Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de pós-graduação em Ciência da Computação– PPGCC

Matheus Ramos de Carvalho

Segmentação Semântica Fracamente Supervisionada para Imagens de *fast-food* usando Arquitetura SegFormer com duplo estudante

São Carlos
Março de 2026

Matheus Ramos de Carvalho

Segmentação Semântica Fracamente Supervisionada para Imagens de *fast-food* usando Arquitetura SegFormer com duplo estudante

Dissertação apresentada ao Programa de pós-graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Visão Computacional

Orientador: Prof. Dr. Ricardo José Ferrari

São Carlos
Março de 2026

Folha de Aprovação

Defesa de dissertação de mestrado do(a) candidato(a) Matheus Ramos de Carvalho, realizada em 02/03/2026

Comissão Julgadora

Prof(a) Dr(a) Ricardo José Ferrari (UFSCar)

Prof(a) Dr(a) Alan Demétrius Baria Valejo (UFSCar)

Prof(a) Dr(a) Bruno Augusto Nassif Travençolo (UFU)

O relatório de defesa assinado pelos membros da comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação

Agradecimentos

Em primeiro lugar, agradeço a Deus, pela força e sabedoria concedidas ao longo dessa jornada, e por me guiar em cada passo. Aos meus pais, Valdécio e Donizete, por seu apoio incondicional, amor e incentivo constantes. Aos meus amigos da Visio.ai, André e Leonardo que me acompanharam de perto durante a pesquisa. Ao meu amigo da pós-graduação, Paulo, pela camaradagem e troca de ideias. Ao meu orientador, Prof. Dr. Ricardo José Ferrari, pela orientação cuidadosa, pelos ensinamentos preciosos, pela paciência, prezando sempre pela qualidade do trabalho e pela ciência.

Resumo

Este estudo foca na Segmentação Semântica (SS) de imagens de montagem de lanches *fast-food*, utilizando técnicas de *Deep Learning* e aprendizado fracamente supervisionado. A Segmentação Semântica, que envolve a classificação pixel a pixel da imagem, é uma tarefa fundamental na visão computacional e relevante para o controle de qualidade da montagem em franquias. Ao lidar com alimentos, surgem desafios como a grande diversidade visual dos ingredientes e o desequilíbrio entre classes; além disso, os modelos existentes frequentemente se baseiam em *datasets* públicos que não refletem cenários como imagens de câmeras de Circuito Fechado de Televisão (CFTV), foco deste trabalho. A automatização do controle de qualidade, distinguindo ingredientes a nível de pixel, contribui para a padronização e a consistência. A anotação manual para treinamento supervisionado é custosa e demorada, o que justifica a exploração de modelos fracamente supervisionados, que requerem apenas rótulos de classificação de imagem. Propõe-se um modelo de SS fracamente supervisionado, baseado em *Class Activation Maps* (CAMs), para identificar ingredientes em imagens de baixa resolução capturadas por CFTV. A arquitetura é leve e eficiente: utiliza o SegFormer-B0 em configuração de duplo estudante, com supervisão mútua por pseudo-rótulos cruzados, e incorpora Aumento no Tempo de Teste (TTA) e Ajuste Dinâmico de Limiar (DTA) para refinamento dos pseudo-rótulos. Os resultados mostram mIoU de 43,9% (comparável ao DuPL, 41,2%), com maior estabilidade entre *folds* (desvio padrão 0,6% versus 6,8% do DuPL). O modelo apresenta ainda eficiência computacional superior, utilizando 33% menos VRAM e *throughput* de 1,5 a 5 vezes maior que o DuPL, tornando-o adequado à implantação em tempo real em ambientes com recursos limitados.

Palavras-chave: Segmentação Semântica. Segmentação de ingredientes. Aprendizado Fracamente Supervisionado. Segmentação em alimentos. *Deep Learning*. *Fast-Food*.

Abstract

This study focuses on Semantic Segmentation (SS) of fast-food sandwich assembly images using Deep Learning and weakly supervised learning techniques. Semantic segmentation, which involves pixel-wise classification of the image, is a fundamental task in computer vision and relevant for quality control in franchise assembly. When dealing with food, challenges arise such as the high visual diversity of ingredients and class imbalance; moreover, existing models often rely on public datasets that do not reflect scenarios such as Closed-Circuit Television (CCTV) imagery, which is the focus of this work. Automating quality control by distinguishing ingredients at the pixel level contributes to standardization and consistency. Manual annotation for supervised training is costly and time-consuming, which justifies exploring weakly supervised models that require only image-level labels. We propose a weakly supervised SS model based on Class Activation Maps (CAMs) to identify ingredients in low-resolution images captured by CCTV. The architecture is lightweight and low-resource: it uses SegFormer-B0 in a dual-student setup with mutual supervision via cross pseudo-labels, and incorporates Test-Time Augmentation (TTA) and Dynamic Threshold Adjustment (DTA) for pseudo-label refinement. Results show mIoU of 43.9% (comparable to DuPL, 41.2%), with greater stability across folds (standard deviation 0.6% versus 6.8% for DuPL). The model also exhibits superior computational efficiency, using 33% less VRAM and achieving 1.5–5× higher throughput than DuPL, making it suitable for real-time deployment in resource-constrained environments.

Keywords: *Semantic Segmentation. Ingredient Segmentation. Weakly Supervised Learning. Food Segmentation. Deep Learning. Fast-Food.*

Lista de ilustrações

Lista de tabelas

Lista de equações

Lista de siglas

ASPP	Atrous Spatial Pooling Pyramid
CAMs	Class Activation Maps
CB	Class Balancing
CFTV	Circuito fechado de televisão
CL	Contrastive Learning
CPM	CAMs-based Prompting Module
DL	Deep Learning
DTA	Ajuste Dinâmico de Limiar
GAP	Global Average Pooling
GMAP	Global Max-Average Pooling
IoU	Interseção sobre União
MIL	Multiple Instances Learning
mIoU	mean Intersection over Union
MLP	Multilayer Perceptron
MLSM	Multi-Label Soft Margin
NCE	Noise Contrastive Estimation
NLP	Natural Language Processing
PCM	Pixel Correlation Module

ResNet	Redes Neurais Residuais
RNAs	Redes Neurais Artificiais
RNCs	Redes Neurais Convolucionais
SAM	Segment Anything Model
SEAM	Self-supervised Equivariant Attention Mechanism
SPP	Spatial Pyramid Pooling
SS	Segmentação Semântica
SSC	SAM-Segment Contrasting
SSDB	Semantic Segmentation Database Network
TTA	Aumento no Tempo de Teste
ViT	Vision Transformers
VSS	Visio Sandwich Segmentation
WSSS	Segmentação Semântica Fracamente Supervisionada

Sumário

Capítulo 1

Introdução

Neste capítulo é apresentado o contexto no qual a pesquisa está inserida, indicando o problema investigado e os objetivos alcançados.

1.1 Contexto e motivação

Segmentação Semântica (SS) é uma das principais tarefas de visão computacional, desempenhando um papel significativo na análise e compreensão de imagens. Recentemente, essa área tem obtido avanços notáveis devido aos métodos de *Deep Learning* (DL) e à ascensão do campo. A tarefa de SS visa classificar regiões ou *pixels* de imagens em classes de objetos ou não-objetos, servindo como base para diversas tarefas em campos de pesquisa (LATEEF; RUICHEK, 2019; SHARMA; ARTACHO; SAVAKIS, 2021).

As aplicações da SS incluem a detecção de ervas daninhas e pragas na agricultura (MILIOTO; LOTTES; STACHNISS, 2018; GONCALVES et al., 2021), o reconhecimento de ambiente para carros autônomos (FENG et al., 2020) e a contagem de calorias em alimentos (MEYERS et al., 2015). Na área de alimentos, além do cálculo de calorias, a SS também é utilizada para o reconhecimento de ingredientes e pratos, geração de receitas e cálculo de volume de ingredientes utilizados (SHARMA; ARTACHO; SAVAKIS, 2021).

