Universidade do Minho
Escola de Engenharia

**Universidade do Minho**
Escola de Engenharia
Mestrado em Engenharia Informática

# Unidade Curricular de Aprendizagem Profunda

Ano Letivo de 2024/2025

## AP - Grupo 12

**José Costa**
PG55970

**Pedro Azevedo**
PG57897

**Ruben Silva**
pg57900

**Rui Pinto**
PG56010

Maio, 2025

| | |
|---|---|
| Data da Receção | |
| Responsável | |
| Avaliação | |
| Observações | |

# AP - Grupo 12

Maio, 2025

# Índice

# 1. Introdução

O avanço das técnicas de Aprendizagem Profunda (Deep Learning) tem impulsionado a automação em áreas críticas como a medicina, permitindo o desenvolvimento de sistemas inteligentes capazes de interpretar dados complexos, como vídeos de procedimentos cirúrgicos.

Neste contexto, surge o desafio internacional **OSS: Open Suturing Skills**, parte do **Endoscopic Vision Challenge 2025**, com o objetivo de promover o desenvolvimento de soluções de visão por computador para análise automática de vídeos de treino de sutura cirúrgica.

A crescente digitalização da medicina, aliada à evolução das técnicas de visão por computador, tem permitido o desenvolvimento de ferramentas avançadas para análise automática de procedimentos clínicos. Em particular, a área de cirurgia assistida por computador tem beneficiado do uso de Aprendizagem Profunda para extrair informação relevante a partir de vídeos operatórios, possibilitando a avaliação objetiva do desempenho técnico de cirurgiões em formação.

Este trabalho visa conceber e implementar uma solução computacional baseada em técnicas modernas de Deep Learning, capaz de identificar fases cirúrgicas, detetar erros técnicos e extrair métricas de desempenho a partir de vídeos anotados.

A solução proposta tem como principais metas:
- A precisão na segmentação temporal do procedimento;
- A deteção robusta de falhas técnicas durante a sutura;
- A geração de outputs interpretáveis que possam auxiliar na formação de cirurgiões.

O projeto segue a metodologia de **challenge-based learning**, promovendo uma abordagem prática, colaborativa e orientada a um problema real, com impacto no domínio da educação médica e cirurgia assistida por computador.

# 2. Trabalhos Relacionados

A aplicação de Aprendizagem Profunda à área da medicina tem vindo a crescer, em especial na análise automática de vídeos médicos. Um dos focos mais relevantes é a avaliação de cirurgias, onde se procura identificar fases do procedimento, erros técnicos e métricas de desempenho do cirurgião.

Para resolver este tipo de problema, os trabalhos existentes recorrem geralmente a modelos de Deep Learning que combinam duas capacidades:

- **Extrair informação visual de imagens ou vídeo** (por exemplo, usando redes convolucionais – CNNs);
- **Capturar padrões ao longo do tempo** (com redes LSTM ou Transformers).

Estudos como o de Lee et al. mostram como redes CNN combinadas com LSTM podem ser eficazes na classificação de fases cirúrgicas a partir de vídeo [1]. Uma revisão sistemática recente reforça essa abordagem, salientando o sucesso das CNNs e LSTMs no modelamento de processos cirúrgicos complexos [2].

Mais recentemente, modelos baseados em **Transformers para vídeo**, como o Trans--SVNet, têm sido propostos com resultados promissores. Estes modelos conseguem capturar relações espaciais e temporais de forma mais eficiente, sendo indicados para vídeos longos e com transições suaves entre fases [3].

Na área médica, é comum trabalhar com **dados limitados e desbalanceados**, e muitos autores adotam técnicas de aumento de dados, validação cruzada e normalização de entradas [2].

Além disso, a interpretabilidade dos modelos é crítica. Técnicas como o **Grad-CAM** têm sido aplicadas com sucesso para tornar as decisões de redes neuronais mais compreensíveis para médicos. Exemplos incluem sua utilização em imagens de ressonância magnética para tumores cerebrais [4] e em mamografias para deteção de cancro da mama [5].

## 2.1. Reconhecimento de Fases Cirúrgicas com CNN--LSTM

O reconhecimento de fases cirúrgicas é uma tarefa crítica na análise de vídeos médicos, com impacto direto na avaliação de desempenho e automatização de sistemas de apoio à decisão clínica. Modelos que combinam Redes Convolucionais (CNNs) com Redes Recorrentes (LSTM) continuam a ser amplamente utilizados, devido à sua capacidade de lidar simultaneamente com características espaciais (de cada frame) e temporais (entre frames).

