのショットが含まれるシーン（$\mathcal{V}_L$ または $\mathcal{V}_R$）の表現ベクトルであり，そのシーンに含まれるすべてのショットの CLS トークンを平均することで算出される．また，$\mathbf{r}_n$ はそのショットとは別のショット，$\bar{\mathbf{r}}_n$ はそのシーンとは別のシーンの表現ベクトルであり，それぞれネガティブサンプルであることを意味する．最終的に，ショット間の類似度は，$\mathcal{S}(\mathbf{x},\mathbf{y}) = \exp((\mathbf{x}^\top \mathbf{y})/(\|\mathbf{x}\|\|\mathbf{y}\|))$ で算出され，SSCL は $\mathcal{L}_C = l_C(\mathbf{r}_{i-m}, \bar{\mathbf{r}}_L) + l_C(\mathbf{r}_{i+m}, \bar{\mathbf{r}}_R)$ という式でまとめられる．

続いて，PBPL は，擬似ラベルを予測するための損失関数であり，クロスエントロピー誤差を用いて下記のように表される．

$$\mathcal{L}_B = -\log(\rho_b(\mathbf{r}_{i^*})) - \log(1 - \rho_b(\mathbf{r}_{\bar{b}})) \tag{4}$$

ここで，$\mathbf{r}_{i^*}$ は，シーン境界の擬似ラベルが付与されたショット s_{i^*} の CLS トークンであり，$\mathbf{r}_{\bar{b}}$ はランダムに選択された，シーン境界ではない擬似ラベルが付与されたショット $s_{\bar{b}}$ の CLS トークンである．また，ρ_b はそのショットがシーン境界である確率を出力する関数である．

最終的に，事前学習における損失関数 $\mathcal{L}_P$ は，SSCL と PBPL を足し合わせて下記で表される．

$$\mathcal{L}_P = \mathcal{L}_C + \mathcal{L}_B \tag{5}$$

追加学習

自己教師あり学習の枠組みで事前学習を行なった後，Ground Truth としてシーン境界であるショットのラベルを持つデータで，モデルの追加学習を行う．損失関数については，事前学習で説明した式（4）と同様にクロスエントロピー誤差を用いる．

2.2 会話部分の抽出

本研究では，会話部分の抽出に，inaSpeechSegmenter[18] を用いる．具体的には，動画コンテンツにおける音声データをウィンドウ幅 Q[ms] およびシフトサイズ T[ms] のスライディングウィンドウ方式で抽出する．続いて，それらに対して Mel-Frequency Cepstral Coefficient（MFCC）を算出することで音響特徴量を取得する．さらに，この音響特徴量を Convolutional Neural Network（CNN）[19] に入力することで，無音部分，会話部分，音楽部分の抽出を行う．

2.3 ミッドロール広告の挿入位置の候補点の決定

本研究の目的は，映像配信サービスにおける視聴者の視聴体験を妨げないために，ミッドロール広告の挿入位置を最適化することである．そのため，**2.1** 節で抽出したシーン境界でありながら，**2.2** 節で抽出した会話部分には該当しない箇所をミッドロール広告の挿入位置の候補点とする．

3. 実　験

本章では，提案手法の有効性を検証するための実験について説明を行う．本実験では，実際のドラマ映像を用いて被験者実験を行う．以降，**3.1** 節で，本実験で用いるデータセットについて説明を行い，**3.2** 節で，実験設定について説明を行う．最後に **3.3** 節で実験結果を示し，考察を行う．

3.1 データセット

本節では，実験で用いるデータセットについて説明を行う．**2** 章における TranS4mer の事前学習および追加学習には，MovieNet[20] を用いる．MovieNet は，1,100 本のドラマや映画を含む大規模なデータセットであり，それぞれの映像にラベルとして，シーンの境界点が付与されている．**2** 章に従い，1,100 本の映像と擬似ラベルにより TranS4mer を自己教師あり学習の枠組みで事前学習する．また，318 本の映像とデータセットに付与されているラベルを使用して，教師あり学習の枠組みで追加学習を行う．

