

Avaliação do Desempenho entre U-Net, TransUNet e Swin-UNet++ com Dataset Térmico Charlotte-ThermalFace

Felipe Bedinotto Fava¹, Mariana Rodrigues Padilha¹, Claudio Schepke¹,
Fabricio Neitzke Ferreira²

¹ Programa de Pós-graduação em Engenharia de Software (PPGES)
Universidade Federal do Pampa (UNIPAMPA), Alegrete, Brazil

²Instituto Federal Sulriograndense (IFSul) - Pelotas, Brasil

{felipefava.aluno, marianarp.aluno, claudioschepke}@unipampa.edu.br

fabricioferreira@ifsul.edu.br

Abstract. *Temperature is a crucial parameter for assessing a patient's health. Its measurement requires physical contact or the generation of infrared images, a non-invasive method, but it requires appropriate treatment to achieve valid clinical accuracy. In this paper, we propose a comprehensive, production-ready, fully automated computer vision pipeline for Thermal Facial Region Segmentation. The project evaluates, trains, and benchmarks three segmentation algorithms — U-Net, TransUNet, and Swin-UNet++ — under identical conditions. We use the Charlotte ThermalFace dataset to run the pipeline of the implemented experiment. The results show that TransUNet achieved the best absolute performance in geometrical fidelity, with a Mean Intersection over Union (mIoU) of 0.6145 and a Dice-Sørensen coefficient of 0.5637; Swin-UNet++ demonstrated total speed supremacy, requiring only 12.55 ms per image inference, which translates to a transfer rate of 79.66 Frames Per Second (FPS); and U-Net established the lowest memory consumption.*

Resumo. *A temperatura é um parâmetro crucial para avaliar a saúde de um paciente. Sua medição é realizada por métodos que requerem contato físico ou através da geração de imagens infravermelhas, um método não invasivo, mas que requer tratamento adequado para a obtenção de precisão clínica válida. Neste artigo, propomos um pipeline de visão computacional abrangente e automatizado para Segmentação Térmica da Região Facial. O projeto avalia, treina e compara o desempenho de três algoritmos de segmentação — U-Net, TransUNet e Swin-UNet++ — sob condições idênticas. Utilizamos o conjunto de dados Charlotte ThermalFace para executar o pipeline de um experimento. Os resultados mostram que TransUNet alcançou o melhor desempenho absoluto em fidelidade geométrica, com uma Interseção Média sobre União (mIoU) de 0,6145 e um Coeficiente de Dice-Sørensen de 0,5637; Swin-UNet++ demonstrou maior vazão de processamento, exigindo apenas 12,55 ms por inferência de imagem, o que se traduz em uma taxa de transferência de 79,66 quadros por segundo (FPS); e U-Net apresentou o menor consumo de memória.*

1. Introdução

A exploração de algoritmos de segmentação de regiões é vasta e abrange diversos setores, com ampla aplicação em problemas reais. Um dos pontos de estudo é a busca por análises clínicas de humanos usando visão computacional. A análise facial e o monitoramento biomédico contínuo por meio de sistemas de visão computacional têm experimentado avanços significativos na última década, impulsionados principalmente pela evolução das Redes Neurais Convolucionais (CNNs) [Robledo-Vega, I. et al. 2025].

Historicamente, detecção, reconhecimento e segmentação, são trabalhados no espectro de luz visível, com imagens RGB [Nowakowski and Kaczmarek 2025]. Embora esse tipo de abordagem tenha alcançado precisão notável em um nível de confiabilidade e adoção significativa em aplicações comerciais, há diversas limitações, principalmente físicas, como iluminação, características do sensor, especificações da lente, etc. A dependência fundamental de fótons refletidos torna o sistema que usa luz visível suscetível a falhas em ambientes com pouca ou muita luz, oclusões causadas por sombras e variações extremas de iluminação [Sheng, D. et al. 2025].

Por outro lado, dados térmicos, especialmente aqueles que operam no espectro infravermelho de onda-longa (LWIR), capturam a radiação eletromagnética emitida diretamente pelos próprios objetos, baseando-se na assinatura térmica e na emissividade superficial de cada objeto [Stanić and Geršak 2025]. A termografia facial, enquanto modalidade de sensoriamento não invasivo, possui implicações profundas, expansivas e multifacetadas, estendendo-se muito além da utilização mais comumente conhecida na vigilância noturna ou no reconhecimento facial à distância, por exemplo [Sheng, D. et al. 2025].

Para que os algoritmos extraiam sinais fisiológicos sutis com a precisão exigida para validação clínica, é imprescindível um isolamento prévio, semântico e altamente preciso das Regiões de Interesse (ROIs) faciais relevantes, como: i) testa, ii) bochechas e iii) olhos [Jerbi and Lee 2024]. A separação de componentes faciais adjacentes, como a transição imperceptível entre o lábio superior e a região inferior adjacente do nariz, ou a delimitação das sobrancelhas isoladas da planície da testa — não é demarcada por gradientes de luminância nítidos, mas sim por variações sutis, graduais e frequentemente corrompidas por ruído térmico nos valores de intensidade dos dados radiométricos brutos.