Uma contribuição notável nesta área é o trabalho de Yengera et al., que propôs uma abordagem auto-supervisionada na qual a rede é treinada para prever o tempo restante

da cirurgia. Esta tarefa auxiliar (regressão do tempo restante) força o modelo a aprender representações temporais ricas, que são depois reutilizadas para reconhecer as fases cirúrgicas com maior precisão, mesmo com menos anotações manuais [6]. Este tipo de pré-treinamento é especialmente útil em contextos onde os dados anotados são escassos, como é comum em medicina.

## 2.2. Transformers na Cirurgia: Trans-SVNet

Nos últimos anos, a arquitetura Transformer, originalmente concebida para processamento de linguagem natural, tem-se revelado promissora também na análise de vídeos. O modelo Trans-SVNet, introduzido por Gao et al., representa um avanço importante ao combinar embeddings espaciais (das imagens) com atenção temporal agregada, o que permite capturar dependências complexas ao longo do tempo de forma mais eficaz do que modelos LSTM tradicionais.

O Trans-SVNet foi desenhado especificamente para reconhecer fases cirúrgicas, integrando dois módulos principais: um encoder espacial baseado em CNNs para extrair features dos frames e um módulo Transformer para agregar essas informações ao longo do tempo. Os resultados demonstraram melhorias substanciais sobre os modelos anteriores, particularmente em cirurgias com maior variação na duração e sequência das fases [7]. Este modelo também abre portas para a integração de atenção multimodal (vídeo + sensores).

## 2.3. Avaliação de Habilidades com Redes 3D

Para além da classificação de fases, a avaliação objetiva da destreza técnica de cirurgiões é um problema relevante e tradicionalmente sujeito a variações entre avaliadores humanos. Neste contexto, Funke et al. propuseram uma abordagem baseada em Redes Convolucionais 3D, que considera não apenas as dimensões espaciais de cada frame, mas também a dimensão temporal diretamente, tratando sequências de vídeo como volumes tridimensionais.

Este tipo de rede é capaz de identificar padrões de movimento e fluidez técnica que distinguem operadores experientes de novatos. No estudo, foram utilizadas gravações de tarefas de laparoscopia, com o modelo alcançando elevado desempenho na predição de scores de habilidade previamente atribuídos por especialistas [8]. A sua aplicabilidade prática está na automatização da avaliação em contexto de treino cirúrgico.

## 2.4. Grad-CAM para Justificação Clínica

Em aplicações médicas, não basta que um modelo produza uma resposta correta — é crucial que a sua decisão possa ser explicada de forma compreensível por humanos, especialmente médicos. A técnica Grad-CAM (Gradient-weighted Class Activation Mapping) tem-se tornado uma ferramenta central neste aspeto.

Grad-CAM gera mapas de calor que sobrepõem as regiões da imagem que mais contribuíram para a decisão do modelo. Isto permite, por exemplo, verificar se uma rede neural

está realmente a focar-se em áreas clinicamente relevantes (como uma lesão ou tumor) em vez de ruídos ou artefactos. Selvaraju et al. demonstraram a utilidade do Grad-CAM em diversos domínios, incluindo a análise de imagens radiológicas e detecção de patologias em exames de imagem [9]. A técnica tem sido amplamente adotada em sistemas médicos com foco em confiabilidade e auditoria.

## 2.5. Outros Trabalhos Relevantes

Outras contribuições incluem:

- **DeepPhase**: reconhecimento de fases em cirurgias de catarata, usando CNNs treinadas para identificar etapas críticas [10].
- **SV-RCNet**: integração entre CNNs e redes recorrentes para modelagem do fluxo de procedimentos laparoscópicos [11].
- **LRTD**: abordagem ativa e com foco em dependências temporais longas, relevante para contextos de anotações escassas e vídeos longos [12].

# 3. Descrição dos Dados

Esta secção descreve detalhadamente o conjunto de dados utilizado para abordar as três tarefas propostas no desafio **Open Suturing Skills (OSS) 2025**. O dataset baseia-se em gravações vídeo de procedimentos de sutura realizados por estudantes e residentes, e contém múltiplas modalidades de anotação relevantes para a avaliação automática de competência cirúrgica.

