

Segmentação de vídeo em cenas - uma abordagem multimodal baseada em *Transformers*

Marcus Vinicius Cerrato
ICMC-USP
mcerrato@usp.br

Wan Song Rocha
ICMC-USP
wansong@usp.br

Rudinei Goularte
ICMC-USP
rudinei@icmc.usp.br

ABSTRACT

Automatic temporal video scene segmentation remains a challenging and open problem in the Multimedia field. While recent state-of-the-art approaches shift toward multimodal paradigms, conventional methods often rely on rigid local fusion strategies that may fail to capture long-range temporal dependencies or handle cross-modal asynchronous perturbations. To alleviate these limitations, this paper proposes a novel multimodal Transformer-based architecture configured with a joint space-temporal self-attention mechanism. The proposed method introduces an adaptive fusion strategy that integrates visual and acoustic streams, represented by dense MFCC embeddings, at the shot level and dynamically regulated by a gating mechanism. Evaluated under a cross-validation protocol on a public dataset, the proposed approach exceeds the average scores from related work.

KEYWORDS

Multimedia, Video, Transformers, Temporal Scene Segmentation, Scene Boundary Detection

1 INTRODUÇÃO

A análise do crescente volume de vídeos digitais demanda soluções eficientes para serviços de recomendação, seleção e sumarização de conteúdo [9]. Contudo, o processamento desses dados é oneroso devido à sua complexidade intrínseca, gerando a necessidade de segmentar os vídeos em unidades temporais menores e semanticamente significativas, como tomadas (*shots*) e cenas [26]. A segmentação temporal automática de cenas permanece um desafio em aberto [11], limitado pela lacuna semântica (*semantic gap* [17]) entre as representações de baixo nível e o significado real do conteúdo, além da subjetividade na definição de cena — comumente estabelecida como um conjunto de tomadas semanticamente relacionadas.

Embora abordagens unimodais seminais analisassem fluxos visual, aural ou textual isoladamente [6], a detecção de transições de cena pode falhar se a pista de transição não estiver presente na modalidade escolhida. Abordagens multimodais, por outro lado, utilizam informações extraídas de duas ou mais modalidades, idealmente de forma complementar.

Assim, técnicas recentes migraram para o paradigma multimodal [11]. No entanto, encontrar similaridades conceituais ao

longo de sequências longas de tomadas continua sendo um obstáculo para fusões multimodais tradicionais. Recentemente, a arquitetura *Transformer* emergiu como uma alternativa promissora ao utilizar mecanismos de atenção para modelar dependências de maior alcance [23]. No caso de segmentação temporal de vídeo em cenas, isso poderia ser útil para extrapolar o escopo intra-tomada. A literatura divide-se em vertentes focadas em alinhamento temporal [14, 18], atenção local [13], atenção espaço-temporal dividida [5] e aprendizado contrastivo [1]. Todavia, esses modelos negligenciam a dinâmica hierárquica e tratam o vídeo como um fluxo contínuo de *tokens*, concentrando-se em detalhes visuais locais. Isso dificulta a distinção entre tomadas comuns e tomadas de transição (fronteiras de cenas).

Para mitigar essas limitações, este trabalho propõe uma arquitetura multimodal baseada em *Transformers* para a segmentação temporal de cenas, utilizando uma representação visual espaço-temporal. Especificamente, os mecanismos de *tubelet embedding* e atenção espaço-temporal conjunta são inspirados na arquitetura ViViT [3] para construir um codificador visual de tomadas, adaptado neste trabalho para a tarefa hierárquica de identificação de fronteiras entre cenas. Complementarmente, do fluxo aural extraem-se sequências densas de características MFCC (*Mel-Frequency Cepstral Coefficients*), posteriormente modeladas temporalmente como *embeddings* aurais que servirão de assinatura acústica. Por fim, um mecanismo de *gating* multimodal adaptativo integra as representações visual e aural para auxiliar na decisão sobre continuidade ou transição entre tomadas.

O restante deste artigo está assim organizado: a Seção 2 apresenta alguns fundamentos sobre segmentação de cenas, e o estado da arte em técnicas ligadas à segmentação temporal de cenas; a Seção 3 descreve a abordagem proposta; a Seção 4 detalha os experimentos realizados e os resultados obtidos; e a Seção 5 apresenta as considerações finais sobre este trabalho.

2 TRABALHOS RELACIONADOS

A evolução das técnicas de segmentação temporal de vídeo em cenas é marcada por uma busca por reduzir a lacuna semântica entre dados de baixo nível e a estrutura narrativa dos vídeos. Abordagens clássicas dependiam de pistas unimodais ou fusões tardias heurísticas [6]. Já as técnicas mais recentes utilizam arquiteturas baseadas em aprendizado profundo e mecanismos de atenção. Contudo, a literatura atual ainda enfrenta desafios na modelagem multimodal e de longo alcance para identificar continuidades contextuais.

Uma vertente pioneira na literatura investigou a transposição dos modelos de linguagem baseados em *Transformers*

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

para o domínio visual, com foco no alinhamento multimodal, destacando-se as arquiteturas VideoBERT [18] e ViLBERT [14]. O VideoBERT foi projetado para tarefas de indexação instrucional e predição de vídeo utilizando a base YouCook2 [25], caracterizada por vídeos de culinária onde há um alinhamento temporal estrito e literal entre as ações visuais e as descrições textuais correspondentes. Por sua vez, o ViLBERT trata tarefas de *Visual Question Answering* (VQA) e recuperação de imagens, sendo treinado sobre grandes volumes de dados de imagens estáticas emparelhadas com legendas (*Datasets Conceptual Captions* [16] e COCO [10]). Ambas arquiteturas dependem dessa forte correlação explícita e síncrona entre o texto e o componente visual para aprender representações. Além disso, dependem de características visuais estáticas pré-computadas - não são extraídas do vídeo sendo segmentado. Isso impede essas redes de perceber a dinâmica fina de movimento contida no interior de cada tomada, o que restringe a aplicação de tais modelos em tarefas de segmentação temporal de vídeo em cenas, especialmente em vídeos de longa duração com tomadas complexas.

Para capturar dependências visuais detalhadas, modelos como o Video Swin Transformer (VST) [13] introduziram mecanismos de atenção local baseados em janelas. O VST utiliza bases como o Kinetics-400, de vídeos curtos, com foco em análise de imagens para detecção de ações. Se por um lado essa estratégia reduz o custo computacional, por outro ela induz a uma "miopia temporal". Em narrativas complexas ou documentários (onde uma cena pode se estender por minutos com variações de planos), a atenção local falha em estabelecer a coerência global, pois o modelo perde a capacidade de correlacionar eventos visuais distantes no tempo.