Para contornar essa escassez de bordas bem definidas e alto contraste, a inteligência artificial médica tem empregado de aprendizado profundo de densidade variável e campos receptivos adaptativos. A arquitetura U-Net, consolidada na literatura desde 2015 [Ronneberger et al. 2015], estabeleceu um padrão inicial utilizando o viés espacial das convoluções para delinear características em imagens biomédicas. Elas são conhecidas por sua estrutura em formato de “U” (codificador-decodificador), que permite uma localização precisa em nível de pixel.

A comparação entre diferentes arquiteturas de redes neurais é frequentemente prejudicada pela falta de padronização dos dados de treinamento, das técnicas de aumento de dados e das métricas de avaliação. Diante disso, o **objetivo** deste trabalho é definir em quais métricas cada algoritmo se sobressai. Para alcançar esse objetivo, torna-se necessário estabelecer um procedimento metodológico consistente que permita avaliar, sob condições controladas e reproduzíveis, o comportamento de diferentes arquiteturas frente ao mesmo conjunto de desafios experimentais.

Este artigo apresenta um sistema metodológico rigoroso, detalhado e extenso para quantificar, qualificar e examinar o desempenho de algoritmos que seguem uma mesma arquitetura, com o desafio específico da segmentação facial térmica multissegmentada. Através da análise dos resultados comparativos derivados de testes empíricos intensivos, esta pesquisa realiza uma avaliação paramétrica da clássica U-Net convolucional pura, da arquitetura híbrida TransUNet [Chen, J. et al 2021] e do Swin-UNet++ [Liu, P. et al. 2021]. A **contribuição** principal deste estudo é estabelecer não apenas qual modelo domina as métricas convencionais de Interseção sobre União Média (mIoU) e o Coeficiente de Similaridade de Dice, mas também elucidar os gargalos teóricos de complexidade espacial alocada, eficiência de inferência temporal (medida rigorosamente em FPS) e alocação volátil de memória de hardware (Peak GPU Memory).

O restante do artigo está estruturado da seguinte forma. A Seção 2 discute os trabalhos relacionados. Na Seção 3 estão detalhados os passos da implementação do experimento proposto, incluindo as variações entre U-Net, TransUNet e Swin-UNet++. A metodologia é detalhada na Seção 4. Os resultados obtidos são apresentados na Seção 5. Por fim, a Seção 6 apresenta as contribuições e trabalhos futuros.

2. Trabalhos Relacionados

A termografia infravermelha tem sido amplamente explorada em aplicações biomédicas devido ao seu caráter não invasivo e à capacidade de capturar variações térmicas associadas a processos fisiológicos. Nesse contexto, o trabalho de [Nowakowski and Kaczmarek 2025] apresenta uma revisão abrangente do uso de inteligência artificial em imagens térmicas médicas, destacando aplicações em classificação e detecção de padrões relacionados a doenças como câncer de mama e pé diabético. Apesar de consolidar o estado da arte, o estudo evidencia desafios relevantes, como baixa qualidade das imagens, escassez de dados e ausência de padronização experimental, além de focar predominantemente em tarefas de classificação global ou segmentação binária, sem abordar a segmentação semântica detalhada de múltiplas regiões faciais.

Complementarmente, [Stanić and Geršak 2025] realizam uma revisão sistemática focada especificamente em termografia facial, enfatizando a importância de protocolos rigorosos para aquisição e análise dos dados. Os autores destacam o uso de regiões de interesse (ROIs) como elemento central na interpretação dos termogramas, bem como fatores críticos como emissividade, condições ambientais e preparo do sujeito. No entanto, a definição dessas regiões é realizada de forma manual ou simplificada, sem o uso de técnicas automatizadas baseadas em aprendizado profundo, o que pode introduzir variabilidade e limitar a reprodutibilidade dos resultados.

No âmbito de aplicações práticas, [Robledo-Vega, I. et al. 2025] propõem um sistema embarcado para detecção de faces em imagens térmicas utilizando modelos baseados em YOLO. Os resultados demonstram alta precisão e viabilidade em tempo real, mesmo em condições adversas de iluminação. Entretanto, a abordagem restringe-se à detecção por meio de caixas delimitadoras, fornecendo apenas a localização aproximada dos rostos, sem detalhamento pixel a pixel das estruturas faciais, o que limita sua aplicação em análises fisiológicas mais refinadas.

Outras abordagens exploram o processamento avançado de imagens térmicas com foco na melhoria visual. Nesse sentido, [Sheng, D. et al. 2025] propõem um modelo ba-

seado em difusão para conversão de imagens infravermelhas em representações pseudo-coloridas, utilizando mecanismos de auto-segmentação para guiar o processo generativo. Embora apresente ganhos significativos em qualidade visual e preservação semântica, a segmentação é empregada apenas como componente auxiliar, não sendo tratada como tarefa principal nem avaliada de forma independente.