No entanto, a SS aplicada à área de alimentos é considerada uma tarefa mais complexa em comparação com sua aplicação em objetos gerais, devido à grande diversidade na aparência dos alimentos e à falta de balanceamento das classes (WU et al., 2021b; LAN et al., 2023).

Alguns modelos foram desenvolvidos como resposta a esse desafio (LAN et al., 2023; SHARMA; ARTACHO; SAVAKIS, 2021). No entanto, devido à limitada variedade de *datasets* públicos disponíveis para a SS de alimentos, esses modelos não possuem dados específicos do processo de montagem de lanches. Como resultado, esses conjuntos de dados não são adequados para cenários como a montagem de lanches de franquias *fast-*

food, capturados por Câmeras de Circuito Fechado de Televisão (CFTV), que é o foco deste projeto.

Além da inadequação em termos de cenário de captura, os *datasets* públicos também apresentam limitações em relação à granularidade de classes. Enquanto *datasets* como FoodSeg-103 (WU et al., 2021b; SALVADOR et al., 2017) e UECFoodPix (MATSUDA; HOASHI; YANAI, 2012; OKAMOTO; YANAI, 2021) segmentam ingredientes em nível grosso, agrupando classes visualmente similares ou funcionalmente relacionadas, o contexto industrial de controle de qualidade requer distinção em nível de granularidade fina. Por exemplo, enquanto *datasets* públicos podem segmentar genericamente "queijo", o contexto industrial requer distinção entre "*cheddar*", "queijo" e "*cream cheese*", cada um com características visuais, texturais, monetárias e funcionais distintas que são relevantes para o controle de qualidade e custo. Da mesma forma, é necessário distinguir tipos específicos de carne (bacon, carnes processadas) uma vez que são classes distintas e com custos distintos e tem relevância na montagem de lanches devido a isso. O *dataset* VSS, proprietário e construído a partir de imagens capturadas em ambiente industrial de montagem de lanches, utilizado neste trabalho, detalhado no Capítulo 4, apresenta 17 classes específicas que refletem esta necessidade de granularidade fina, permitindo controle de qualidade mais preciso e adequado às necessidades operacionais de franquias *fast-food*. Adicionalmente, as imagens capturadas por câmeras CFTV apresentam características distintas dos *datasets* públicos, incluindo baixa resolução, condições de iluminação variáveis, oclusões frequentes e artefatos de compressão, tornando ainda mais desafiador o uso direto de modelos treinados em *datasets* acadêmicos.

Os *datasets* padrão para SS de alimentos são o FoodSeg-103, que utiliza imagens de alimentos disponíveis online em sites de receitas para anotação em nível de ingrediente (WU et al., 2021b; SALVADOR et al., 2017), e o UECFoodPix *Complete*, que se baseia em imagens coletadas do *dataset* UEC-Food100, apresentado em (MATSUDA; HOASHI; YANAI, 2012; OKAMOTO; YANAI, 2021) com imagens coletadas pelo aplicativo FoodCam. Além desses, existe o *dataset* de SS de comidas brasileiras *MyFood* (FREITAS; CORDEIRO; MACARIO, 2020). A Figura 1 ilustra exemplos de imagens de cada um desses *datasets* de alimentos. Diferente dos *datasets* públicos, o *dataset* das imagens de montagem de lanches *fast-food* denominado *Visio Sandwich Segmentation* (VSS), que é o *dataset* principal utilizado neste trabalho, apresenta desafios como a oclusão de ingredientes, tanto pela mão da pessoa que está preparando quanto por outros ingredientes. Adicionalmente, a resolução das imagens adquiridas varia conforme a distância da câmera e suas configurações, e as condições de iluminação são não controladas, havendo também diferentes perspectivas para os lanches. Embora exista um certo padrão na ordem e disposição dos ingredientes no Brasil, onde foram coletadas as imagens, onde queijo e proteína são colocados no início da montagem para serem aquecidos junto com o pão, a ordem dos demais ingredientes é aleatória, mantendo apenas o molho como potencial último ingre-

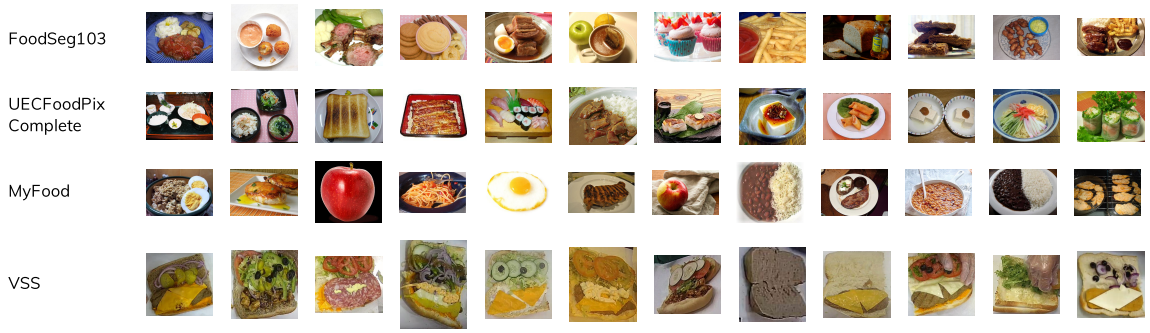

Figura 1 – Amostras de imagens de *datasets* de alimentos. Adaptado de (WU et al., 2021b; OKAMOTO; YANAI, 2021; FREITAS; CORDEIRO; MACARIO, 2020).

diente, ainda assim existem divergências nas aparências de ingredientes da mesma classe e coloração incomum de algumas classes, tornando o *dataset* desafiador até mesmo para a anotação humana.

O uso de SS para o controle de qualidade na montagem de lanches *fast-food* é motivado pela necessidade de automatizar a verificação da padronização da montagem dos alimentos, abordando problemas como excesso, escassez e distribuição inadequada dos insumos. Esse processo exige a distinção dos ingredientes em nível de pixel, permitindo a análise da distribuição e da quantidade dos ingredientes utilizados. A inspeção visual automatizada em ambientes de montagem de lanches *fast-food* é uma tarefa desafiadora e pouco explorada, sendo crítica para garantir a consistência do produto, a satisfação do cliente, a redução de desperdícios e a eficiência da produção (RAJU; IMTIAZ; SAZONOV, 2023; ZHU et al., 2021). Abordagens tradicionais de inspeção que dependem de características manuais ou heurísticas baseadas em cor frequentemente falham sob variabilidade visual complexa, incluindo mudanças de iluminação, deformação de ingredientes, oclusão e texturas heterogêneas (WU et al., 2021b).

Além disso, as franquias de *fast-food* enfrentam dificuldades ao treinar seus funcionários no correto porcionamento adequado dos ingredientes. A aplicação da SS pode auxiliar na amostragem de exemplos que seguem ou não o padrão, além de contribuir para o reconhecimento do tipo de lanche montado. Isso destaca a relevância de desenvolver soluções robustas e adaptadas para esse tipo de aplicação, especialmente considerando que ambientes industriais frequentemente fornecem imagens de baixa resolução, dados ruidosos e desbalanceamento de classes, fatores que dificultam a adoção direta de métodos gerais de SS fracamente supervisionada (CAI; ABHAYARATNE, 2023; RAJU; IMTIAZ; SAZONOV, 2023; ZHU et al., 2021).