Por outro lado, arquiteturas voltadas ao processamento global de vídeo, como o TimeSformer [5], buscaram otimizar a atenção ao propor eixos espaciais e temporais divididos (*Divided Attention*). O TimeSformer também tem foco em detecção de ações em vídeos curtos, usando bases como Kinetics-400 e Something-Something v2. Embora eficiente para o reconhecimento de ações em vídeos curtos (de poucos segundos), essa fatoração precoce introduz uma separação artificial entre "o que muda" e "onde muda". Para a segmentação de cenas, onde movimentos de câmera complexos alteram a semântica do quadro continuamente, a aproximação fatorada pode perder correlações não-lineares cruzadas.

No contexto específico da segmentação multimodal de cenas, Rao et al. [15] propuseram uma abordagem *Local-to-Global* que integra informações multimodais em diferentes níveis temporais, abrangendo *clip*, segmento e filme. São explorados a estrutura hierárquica das narrativas e o contexto de longo alcance para orientar a segmentação de cenas. Os autores exploram um *dataset* próprio e métricas de avaliação generalistas, em vez de métricas clássicas para segmentação.

Abordagens recentes em estado da arte evoluíram para analisar tomadas vizinhas mais distantes. Nesse contexto, três trabalhos se destacam na literatura especializada, em particular por abordarem relações de longo alcance em vídeos longos. Liu et al. [12] resolvem o problema da "miopia" criando redes de memória de longo alcance para ligar tomadas distantes.

Porém, depende de *features* visuais estáticas pré-computadas, incorrendo nos mesmos problemas de abordagens como VideoBERT e ViLBERT. Tseng et al. [20] usa um extrator 3D CNN para aprender relações entre tomadas com um *Transformer Encoder*, contudo, ele é um modelo estritamente unimodal (visual). Por outro lado, Ye et al. [24] propõem o SFGNet, utilizando um grafo onde cada tomada é um nó e as arestas representam as relações de dependência não-lineares. Apesar de conseguirem mapear relações hierárquicas complexas de um vídeo, o áudio é tratado como uma característica complementar concatenada no nível de fusão semântica tardia.

Diferente desses trabalhos, a abordagem proposta (Seção 3) introduz uma arquitetura multimodal baseada em *Transformer* para construir uma representação visual espaço-temporal enriquecida. Expandindo o contexto de entrada para duas tomadas consecutivas, o mecanismo de atenção conjunta passa a operar sobre a interface de transição, capturando dependências inter-tomadas. Paralelamente, a modalidade aural extrai e sintetiza o contexto acústico contínuo da tomada por meio de uma rede BiLSTM — como uma âncora acústica. Por fim, as representações resultantes (visual e aural) são fundidas dinamicamente através de um mecanismo de gating adaptativo, que pondera a relevância de cada modalidade em função das características contextuais presentes.

3 ABORDAGEM PROPOSTA

Esta seção descreve os detalhes da abordagem proposta, começando por sua formalização (Seção 3.1), passando pelas modelagens aural (Seção 3.2) e visual (Seção 3.3), detalhando como foi realizada a fusão (Seção 3.4), e, por fim, a predição de fronteiras (Seção 3.5).

3.1 Formalização e Tokenização Espaço-Temporal

Seja um vídeo V segmentado em uma sequência de N tomadas $S = \{s_1, s_2, \dots, s_N\}$. Cada tomada s_i é composta por um fluxo visual $\mathcal{X}_v \in \mathbb{R}^{T \times H \times W \times C}$ e um fluxo aural correspondente $\mathcal{X}_a$. No fluxo visual, T denota a dimensão temporal (sequência de quadros existentes na representação da tomada), H denota a altura do quadro, W denota a largura do quadro e C denota a quantidade de canais de um quadro.

Para mapear a dinâmica espaço-temporal intrínseca ao fluxo visual da tomada, em vez de tratar o volume como uma sequência de imagens isoladas, adotou-se o mecanismo de incorporação por *tubelets* (*Tubelet Embedding*) introduzido por Arnab et al. [3] na arquitetura ViViT e empregado neste trabalho como representação de entrada do codificador visual. O volume de vídeo $\mathcal{X}_v$ é decomposto em cubos tridimensionais não sobrepostos de dimensão $t \times h \times w$, onde t representa a extensão temporal e $h \times w$ denota a resolução espacial do *tubelet*, conforme ilustrado na Figura 1. Por exemplo, o *tubelet* x_1 possui dimensões $t = 5$, $h = \frac{H}{3}$ e $w = \frac{W}{3}$.

Os *tubelets* são submetidos a uma operação de projeção linear (implementada computacionalmente através de uma convolução 3D) e mapeados para um espaço latente de dimensão D . A sequência resultante de tokens espaço-temporais (elipses amarelas e vermelhas após a seta na Figura 1) é

estendida com *embeddings* posicionais [3] para preservar a cronologia do fluxo e a topologia geométrica da tomada, gerando o vetor de entrada unificado $\mathbf{z}_v \in \mathbb{R}^{M \times D}$. O número total de *tokens* M gerados para o volume da tomada é determinado por:

$$M = \left\lfloor \frac{T}{t} \right\rfloor \times \left\lfloor \frac{H}{h} \right\rfloor \times \left\lfloor \frac{W}{w} \right\rfloor \quad (1)$$

Esta modelagem funde os dados temporais e espaciais locais logo na camada de entrada da rede, antes mesmo de qualquer camada de atenção processar os dados.

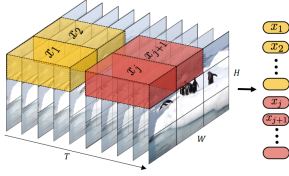


Figura 1: Tubelet embedding. Fonte: Arnab et al. (2021) [3].

3.2 Representação Aural e Modelagem Temporal

O fluxo aural correspondente à tomada s_i é inicialmente decomposto em uma sequência temporal de vetores de características espectrais $\mathcal{X}_a = \{\mathbf{x}_{a,1}, \mathbf{x}_{a,2}, \dots, \mathbf{x}_{a,T_a}\}$, em que $\mathbf{x}_a \in \mathbb{R}^{T_a \times D_{mfcc}}$. Cada vetor $\mathbf{x}_{a,t}$ encapsula os coeficientes MFCC extraídos em um determinada janela temporal t (Seção 4.2.1), totalizando T_a passos de tempo por tomada.