A arquitetura U-Net [Ronneberger et al. 2015] consolidou-se como um dos principais modelos no contexto da segmentação de imagens biomédicas, sendo amplamente utilizada como baseline devido à sua capacidade de produzir segmentações precisas mesmo com poucos dados anotados. No entanto, sua dependência de operações convolucionais limita a modelagem de dependências de longo alcance, o que pode comprometer o desempenho em domínios com baixo contraste e fronteiras difusas, como imagens térmicas.

Para superar essas limitações, [Chen, J. et al 2021] propõem a TransUNet, uma arquitetura híbrida que combina CNNs e Transformers, permitindo capturar simultaneamente características locais e relações globais. Essa abordagem apresenta ganhos consistentes em métricas como *Dice Score* e *Hausdorff Distance* em tarefas de segmentação médica, indicando maior robustez em regiões complexas.

Mais recentemente, [Liu, P. et al. 2021] introduzem a Swin-UNet++, baseada em Transformers hierárquicos, que utilizam mecanismos de atenção em janelas deslocadas para melhorar a eficiência computacional e a modelagem de estruturas complexas. Apesar de ter sido proposta em um contexto distinto, seus resultados evidenciam a capacidade de lidar com fronteiras pouco definidas, característica relevante para imagens térmicas.

Apesar dos avanços apresentados, observa-se que a maioria dos trabalhos aborda tarefas distintas, como classificação, detecção ou melhoria visual, ou ainda avalia arquiteturas sob condições experimentais heterogêneas, dificultando comparações diretas. Diante desse cenário, o presente trabalho propõe uma avaliação sistemática de arquiteturas de segmentação (U-Net, TransUNet e Swin-UNet++) sob um protocolo experimental padronizado, utilizando um dataset anotado com múltiplas regiões anatômicas faciais. Diferentemente dos trabalhos existentes, esta abordagem foca na segmentação pixel a pixel como objetivo central, além de analisar explicitamente métricas de eficiência computacional, como tempo de inferência, FPS e uso de memória, contribuindo para o desenvolvimento de soluções mais robustas e aplicáveis em cenários reais.

3. Processo de Desenvolvimento Aplicado

A arquitetura experimental deste artigo foi delineada visando quantificar a eficiência teórica, espacial e temporal de três modelos em disputa. O projeto desenvolvido em Python submeteu esses modelos a um processo intensivo, utilizando um fluxo de processamento de dados térmicos estritamente controlado.

Um processo automatizado foi criado para o processamento e exploração dos experimentos¹. A Figura 1 apresenta os passos desenvolvidos para a execução do pipeline. O pipeline que conduz o experimento é composto por uma etapa de autoinstalação, onde as dependências são controladas na execução do projeto.

Uma vez otimizadas e instaladas as ferramentas computacionais, o conjunto de dados é descompactado. Na sequência é executado a etapa de tratamento e validação dos

¹<https://github.com/bedinotto/UBench>

dados para uso no pipeline para cada um dos três algoritmos de redes convolucionais escolhidos. O conjunto de dados para os experimentos é importado do repositório GitHub² [Ashrafi et al. 2022] e para o armazenamento no repositório deste projeto, é utilizado um esquema de *Git Large File Storage* (LFS).

A ferramenta dimensiona dinamicamente, modo autônomo, o tamanho do lote (*batch size*) e a precisão mista, conforme as capacidades da GPU detectada no ambiente de execução. Os requisitos de hardware mínimos consistem em: i) GPU NVIDIA com pelo menos 6 GB de VRAM; ii) Sistema Operacional com suporte ao CUDA 11.8 ou 12.1; iii) Python 3.8 a 3.12 instalado.

4. Metodologia

Considerando os processos técnicos descritos no capítulo anterior, o projeto propõe-se a automatizar e equalizar o ambiente de testes dos três modelos, a fim de torná-lo o mais adequado possível para uma comparação de resultados, partindo desde a padronização dos dados brutos, anotações e informações auxiliares até as ferramentas e bibliotecas usadas para o treinamento dos algoritmos.

O conjunto de dados utilizado, denominado Charlotte ThermalFace [Ashrafi et al. 2022], criado por pesquisadores da *University of North Carolina at Charlotte* é composto por 10.000 imagens térmicas infravermelhas, em formato *.tiff*, de 10 indivíduos em diferentes condições térmicas, a várias distâncias da câmera e com diferentes posições da cabeça, como: frontal, lado direito, lado esquerdo, entre outras. De acordo com as anotações feitas pelos autores do dataset e presentes no conjunto de dados, para este artigo foram criadas diferentes classes fisiológicas, de acordo com a seguinte organização, também semelhante a outros conjuntos de dados:

- *Background* (espaço de fundo inerte e assinaturas de corpo periférico).
- Contorno inferior do rosto (linha da mandíbula térmica e convexidade do queixo).
- Sobrancelha esquerda (aglomeração folicular térmica).
- Sobrancelha direita (aglomeração folicular térmica).
- Nariz (pirâmide nasal e narinas termicamente voláteis devido ao fluxo de ar).
- Olho esquerdo (globo ocular, fenda palpebral de alta emissividade e canto interno).
- Olho direito (globo ocular e anexos perioculares).
- Boca (borda oral interna e cavidade de exalação gasosa).