## 3.1. Composição do Conjunto de Dados

O dataset principal é composto por vídeos com aproximadamente cinco minutos de duração, captados em ambiente controlado de treino com uma câmara colocada em perspetiva top-down. Esta configuração facilita a visibilidade das mãos e instrumentos durante o procedimento.

Cada vídeo é identificado de forma única e está associado a um conjunto de anotações disponibilizadas em ficheiros `.csv`, provenientes de três avaliadores humanos independentes. As anotações dividem-se em:

- **Global Rating Score (GRS)**: pontuação global da performance cirúrgica (escala 8–40), usada na Task 1.
- **OSATS**: avaliação técnica objetiva com base em oito critérios, cada um avaliado numa escala de 1 a 5. Estes dados suportam a Task 2.

## 3.2. Dados Complementares para a Task 3

Para a Task 3, foi fornecido um subconjunto adicional com anotações específicas para **tracking**. Estas incluem:

- **Keypoints** de mãos e ferramentas (ex. articulações e pontas dos instrumentos)
- **Máscaras de segmentação semântica** para classes como mão esquerda, porta-agulhas, pinça, entre outros
- **Flags de visibilidade**, com os valores:
  - 0: fora de frame
  - 1: oculto
  - 2: visível

Este conjunto é essencial para a geração de métricas cinemáticas e para a modelação dos padrões de movimento ao longo do tempo.

# 4. Metodologia

Esta secção descreve a abordagem seguida no desenvolvimento da solução para o desafio OSS – Open Suturing Skills, estruturada em três tarefas principais:

- Classificação do nível global de competência cirúrgica (GRS)
- Avaliação detalhada com base nos critérios OSATS,
- Rastreamento de mãos e instrumentos.

A metodologia inclui o pré-processamento dos dados, extração de métricas cinemáticas, construção e treino de modelos de aprendizagem profunda e os métodos de avaliação utilizados.

## 4.1. Pré-processamento

O conjunto de dados fornecido consiste em vídeos de aproximadamente cinco minutos de duração, captados em perspetiva de cima para baixo, que documentam sessões de treino de sutura realizadas por estudantes e residentes. Estes vídeos estão acompanhados de anotações em ficheiro tabular, contendo os valores das métricas OSATS e do Global Rating Score (GRS), avaliados por três classificadores independentes.

Inicialmente, os dados foram organizados de forma a garantir consistência entre vídeos e anotações:

- Os valores foram agregados por média aritmética, como sugerido pelo desafio.

### 4.1.1. Extração de Frames

Para possibilitar a análise visual frame-a-frame e posterior rastreamento de keypoints, cada vídeo foi segmentado em frames utilizando uma taxa de amostragem de 1 frame por segundo. Esta escolha visa equilibrar a riqueza temporal com a viabilidade computacional durante a fase de tracking.

### 4.1.2. Pré-processamento Adicional dos Frames

De modo a garantir consistência e qualidade nas imagens processadas pelos modelos de aprendizagem profunda, foi aplicado um conjunto adicional de transformações aos frames extraídos:

- Redimensionamento: todos os frames foram redimensionados para uma resolução fixa (256×256 píxeis), assegurando compatibilidade com arquiteturas como ResNet, HRNet ou Transformers visuais.
- Remoção de Ruído e Aumento de Contraste: foram aplicados filtros de suavização, como Gaussian Blur, bem como técnicas de aumento de contraste local como CLAHE.

- Normalização RGB: os valores dos píxeis foram normalizados por canal (R, G, B) usando a fórmula `(x - média) / desvio-padrão` para facilitar o treino.

#### 4.1.3. Normalização de Labels

As anotações de GRS foram reclassificadas em quatro classes:
- Classe 0: GRS de 8 a 15 (Novice)
- Classe 1: GRS de 16 a 23 (Intermediate)
- Classe 2: GRS de 24 a 31 (Proficient)
- Classe 3: GRS de 32 a 40 (Expert)

Para OSATS, os valores de 1–5 foram convertidos para 0–4 e agregados por média.

#### 4.1.4. Divisão dos Dados

- 70% para treino
- 15% para validação
- 15% para teste

A divisão foi estratificada por GRS e OSATS, comum a todas as tarefas.

## 4.2. Tracking de Mãos e Ferramentas

A Task 3 do desafio OSS requer o desenvolvimento de uma solução de rastreamento capaz de detetar mãos e instrumentos cirúrgicos em vídeos de treino de sutura. Esta tarefa é essencial para a extração de métricas cinemáticas e para compreender os padrões de movimento associados a diferentes níveis de competência.