Para capturar a continuidade acústica e as propriedades dinâmicas do ambiente sonoro, a sequência $\mathcal{X}_a$ é processada por uma camada recorrente do tipo LSTM Bidirecional (BiLSTM) [7]. A BiLSTM processa o sinal simultaneamente nos sentidos direto ($\rightarrow$) e inverso ($\leftarrow$), gerando, para cada instante t , os estados ocultos correspondentes:

$$\vec{\mathbf{h}}_t = \text{LSTM}_{\text{fwd}}(\mathbf{x}_{a,t}, \vec{\mathbf{h}}_{t-1}) \quad (2)$$

$$\overleftarrow{\mathbf{h}}_t = \text{LSTM}_{\text{bwd}}(\mathbf{x}_{a,t}, \overleftarrow{\mathbf{h}}_{t+1}) \quad (3)$$

Os estados ocultos de ambas as direções são concatenados para compor a representação temporal unificada de cada instante, definida por $\mathbf{h}_t = \vec{\mathbf{h}}_t \parallel \overleftarrow{\mathbf{h}}_t$; $\mathbf{h}_t \in \mathbb{R}^{2d}$, em que d denota a dimensão oculta de cada LSTM física.

Como a duração das tomadas e a relevância de determinados eventos sonoros variam ao longo do tempo, aplicar um *pooling* simples (como a média) sobre a sequência de estados ocultos $\mathcal{H}_a = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{T_a}\}$ poderia diluir pistas acústicas. Assim, introduziu-se um mecanismo de atenção temporal baseado em produto escalar com contexto para computar uma média ponderada adaptativa de toda a sequência.

Primeiramente, cada estado oculto $\mathbf{h}_t$ é projetado linearmente através de uma camada totalmente conectada para gerar uma representação latente $\mathbf{u}_t \in \mathbb{R}^{d_u}$:

$$\mathbf{u}_t = \tanh(\mathbf{W}_u \mathbf{h}_t + \mathbf{b}_u) \quad (4)$$

em que $\mathbf{W}_u \in \mathbb{R}^{d_u \times 2d}$ e $\mathbf{b}_u \in \mathbb{R}^{d_u}$ são os parâmetros de projeção aprendíveis. A importância relativa de cada instante

t é mensurada calculando-se o alinhamento (produto escalar) entre o vetor $\mathbf{u}_t$ e um vetor de contexto aural compartilhado $\mathbf{u}_a \in \mathbb{R}^{d_u}$, inicializado aleatoriamente e otimizado via retropropagação. Os pesos de atenção normalizados α_t são obtidos através de uma função *softmax*:

$$\alpha_t = \frac{\exp(\mathbf{u}_t^T \mathbf{u}_a)}{\sum_{k=1}^{T_a} \exp(\mathbf{u}_k^T \mathbf{u}_a)} \quad (5)$$

Esses pesos atuam como filtros dinâmicos de relevância, mapeando o foco do modelo sobre os quadros sonoros mais significativos. A representação condensada da tomada, denotada por $\mathbf{z}_a \in \mathbb{R}^{D_a}$ (onde $D_a = 2d$), é gerada pela soma ponderada de todos os estados ocultos da BiLSTM:

$$\mathbf{z}_a = \sum_{t=1}^{T_a} \alpha_t \mathbf{h}_t \quad (6)$$

Por fim, para harmonizar as dimensões latentes e viabilizar a fusão multimodal subsequente com o fluxo visual, o vetor condensado $\mathbf{z}_a$ é projetado linearmente para o mesmo espaço dimensional D ocupado pelos tokens espaço-temporais (visuais), utilizando uma função de ativação não-linear:

$$\mathbf{a} = \text{GELU}(\mathbf{W}_a \mathbf{z}_a + \mathbf{b}_a) \quad (7)$$

em que $\mathbf{W}_a \in \mathbb{R}^{D \times D_a}$ e $\mathbf{b}_a \in \mathbb{R}^D$. A assinatura acústica resultante $\mathbf{a} \in \mathbb{R}^D$ captura tanto a identidade espectral imediata quanto a dinâmica de longa duração do som da tomada, permitindo ao modelo rastrear com precisão as persistências e rupturas de ambiente que delimitam uma cena.

3.3 Codificador Visual de Atenção Conjunta

Para processar as interações espaciais e a dinâmica de movimento de forma integrada na interface de transição visual, a sequência de tokens espaço-temporais $\mathbf{z}_v \in \mathbb{R}^{M \times D}$ (obtida via *tubelets* na Subseção 3.1 a partir de um volume de 32 quadros abrangendo duas tomadas consecutivas, s_i e s_{i+1}) é submetida a um módulo de codificação visual estruturado segundo os princípios de Atenção Espaço-Temporal Conjunta, propostos na arquitetura seminal ViViT [3]. Essa escolha permite aplicar a autoatenção de forma simultânea aos *tokens* espaciais e temporais ao longo de todo o volume, capturando tanto a coerência visual interna de cada tomada quanto as dependências inter-tomadas na fronteira de corte.

Formalmente, a sequência de entrada $\mathbf{z}_v$ é projetada linearmente em três matrizes distintas de representação latente, correspondentes aos espaços de *Query* (Q), *Key* (K) e *Value* (V):

$$Q = \mathbf{z}_v \mathbf{W}_Q, \quad K = \mathbf{z}_v \mathbf{W}_K, \quad V = \mathbf{z}_v \mathbf{W}_V \quad (8)$$

em que $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times D}$ são as matrizes de pesos aprendíveis do codificador. Para capturar diferentes padrões de dependência, o mecanismo utiliza múltiplas cabeças de atenção em paralelo (*Multi-Head Attention* [21]). Para uma cabeça individual i , a matriz de afinidade estendida ao volume espaço-temporal completo é computada diretamente por:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (9)$$

em que Q_i, K_i, V_i representam as fatias projetadas para a i -ésima cabeça e $d_k = DH_c$, sendo H_c o número total de cabeças de atenção.

Diferente de arquiteturas que segregam o tempo e o espaço por aproximações fatoradas (gerando matrizes isoladas para eixos geométricos e cronológicos), a matriz de similaridade gerada por esta formulação estende o cálculo do produto escalar concorrentemente sobre os M tokens tridimensionais. As saídas de todas as cabeças são concatenadas e projetadas novamente para compor o fluxo refinado:

$$\text{MHA}(Q, K, V) = [\text{head}_1; \text{head}_2; \dots; \text{head}_{H_c}] \mathbf{W}_O \quad (10)$$

em que $\mathbf{W}_O \in \mathbb{R}^{D \times D}$ denota a projeção linear de saída.

Como a arquitetura proposta estende o uso desta atenção para a tarefa hierárquica de delimitação de fronteiras (e não para a classificação de ações em cliques isolados), a assinatura semântica conjunta da interface é extraída por meio do token de classificação ($[CLS]$) localizado na última camada do codificador. Esse vetor unificado de características visuais, denotado por $\mathbf{v} \in \mathbb{R}^D$, encapsula a topologia geométrica, a continuidade do movimento e a eventual ruptura visual entre o par de tomadas s_i, s_{i+1} , servindo como uma das âncoras para o mecanismo de fusão adaptativa.