²<https://github.com/TeCSAR-UNCC/UNCC-ThermalFace>

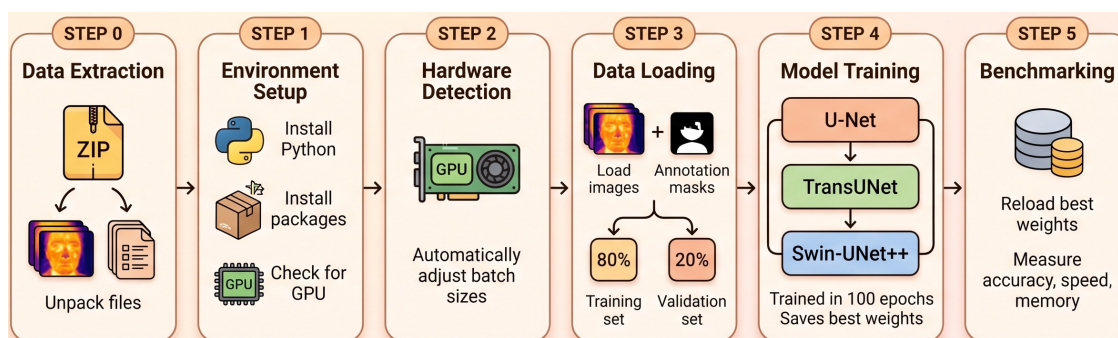


Figura 1. Passos de execução do *pipeline* do projeto

- Lábios (membranas mucosas térmicas em constante flutuação e resfriamento por evaporação).
- Testa (planície superior frontal).

Foram avaliados dois valores de época: 50 e 100, com o objetivo de determinar uma configuração adequada para o treinamento do modelo. A literatura costuma adotar valores próximos de 100 épocas em tarefas semelhantes [Pedregosa, F. et al. 2011]. Após as avaliações, estabeleceu-se o uso de 100 épocas como configuração padrão para a geração de todos os resultados apresentados, conforme descrito na Seção 5.

Os modelos foram submetidos ao mesmo regime de otimização de pesos. Utilizou-se um escalonador de taxa de aprendizado (*learning-rate scheduler*) baseado em platôs (*plateau*), reduzindo a taxa de atualização quando a precisão de validação estagnava. Para lidar com os desafios intrínsecos do baixo contraste térmico e a alta variação na contagem de pixels entre as classes (o background domina a imagem em contraste com regiões minúsculas como os lábios), a otimização minimizou uma função de perda combinada composta por *Cross-Entropy* e *Dice Loss*. Os pesos dessa função, bem como os parâmetros de suavização da perda geométrica (*smoothing*), puderam ser ajustados através do arquivo *config.yaml* parametrizado pelo framework.

O pipeline experimental foi estruturado com um protocolo rigoroso de separação de dados para mitigar vieses de particionamento. Adotou-se a estratégia de validação cruzada (*k-fold cross-validation*, com $k = 5$). As sementes globais (*seeds*) dos geradores pseudoaleatórios das bibliotecas computacionais e do hardware acelerador foram fixadas em valores constantes.

Os testes foram executados em um computador de bancada, composto por um processador AMD Ryzen 7 3700x, com 16 GB de memória RAM e uma GPU com 6 GB de memória (série GTX 1660 Ti). Monitorou-se a eficiência computacional por meio do Tempo de Inferência (ms), Quadros por Segundo (FPS), número total de parâmetros, tamanho do modelo (MB) e o Pico de Memória em GPU (VRAM) alocada (MB). Adicionalmente, algumas métricas de acurácia espacial foram coletadas:

- A Interseção sobre a União Média (*Mean Intersection over Union* - mIoU) é uma métrica usada para avaliar o desempenho de modelos de segmentação semântica e detecção de objetos, que calcula a intersecção entre a área prevista e a área real, dividida pela área da união entre elas. A métrica mIoU é adotada porque fornece uma pontuação única para comparar a qualidade de diferentes modelos.
- O Coeficiente de Similaridade de Dice ou pontuação de Dice (*Dice Score*) é uma abordagem estatística usada para avaliar a similaridade entre duas amostras, variando de 0 a 1 (sobreposição perfeita).
- Perda de Treinamento: é a média do erro do modelo calculado sobre o conjunto de dados de treinamento durante cada época. Menos perda é melhor.
- Perda de Validação: é a média do erro do modelo calculada sobre um conjunto de dados separado e nunca visto pelo modelo durante o treinamento. Quanto menor a perda, melhor o resultado. A Perda de Validação deve se aproximar da Perda de Treinamento, indicando uma acurácia do modelo ao interagir com os dados.