3.2 実験設定

本節では，実験設定としてハイパーパラメータ，実験方法および比較手法について説明を行う．まず，ハイパーパラメータについて説明する．TransNet V2 でのショット境界の抽出においては，k=50 としてそれぞれのフレームの前後 101 フレームを参照し，θ=0.5 とする．また，TranS4mer に入力されるショットの集合の要素数は，文献 [11] に従い N=25 とする．さらに，それぞれのショットから抽出されるフレーム数は K=3 として，ショットにおける先頭のフレーム，中間のフレーム，最終のフレームの 3 枚を抽出する．映像のフレームは，H=224，W=224 となるようにリサイズを行う．なお，パッチ分割は p=32 のサイズで行い，線形層では D=384 次元へ変換を行い，S4A Block の数は B=12 とする．また，事前学習および追加学習時のバッチサイズは 256 とし，事前学習時は学習率を 0.3 として 10 エポック，追加学習時は学習率を 10^{-6} として 20 エポックでモデルの学習を行なう．また，最適化アルゴリズムには，モメンタムが 0.9，重み減衰が 10^{-6} の Adam[21] を用いる．最後に，inaSpeechSegmenter においては，ウィンドウ幅 Q=25[ms] およびシフトサイズ T=10[ms] として処理を行う．なお，本実験における学習および推論は，Amazon Web Services (AWS) EC2 の g4dn.4xlarge インスタンスを 1 つ使用して行う．

続いて，提案手法の有効性を検証するために用いる比較手法について説明する．具体的には，以下に示す 3 種類の比較手法 (CM1–3) を用いる．

CM1: 動画を均等に 3 分割し，それら分割点を広告挿入位置とする手法．

CM2: 映像編集者が手動で広告挿入位置を決定する手法．

CM3: 提案手法で音声情報の考慮を行わない手法．

以降では，実験方法について説明を行う．本実験では，提案手法の有効性を検証するために被験者実験を行う．まず，実際に映像配信サービスで配信されているドラマ 8 本（動画 A–H）に対して，提案手法（PM）および CM1–3 を用いて広告挿入位置の決定を行う．ここで，1 本の動画は約 50 分のドラマ 1 話分であり，1 本の動画に対してそれぞれ

Two ad insertion points are determined using this method . The number of ad insertion points is set to match the actual video distribution service. In this case, the proposed method selects two ad insertion points that are more than 10 minutes apart from the final extracted candidate points, in order of the highest probability of being a scene boundary output by TranS4mer in Section 2.1.2 . The subjects then perform the following two evaluations of the ad positions in each video. Evaluation 1: Was this ad position in a position that did not interfere with your viewing experience when viewing the video?

Rating 2: This ad position is in the context of the video.

I guess it was a dividing line between us.

Each evaluation was done on a three- point scale: "1 Not," "2 Neither," or "3 Yes," with a higher number indicating a better evaluation. The subjects in this experiment were 15 people, 3 women and 12 men , aged 26–49 .

Finally, we measure the processing time when applying the proposed method to videos A–H on an AWS EC2 g4dn.4xlarge instance . By comparing the results with the time and cost required to manually determine the ad insertion positions, we consider the feasibility of applying the proposed method to practical applications.

3.3 Experimental

Results This section shows the results of the subject experiments. Tables 1 and 2 show the average evaluation results of CM1-3 and PM for Evaluation 1 and Evaluation 2. First, the average evaluation results for PM were the highest for both Evaluation 1 and Evaluation 2 .

The effectiveness of the proposed method was confirmed by these results. Furthermore, the accuracy of CM3 was comparable to that of CM2 , confirming the effectiveness of considering video context based on Transformer. Furthermore , the accuracy of PM exceeded that of CM3 , confirming the effectiveness of considering audio information. Specifically , CM3 occasionally determined ad insertion locations in scenes where the visual scene changed but where narration or character conversation continued. On the other hand, PM, which includes a process to exclude such conversational parts from ad insertion locations (Section 2.3), suggests that it was possible to determine ad insertion locations that do not disrupt the viewer's viewing experience. Furthermore , the accuracy of PM surpassed that of CM2 , a method in which video editors manually determine ad insertion locations . This is because PM can objectively and consistently analyze scene development and audio information, whereas video editors' judgments depend on experience and intuition and may be influenced by subjectivity and bias. In addition, for video F , the evaluation results for CM2 were