3.4 Mecanismo de Fusão Multimodal Adaptativa

Uma vez obtidos o vetor visual global $\mathbf{v}$ e o vetor aural projetado $\mathbf{a}$, o modelo realiza a integração multimodal por meio de um mecanismo de *gating* adaptativo. A formulação adotada segue o princípio de fusão multimodal por gates multiplicativos *Gated Multimodal Units* (GMU) [2], adaptado neste trabalho para integrar as representações visual e aural no contexto da segmentação temporal de cenas.

Primeiramente, as representações são normalizadas independentemente para reduzir discrepâncias de escala entre os espaços latentes originais. A adaptabilidade do sistema é governada por um vetor de ponderação não-linear $\mathbf{g}_i \in [0, 1]^D$, computado especificamente para cada tomada s_i . As representações mapeadas são combinadas via operador de concatenação, $[\mathbf{v}; \mathbf{a}] \in \mathbb{R}^{2D}$, e submetidas a uma projeção linear modulada por uma função sigmoide contendo um fator de escala τ (parâmetro de temperatura):

$$\mathbf{g}_i = \sigma \left(\frac{\mathbf{W}_g [\mathbf{v}; \mathbf{a}] + \mathbf{b}_g}{\tau} \right) \quad (11)$$

em que $\mathbf{W}_g \in \mathbb{R}^{D \times 2D}$ e $\mathbf{b}_g \in \mathbb{R}^D$ são os parâmetros sinápticos aprendíveis da camada de fusão.

O fator de escala $\tau \in \mathbb{R}$ atua regulando a sensibilidade do mecanismo de ponderação. Valores reduzidos de τ produzem ativações mais abruptas da função sigmoide, aproximando-a de uma escolha seletiva e rígida entre modalidades, enquanto valores elevados promovem uma combinação mais suave e distribuída entre as representações visual e aural.

Como o vetor $\mathbf{g}_i$ é gerado dinamicamente a partir dos dados de entrada de cada tomada, o modelo ganha a capacidade de alternar seu foco de maneira autônoma. A fusão final das características da tomada é computada através de

uma interpolação ponderada elemento a elemento (*Hadamard product*), gerando o vetor multimodal unificado $\mathbf{h}_i \in \mathbb{R}^D$ que condensa a assinatura espaço-temporal e acústica da tomada s_i :

$$\mathbf{h}_i = \mathbf{g}_i \odot \mathbf{v} + (\mathbf{1} - \mathbf{g}_i) \odot \mathbf{a} \quad (12)$$

em que $\mathbf{1}$ representa um vetor unitário de "uns" com dimensão D .

Esta formulação garante que descontinuidades visuais enganosas (como cortes de câmera intra-cena) sejam mitigadas caso o ambiente sonoro permaneça estável, situação em que o *gate* atenua os valores de $\mathbf{g}_i$ e força a rede a se apoiar na assinatura acústica $\mathbf{a}$ para preservar a coerência inter-tomadas.

3.5 Contextualização entre Tomadas e Predição de Fronteiras

A etapa final do pipeline consiste em classificar a existência de uma transição de cena na interface delimitada pelo volume multimodal de entrada. Como a representação latente integrada $\mathbf{h} \in \mathbb{R}^D$ (obtida na Subseção 3.4) já encapsula conjuntamente as informações visuais de atenção inter-tomadas de s_i, s_{i+1} e a âncora de contexto aural, ela é submetida diretamente a um Perceptron Multicamadas (MLP) de classificação.

Formalmente, a probabilidade $\hat{y}$ de ocorrência de uma transição no ponto de corte entre as tomadas é dada por:

$$\hat{y} = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{h} \mathbf{b}_1 \mathbf{b}_2)) \quad (13)$$

em que $\mathbf{W}_1 \in \mathbb{R}^{D_{\text{hidden}} \times D}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times D_{\text{hidden}}}$, $\mathbf{b}_1, \mathbf{b}_2$ são os pesos e vieses aprendíveis da rede, e σ representa a função de ativação Sigmóide.

Durante o treinamento, a otimização dos parâmetros de toda a rede (codificadores visual, aural, módulo de gating e MLP de predição) é realizada de forma *end-to-end* minimizando a função de perda de Entropia Cruzada Binária (*Binary Cross-Entropy* - BCE):

$$\mathcal{L}_{\text{BCE}} = -[y \log \hat{y} + 1 - y \log 1 - \hat{y}] \quad (14)$$

em que $y \in \{0, 1\}$ indica a rotulagem de referência (*ground truth*) para a presença de transição na interface analisada.

4 EXPERIMENTOS

Esta seção tem o objetivo de apresentar os experimentos realizados, desde a descrição do *dataset* utilizado (Seção 4.1), o pré-processamento adotado (Seção 4.2), os treinamentos realizados (Seção 4.3), as métricas de avaliação adotadas (Seção 4.4) e, por fim, os resultados obtidos (Seção 4.5).

4.1 Descrição da Base de Dados

Para a validação experimental da arquitetura utilizou-se o *dataset BBC Planet Earth* [4], amplamente adotada na literatura de segmentação de vídeos. A base é composta por 11 documentários sobre diferentes regiões do Planeta Terra. Os vídeos possuem algumas características em comum que auxiliam análises comparativas: um narrador ao longo do vídeo; uma trilha sonora misturada com sons ambiente; e mesma quantidade aproximada de cenas, tomadas e quadros. Os detalhes estão sintetizados na Tabela 1.

Tabela 1: Estatísticas dos vídeos da base *BBC Planet Earth*.

ID	Vídeo	Cenas	Tomadas	Quadros	Duração
01	From Pole to Pole	66	450	73875	49:15
02	Mountains	53	395	72115	48:04
03	Ice Worlds	62	425	73925	49:17
04	Great Plains	71	473	73575	49:03
05	Jungles	65	461	73850	49:14
06	Seasonal Forests	65	529	73985	49:19
07	Fresh Water	62	533	73935	49:17
08	Ocean Deep	53	418	73850	49:14
09	Shallow Seas	61	367	73850	49:14
10	Caves	57	393	73376	48:55
11	Deserts	55	469	73492	48:59

A base de dados fornece, além dos fluxos de mídia, os arquivos de anotação (*Ground Truth*) para cada episódio. As anotações estabelecem a hierarquia narrativa estrutural do vídeo, mapeando os índices absolutos de quadros que delimitam o início e o fim de cada tomada, bem como os índices sequenciais de tomadas que definem cada cena.

begin, end 1, 1 2, 9 10, 15	begin, end 1, 633 651, 771 772, 892
--------------------------------------	--

Figura 2: Trechos do *groundtruth* do vídeo 1: cenas (esquerda) e tomadas (direita). Fonte: próprio autor.