Neste trabalho, foram utilizadas ferramentas de Inteligência Artificial Generativa, auxiliando os autores em demandas dentro do código e do texto do projeto. O ambiente de desenvolvimento integrado utilizado foi o *Antigravity* da Google, onde foram usados



Figura 2. Modelo U-Net

os modelos Claude Sonnet 4.6 em modo “Thinking” e Gemini 3.1 Pro na versão (*High*), ambos com a funcionalidade “Plan” ativa para auxílio na estruturação e documentação dos códigos, bem como na análise da consistência do projeto durante diferentes melhorias e correções de erros. A Figura 1 foi gerada com auxílio de inteligência artificial utilizando o modelo Gemini 3.1 Pro, sendo empregada exclusivamente para fins ilustrativos.

5. Resultados Experimentais

A Figura 2, Figura 3 e Figura 4 mostram as perdas de treinamento e validação, a validação mIoU e Dice Score, e o tempo de inferência por imagem. A observação das curvas de aprendizado ao longo das 100 épocas de treinamento endossa as discrepâncias de estabilidade de cada topologia. Embora todas as arquiteturas apresentem uma redução contínua da função de perda global, a U-Net (Figura 2) possibilitou observar micro-flutuações e instabilidades nas curvas de mIoU e Dice Score de validação durante os terços médio e final do treinamento. Os modelos baseados em atenção TransUNet (Figura 3) e Swin-UNet++ (Figura 4) exibiram trajetórias de convergência mais limpas, robustas e estáveis em suas métricas de validação. Isso demonstra que a integração do contexto semântico de longa distância (*long-range dependencies*) confere ao modelo uma capacidade de generalização superior quando testado em dados invisíveis do mesmo domínio térmico, mitigando a sensibilidade excessiva a ruídos radiométricos experimentada pela convolução pura.

A Figura 5 possibilita uma comparação da acurácia dos modelos, considerando os valores médios de IoU (mIoU) e Dice. A TransUNet alcançou o melhor desempenho absoluto em fidelidade geométrica, com um mIoU de 0,6145 e Dice de 0,5637. A Swin-UNet++ estabeleceu-se como um meio-termo robusto (mIoU 0,6022), enquanto a U-Net tradicional apresentou as métricas globais mais baixas (mIoU 0,5871).