Como um problema de domínio ainda não abordado, a aquisição de anotações para o treinamento dos modelos é uma tarefa necessária. No entanto, anotações de nível de pixel, exigidas para um treinamento supervisionado de SS, além de demandarem muito tempo e esforço humano, também acarretam em altos custos financeiros (HAO; ZHOU; GUO,

2020). Como um exemplo, cada imagem do *Cityscapes*, um *dataset* de SS de contexto urbano, demora por volta de 1 hora e meia para ser anotada (CORDTS et al., 2016). No contexto do *dataset* VSS utilizado neste trabalho, a anotação de nível de pixel demandou aproximadamente 35 minutos por imagem, enquanto anotações em nível de imagem levaram cerca de 40 segundos por imagem, representando uma redução de aproximadamente 85% no tempo de anotação manual. Esse tipo de problema já é conhecido, e existem soluções que necessitam de anotações em nível menor, como anotações de nível de imagem e *bounding boxes*, denominadas modelos fracamente supervisionados (HAO; ZHOU; GUO, 2020). A adoção de aprendizado fracamente supervisionado oferece uma via promissora para inspeção visual industrial, onde o custo de anotação e a escalabilidade são tão críticos quanto a precisão da segmentação (ZHOU et al., 2016; AHN; KWAK, 2018; LI et al., 2022).

É importante destacar que a viabilidade de implantação industrial (*industrial deployability*) não se resume apenas a métricas de desempenho isoladas como FPS (*frames per second*), mas envolve múltiplos aspectos críticos que determinam a adequação prática de um método para ambientes de produção. Estes aspectos incluem: (i) custo de anotação, que determina a escalabilidade e adaptação rápida a novos domínios; (ii) estabilidade entre *folds* de validação cruzada, que indica robustez a mudanças na distribuição de dados e previsibilidade operacional; (iii) consumo de memória (VRAM), que determina viabilidade de implantação em *hardware* acessível e capacidade de processamento paralelo; e (iv) viabilidade de treinamento contínuo (*continuous training*), essencial para adaptação em produção onde condições operacionais variam ao longo do tempo. O modelo proposto neste trabalho foi desenvolvido especificamente para navegar estes trade-offs, priorizando adequação para aplicações industriais sobre maximização isolada de métricas de desempenho.

O modelo proposto neste trabalho utiliza uma arquitetura de duplo estudante inspirada no DuPL (Dual Student with Trustworthy Progressive Learning) (WU et al., 2024), que explora o espaço de representações através da perda de discrepância, garantindo diversidade de características entre as duas redes estudante e reduzindo o viés de confirmação. Esta abordagem, que induz representações distintas entre as redes sem aprendizado contrastivo explícito, difere de métodos como o Multi-Label Supervised Contrastive Learning (MulSupCon) (ZHANG; WU, 2024) e outras abordagens contrastivas supervisionadas (JAISWAL et al., 2021; ZAIGRAJEW; ZIEBA, 2022), que requerem tamanhos de *batch* substancialmente grandes (tipicamente 64 ou mais) para funcionar efetivamente. Como discutido em detalhes no Capítulo 2, a limitação de *batch size* imposta por restrições de *hardware* torna métodos de aprendizado contrastivo explícito inviáveis, motivando a adoção de estratégias alternativas. A perda de discrepância, detalhada no Capítulo 4, promove diversidade de características minimizando a similaridade cosseno entre as representações das duas redes estudante, reduzindo o viés de confirmação e diversificando

as representações (*features*) entre as redes; isso permite que o modelo explore o espaço de representações de forma eficiente mesmo com restrições de memória.

1.2 Contribuições do Trabalho

As principais contribuições deste trabalho são:

1. ***Framework* WSSS para controle de qualidade em *fast-food*:** Introdução de um *framework* de segmentação semântica fracamente supervisionada adaptado para controle de qualidade de sanduíches *fast-food*, enfatizando eficiência de anotação e desempenho em tempo real.

2. ***Benchmark* de WSSS baseado em Transformers:** Fornecimento de um *benchmark* abrangente de WSSS baseado em Transformers em um *dataset* industrial de alimentos desafiador e de baixa resolução, demonstrando a adequação do *backbone* SegFormer e dos componentes propostos neste domínio.

3. **Análise de custo-desempenho:** Realização de uma análise abrangente de custo-desempenho que quantifica que o método proposto alcança 89% do desempenho totalmente supervisionado enquanto é mais de 6 vezes mais eficiente em termos de mIoU por hora de anotação, e opera a mais de 1.100 FPS.

1.3 Objetivos

A seguir são descritos os objetivos geral e específicos que nortearão esta pesquisa.

1.3.1 Objetivo Geral

Desenvolver um modelo de Segmentação Semântica de imagens de montagem de lanches *fast-food* usando *Deep Learning* e aprendizado fracamente supervisionado.

1.3.2 Objetivos Específicos

1. Avaliar quantitativamente os principais modelos supervisionados e fracamente supervisionados aplicados na segmentação de ingredientes em imagens de *fast-food*. Neste estudo avaliativo, foram observados e relatados os pontos fracos e fortes de cada abordagem visando o desenvolvimento de um novo modelo;

2. Organizar uma base de imagens para o desenvolvimento do modelo de Segmentação Semântica;

3. Investigar estratégias específicas de aumento de dados (*data augmentation*) para o problema em estudo;

4. Propor um modelo de *Deep Learning* para Segmentação Semântica de imagens de *fast-food*;

5. Entender a influência do tamanho e balanceamento dos conjuntos de imagens de treinamento no desempenho do modelo proposto;

6. Avaliar o desempenho de segmentação do modelo desenvolvido usando métricas quantitativas, como média de Intersection over Union (mIoU), precisão, revocação e métrica F1 (F1-Score).

1.4 Organização do trabalho

Este trabalho está dividido em sete capítulos, sendo eles:

- Capítulo 1: aborda as informações sobre as motivações e contexto para a pesquisa.
- Capítulo 2: apresenta a fundamentação teórica sobre as técnicas utilizadas neste trabalho.
- Capítulo 3: contém a revisão bibliográfica do estado da arte relacionado à SS de ingredientes em imagens de alimentos.
- Capítulo 4: apresenta os materiais e métodos utilizados, incluindo a descrição do *dataset*, a arquitetura do modelo proposto e as estratégias de treinamento.
- Capítulo 5: apresenta os resultados experimentais obtidos, incluindo comparações com métodos do estado da arte, estudos de ablação e análises quantitativas e qualitativas.
- Capítulo 6: discute os resultados obtidos, as limitações do método e as implicações para aplicações industriais.
- Capítulo 7: apresenta as conclusões do trabalho, destacando as contribuições principais e direções futuras.

1.5 Trabalho submetido para publicação

Parte dos resultados desta dissertação foi submetida à revista *IEEE Access* na forma de artigo científico intitulado “Weakly Supervised Semantic Segmentation for Fast-Food Quality Control: A Dual-Student Approach with SegFormer”. O trabalho descreve uma abordagem de Segmentação Semântica Fracamente Supervisionada (WSSS) para controle de qualidade na montagem de lanches *fast-food*, utilizando uma arquitetura dual-student

com backbone SegFormer, redução do custo de anotação pixel a pixel e avaliação em *dataset* industrial adquirido por câmeras de circuito fechado de televisão (CFTV).

As principais contribuições do artigo são: (i) um *framework* WSSS voltado ao controle de qualidade em montagem de sanduíches, com ênfase em eficiência de anotação e desempenho em tempo real; (ii) um *benchmark* de WSSS baseado em transformadores em *dataset* industrial de alimentos de baixa resolução; e (iii) uma análise custo–desempenho mostrando que o método atinge cerca de 89% do mIoU do modelo totalmente supervisionado, com redução de custo de anotação superior a 85% e eficiência de mais de 6× em mIoU por hora de anotação, além de inferência a mais de 1.100 FPS. No *dataset* VSS, o modelo proposto é comparável ao DuPL com 2,7% de mIoU de diferença no valor médio, com desvio padrão entre *folds* substancialmente menor, evidenciando maior robustez.

O artigo (CARVALHO; SACILOTTI; FERRARI, 2025) encontra-se **em análise após 1ª revisão** na *IEEE Access*.