#### 4.2.1. Conjunto de Dados Anotados

Foi disponibilizado um conjunto de dados com anotações detalhadas de keypoints e máscaras de segmentação. O conjunto de treino consiste em frames a 1 frame por minuto; o de validação a 1 frame por segundo.

As anotações incluem:
- Segmentação semântica de 6 classes (mãos e ferramentas)
- Keypoints das mãos (polegar, dedos, dorso)
- Keypoints das ferramentas (ponta da agulha, juntas, etc.)
- Flags de visibilidade (0: fora de frame, 1: oculto, 2: visível)

#### 4.2.2. Arquitetura Utilizada

Foi treinado um modelo baseado em HRNet, ajustado a partir de pesos pré-treinados em COCO/MPII. A tarefa foi tratada como regressão de heatmaps para cada keypoint.

#### 4.2.3. Pós-processamento dos Keypoints

- Extração via `argmax` dos heatmaps
- Normalização para coordenadas relativas
- Interpolação temporal para lidar com visibilidade parcial

- Suavização com filtro Gaussiano 1D

#### 4.2.4. Avaliação do Modelo

A métrica HOTA (Higher Order Tracking Accuracy) adaptada à deteção de keypoints foi usada, penalizando erros de localização e identidade.

#### 4.2.5. Output e Formato

Os resultados foram armazenados no formato MOTChallenge modificado:
- ID do vídeo
- Frame
- Tipo de entidade
- Coordenadas dos keypoints
- Flags de visibilidade

## 4.3. Extração de Features Cinemáticas

Com base nos keypoints obtidos através do rastreamento de mãos e instrumentos, foi possível gerar um conjunto de características (features) que descrevem os movimentos realizados ao longo dos vídeos. Estas características foram utilizadas como entrada nos modelos de classificação das Tarefas 1 e 2.

#### 4.3.1. Objetivo

Transformar sequências de coordenadas espaciais em vetores numéricos fixos que capturam fluidez, coordenação e precisão.

#### 4.3.2. Features Extraídas

A partir das coordenadas (x, y) dos keypoints normalizados ao longo do tempo, foram extraídas:

- Velocidade média dos movimentos
- Aceleração média
- Distância total percorrida
- Número de paragens (frames com baixa velocidade)
- Variação da velocidade (desvio padrão)
- Índice de fluidez (mudanças angulares)
- Aproximações entre mãos e instrumentos
- Duração média do contacto mão-instrumento
- Correlação entre os movimentos das duas mãos

#### 4.3.3. Normalização Temporal

As sequências foram interpoladas para 300 frames por vídeo para uniformizar a estrutura dos dados.

#### 4.3.4. Vetor Final

Cada vídeo foi representado por um vetor com todas as métricas acima, servindo como entrada nos classificadores de GRS e OSATS.

## 4.4. Classificação do Global Rating Score (GRS)

#### 4.4.1. Arquitetura do Modelo

Foram exploradas várias arquiteturas de redes neurais para a tarefa de classificação do Global Rating Score (GRS), utilizando sequências de imagens como entrada. As arquiteturas incluíram redes convolucionais simples, modelos híbridos CNN-LSTM e variantes com Transformers e EfficientNet.

Os modelos recebiam como entrada uma sequência de frames RGB extraídos dos vídeos de sutura, que eram processados para prever uma das quatro classes de competência cirúrgica.

#### 4.4.2. Estratégia de Treino

- Função de perda: `CrossEntropyLoss`
- Otimizador: `Adam` com taxa de aprendizagem de `0.001`
- Batch size: `4`
- Early stopping com base no F1-score de validação
- Divisão dos dados: treino (70%), validação (15%), teste (15%)

#### 4.4.3. Avaliação do Desempenho

O desempenho foi avaliado com base no F1-score macro e na matriz de confusão. As classes centrais (1 e 2) obtiveram melhores resultados, com maior dificuldade em distinguir os extremos.