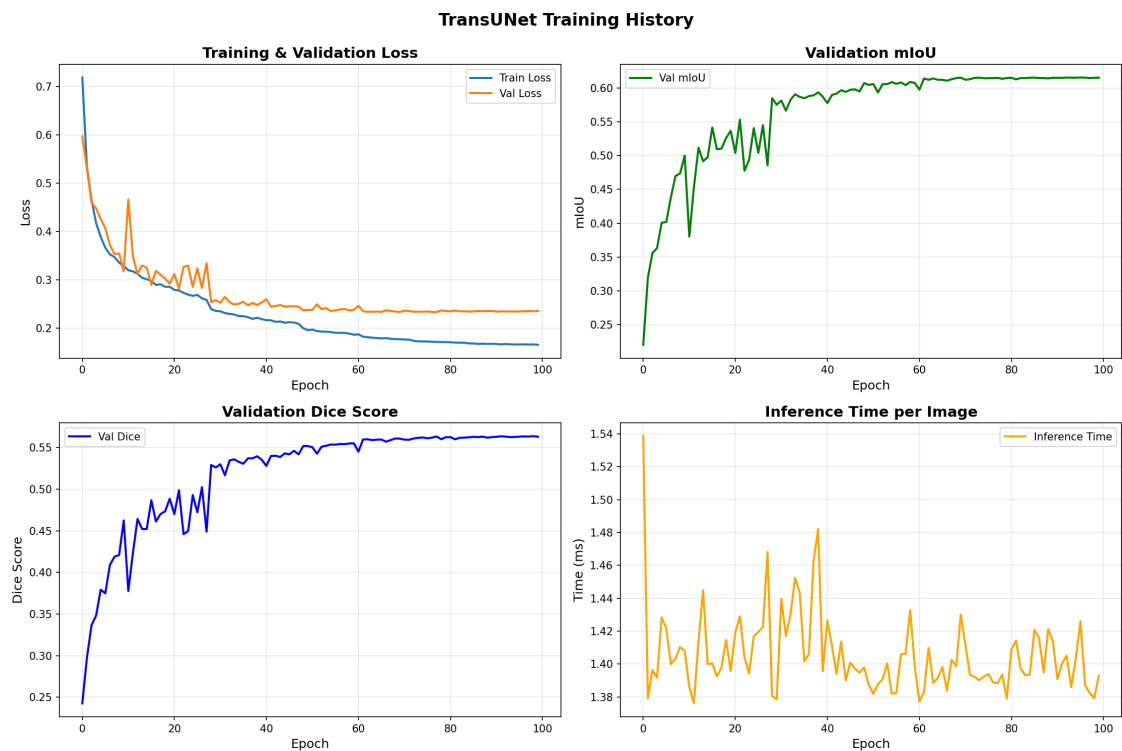


Figura 3. Modelo TransUNet



Figura 4. Modelo Swin-Unet++

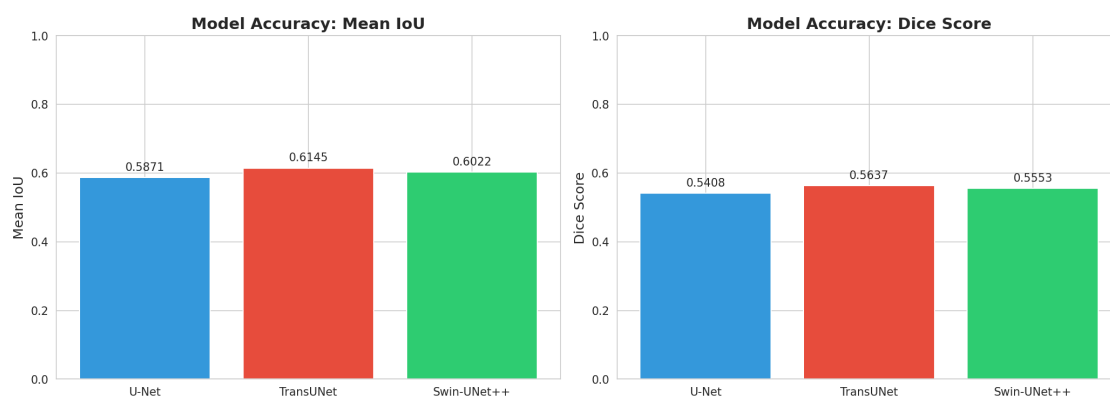


Figura 5. Acurácia dos Modelos

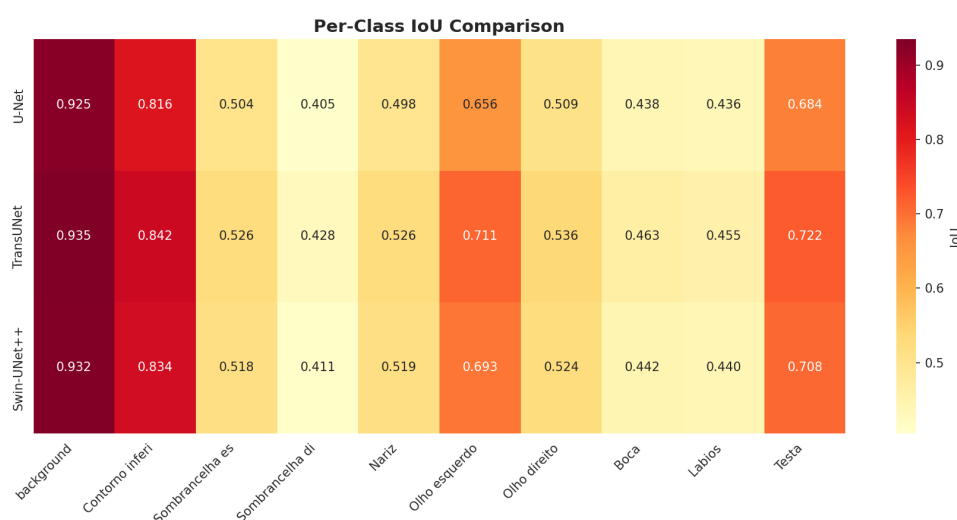


Figura 6. Acurácia por classe dos Modelos

Uma análise mais detalhada revela padrões importantes no mapa de calor (*heatmap*) de mIoU por classe, conforme a acurácia apresentada na Figura 6. Todas as abordagens desempenharam excepcionalmente bem em classes de área ampla e emissividade estável, como o “Background” (IoU > 0,92) e o “Contorno inferior do rosto” (IoU > 0,81). Em contrapartida, regiões faciais suscetíveis a variações termodinâmicas agudas e resfriamento evaporativo induzido pela respiração, notadamente a “Boca” e os “Lábios”, impuseram os maiores desafios. Nessas regiões, a precisão decaiu drasticamente para médias entre 0,43 e 0,46 em todas as redes. Contudo, a TransUNet manteve uma leve vantagem nessas classes difusas, evidenciando que seus mecanismos de atenção global conseguem inter-relacionar estruturas fixas da face para inferir bordas térmicas altamente voláteis que carecem de gradientes nítidos.

A análise por classe (Per-Class IoU) revela que a detecção do contorno inferior do rosto e do fundo anatômico é altamente precisa em todos os modelos. No entanto, classes menores apresentam maior dificuldade. A classe “Sombrancelha direita”, por exemplo, registrou um IoU de apenas 0,4048 na U-Net e 0,4285 na TransUNet. Por outro lado, o “Olho esquerdo” obteve um IoU expressivo de 0,7114 utilizando o modelo TransUNet.