Como exemplificado na Figura 2, do lado esquerdo percebe-se o formato do *ground truth* de transição de cenas para o primeiro vídeo, em que os índices indicam o intervalo de tomadas pertencentes a uma mesma unidade semântica. Analogamente, o lado direito expõe o mapeamento de tomadas, cujos valores numéricos delimitam os intervalos de quadros físicos (*frames*) obtidos via detecção de cortes simples.

A partir das delimitações de intervalos fornecidas pelo *ground truth* de tomadas, os fluxos visuais e aurais equivalentes a cada segmento foram extraídos de forma automatizada, isolando as tomadas que servem como unidades básicas de processamento para os codificadores da arquitetura proposta.

4.2 Pré-Processamento

Esta seção apresenta os procedimentos de preparação e transformação aplicados aos fluxos brutos do *dataset*. O *pipeline* compreende a parametrização para extração das assinaturas espectrais e visuais em nível de tomada (Seção 4.2.1), seguida pela estratégia de *Data Augmentation* projetada para mitigar a disparidade entre as classes (Seção 4.2.2).

4.2.1 Extração de Características. Para cada tomada, gerou-se uma representação aural e uma visual, seguindo o protocolo detalhado a seguir.

Do fluxo aural extraiu-se áudio em formato *wav* de cada tomada, visando obtenção de características espectrais *MFCC*, reconhecidas pela capacidade de representar propriedades perceptualmente relevantes do espectro de frequência. A extração utilizou janelas de 2048 amostras com deslocamento de 512 amostras entre blocos consecutivos. Considerando a taxa de amostragem de 16 kHz, essa configuração equivale a janelas temporais de 128 ms com taxa de atualização de 32 ms (75% de sobreposição). A janela expandida (128 ms),

superior aos padrões de reconhecimento de fala, justifica-se pela necessidade de capturar nuances do ambiente acústico (como trilhas sonoras e ruídos de fundo), evitando falsos positivos na detecção de fronteiras. A sequência resultante alimenta a camada recorrente da arquitetura (Seção 3.2).

Para o fluxo visual, implementou-se uma amostragem temporal uniforme para fixar a extensão dos dados independentemente da duração real da tomada. Partindo do primeiro quadro físico, extraiu-se um total de 16 quadros (segundo [3]) distribuídos proporcionalmente. Em tomadas com extensão inferior a 16 quadros, realizou-se o preenchimento (*padding*) via repetição do último quadro disponível. Como o classificador avalia a interface de transição entre duas tomadas consecutivas (s_i e s_{i+1}), a janela de contexto de entrada totaliza 32 quadros processados concorrentemente.

Os quadros extraídos foram organizados em tensores estruturados no formato $\mathcal{X}_v \in \mathbb{R}^{T \times H \times W \times C}$, preservando a ordem cronológica nativa para a atenção espaço-temporal conjunta. Os quadros foram inicialmente carregados utilizando a biblioteca OpenCV, que emprega a ordenação BGR para os canais de cor, e posteriormente convertidos explicitamente para RGB antes de serem processados pelo codificador. Em seguida, foram redimensionados para a resolução de 224×224 *pixels*, mantendo profundidade de cor de 24 *bits*. Esse processamento, que inclui a amostragem temporal, o redimensionamento dos quadros e sua preparação para entrada no codificador visual, é descrito em maior detalhe na Seção 3.1 e compõe o representante visual de cada tomada.

4.2.2 Data Augmentation. Conforme observado na Seção 3.5, o problema é modelado como uma tarefa de classificação binária entre tomadas de transição e não-transição. Com o objetivo de ampliar a representatividade da classe de transição, numericamente inferior, foram geradas novas amostras a partir das tomadas originais por meio de uma técnica de deslocamento temporal amostral (*temporal shift sampling*), aplicada exclusivamente às tomadas de transição.

Essas novas amostras preservam a estrutura temporal da tomada, mas utilizam uma representação alternativa dos quadros, iniciando-se no quadro seguinte ao da sequência original e mantendo o mesmo espaçamento temporal. Dessa forma, para uma sequência representada originalmente pelos quadros $1, 1+n, 1+2n, \dots$, constrói-se uma nova representação dada por $2, 2+n, 2+2n, \dots$.

Esta etapa foi realizada como pré-processamento, antes do treinamento e da inferência. Em termos numéricos, o conjunto original é composto por 4.902 amostras, sendo 4.243 exemplos de não-transição e 659 exemplos de transição. Foram então geradas 3.584 novas amostras, todas pertencentes à classe de transição, resultando em um conjunto com 8.486 amostras, com 4.243 exemplos em cada classe.

4.3 Treinamento

A proposta foi desenvolvida em linguagem Python, utilizando o framework PyTorch. A etapa experimental apresentada neste trabalho foi conduzida em um computador equipado com um processador Intel Core i7-7700 de 3,60 GHz, com

quatro núcleos físicos e oito *threads*, 22 GB de memória RAM e uma GPU NVIDIA GeForce RTX 3070 com 8 GB de memória VRAM. O ambiente de software utilizado empregou Python 3.12.11, PyTorch 2.7.1 com CUDA 12.6, Transformers 4.52.4, Torchvision 0.22.1, NumPy 2.4.6 e OpenCV 4.11.0. O processo de otimização foi realizado durante 20 épocas, limite determinado a partir de testes empíricos preliminares que evidenciaram a estabilização da perda de validação.

A configuração do mecanismo de *gating* foi estabelecida de modo que os pesos multimodais fossem aprendidos diretamente durante o treinamento. O fator de escala aplicado aos *logits* do mecanismo de *gating* (Equação 11) foi definido em $\tau = 2,0$, valor que não privilegia nenhuma modalidade.

Para a modalidade visual, o codificador de atenção conjunta foi inicializado com os pesos pré-treinados do modelo ViViT [3]. Empregou-se a estratégia de *fine-tuning* associada a uma taxa de aprendizado diferencial (*differential learning rate*). Assim, os parâmetros das camadas nativas do *backbone* visual foram otimizados com uma taxa de aprendizado reduzida, equivalente a 10% do valor base, restringindo modificações abruptas nas representações de baixo nível. Em contrapartida, as novas camadas adaptativas e a MLP de predição de fronteiras operaram sob a taxa de aprendizado de 10^{-5} . Desse modo, procura-se evitar alterações abruptas em representações previamente consolidadas, ao mesmo tempo em que favorece a adaptação das novas camadas ao domínio específico da segmentação temporal de cenas.