Capítulo 2

Fundamentação teórica

Neste capítulo é apresentado o embasamento teórico das principais técnicas e métodos relatados nesta pesquisa e modelos de Aprendizado de Máquina utilizados.

Este capítulo apresenta os conceitos fundamentais, técnicas e ferramentas utilizadas neste trabalho, fornecendo a base teórica necessária para compreender os modelos e métodos aplicados. Diferentemente do capítulo seguinte, que apresenta trabalhos relacionados ao problema específico, este capítulo foca nos fundamentos teóricos gerais das técnicas de Aprendizado de Máquina, Redes Neurais, Transformers e aprendizado fracamente supervisionado que fundamentam a pesquisa.

2.1 Redes Neurais Convolucionais

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados no sistema nervoso biológico (YEGNANARAYANA, 2009), que têm se popularizado como uma abordagem eficaz no campo do Aprendizado de Máquina. Esta inspiração biológica se manifesta na estrutura das redes através de unidades de processamento (neurônios artificiais) interconectadas que processam informações de forma distribuída, similar ao funcionamento de sinapses neurais. O desenvolvimento do algoritmo *backpropagation* (RUMELHART; HINTON; WILLIAMS, 1986) foi fundamental para o avanço das RNAs, permitindo o treinamento eficiente de redes profundas através da propagação reversa do erro e ajuste dos pesos via gradiente descendente. Este mecanismo permite que a rede aprenda representações hierárquicas de dados através da atualização iterativa dos parâmetros, minimizando uma função de perda que mede a discrepância entre predições e valores esperados.

Diferentes tipos de RNAs foram desenvolvidas para diferentes tipos de dados. Os Perceptrons de Multicamada, do inglês *Multi-Layer Perceptron* (MLP) (RUMELHART; HINTON; WILLIAMS, 1986), permitiram modelar padrões não lineares através de múltiplas camadas interconectadas. As camadas ocultas dos MLPs permitem que a rede aprenda representações intermediárias dos dados, capturando relações complexas que não seriam possíveis com uma única camada de processamento. No entanto, para processar dados de natureza espacial, como imagens, foi necessário desenvolver arquiteturas especializadas. As Redes Neurais Convolucionais (RNCs) (LECUN et al., 1998) surgiram como uma evolução das RNAs MLP para lidar com dados de natureza espacial, aplicadas principalmente a problemas de Visão Computacional, como reconhecimento de imagens, detecção de objetos e SS. A principal inovação das RNCs reside na capacidade de capturar padrões locais através de operações de convolução, preservando informações espaciais que são perdidas quando imagens são achatadas em vetores unidimensionais, como ocorre nos MLPs tradicionais. A primeira aplicação bem-sucedida foi no trabalho LeNet (LECUN et al., 1998), que demonstrou a viabilidade prática das RNCs para reconhecimento de dígitos manuscritos. A partir de 2012, com o excelente resultado obtido no desafio da *ImageNet* (DENG et al., 2009) usando a arquitetura *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2017), esse tipo de arquitetura começou a se popularizar, marcando o início da era moderna de aprendizado profundo em visão computacional e estabelecendo as RNCs como o padrão dominante para tarefas de processamento de imagens.

As RNCs são baseadas em três princípios fundamentais (LECUN et al., 1989): campos receptivos locais para extrair padrões locais, pesos compartilhados para reduzir custo computacional, e *subsampling* (ou *pooling*) para reduzir dimensionalidade e aumentar invariância. A camada *Global Average Pooling* (GAP) é particularmente relevante neste trabalho, pois simplifica a correspondência entre mapas de características e categorias de classificação, facilitando a interpretação dos resultados como mapas de confiança (LIN, 2013).

2.1.1 Principais Arquiteturas

Entre as muitas arquiteturas propostas para RNCs, algumas famílias tornaram-se referência em visão computacional e aparecem de forma explícita ou implícita nos modelos e no pré-processamento adotados neste trabalho. A seguir é apresentada a ResNet, cuja ideia de conexões residuais influenciou redes profundas subsequentes e, aqui, serve de base para extração de representações e para *backbones* comparados no estado da arte.

2.1.1.1 ResNet

As Redes Neurais Residuais (ResNet) têm como princípio o uso de conexões de atalho (*shortcut connections*) que permitem que a rede neural se especialize em funções residuais,

facilitando a otimização e permitindo que redes mais profundas sejam treinadas de forma eficaz, mitigando problemas como desaparecimento de gradientes e degradação de desempenho (HE et al., 2016). Neste trabalho, uma ResNet50 pré-treinada foi utilizada para extrair *embeddings* de imagens na etapa de pré-processamento, permitindo a remoção de imagens duplicadas através da comparação de similaridade cosseno entre os *embeddings*. Além disso, a arquitetura ResNet-38 é utilizada como *backbone* no modelo SSDB-II (CAI; ABHAYARATNE, 2023), que é um dos modelos do estado da arte comparados neste trabalho.

2.2 *Vision Transformers*

Os modelos baseados em *Vision Transformers* (ViT) constituem uma abordagem arquitetural alternativa às redes neurais convolucionais para tarefas de visão computacional. Esses modelos foram propostos a partir da adaptação da arquitetura *Transformer* (VASWANI et al., 2017), originalmente desenvolvida para Processamento de Linguagem Natural, do inglês *Natural Language Processing* (NLP), para o domínio de processamento de imagens (DOSOVITSKIY et al., 2021). A arquitetura ViT demonstra desempenho competitivo quando comparada às redes convolucionais tradicionais, particularmente quando treinada com grandes volumes de dados (DOSOVITSKIY et al., 2021; ZHAI et al., 2022).

Diferente das RNCs que capturam padrões locais da imagem através de operações convolucionais com campos receptivos limitados, o ViT lida com as imagens de forma fundamentalmente distinta. Inicialmente, cada imagem é dividida em pequenos "*patches*" (por exemplo, blocos de 16×16 pixels) que são tratados como *tokens* sequenciais, análogos a palavras em problemas de NLP. Cada *patch* é linearizado em um vetor unidimensional e projetado em um espaço de *embeddings* de dimensão fixa através de uma camada linear de projeção, gerando os chamados *patch embeddings* (DOSOVITSKIY et al., 2021).

Para preservar a informação espacial que é perdida na linearização, vetores de *positional embeddings* são adicionados aos *patch embeddings*. Esses vetores codificam a posição espacial original de cada *patch* na imagem, permitindo que o modelo aprenda relações espaciais entre diferentes regiões (DOSOVITSKIY et al., 2021). A combinação dos *patch embeddings* com os *positional embeddings* resulta em uma sequência de *tokens* enriquecidos que mantém tanto informação semântica quanto espacial.

Os *tokens* são então processados por múltiplas camadas de mecanismo de atenção auto-referencial (*self-attention*), conforme representado na Figura 2. O mecanismo de atenção calcula relações de similaridade entre todos os pares de *tokens* na sequência, independentemente da distância espacial entre eles na imagem original (VASWANI et al., 2017). A atenção é computada através de três matrizes: *Query* (Q), *Key* (K) e *Value* (V), geradas a partir dos *embeddings* através de transformações lineares. A similaridade entre