A viabilidade operacional em tempo real requer a otimização conjunta da latência

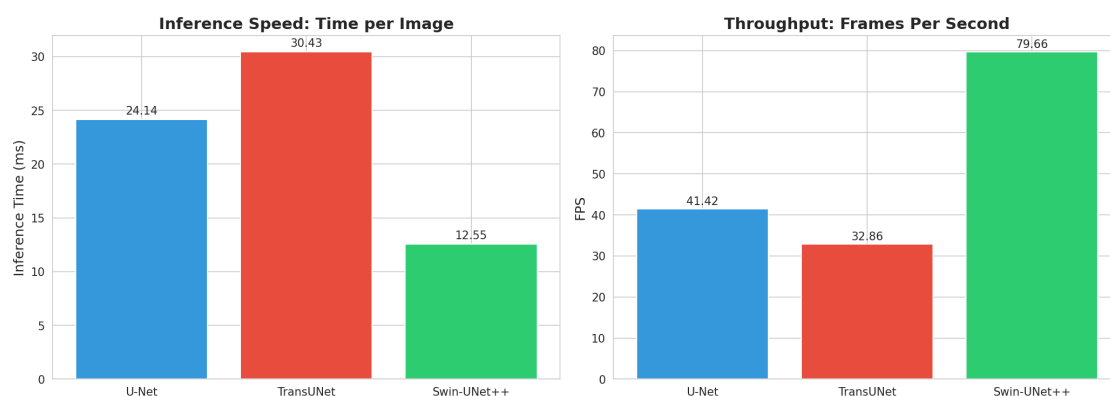


Figura 7. Comparação da velocidade de inferência e vazão dos Modelos

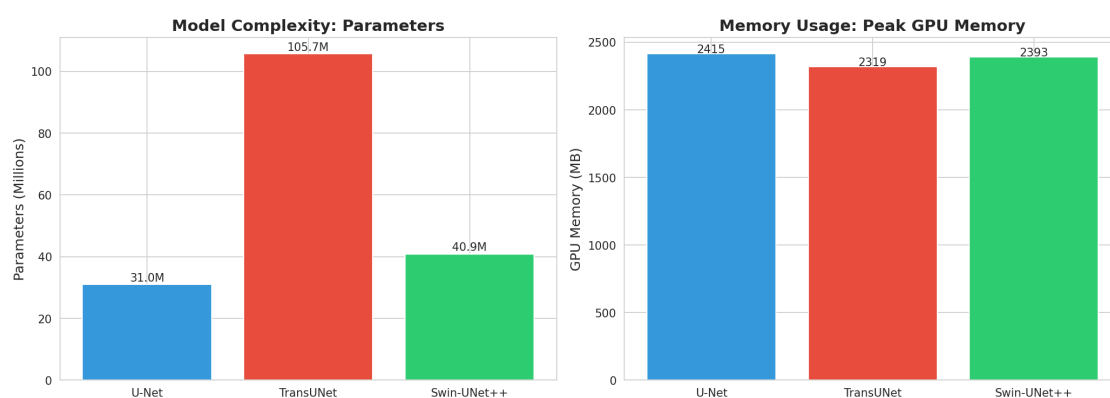


Figura 8. Comparação da Complexidade dos Modelos

temporal e da alocação de hardware, conforme evidenciada na Figura 7. Nesse escopo, a Swin-UNet++ demonstrou maior vazão, requerendo apenas 12,55 ms por inferência de imagem, o que se traduz em uma taxa de transferência de 79,66 quadros por segundo (FPS). Esta vazão é mais do que o dobro visto na TransUNet (32,86 FPS a 30,43 ms por inferência), tornando a Swin-UNet++ ideal para implantações em computação de borda.

A análise de complexidade estrutural, Figura 8, revelou uma relação contraintuitiva entre a contagem de parâmetros e o teto de alocação de memória volátil (VRAM). TransUNet possui um tamanho de 105,71 Milhões de parâmetros. A U-Net, em contraste, abriga apenas 31,04 Milhões de parâmetros e é fisicamente o menor modelo em disco. Apesar de ser a menor rede, a U-Net consumiu o pico máximo de VRAM durante a inferência ativa (2414,82 MB). Por outro lado, a TransUNet operou com menor consumo de memória (2319,11 MB). A justificativa reside na dimensionalidade espacial: enquanto a U-Net é forçada a carregar e propagar mapas de características convolucionais tridimensionais de altíssima resolução ao longo de todas as suas camadas, a TransUNet aplica um achatamento vetorial (*flattening*) precoce das *features* em blocos estritos de 16×16 *patches*, reduzindo a dimensão latente transitória necessária para armazenar as matrizes de gradiente ativo na placa de vídeo.