O processo de atualização dos gradientes foi gerenciado pelo otimizador *Adam*, configurado com decaimento de pesos (*weight decay*) de 10^{-2} para penalizar a complexidade estrutural e coibir o *overfitting*. Como salvaguarda de regularização adicional, aplicou-se uma taxa de *dropout* de 0,2 distribuída tanto nas projeções internas quanto nas camadas de atenção do *Transformer*. O teto amostral utilizado foi de 5000 transições por época. Durante a inferência, a probabilidade estimada para a classe de transição foi convertida em uma decisão binária utilizando um limiar de 0,5.

Os experimentos foram conduzidos utilizando o protocolo de validação cruzada do tipo *Leave-One-Video-Out* (LOVO), na qual um vídeo completo é integralmente removido do treinamento e reservado exclusivamente para inferência e avaliação. Nesta configuração, o modelo é treinado apenas com os vídeos remanescentes, sendo estes divididos estratificadamente entre treinamento e validação, utilizando uma proporção de 80% e 20%, respectivamente.

4.4 Métricas de Avaliação

As métricas tradicionais de classificação binária Precisão (*Precision*) e Abrangência (*Recall*) são excessivamente rígidas quando aplicadas à segmentação temporal de vídeo em cenas. Se o modelo errar o limite real por uma única tomada de diferença, essas métricas penalizam a predição com um falso positivo e um falso negativo simultâneos, desconsiderando a proximidade temporal do acerto [19, 22].

Para contornar essa limitação, adotou-se neste trabalho as métricas Cobertura (*Coverage*) e Transbordamento (*Overflow*) propostas originalmente por Vendrig e Worring [22] e estendidas por Han et al. [8]. Tais métricas avaliam a sobreposição volumétrica entre as tomadas, sendo mais justas em tarefas de segmentação temporal [omitido para revisão cega].

Seja o conjunto de cenas previstas e de cenas reais respectivamente $S = \{s_1, \dots, s_{|S|}\}$ e $\tilde{S} = \{\tilde{s}_1, \dots, \tilde{s}_{|\tilde{S}|}\}$; e seja $|s_i|$ o número de tomadas em uma cena s_i , pode-se definir a Cobertura C_t de uma cena $\tilde{s}_t$ como:

$$C_t = \frac{\max_{i=1, \dots, |S|} |s_i \cap \tilde{s}_t|}{|\tilde{s}_t|} \quad (15)$$

Para medir quantas tomadas foram previstas além das fronteiras, usa-se a métrica Transbordamento, segundo a qual melhores predições terão menor valor, que em sua reformulação [8] (*New Overflow*) pode ser definida por:

$$NO_t = 1 - \frac{|\tilde{s}_t|}{\sum_{i=1}^{|S|} |s_i| \times \min(1, |s_i \cap \tilde{s}_t|)}, \quad (16)$$

A partir dessas métricas, a F_1 *score* entre ambas, denominada aqui como F_1^{CNO} , é obtida através da média harmônica entre C e $1 - NO$:

$$F_1^{CNO} = \frac{2 \cdot C \cdot (1 - NO)}{C + (1 - NO)} \quad (17)$$

Valores elevados de F_1^{CNO} indicam segmentações mais consistentes, bem alinhadas às cenas de referência e menos suscetíveis à fragmentação excessiva (baixo transbordamento).

4.5 Resultados dos Experimentos

Os experimentos foram realizados considerando três configurações distintas: (i) apenas características aurais, (ii) apenas características visuais e (iii) a abordagem multimodal proposta baseada na fusão adaptativa de ambas as modalidades, conforme detalhado na Subseção 3.2.

Os resultados foram obtidos aplicando-se o protocolo LOVO em 11 rodadas - para cada rodada a arquitetura foi treinada iterativamente (Seção 4.3) utilizando 10 vídeos, avaliando estritamente o vídeo remanescente. As métricas de F_1^{CNO} , dispostas individualmente nas tabelas desta seção, refletem o desempenho obtido em cada rodada isolada do LOVO. Por sua vez, as linhas denominadas "Média" correspondem à média aritmética simples calculada sobre os resultados consolidados das 11 rodadas independentes.

Os experimentos possuem duas etapas principais. Na primeira, realizou-se um estudo de ablação entre as configurações internas, buscando-se averiguar se a versão multimodal de fato apresenta vantagens sobre as versões monomodais e mapear como cada modalidade contribuiu para a tarefa (Subseção 4.5.1). Na segunda etapa (Subseção 4.5.2), comparam-se os resultados da arquitetura proposta com os de três trabalhos no estado da arte: Liu et al. [12], Tseng et al. [20] e Ye et al. [24]. Esses trabalhos foram selecionados por se alinharem à presente proposta ao abordarem a segmentação de cenas

com dependências além de uma tomada e por utilizarem o mesmo *dataset*, viabilizando uma comparação justa.

Tabela 2: Comparação quantitativa entre estratégias mono e multimodais para segmentação de cenas. Valores em F_1^{CNO} .

Vídeo	Aural	Visual	Visual + Aural
From Pale to Pole (01)	0.70	0.73	0.83
Mountains (02)	0.69	0.72	0.84
Ice Worlds (03)	0.70	0.73	0.82
Great Plains (04)	0.71	0.74	0.82
Jungles (05)	0.70	0.73	0.81
Seasonal Forests (06)	0.72	0.75	0.82
Fresh Water (07)	0.71	0.74	0.82
Ocean Deep (08)	0.70	0.74	0.83
Shallow Seas (09)	0.72	0.73	0.83
Caves (10)	0.73	0.74	0.82
Deserts (11)	0.74	0.75	0.83
Média	0.71	0.74	0.82

4.5.1 Comparação Monomodal x Multimodal. Os resultados da avaliação comparativa entre as versões monomodais e a multimodal estão na Tabela 2, revelando um comportamento consistente da proposta ao longo dos 11 vídeos do *dataset*.

A vertente monomodal Aural registrou média global de 0,71 no índice F_1^{CNO} . Esse desempenho indica que, embora transições no espectro sonoro (como variações de trilha e efeitos ambientais) ofereçam pistas para identificar mudanças de cena, elas se mostram insuficientes se processadas isoladamente. Por sua vez, a vertente monomodal Visual atingiu média de 0,74. O desempenho ligeiramente superior reforça a relevância das assinaturas espaço-temporais para determinar as fronteiras, porém evidencia vulnerabilidade diante de cortes de tomadas visualmente heterogêneas executados dentro de uma mesma cena. O ganho obtido pela abordagem multimodal em relação ao modelo visual puro evidencia a contribuição corretiva do canal aural.