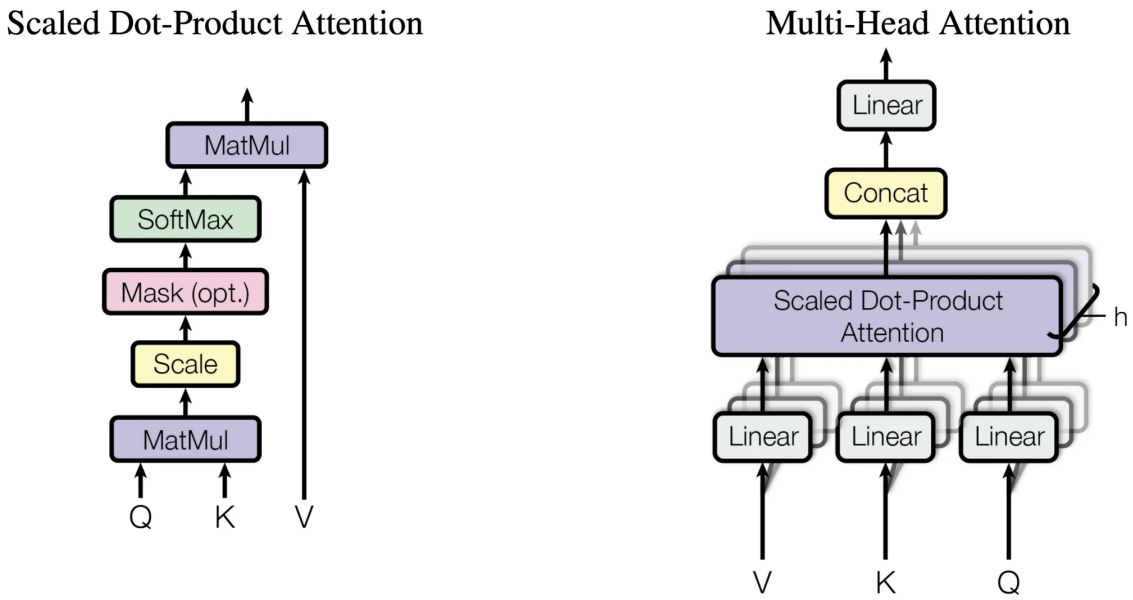

Figura 2 – Arquitetura de atenção e atenção multi-cabeças (VASWANI et al., 2017).

tokens é medida pelo produto escalar entre Q e K, normalizado por uma função *softmax*, e os valores resultantes são utilizados para ponderar as representações em V (VASWANI et al., 2017).

A arquitetura empregada tipicamente utiliza atenção multi-cabeças (*multi-head attention*), onde múltiplas instâncias do mecanismo de atenção operam em paralelo, cada uma aprendendo diferentes tipos de relações entre os *tokens* (VASWANI et al., 2017). Esta abordagem permite que o modelo capture simultaneamente diferentes aspectos das dependências espaciais e semânticas na imagem. A capacidade de modelar relações globais entre quaisquer regiões da imagem, independentemente da distância, confere aos ViTs uma vantagem sobre RNCs tradicionais na captura de contextos de longo alcance e dependências não-locais, características essenciais para tarefas que requerem compreensão holística da cena (DOSOVITSKIY et al., 2021).

No entanto, a principal limitação dos *Vision Transformers* era a necessidade de bases de dados volumosas para obter bons desempenhos. Para lidar com isso, foi proposto o uso de transferência de aprendizado, usando como base modelos com pré-treinamento em bases de dados maiores, como a *Laion-5B* (SCHUHMANN et al., 2022), e o JFT-3B (ZHAI et al., 2022), que possuem bilhões de dados rotulados de forma semiautomática e com presença de ruídos. Além disso, algumas arquiteturas propuseram formas de realizar treinamentos mais eficientes em termos de quantidade de dados, como DeIT (TOUVRON et al., 2021) e os *Transformers* híbridos (WU et al., 2021a).

A arquitetura do ViT, representada na Figura 3, segue uma estrutura sequencial composta por três componentes principais (DOSOVITSKIY et al., 2021). O primeiro componente consiste em uma camada de projeção linear (*linear projection*) que recebe os *patches* linearizados da imagem e os mapeia para um espaço de *embeddings* de dimensão

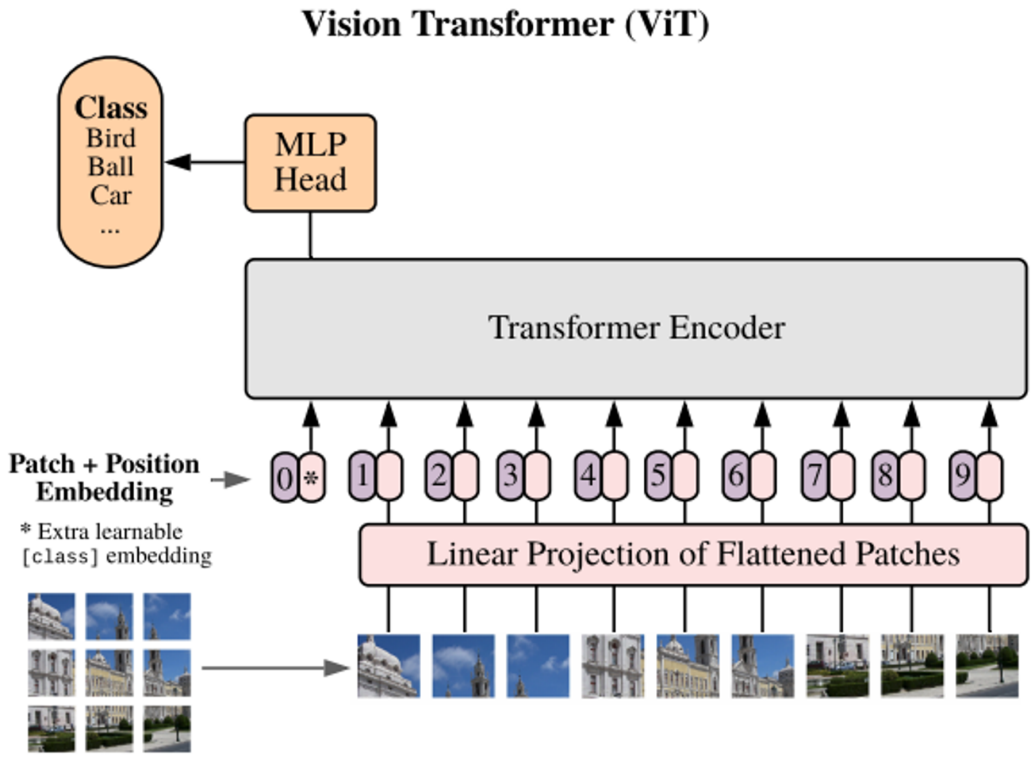

Figura 3 – Arquitetura de um Vision Transformer (DOSOVITSKIY et al., 2021).

fixa, gerando os *patch embeddings*. O segundo componente é um codificador *Transformer* composto por N blocos idênticos, cada um contendo camadas de atenção multi-cabeças e redes *feed-forward* multicamadas (MLP), intercaladas por normalização de camada (*Layer Normalization*) e conexões residuais. O terceiro componente é uma cabeça de classificação implementada como uma MLP que processa o *token* especial de classificação (*class token*) extraído da saída do codificador, gerando as probabilidades de classificação sobre as classes do problema.

2.2.1 SegFormer

O SegFormer (XIE et al., 2021) é uma arquitetura de segmentação semântica baseada em Transformers que combina um codificador hierárquico eficiente com um decodificador MLP leve. Diferentemente dos ViTs clássicos que empregam um codificador monolítico composto por blocos idênticos operando sobre uma única escala de resolução, o SegFormer utiliza um codificador hierárquico denominado Mix Transformer (MiT), que gera representações multi-escala através de múltiplos estágios progressivos. Esta distinção arquitetural é fundamental para tarefas de segmentação semântica, uma vez que a natureza monolítica dos ViTs convencionais limita sua capacidade de capturar informações em diferentes níveis de granularidade espacial, essenciais para a segmentação precisa de objetos

com dimensões variadas. A arquitetura hierárquica do SegFormer permite que o modelo processe simultaneamente características de baixo nível (detalhes finos e bordas) e alto nível (contexto semântico e relações espaciais), resultando em melhorias significativas tanto na precisão de segmentação quanto na eficiência computacional quando comparado a abordagens baseadas em ViTs monolíticos adaptados para segmentação.