Na Figura 9 é possível comparar a segmentação realizada entre os três modelos resultantes após a execução do *pipeline*, aplicados em três imagens em distintas posições e distâncias, onde se pode identificar a diferença entre a U-Net pura e a capacidade de



Figura 9. Comparação de detecção de regiões entre os três modelos

Tabela 1. Resumo da Comparação dos Modelos

Modelo	mIoU	Dice	Inf. (ms)	FPS	Params (M)	Size (MB)	GPU (MB)
U-Net	0.587	0.540	24.14	41.42	31.04	118.42	2414.82
TransUNet	0.614	0.564	30.43	32.86	105.71	403.25	2319.11
Swin-UNet++	0.602	0.555	12.55	79.66	40.92	156.10	2393.13

contexto global dos modelos baseados em *Transformers*. Enquanto TransUNet e Swin-UNet++ superam em poses com distorções geométricas (Linhas 1 e 3) devido aos mecanismos de autoatenção, a U-Net demonstra maior robustez em maior resolução (Linha 2). Sob tais mudanças bruscas de perspectiva, os componentes de atenção exibem sensibilidade ao alinhamento de *patches*, resultando em desfoque nas bordas das classes e em queda nas métricas de mIoU e Dice. A Tabela 1 resume os resultados médios obtidos pelos três modelos, aplicando a abordagem de avaliação proposta.

6. Conclusão e Trabalhos Futuros

Este artigo avaliou e comparou o desempenho dos algoritmos de segmentação de imagens U-Net, TransUNet e SwinUnet++ através da criação de um pipeline automatizado de experimentos. A execução ocorreu sobre um conjunto de imagens térmicas da região facial. O objetivo deste trabalho foi caracterizar os aspectos de cada um dos algoritmos, buscando evidenciar qual possui melhores resultados nos aspectos considerados.

Na avaliação do desempenho, são indicados alguns *trade-offs* empíricos claros entre as abordagens de redes neurais profundas para o processamento de radiometria termográfica. **TransUNet**: estabelece o teto de precisão morfológica, superando falhas em regiões difusas, como os lábios, graças à auto-atenção global. O algoritmo pode ser utilizado para análise clínica minuciosa, sem restrições severas de latência de quadros. **Swin-UNet++**: entrega acurácia competitiva com um consumo de memória equilibrado e

velocidades de inferência excelentes (cerca de 80 FPS), sendo estritamente recomendada para biometria ininterrupta e monitoramento em dispositivos de borda. É a abordagem ótima para implementações práticas, de acordo com os resultados encontrados. **U-Net**: opera como uma forte base metodológica, necessitando um mínimo de espaço em disco estático (apenas 118 MB), mas exigindo mais memória RAM gráfica, além de sofrer com a perda de contexto global em classes de baixa resolução térmica.

A perspectiva com os resultados obtidos é promover melhorias nas medições de temperatura corporal, que até o momento apresentam: i) medições com sensores invasivos desconfortáveis; ii) sensores não invasivos não são utilizados em larga escalas; iii) medições atuais tem potencial de alto risco de contaminações; iv) medições são feitas em intervalos de tempo sem identificação do perfil térmico dos pacientes.

Para expandir e validar ainda mais este pipeline, avaliações futuras demandam de mais testes em outras bases de dados, como as que apresentam exalação respiratória em tempo real e avaliações frente a diferentes espectros de emissividade da pele.

Referências

- Ashrafi, R., Azarbayjani, M., and Tabkhi, H. (2022). Charlotte-ThermalFace: A Fully Annotated Thermal Infrared Face Dataset with Various Environmental Conditions and Distances. *Infrared Physics & Technology*, 124:104209.
- Chen, J. et al (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *Johns Hopkins University*.
- Jerbi, O. and Lee, S. (2024). Thermal Imaging Body Temperature Estimation Independent of Environmental Conditions and Individual Variations. In *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*.
- Liu, P. et al. (2021). Swin-UNet++: A Nested Swin Transformer Architecture for Location Identification and Morphology Segmentation of Dimples on 2.25Cr1Mo0.25V Fractured Surface. *Materials (Basel)*, v. 14, n. 24, p. 7504, 7 dez. 2021.
- Nowakowski, A. Z. and Kaczmarek, M. (2025). Artificial Intelligence in IR Thermal Imaging and Sensing for Medical Applications. *Sensors*, 25(3).
- Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Robledo-Vega, I. et al. (2025). Embedded Vision System for Thermal Face Detection Using Deep Learning. *Sensors*, 25(10).
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.
- Sheng, D. et al. (2025). Region-Based Self-Segmentation Guided Diffusion Model for Thermal Infrared to Pseudo-Color Visible Light Image Conversion. *IEEE Access*, v. 13, p. 47860–47873, 2025.
- Stanić, V. and Geršak, G. (2025). Facial thermal imaging: A systematic review with guidelines and measurement uncertainty estimation. *Measurement*, v. 242, p. 115879.