Esse comportamento justifica-se pelo perfil do *dataset* de documentários, caracterizado pelo uso sistemático de tomadas de cobertura para ilustrar a narrativa conduzida pelo locutor. Nesses cenários, enquanto a alternância abrupta de imagens gera variações na matriz de *pixels* — induzindo o codificador visual a sinalizar falsas transições (sobre-segmentação) — a modalidade aural preserva a continuidade pela regularidade da voz do narrador e da trilha sonora. Ao integrar ambos os fluxos, o mecanismo de fusão soluciona o conflito de sinais, utilizando a estabilidade acústica como âncora semântica para inibir a fragmentação da cena. É importante notar que esse contexto é comum em outros domínios audiovisuais, como filmes, telejornais, aulas e tutoriais.

Ao unificar os fluxos sob o mecanismo de *gating* adaptativo parametrizado com fator de escala $\tau = 2,0$, a variante multimodal atinge média de 0,82. O ganho observado em todos os vídeos individualmente indica que a rede realiza uma fusão sinérgica e complementar, em vez de uma escolha binária simples ou submissão à modalidade estatisticamente dominante. Esse comportamento ratifica a eficácia da abordagem multimodal, demonstrando que o modelo explora pistas contextuais latentes de ambas as modalidades de forma complementar, garantindo uma segmentação de cenas robusta.

Tabela 3: Comparação dos valores em F_1^{CNO} obtidos com trabalhos em estado da arte. Os resultados de ACRNet, TELNet e SFGNet foram extraídos dos respectivos trabalhos originais e os resultados da abordagem proposta foram obtidos nos experimentos realizados neste trabalho.

Vídeo	ACRNet [12]	TELNet [20]	SFGNet [24]	Abordagem Proposta
01	0.83	0.77	0.69	0.83
02	0.82	0.68	0.71	0.84
03	0.80	0.69	0.66	0.82
04	0.70	0.75	0.62	0.82
05	0.73	0.74	0.70	0.81
06	0.75	0.75	0.54	0.82
07	0.77	0.74	0.57	0.82
08	0.71	0.76	0.69	0.83
09	0.80	0.70	0.65	0.83
10	0.73	0.77	0.65	0.82
11	0.70	0.77	0.64	0.83
Média	0.76	0.74	0.64	0.82

4.5.2 Comparação com o Estado da Arte. A comparação com os resultados reportados pelos trabalhos em estado da arte podem ser observada na Tabela 3. A análise dos dados expostos nesta tabela comprova que a abordagem proposta supera os principais métodos encontrados na literatura para o *dataset* utilizado. O modelo proposto atingiu a maior pontuação de F_1^{CNO} em 10 dos 11 vídeos avaliados, empatando com o ACRNet [12] no Vídeo 01 com um índice de 0,83.

Um aspecto importante está na consistência dos resultados inter-vídeos. Enquanto os métodos comparados apresentam oscilações - como o ACRNet [12], que decai para 0,70 nos Vídeos 04 e 11, e o TELNet [20], que registra 0,68 no Vídeo 02, a arquitetura desenvolvida neste trabalho manteve-se estável, operando com um piso de desempenho de 0,81 (Vídeo 05).

Também é interessante analisar os resultados obtidos pela abordagem proposta frente cada trabalho relacionado. O modelo SFGNet [24], que obteve a menor média global (0,64), baseia-se no agrupamento via grafos de relações semânticas abstratas. No domínio de documentários, essa estratégia depara-se com o problema da esparsidade semântica: elementos de fundo visual semelhantes (como vegetação, água ou céu) persistem ao longo de múltiplos blocos narrativos distintos. Sem uma ancoragem temporal estrita e o suporte de representações acústicas, o SFGNet [24] falha em discriminar pontos de transição semântica complexos, culminando em instabilidades como o índice de 0,54 do Vídeo 06.

O modelo TELNet [20], que obteve média F_1^{CNO} de 0,74, é uma arquitetura dependente de assinaturas puramente visuais (CNNs 3D e *Transformers* de tomadas). Essa característica prejudica sua capacidade preditiva em situações de tomadas de cobertura. Isso fica evidente ao se comparar com a variante monomodal aural desenvolvida neste trabalho (F_1^{CNO} 0,71), que obteve média geral muito próxima ao patamar do TELNet. Esse fato demonstra que rastrear a continuidade da trilha sonora e da locução em um sinal unidimensional é uma estratégia semanticamente mais estável para o gênero documental do que computar relações visuais complexas sujeitas às ambiguidades da edição.

Por fim, o modelo ACRNet [12] obteve F_1^{CNO} de 0,76, sendo o concorrente mais competitivo. Embora mapeie tomadas de longo alcance avaliando aspectos visuais e auditivos, como na arquitetura proposta, ele usa características visuais pré-computadas, em vez de características dos quadros reais da tomada. Além disso, faz uma fusão rígida por concatenação com as características aurais (MFCC também), ambas obtidas do mesmo instante (janela) temporal. Tal rigidez atrapalha ao processar vídeos onde existem transições rápidas e assincronias entre imagem e som, podendo contribuir para suas quedas de desempenho nos vídeos 04 e 11. Diferentemente, a arquitetura proposta faz uma fusão adaptativa: se o componente visual sofrer uma alteração drástica local, o mecanismo de *gating* percebe, por meio do contexto global, que o fluxo aural permanece contínuo, então, ele altera dinamicamente os pesos (atenuando o visual) para manter a coerência da cena. Esse comportamento está refletindo na estabilidade dos resultados multimodais da arquitetura proposta, assim como em sua média superior (0.82) de F_1^{CNO} .

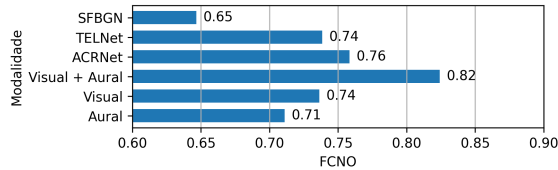


Figura 3: Comparação geral entre as variantes. Valores em F_1^{CNO} .

O panorama comparativo completo entre as variantes monomodais (áudio e vídeo isolados) e a abordagem multimodal proposta, confrontadas com os métodos do estado da arte, pode ser visualizado no gráfico da Figura 3. Observa-se que a arquitetura desenvolvida estabelece um incremento absoluto de 6% no índice F_1^{CNO} médio em relação ao melhor competidor: ACRNet [12]. Esse avanço quantitativo evidencia empiricamente a competitividade do mecanismo de atenção espaço-temporal conjunta associado ao *gating* adaptativo. Adicionalmente, os resultados demonstram a capacidade da proposta em explorar a complementaridade *cross-modal* para amenizar as oscilações estéticas visuais e assincronias que prejudicam as abordagens da literatura.