It was confirmed that this was significantly higher than that of PM and CM3 . This is thought to be because video F contained many flashback scenes. In CM3, the ad insertion position is determined to be the scene where the scene changes visually, while in PM, the ad insertion position is determined to be the part where no conversation occurs. Therefore, there were cases where the moment of transition to a flashback scene was also determined to be the ad insertion position. In flashback scenes, it is important to have a narrative connection with the scenes before and after. Therefore, in order not to disrupt the viewer's viewing experience, it is desirable not to judge such transitions to flashback scenes as ad insertion positions. To deal with such cases, it is necessary to consider the video context in more detail in the future, such as by evaluating the relevance of scenes by analyzing the content of conversations in the video content.

Table 3 also shows the correlation coefficient between the evaluation results of Evaluation 1 and Evaluation 2. Table 3 confirms that there is a high correlation between Evaluation 1 and Evaluation 2. This confirms the assumption that the ad position that does not disrupt the viewing experience is the break between scenes in the video. It was suggested that this is

Finally, we consider the feasibility of the proposed method for practical use. Table 4 shows the video durations of videos A–H and the processing time required for inference in the experimental environment. Table 4 shows that the processing time for each video was approximately half the video duration. This suggests that the proposed method can determine ad insertion positions faster than manually determining ad insertion positions by actually watching the video. Furthermore, the hourly on-demand cost of an AWS EC2 g4dn.4xlarge instance in the Asia Pacific (Tokyo) region is $ 1.204 (as of April 2025) * , which is less than typical labor costs. Note that when building a system using AWS in this use case, the cost of the EC2 instance dominates, so other costs can be ignored. From the above, we conclude that the proposed method is feasible for practical use because it is believed to determine ad insertion positions faster and more inexpensively than manual methods.

Table **1.** Average evaluation results for CM1–3 and PM for Evaluation 1 .

	CM1	CM2	CM3	PM
動画 A	1.20	2.10	2.00	**2.20**
動画 B	1.67	2.23	2.13	**2.67**
動画 C	1.33	**2.77**	2.33	2.67
動画 D	1.37	2.87	2.67	**3.00**
動画 E	1.13	**2.77**	2.44	2.73
動画 F	1.17	**2.70**	1.77	2.30
動画 G	1.50	1.73	2.10	**2.50**
動画 H	1.40	1.77	1.17	**1.93**
平均	1.35	2.37	2.08	**2.50**

Table **2.** Average evaluation results for CM1–3 and PM for Evaluation 2 .

	CM1	CM2	CM3	PM
動画 A	1.13	2.10	1.90	**2.40**
動画 B	1.57	2.30	2.27	**2.67**
動画 C	1.27	**2.83**	2.40	**2.83**
動画 D	1.13	2.97	2.70	**3.00**
動画 E	1.03	2.90	2.67	**2.93**
動画 F	1.03	**2.80**	2.27	2.27
動画 G	1.50	1.77	1.77	**2.70**
動画 H	1.27	1.90	1.77	**1.93**
平均	1.24	2.45	2.22	**2.59**

* https://aws.amazon.com/jp/ec2/pricing/on-demand/

の手法で 2 つずつ広告挿入点を決定する．なお，広告挿入点の数は実際の映像配信サービスに合わせて設定した．この際，提案手法においては，最終的に抽出された候補点のうち，**2.1.2** 節で TranS4mer により出力されるシーン境界である確率が高いものから順に，それぞれが 10 分以上離れるように 2 つ抽出する．ここで，被験者は，それぞれの映像における広告位置に対して，下記の 2 つの評価を行う．

評価 1： この広告位置は，本動画を閲覧する上であなたの視聴体験を妨げない位置であったか．

評価 2： この広告位置は，本動画における文脈すなわちの区切りになっていたか．

それぞれの評価は，“1 そうでない”，“2 どちらでもない”，“3 そうである” の 3 段階で行い，数値が高い方が評価が良いことを表す．なお，本実験における被験者は 15 名であり，女性 3 名，男性 12 名，年齢は 26–49 歳である．