O codificador MiT do SegFormer é composto por múltiplos estágios, cada um reduzindo progressivamente a resolução espacial enquanto aumenta a dimensão das características. Cada estágio contém blocos Mix Transformer que combinam operações de convolução e atenção, permitindo capturar tanto padrões locais quanto dependências globais. A estrutura hierárquica permite que o modelo processe informações em diferentes escalas, essencial para segmentação precisa de objetos de tamanhos variados.

O decodificador do SegFormer apresenta uma arquitetura minimalista, constituído exclusivamente por camadas MLP responsáveis pela agregação das características multi-escala provenientes do codificador hierárquico (XIE et al., 2021). Esta arquitetura simplificada diverge substancialmente de decodificadores empregados em outras arquiteturas de segmentação semântica, que frequentemente incorporam mecanismos mais complexos, tais como convoluções atrous (*atrous convolutions*) empregadas no DeepLab (CHEN et al., 2017; CHEN et al., 2018) ou módulos de atenção mais elaborados. O decodificador MLP do SegFormer realiza a fusão das características extraídas dos diferentes estágios do codificador mediante operações de concatenação seguida de projeção linear, culminando em uma camada de classificação por pixel que produz os mapas de segmentação semântica. Esta abordagem arquitetural minimalista permite manter a eficiência computacional do modelo, ao mesmo tempo em que preserva a capacidade de modelagem de contextos globais conferida pelo codificador hierárquico, demonstrando que a complexidade do decodificador pode ser reduzida quando o codificador é adequadamente projetado para capturar informações multi-escala (XIE et al., 2021).

O SegFormer-B0 constitui a variante mais compacta da família SegFormer, tendo sido especificamente concebido para atender aplicações que demandam eficiência computacional sem comprometer substancialmente a precisão de segmentação (XIE et al., 2021). Esta variante arquitetural emprega uma configuração reduzida de blocos Transformer e dimensões de características (*feature dimensions*) significativamente menores quando comparadas às variantes de maior capacidade (B1-B5), resultando em uma arquitetura adequada para implantação em ambientes com restrições computacionais, tais como sistemas de inspeção industrial que requerem processamento em tempo real (XIE et al., 2021).

A Tabela 1 apresenta uma análise comparativa das variantes SegFormer B0-B5, contemplando métricas de complexidade computacional (número de parâmetros e FLOPs) e desempenho em *datasets* de referência para segmentação semântica (XIE et al., 2021). A análise dos resultados evidencia um *trade-off* fundamental entre eficiência computacional

e precisão de segmentação: as variantes de maior capacidade (B3-B5) alcançam valores superiores de mIoU (*mean Intersection over Union*), porém demandam incrementos substanciais em termos de parâmetros treináveis e operações de ponto flutuante. Para o presente trabalho, o SegFormer-B0 foi selecionado em virtude de seu equilíbrio otimizado entre eficiência computacional e capacidade de modelagem, sendo particularmente adequado para aplicações industriais que impõem requisitos de processamento em tempo real sob restrições de recursos computacionais, ao mesmo tempo em que preserva capacidade suficiente para extração de características semânticas relevantes para a geração de mapas de ativação de classe (CAMs) de qualidade adequada.

Variante	Parâmetros (M)	GFLOPs	ADE20K mIoU (%)	Cityscapes mIoU (%)
SegFormer-B0	3,7	8,4	37,4	76,2
SegFormer-B1	13,2	15,9	42,2	78,5
SegFormer-B2	27,4	62,4	46,5	81,0
SegFormer-B3	47,3	79,0	49,4	81,7
SegFormer-B4	64,1	95,7	50,3	82,3
SegFormer-B5	84,7	121,1	51,0	82,4

Tabela 1 – Comparação das variantes SegFormer B0-B5 (XIE et al., 2021)

Uma característica arquitetural distintiva do SegFormer, que o diferencia substancialmente de outras arquiteturas baseadas em Vision Transformers, consiste na preservação explícita da estrutura espacial bidimensional ao longo das representações intermediárias geradas pelas camadas do codificador. Em contraste com ViTs convencionais, que submetem os *patches* da imagem a um processo de achatamento (*flattening*) que os transforma em uma sequência unidimensional de *tokens*, o SegFormer mantém a organização espacial bidimensional mediante sua arquitetura hierárquica multi-estágio, na qual cada estágio do codificador Mix Transformer (MiT) preserva a estrutura espacial das características através de operações que mantêm invariantes as dimensões espaciais $H \times W$ das representações (XIE et al., 2021). Esta característica arquitetural contribui de forma significativa para a geração de segmentações semânticas com maior precisão, uma vez que preserva informações de localização espacial que são essenciais para a tarefa de classificação por pixel, permitindo que o modelo mantenha correspondência espacial entre as características extraídas e as regiões originais da imagem de entrada.

2.3 Arquiteturas de Redes Neurais Convolucionais para SS

Uma das famílias de arquiteturas mais populares em SS, a DeepLab (CHEN et al., 2017), trouxe novos módulos que melhoraram a capacidade de segmentação dos modelos. Dentre esses módulos, podemos destacar o *Atrous Spatial Pooling Pyramid* (ASPP),

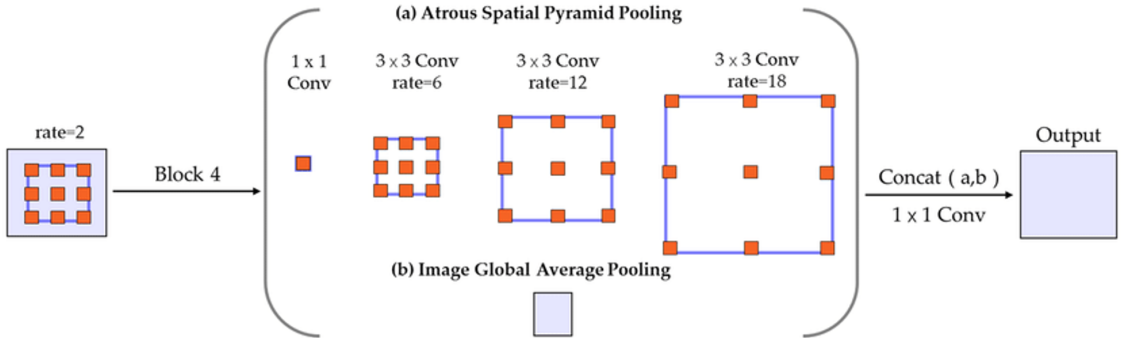

Figura 4 – *Atrous Spatial Pooling Pyramid* (CHEN et al., 2017).

responsável por aumentar o campo de visualização da convolução, incorporando maiores contextos espaciais conforme representado na Figura 4 e o *Spatial Pyramid Pooling* (SPP), que introduz um mecanismo que permite com que as RNCs não precisem de um tamanho fixo de imagem de entrada;

A evolução das versões do *DeepLab* também introduziu melhorias significativas. O *DeepLab* V2, por exemplo, aprimorou o ASPP, permitindo que a rede representasse objetos em múltiplas escalas, combinando mapas de pontuação ou características de diferentes versões redimensionadas da mesma imagem e melhorando a precisão da segmentação ao considerar a variabilidade de escala dos objetos (OUASSIT et al., 2022).

O *DeepLab* V3 adotou um modelo de codificador-decodificador que utilizava a Convolução *Atrous Separable*. Essa arquitetura, não apenas melhorou a captura de informações locais, mas também ajudou a definir melhor as bordas dos objetos, resultando em segmentações mais nítidas. A introdução de um módulo decodificador simples, como visto no *DeepLab* V3+, trouxe melhorias significativas nos resultados de segmentação (OUASSIT et al., 2022).

2.4 Aumento de dados

O aumento de dados (*data augmentation*) é uma técnica utilizada para melhorar o desempenho de modelos de *DL*, ao expandir conjuntos de dados limitados e, assim, ajudando na generalização dos modelos e reduzindo a chance de *overfitting*. Essa técnica consiste na aplicação de transformações e modificações nas imagens de treinamento, como rotações, espelhamentos, translações e alterações de cor, criando novas versões das imagens originais. Além das transformações básicas, também há transformações mais complexas, que podem incorrer da substituição de partes da imagem, distorção da imagem, entre outros (SHORTEN; KHOSHGOFTAAR, 2019).