Para certificar a relevância estatística desses resultados, aplicou-se o teste não-paramétrico de Wilcoxon (*Wilcoxon signed-rank test*) bicaudal sobre as distribuições de F_1^{CNO} dos 11 episódios entre a arquitetura proposta e o ACRNet [12]. O teste apontou significância estatística relevante com $p < 0,05$ ($p = 0,005$), rejeitando a hipótese nula de igualdade e confirmando que a superioridade da abordagem multimodal desenvolvida é consistente, sem variações amostrais aleatórias.

5 CONSIDERAÇÕES FINAIS

Este trabalho propôs uma abordagem multimodal baseada em *Transformers* para a segmentação temporal de vídeos em

cenas. O método introduz uma estratégia de fusão adaptativa que integra características aurais e visuais em nível de tomada, coordenada por um mecanismo de *gating*. A delimitação de fronteiras de cena foi modelada como uma classificação binária, mapeando a probabilidade de transição para cada tomada do fluxo audiovisual.

Os experimentos, conduzidos via validação cruzada *Leave-One-Video-Out* (LOVO), comprovaram a superioridade da técnica multimodal sobre as variantes monomodais. A fusão adaptativa pondera a relevância de cada fluxo, mitigando perturbações visuais locais (como tomadas de cobertura) ao adotar a assinatura acústica (MFCC) como âncora de continuidade narrativa — evidenciando que essa estratégia é importante em domínios como documentários. Corroborando essa evidência (Seção 4.5.1): 1) o modelo aural isolado operou próximo ao concorrente visual TELNet e impulsionou o desempenho na variante multimodal; 2) comprovou-se que a proposta explora pistas contextuais latentes de ambas as modalidades de forma complementar.

A arquitetura superou consistentemente os principais modelos do estado da arte (ACRNet, TELNet e SFGNet), atingindo o maior índice F_1^{CNO} em 10 dos 11 vídeos e média global de 0,82 — um ganho absoluto de 6% sobre o melhor competidor da literatura. O teste não-paramétrico de Wilcoxon confirmou a significância estatística e a consistência dos resultados ($p = 0,005$).

Como trabalhos futuros, vislumbra-se: (i) a incorporação da modalidade textual via processamento de linguagem natural; (ii) a avaliação da arquitetura em bases de dados de outros domínios; (iii) a experimentação com novos extratores de características espaço-temporais profundas; e (iv) uma análise comparativa de eficácia e eficiência substituindo a atenção conjunta por atenção fatorada sobre o mesmo conjunto de dados.

6 CUIDADOS ÉTICOS

Não se aplica.

7 USO DE FERRAMENTAS DE IA

Os autores declaram que utilizaram tecnologias de IA generativa somente como auxílio para correção de *bugs* de programação e para correção gramatical e ortográfica do texto.

8 AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001 – e dos recursos computacionais do *cluster Euler*, do Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI/ICMC/USP).

REFERÊNCIAS

- [1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang,

- and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 24206–24221. https://proceedings.neurips.cc/paper_files/paper/2021/file/cb3213ada48302953cb0f166464ab356-Paper.pdf
- [2] John Edison Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2017. Gated Multimodal Units for Information Fusion. In *International Conference on Learning Representations*.
 - [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6836–6846.
 - [4] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. A Deep Siamese Network for Scene Detection in Broadcast Videos. In *Proceedings of the 23rd ACM International Conference on Multimedia (Brisbane, Australia) (MM '15)*. Association for Computing Machinery, New York, NY, USA, 1199–1202. doi:10.1145/2733373.2806316
 - [5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 813–824. <https://proceedings.mlr.press/v139/bertasius21a.html>
 - [6] Manfred Del Fabro and Laszlo Böszörményi. 2013. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems* 19, 5 (oct 2013), 427–454. doi:10.1007/s00530-013-0306-4
 - [7] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610. doi:10.1016/j.neunet.2005.06.042
 - [8] Bo Han and Weiguo Wu. 2011. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In *2011 IEEE International Conference on Multimedia and Expo*. 1–6. doi:10.1109/ICME.2011.6012001
 - [9] Rodrigo Mitsuo Kishi, Tiago Henrique Trojahn, and Rudinei Goularte. 2019. Correlation based feature fusion for the temporal video scene segmentation task. *Multimedia Tools and Applications* 78, 11 (01 Jun 2019), 15623–15646. doi:10.1007/s11042-018-6959-4
 - [10] Tsung-Yi Lin, Serge Maire, Michael Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13. Springer, 740–755.
 - [11] Siva Jyothi Lingam and Sudha Pelluri. 2025. Multi-modal Video Summarization: A Comprehensive Survey. In *2025 International Conference on Robotics and Mechatronics (ICRM)*. 1–6. doi:10.1109/ICRM66809.2025.11349049
 - [12] Dong Liu, Nagendra Kamath, Subhabrata Bhattacharya, and Rohit Puri. 2020. Adaptive context reading network for movie scene detection. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 9 (2020), 3559–3574.
 - [13] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3202–3211.
 - [14] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViL-BERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf
 - [15] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10143–10152. doi:10.1109/CVPR42600.2020.01016
 - [16] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565. doi:10.18653/v1/P18-1238
 - [17] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1349–1380. doi:10.1109/34.895972
 - [18] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
 - [19] Tiago H. Trojahn, Rodrigo M. Kishi, and Rudinei Goularte. 2018. A new multimodal deep-learning model to video scene segmentation. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web (Salvador, BA, Brazil) (WebMedia '18)*. ACM, New York, NY, USA, 205–212. doi:10.1145/3243082.3243108
 - [20] Shu-Ming Tseng, Zhi-Ting Yeh, Chia-Yang Wu, Jia-Bin Chang, and Mehdi Norouzi. 2023. Video Scene Detection Using Transformer Encoding Linker Network (TELNet). *Sensors* 23, 16 (2023). doi:10.3390/s23167050
 - [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. 5998–6008. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
 - [22] Jeroen Vendrig and Marcel Worring. 2002. Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia* 4, 4 (Dec. 2002), 492–499. doi:10.1109/TMM.2002.802021
 - [23] Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12113–12132. doi:10.1109/TPAMI.2023.3275156
 - [24] Tian Ye, Yang Liu, Mengyu Yang, LanShan Zhang, and Zhigang Li. 2024. Semantic Fusion Based Graph Network for Video Scene Detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
 - [25] Luowei Zhou, Changxu Xu, and Jason J Corso. 2018. Towards Automatic Learning of Procedures from Web Instructional Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7054–7063.
 - [26] Tianfei Zhou, Fatih Porikli, David J. Crandall, Luc Van Gool, and Wenguan Wang. 2023. A Survey on Deep Learning Technique for Video Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7099–7122. doi:10.1109/TPAMI.2022.3225573