最後に，AWS EC2 の g4dn.4xlarge インスタンスにおいて，映像 A–H に対して提案手法を適用した場合の処理時間を計測する．その結果と手動で広告挿入位置を決定する場合にかかる時間や費用を比較することで，提案手法を実用化する上での妥当性について考察を行う．

3.3 実験結果

本節では，被験者実験の結果を示す．表 **1** および表 **2** に，評価 1 および評価 2 に対する CM1–3 および PM の評価結果の平均値を示す．まず，PM における評価結果の平均値が評価 1 および評価 2 のいずれにおいても最も高いことから，提案手法の有効性が確認された．また，CM3 の精度が CM2 のそれに匹敵していることから，Transformer に基づく映像文脈の考慮を行うことの有効性が確認された．さらに，PM の精度が CM3 のそれを上回っていることから，音声情報の考慮を行うことの有効性が確認された．具体的には，CM3 では，視覚的にはシーンが切り替わる場面であるもののナレーションや登場人物の会話が続いている場面を広告挿入位置と決定するケースが散見された．一方で，PM では，2.3 節においてそのような会話部分を広告挿入位置から除外する処理が含まれるため，視聴者の視聴体験を妨げない広告挿入位置の決定が実現したことが示唆される．また，PM の精度が映像編集者が手動で広告挿入位置を決定する手法である CM2 の精度を上回っているのは，PM がシーン展開や音声情報を客観的かつ一貫して解析することができるためであり，これに対して映像編集者の判断は経験や感覚に依存し，主観やバイアスが影響を及ぼす可能性があるからと考察する．なお，動画 F においては，評価 1 および評価 2 いずれの場合においても CM2 の評価結果が PM や CM3 のそれを大きく上回っていることが確認された．これは，動画 F には回想シーンが多く含まれていたことが原因であると考えられる．CM3 では視覚的にシーンが切り替わる場面を，PM ではその上で会話が発生しない部分を広告挿入位置と決定するため，回想シーンへの転換する瞬間も広告挿入位置として判定してしまうケースが存在した．回想シーンは，その前後のシーンとの物語的な繋がりが重要である．そのため，視聴者の視聴体験を妨げないためには，このような回想シーンへの転換を広告挿入位置として判定しないことが望ましい．このようなケースに対応するために，今後は，映像コンテンツ中の会話の内容を解析することでシーン同士の関連性を評価するなど，より詳細に映像文脈を考慮する必要があると考えられる．

また，表 **3** に評価 1 と評価 2 の評価結果における相関係数を示す．表 **3** より，評価 1 と評価 2 に高い相関があることが確認された．このことから，視聴体験を妨げない広告位置は動画におけるシーンの切れ目であるという仮定が正しいことが示唆された．

最後に，提案手法を実用化する上での妥当性について考察する．表 **4** に動画 A–H の動画時間長および実験環境上での推論にかかる処理時間を示す．表 **4** より，各映像の処理時間は動画時間長の半分以下程度となっていることがわかる．これより提案手法では，実際に映像を視聴して手動で広告挿入位置を決定する場合よりも速く広告挿入位置の決定が行えると考えられる．また，AWS EC2 の g4dn.4xlarge インスタンスは，アジアパシフィック（東京）リージョンにおいてオンデマンドの時間単価が 1.204 ドル（2025 年 4 月現在）である * ことから，一般的な人件費よりも安価であることがわかる．なお，今回のユースケースにおいて AWS を利用してシステムを構築する際，EC2 インスタンスの利用料金が支配的であるため，その他の要素にかかる費用は無視できるものとする．以上より，人手での作業より速くかつ安価に広告挿入位置の決定ができると考えられるため，提案手法は実用化する上での妥当性があると結論づけられる．

表 1 評価 1 に対する CM1–3 および PM の評価結果の平均値．

	CM1	CM2	CM3	PM
動画 A	1.20	2.10	2.00	**2.20**
動画 B	1.67	2.23	2.13	**2.67**
動画 C	1.33	**2.77**	2.33	2.67
動画 D	1.37	2.87	2.67	**3.00**
動画 E	1.13	**2.77**	2.44	2.73
動画 F	1.17	**2.70**	1.77	2.30
動画 G	1.50	1.73	2.10	**2.50**
動画 H	1.40	1.77	1.17	**1.93**
平均	1.35	2.37	2.08	**2.50**