#### 4.4.4. Modelos Testados

- **CNNModel_1 a CNNModel_4**: Arquiteturas CNN simples com 2 a 3 camadas convolucionais + pooling, seguidas de MLP. Os frames são processados individualmente e agregados por média. As ativações usam ReLU e Softmax para saída.
- **CNNLSTMClassifier**: CNN usada para extrair embeddings espaciais, seguida por um LSTM que processa a sequência de embeddings no tempo. A saída do LSTM alimenta uma camada `Linear → Softmax` para prever a classe GRS.
- **EfficientNetLSTM**: Usa `EfficientNet-b0` (pré-treinada) como backbone de extração visual. Cada frame é passado pela rede e os embeddings são agregados por um LSTM. A saída do LSTM vai para uma `Linear → Softmax`.
- **ViTLSTM**: Usa Vision Transformer (`vit_b_16`) pré-treinado para cada frame, gerando uma sequência de embeddings que é processada por um LSTM. A previsão final é feita por um MLP.

## 4.5. Classificação dos Scores OSATS

### 4.5.1. Arquitetura Utilizada

Foram testadas diversas arquiteturas para a tarefa de previsão dos critérios OSATS. Os modelos incluíram desde redes convolucionais simples (CNN), arquiteturas com ResNet-18 para extração de características visuais, até combinações com LSTM ou GRU para captura de dependências temporais.

Adicionalmente, foram avaliadas variantes com Vision Transformers (ViT). As saídas foram modeladas como uma regressão multi-saída discreta, com 8 neurónios correspondentes aos critérios OSATS.

### 4.5.2. Estratégia de Treino

- Função de perda: soma das `CrossEntropyLoss` individuais
- Otimizador: `Adam`, com taxa de aprendizagem `0.0005`
- Batch size: `4`
- Early stopping com base no F1-score macro médio
- Targets normalizados com z-score

### 4.5.3. Avaliação

O modelo foi avaliado com F1-score por critério e expected cost. Melhor desempenho foi obtido nos critérios motores (e.g., MOTION, INSTRUMENT), com menor desempenho nos critérios subjetivos como KNOWLEDGE.

### 4.5.4. Modelos Testados

- **CNNModel_1 a CNNModel_4**: CNNs com duas camadas convolucionais e pooling. A saída é passada por um MLP com uma cabeça de 8 neurónios. A função de perda é a soma das CrossEntropyLoss para cada critério OSATS.
- **ResNet18 + MLP**: Usa `ResNet-18` da `torchvision`, com a camada final substituída por uma MLP com 8 saídas (uma por critério). Os pesos foram ajustados para previsões multi-saída discretas.
- **ResNet18 + GRU**: Após a extração de features pela ResNet, os embeddings de cada frame passam por uma GRU que aprende a progressão temporal. A GRU é seguida por camadas densas que mapeiam para os 8 scores.
- **ViT + MLP**: Implementa `ViT` como backbone. Cada frame é passado por um Vision Transformer para obter embeddings, que são concatenados e processados por um MLP com 8 cabeças. Usado para classificação ordinal por critério.

## 4.6. Resultados

Esta secção apresenta os resultados obtidos nas tarefas de classificação de competência cirúrgica (GRS) e previsão dos critérios OSATS. Foram testados vários modelos, com

foco na comparação de desempenho em métricas padronizadas como F1-score, matriz de confusão e accuracy.

#### 4.6.1. Resultados da Tarefa 1: Classificação GRS

Os modelos foram treinados para classificar os participantes em quatro níveis de competência.

Os melhores desempenhos foram registados com:
- `EfficientNetLSTM` — melhor trade-off entre complexidade e performance.
- `CNNLSTMClassifier` — desempenho competitivo.
- `ViTLSTM` — resultados promissores, mas mais sensível ao overfitting.

#### 4.6.2. Resultados da Tarefa 2: Classificação OSATS

Nesta tarefa, os modelos preveram oito critérios técnicos OSATS. O desempenho foi avaliado com F1-score macro por critério e `Expected Cost`, conforme as métricas oficiais do desafio.

As arquiteturas com `ResNet18` e `EfficientNet` seguidas de MLP ou GRU destacaram-se pela robustez e generalização. O modelo `ResNet18 + MLP` alcançou os melhores resultados médios.

#### 4.6.3. Comparação Global

Ambas as tarefas evidenciaram que arquiteturas híbridas CNN + LSTM ou CNN + GRU proporcionam uma boa captação da dinâmica temporal sem perda de estrutura visual.

Transformers (ViT) mostraram potencial, mas requerem datasets maiores para generalizar adequadamente.