2.5 Transferência de aprendizado

A transferência de aprendizado (*transfer learning*) é uma técnica muito utilizada no contexto de *DL*, pois dado a transferência de pesos entre modelos é possível utilizar o conhecimento adquirido por um modelo treinado previamente em grandes *datasets* para um contexto ou tarefa nova, mitigando a necessidade de um grande *dataset* para treinar o modelo no problema alvo (TORREY; SHAVLIK, 2010).

A transferência de aprendizado é especialmente valiosa em cenários onde os dados são escassos ou difíceis de obter. Ao utilizar um modelo que já aprendeu características relevantes de um domínio relacionado, é possível acelerar o processo de aprendizado e melhorar a performance do modelo em tarefas específicas. Essa abordagem é reconhecida como uma forma de tornar o Aprendizado de Máquina mais eficiente, semelhante ao que ocorre no aprendizado humano, em que as experiências anteriores são aplicadas a novas situações (TORREY; SHAVLIK, 2010).

Além da transferência de conhecimento entre modelos de arquiteturas semelhantes, existe o chamado *knowlege distillation*, ou destilação de conhecimento, que consiste no uso de um modelo robusto para treinar um modelo simples. O modelo utilizado para guiar o conhecimento é chamado de *teacher*, ou professor, enquanto o modelo menor a ser treinado é chamado de *student*, ou estudante. Essa transferência de conhecimento pode ocorrer de várias maneiras, incluindo a replicação das respostas do professor, a imitação das características intermediárias do modelo professor e a captura das relações entre as saídas (GOU et al., 2021).

2.6 SS fracamente supervisionada

Neste trabalho, foram utilizadas técnicas de aprendizado fracamente supervisionado e autotreinamento. Portanto, esta Seção apresenta os conceitos teóricos essenciais para a compreensão e aplicação dessas abordagens.

2.6.1 *Class Activation Maps*

RNCs são excelentes ferramentas para a solução de problemas em visão computacional; entretanto, é desafiador compreender seu funcionamento interno e identificar quais partes de uma imagem influenciam suas predições (JUNG; OH, 2021). Zhou *et al.* (ZHOU et al., 2016) propuseram um método denominado *Class Activation Maps* (CAMs), que visa fornecer uma explicação visual das regiões importantes de uma imagem de onde foram extraídas informações relevantes para a predição de classes, como ilustrado na Figura 5.

No entanto, a CAM originalmente proposta só podia ser extraída de camadas de GAP seguidas da função *softmax*, o que restringia seu uso a redes com essa estrutura. Posteriormente, foi desenvolvida a Grad-CAM, que utiliza dados de gradiente para obter CAMs

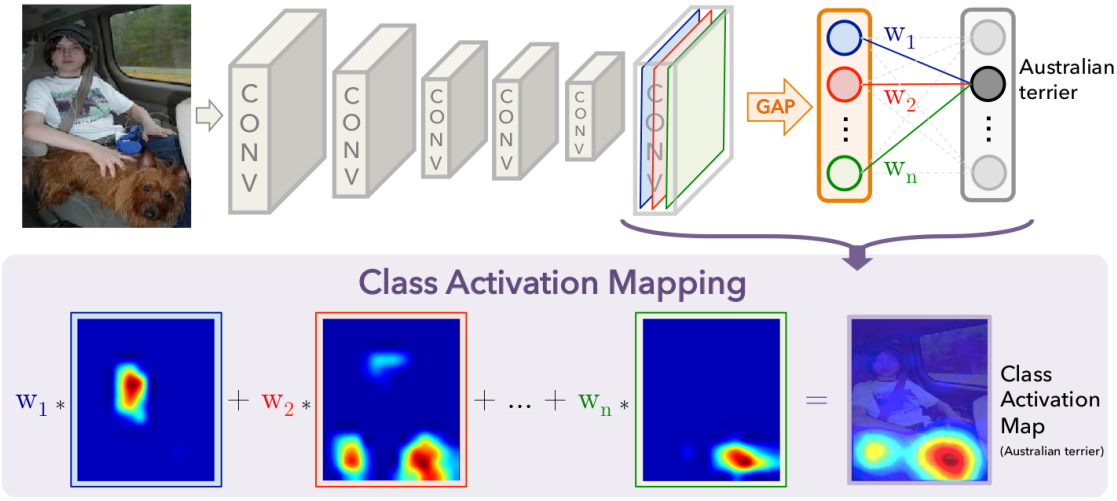

Figura 5 – Estrutura de aquisição padrão das CAMs (WANG et al., 2020).

em RNCs sem a necessidade de uma camada GAP seguida de uma *softmax* (SELVARAJU et al., 2020). Além de permitir a visualização de regiões importantes para a predição, as CAMs também demonstraram eficácia em tarefas de detecção fracamente supervisionada (SELVARAJU et al., 2020).

Para gerar as CAMs com a estratégia Grad-CAM, utilizam-se os gradientes calculados da classificação em relação às ativações da última camada convolucional, pois eles indicam a importância de cada neurônio dessa camada para a predição da classe em questão. Posteriormente, aplica-se uma camada de GAP aos gradientes obtidos, o que resulta em um vetor de pesos que representa a relevância de cada canal de ativação na camada convolucional final em relação à classe de interesse. Ao multiplicar as ativações de cada canal pelos pesos correspondentes, obtém-se um mapa de ativação, que é então processado por uma função ReLu para eliminar valores negativos, gerando as CAMs finais. Essas CAMs destacam as regiões da imagem que mais influenciaram a decisão do modelo (SELVARAJU et al., 2020), como ilustrado na Figura 6.

2.6.1.1 Grad-CAM para Arquiteturas Transformer

O método Grad-CAM original foi desenvolvido especificamente para redes neurais convolucionais, onde as ativações mantêm naturalmente a estrutura espacial bidimensional através de mapas de características (SELVARAJU et al., 2020). A aplicação de Grad-CAM em arquiteturas baseadas em Transformers apresenta desafios adicionais fundamentais, uma vez que estas arquiteturas processam imagens de forma fundamentalmente distinta: enquanto RNCs operam diretamente sobre representações espaciais bidimensionais, Transformers processam imagens como sequências de *tokens*, requerendo adaptações específicas para extrair CAMs espaciais significativas (ZHANG et al., 2022).

Em arquiteturas Transformer clássicas como o ViT (DOSOVITSKIY et al., 2021), os

patches da imagem são achatados em uma sequência unidimensional de *tokens*, perdendo a estrutura espacial bidimensional original durante o processo de *flattening*. Para aplicar Grad-CAM em tais arquiteturas, é necessário reconstruir a estrutura espacial a partir das ativações dos *tokens*, tipicamente através do rearranjo dos *tokens* em um grid bidimensional baseado em suas posições originais na imagem conforme codificadas pelos *positional embeddings*. O processo adaptado envolve calcular gradientes dos *logits* de classificação em relação às ativações dos *tokens*, agregar esses gradientes através de média global para obter pesos de importância, e então multiplicar as ativações pelos pesos correspondentes para gerar o mapa de ativação espacial após reconstrução bidimensional.