表 2 評価 2 に対する CM1–3 および PM の評価結果の平均値．

	CM1	CM2	CM3	PM
動画 A	1.13	2.10	1.90	**2.40**
動画 B	1.57	2.30	2.27	**2.67**
動画 C	1.27	**2.83**	2.40	**2.83**
動画 D	1.13	2.97	2.70	**3.00**
動画 E	1.03	2.90	2.67	**2.93**
動画 F	1.03	**2.80**	2.27	2.27
動画 G	1.50	1.77	1.77	**2.70**
動画 H	1.27	1.90	1.77	**1.93**
平均	1.24	2.45	2.22	**2.59**

* https://aws.amazon.com/jp/ec2/pricing/on-demand/

Table 3: Correlation coefficients between the evaluation results of Rating 1 and Rating 2.

	A	B	C	D	E	F	G	H	
相関係数	0.74	0.78	0.91	0.89	0.71	0.67	0.64	0.76	0.76

Table 4: Video duration and processing time in the experimental environment for videos A–H .Note that the units are [minutes]:[seconds].

動画	A	B	C	D	E	F	G	H
動画時間長	44:33	45:21	46:28	46:51	46:23	52:30	53:28	53:57
処理時間	17:14	20:05	21:47	20:28	19:53	23:50	25:10	26:46

4. Conclusion

In this study, we proposed a method to optimize the insertion position of mid-roll ads in video streaming services without disrupting the viewer's viewing experience. Our proposed method utilizes multimodal data and considers the video context to accurately extract candidate mid-roll ad positions. Experiments confirmed the effectiveness of the proposed method for long videos such as dramas. Specifically, the proposed method extracted ad insertion positions that do not disrupt the viewer's viewing experience, compared to manual insertion positions selected by a video editor. However, the proposed method does not consider the content of the mid-roll ads to be inserted, and we have not yet verified its effectiveness for documentaries, animations, and other media that are not included in the training data. These are issues that remain to be addressed. Therefore, in addition to applying our method to various genres, we will work on extracting insertion positions that take the content of mid-roll ads into account and automatically selecting mid-roll ads that match the video context before and after the insertion position.

[文献]

1) 総務省: “情報通信白書令和 6 年版, ”https://www.soumu.go.jp/johotsusintokei/whitepaper/r06.html (2024)
2) 株式会社サイバーエージェント: “2023 年国内動広告の市場調査,”https://www.cyberagent.co.jp/news/detail/id=29827 (2024)
3) U. Bulkan, T. Dagiuklas, and M. Iqbal: “Modelling quality of experience for online video advertisement insertion,” *IEEE Transactions on Broadcasting*, **66**, 4, pp. 835–846 (2020)
4) 齊藤 義仰, 畠山 智裕, 西岡 大: “スマートフォンを用いた動画広告挿入タイミング決定アルゴリズムの提案,” マルチメディア, 分散, 協調とモバイルシンポジウム *(DICOMO)*, **1**, pp. 1089–1095 (2017)
5) 遠藤 茉樹, 加藤 拓巳: “YouTube における動画広告挿入のタイミングによる記憶と購入意向への影響,” マーケティングレビュー, vol. 4 (2022)
6) Y. Saito: “A Proposal of a method for video advertisement insertion on smartphone,” *International Journal of Informatics Society*, **11**, 1, pp. 55–62 (2019)
7) U. Bulkan, T. Dagiuklas, and M. Iqbal: “Supereye: Smart advertisement insertion for online video streaming,” *Multimedia Tools and Applications (MTA)*, **82**, 6, pp. 9361–9379 (2023)
8) JL. Elman: “Finding structure in time,” *Cognitive Science*, **14**, 2, pp. 179–211 (1990)
9) S. Hochreiter: “Long short-term memory,” *in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, **9**, 8, pp. 1735–1780 (1997)
10) A. Vaswani, N. Shazeer, et al: “Attention is all you need,” *in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, **30** (2017)
11) M. Islam, M. Hasan, et al: “Efficient movie scene detection using state-space transformers,” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18749–18758 (2023)
12) T. Souček and J. Lokoc: “Transnet v2: An effective deep network architecture for fast shot transition detection,” *in Proceedings of ACM International Conference on Multimedia (ACMMM)*, pp. 11218–11221 (2024)
13) J. Lokoc, G. Kovalčík, et al: “A framework for effective known-item search in video,” *in Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pp. 1777–1785 (2019)
14) J. Donahue, Y. Jia, et al: “Decaf: A deep convolutional activation feature for generic visual recognition,” *in Proceedings of the International Conference on Machine Learning (ICML)*, pp. 647–655 (2014)
15) A. Dosovitskiy, L. Beyer, et al: “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929* (2020)
16) H. Mehta, A. Gupta, et al: “Long range language modeling via gated state spaces,” *arXiv preprint arXiv:2206.13947* (2022)
17) S. Hiroaki and S. Chiba: “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **26**, 1, pp. 43–49 (1978)
18) D. Doukhan, E. Lechapt, et al: “Ina’s mirex 2018 music and speech detection system,” *in Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)* (2018)
19) Y. Lecun, L. Bottou, et al: “Gradient-based learning applied to document recognition,” *Proceedings of The IEEE*, **86**, 11, pp. 2278–2324 (1998)
20) Q. Huang, Y. Xiong, et al: “Movienet: A holistic dataset for movie understanding,” *in Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 709–727 (2020)
21) D. Kingma and J. Ba: “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980* (2014)