O resumo dos melhores modelos por tarefa é:

- GRS: `EfficientNetLSTM` com F1-macro mais alto
- OSATS: `ResNet18 + MLP` com menor Expected Cost e maior estabilidade

Estes resultados suportam a adoção de pipelines multimodais e supervisionados com rastreamento de keypoints como base para avaliação automática de competências técnicas em cirurgia.

# 5. Conclusão

Neste trabalho, foi desenvolvida uma solução baseada em aprendizagem profunda para avaliação de competência cirúrgica a partir de vídeos de treino de sutura, no contexto do desafio OSS 2025.

Foram abordadas duas tarefas principais:
- A classificação do nível global de competência (GRS);
- A previsão detalhada dos critérios técnicos OSATS.

A abordagem combinou técnicas modernas de extração de keypoints, processamento de vídeo e modelos multimodais, incluindo CNNs, LSTMs, GRUs, Transformers e variantes híbridas. O pipeline completo abrangeu desde o pré-processamento de dados até à avaliação com métricas especializadas (F1-score, matriz de confusão, Expected Cost).

Os resultados demonstraram que:
- Arquiteturas CNN+LSTM e CNN+GRU mostraram-se eficazes em ambas as tarefas.
- Modelos baseados em Vision Transformers revelaram potencial, mas exigem maior volume de dados para generalização.

### 5.0.1. Trabalho Futuro

A partir do trabalho desenvolvido, identificam-se várias direções possíveis para aprofundamento e continuidade do projeto, tanto do ponto de vista científico como educacional:

- **Aprofundamento de Modelos e Estratégias de Treino**: Estudar o impacto de diferentes estratégias de normalização, regularização e loss functions em tarefas multiclasse e multisaída. Avaliar variantes de treino multitarefa e transfer learning para melhorar a generalização.
- **Avaliação da Variabilidade Interavaliador**: Estudar o efeito da subjetividade das anotações humanas (especialmente em OSATS) e propor estratégias para calibrar ou suavizar essa variabilidade, como consensus learning.
- **Estudo da Explicabilidade dos Modelos**: Analisar as decisões dos modelos com ferramentas de explicabilidade (e.g., saliency maps, Grad-CAM) para verificar se os modelos estão a focar nas regiões anatómicas e instrumentos relevantes.

# Bibliografia


[1] H. Lee, H. Yoon, e S. Kang, «A Multimodal Transformer Model for Recognition of Images from Surgical Videos», *Journal of Biomedical Informatics*, 2023, [Online]. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11011728/

[2] A. Shukla e others, «Deep learning in surgical process modeling: A systematic review», *Journal of Biomedical Informatics*, 2025, [Online]. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1532046425000085

[3] X. Gao, Y. Wang, e S. Liu, «Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer», *arXiv preprint arXiv:2103.09712*, 2021, [Online]. Disponível em: https://arxiv.org/abs/2103.09712

[4] A. Bansal e R. Singh, «Enhancing brain tumor detection in MRI images through explainable deep learning», *BMC Medical Imaging*, 2024, [Online]. Disponível em: https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-024-01292-7

[5] A.-K. Balve e P. Hendrix, «Interpretable breast cancer classification using CNNs on mammographic images», *arXiv preprint arXiv:2408.13154*, 2024, [Online]. Disponível em: https://arxiv.org/abs/2408.13154

[6] G. Yengera, D. Mutter, J. Marescaux, e N. Padoy, «Less is More: Surgical Phase Recognition with Less Annotations through Self-Supervised Pre-training of CNN--LSTM Networks», *arXiv preprint arXiv:1805.08569*, 2018.

[7] X. Gao, Y. Jin, Y. Long, Q. Dou, e P.-A. Heng, «Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer», em *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 593–603.

[8] I. Funke, S. T. Mees, J. Weitz, e S. Speidel, «Video-based surgical skill assessment using 3D convolutional neural networks», *arXiv preprint arXiv:1903.02306*, 2019.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, e D. Batra, «Grad--CAM: Visual Explanations from Deep Networks via Gradient-based Localization», *arXiv preprint arXiv:1610.02391*, 2017.

[10] U. Author, «DeepPhase: Surgical Phase Recognition in CATARACTS Videos», *MICCAI*, 2018.

[11] U. Author, «SV-RCNet: Workflow Recognition from Surgical Videos using Recurrent Convolutional Network», *TMI*, 2018.

[12] X. Shi, Y. Jin, Q. Dou, e P.-A. Heng, «LRTD: Long-range Temporal Dependency based Active Learning for Surgical Workflow Recognition», *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 1573–1584, 2020.