春山 知生 2021 年 3 月，北海道大学大学院情報科学院 メディアネットワーク専攻 修士課程 修了．2021 年 4 月，株式会社 NTT ドコモ 入社．NTT ドコモにて動画像認識技術の，スポーツやスマートモビリティ分野への社会実装業務に従事，現在に至る．正会員．

塚谷 俊介 2016 年 3 月 東京大学大学院情報理工学系研究科電子情報学専攻 修士課程 修了．2016 年 4 月 日本電信電話株式会社 入社．2023 年 10 月 株式会社 NTT ドコモに転籍，現在に至る．

北出 卓也 2013 年，慶應義塾大学大学院理工学研究科開放環境科学専攻 修士課程 修了．2013 年，株式会社 NTT ドコモ 入社．2023 年，慶應義塾大学大学院理工学研究科開放環境科学専攻 博士課程 入学．NTT ドコモにて画像認識及びロボティクス技術の，農業やスマートシティの領域への社会実装業務に従事，現在に至る．

表 3 評価 1 と評価 2 の評価結果における相関係数.

動画	A	B	C	D	E	F	G	H	平均
相関係数	0.74	0.78	0.91	0.89	0.71	0.67	0.64	0.76	0.76

表 4 動画 A–H の動画時間長および実験環境上での処理時間．なお，単位は [分]:[秒] とする．

動画	A	B	C	D	E	F	G	H
動画時間長	44:33	45:21	46:28	46:51	46:23	52:30	53:28	53:57
処理時間	17:14	20:05	21:47	20:28	19:53	23:50	25:10	26:46

4. むすび

本研究では，映像配信サービスにおける視聴者の視聴体験を妨げずに，ミッドロール広告の挿入位置を最適化する手法を提案した．提案手法では，マルチモーダルデータを活用しながら，映像文脈の考慮を行うことで，ミッドロール広告位置の候補点の抽出を高精度に実現した．実験では，ドラマのような長尺映像に対して，提案手法の有効性を確認した．具体的には，提案手法では，映像編集者が手動で広告挿入位置よりも，視聴者の視聴体験を妨げない広告挿入位置の抽出を実現した．その一方で，提案手法では，挿入するミッドロール広告の内容までは考慮できていないことや，学習データに含まれないドキュメンタリーやアニメなどに対する有効性の検証はできていないことは今後の課題である．したがって今後は，様々なジャンルへの適用に加えて，ミッドロール広告の内容を考慮した挿入位置の抽出や，挿入位置の前後の動画文脈に合わせたミッドロール広告の自動選択などに取り組んでいく．

〔文　献〕

1）総務省: “情報通信白書令和 6 年版, ”https://www.soumu.go.jp/johotsusintokei/whitepaper/r06.html (2024)
2）株式会社サイバーエージェント: “2023 年国内動広告の市場調査,”https://www.cyberagent.co.jp/news/detail/id=29827 (2024)
3）U. Bulkan, T. Dagiuklas, and M. Iqbal: “Modelling quality of experience for online video advertisement insertion,” *IEEE Transactions on Broadcasting*, **66**, 4, pp. 835–846 (2020)
4）齊藤 義仰, 畠山 智裕, 西岡 大: “スマートフォンを用いた動画広告挿入タイミング決定アルゴリズムの提案,” マルチメディア，分散，協調とモバイルシンポジウム *(DICOMO)*, **1**, pp. 1089–1095 (2017)
5）遠藤 茉樹, 加藤 拓巳: “YouTube における動画広告挿入のタイミングによる記憶と購入意向への影響,” マーケティングレビュー, vol. 4 (2022)
6）Y. Saito: “A Proposal of a method for video advertisement insertion on smartphone,” *International Journal of Informatics Society*, **11**, 1, pp. 55–62 (2019)
7）U. Bulkan, T. Dagiuklas, and M. Iqbal: “Supereye: Smart advertisement insertion for online video streaming,” *Multimedia Tools and Applications (MTA)*, **82**, 6, pp. 9361–9379 (2023)
8）JL. Elman: “Finding structure in time,” *Cognitive Science*, **14**, 2, pp. 179–211 (1990)
9）S. Hochreiter: “Long short-term memory,” *in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, **9**, 8, pp. 1735–1780 (1997)
10）A. Vaswani, N. Shazeer, et al: “Attention is all you need,” *in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, **30** (2017)
11）M. Islam, M. Hasan, et al: “Efficient movie scene detection using state-space transformers,” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18749–18758 (2023)
12）T. Souček and J. Lokoc: “Transnet v2: An effective deep network architecture for fast shot transition detection,” *in Proceedings of ACM International Conference on Multimedia (ACMMM)*, pp. 11218–11221 (2024)
13）J. Lokoc, G. Kovalčík, et al: “A framework for effective known-item search in video,” *in Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pp. 1777–1785 (2019)
14）J. Donahue, Y. Jia, et al: “Decaf: A deep convolutional activation feature for generic visual recognition,” *in Proceedings of the International Conference on Machine Learning (ICML)*, pp. 647–655 (2014)
15）A. Dosovitskiy, L. Beyer, et al: “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929* (2020)
16）H. Mehta, A. Gupta, et al: “Long range language modeling via gated state spaces,” *arXiv preprint arXiv:2206.13947* (2022)
17）S. Hiroaki and S. Chiba: “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **26**, 1, pp. 43–49 (1978)
18）D. Doukhan, E. Lechapt, et al: “Ina’s mirex 2018 music and speech detection system,” *in Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)* (2018)
19）Y. Lecun, L. Bottou, et al: “Gradient-based learning applied to document recognition,” *Proceedings of The IEEE*, **86**, 11, pp. 2278–2324 (1998)
20）Q. Huang, Y. Xiong, et al: “Movienet: A holistic dataset for movie understanding,” *in Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 709–727 (2020)
21）D. Kingma and J. Ba: “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980* (2014)

春山　知生（はるやま　ともき） 2021 年 3 月，北海道大学大学院情報科学院 メディアネットワーク専攻 修士課程 修了．2021 年 4 月，株式会社 NTT ドコモ 入社．NTT ドコモにて動画像認識技術の，スポーツやスマートモビリティ分野への社会実装業務に従事，現在に至る．正会員．

塚谷　俊介（つかたに　しゅんすけ） 2016 年 3 月 東京大学大学院情報理工学系研究科電子情報学専攻 修士課程 修了．2016 年 4 月 日本電信電話株式会社 入社．2023 年 10 月 株式会社 NTT ドコモに転籍，現在に至る．

北出　卓也（きたで　たくや） 2013 年, 慶應義塾大学大学院理工学研究科開放環境科学専攻 修士課程 修了．2013 年，株式会社 NTT ドコモ 入社．2023 年，慶應義塾大学大学院理工学研究科開放環境科学専攻 博士課程 入学．NTT ドコモにて画像認識及びロボティクス技術の，農業やスマートシティの領域への社会実装業務に従事，現在に至る．