Trabalhos recentes têm explorado adaptações específicas de métodos de explicação para arquiteturas Transformer, reconhecendo as particularidades arquiteturais destes modelos. Zhang *et al.* (ZHANG et al., 2022) propuseram o TransCAM, que utiliza mecanismos de atenção do Transformer para gerar CAMs, explorando diretamente os pesos de atenção ao invés de depender exclusivamente de gradientes. Esta abordagem aproveita a natureza inerentemente interpretável dos mecanismos de atenção em Transformers, onde os pesos de atenção já codificam relações espaciais entre *tokens*, oferecendo uma alternativa aos métodos baseados em gradientes. A escolha entre métodos baseados em gradientes (como Grad-CAM) e métodos baseados em atenção (como TransCAM) para extração de CAMs em Transformers depende de fatores como a arquitetura específica utilizada e os objetivos da aplicação. Métodos baseados em gradientes são mais genéricos e aplicáveis a diferentes arquiteturas, enquanto métodos baseados em atenção exploram características específicas dos Transformers, mas podem ser menos transferíveis para outras arquiteturas (ZHANG et al., 2022).

O SegFormer apresenta uma vantagem significativa para extração de CAMs em relação a outros ViTs devido à sua arquitetura hierárquica que preserva estrutura espacial (XIE et al., 2021). Diferentemente do ViT clássico que achatam os *patches* em sequência, o SegFormer mantém representações espaciais bidimensionais em múltiplos estágios do codificador através de sua arquitetura Mix Transformer (MiT). Especificamente, o SegFormer utiliza camadas de normalização (LayerNorm) que preservam a estrutura espacial bidimensional, permitindo que as ativações sejam diretamente interpretadas como mapas espaciais sem necessidade de reconstrução complexa. Esta preservação estrutural constitui um aspecto fundamental para a aplicabilidade direta de métodos de explicação baseados em gradientes, tais como Grad-CAM, na medida em que elimina a necessidade de reconstrução espacial explícita dos *tokens*, processo que é requerido em ViTs de arquitetura sequencial (RHEUDE et al., 2024).

No contexto deste trabalho, a extração de CAMs do SegFormer-B0 é realizada através do hook na camada de normalização do quarto estágio do codificador (`layer_norm[3]`). Esta camada de LayerNorm opera sobre características que já foram processadas pelos blocos Transformer do estágio, mantendo a estrutura espacial bidimensional das ativações.

A escolha desta camada específica é motivada por três fatores principais: (1) ela captura características de alto nível semântico após processamento por múltiplos blocos de atenção, (2) mantém resolução espacial adequada (tipicamente $H/32 \times W/32$ para o quarto estágio) para geração de mapas de ativação detalhados, e (3) as ativações após normalização apresentam distribuição estável que facilita o cálculo de gradientes significativos. O processo de extração segue os seguintes passos:

1. As ativações da camada de normalização são capturadas através de um *forward hook*, preservando a estrutura espacial bidimensional das características.

2. Os gradientes dos *logits* de classificação são calculados em relação às ativações capturadas, indicando a sensibilidade de cada posição espacial para a predição de cada classe.

3. Os gradientes são agregados através de média ao longo da dimensão espacial, resultando em um vetor de pesos que representa a importância de cada dimensão de características para a classe de interesse.

4. As ativações são multiplicadas pelos pesos correspondentes, gerando um mapa de ativação espacial que destaca as regiões da imagem mais relevantes para a predição.

5. O mapa resultante é redimensionado para a resolução original da imagem através de interpolação bilinear, permitindo visualização e utilização em tarefas de segmentação.

Esta abordagem é viável no SegFormer devido à preservação de estrutura espacial nas camadas intermediárias, diferentemente de ViTs clássicos que requerem reconstrução espacial explícita. O mesmo método pode ser aplicado a outros ViTs hierárquicos que mantêm estrutura espacial em suas representações intermediárias, mas não é diretamente aplicável a ViTs puramente sequenciais sem adaptações adicionais para reconstrução espacial dos *tokens*.

2.6.2 *Contrastive Learning*

O aprendizado contrastivo, em inglês *Contrastive Learning* (CL), baseia-se na manipulação da representação de imagens no espaço de características. Esse método busca aproximar características semelhantes e repelir características distintas nesse espaço, permitindo assim o aprendizado de representações úteis a partir de dados não rotulados (JAISWAL et al., 2021).

A abordagem central do CL consiste em criar pares de amostras, onde uma amostra original (a "âncora") é comparada com uma versão aumentada ou transformada dela mesma (amostra positiva) e com outras amostras do *dataset* (amostras negativas). O

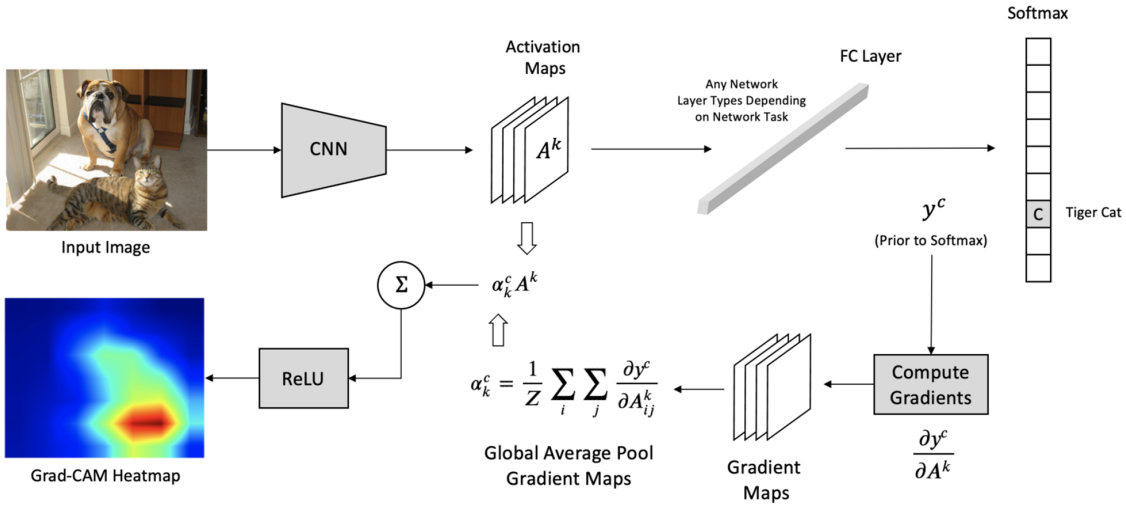

Figura 6 – Estrutura de aquisição de CAMs pelo Grad-CAM utilizando informação de gradiente (K., 2023).

objetivo é aumentar a similaridade entre a âncora e a amostra positiva, ao mesmo tempo em que se reduz a similaridade entre a âncora e as amostras negativas. Essa estratégia permite que o modelo aprenda a distinguir entre diferentes classes ou categorias, mesmo sem a necessidade de rótulos explícitos (JAISWAL et al., 2021).

A fórmula da perda contrastiva pode ser expressa de várias formas, dependendo do método específico adotado. Uma representação comum é a *Noise Contrastive Estimation* (NCE), descrita pela Equação 1 (JAISWAL et al., 2021):

$$L_{NCE} = -\log\left(\frac{e^{\mathrm{sim}(q,k_+)/\tau}}{e^{\mathrm{sim}(q,k_+)/\tau} + e^{\mathrm{sim}(q,k_-)/\tau}}\right), \tag{1}$$

sendo que q a amostra original (âncora), k_+e k_-as amostras positiva (versão aumentada da âncora) e negativa, "sim" é a função de similaridade (comumente similaridade de cosseno) e τ é o coeficiente de temperatura, que controla a penalidade das amostras negativas.

Um dos métodos mais influentes de aprendizado contrastivo auto-supervisionado é o SimCLR (CHEN et al., 2020), que simplifica abordagens contrastivas anteriores sem exigir arquiteturas especializadas ou bancos de memória. O SimCLR demonstrou que a composição de aumentos de dados, transformações não lineares aprendíveis e *batch sizes* maiores são componentes críticos para o aprendizado efetivo de representações contrastivas. Este método estabeleceu novos marcos em aprendizado auto-supervisionado e semi-supervisionado, alcançando desempenho competitivo com modelos supervisionados em tarefas de classificação de imagens.

No entanto, tal perda não aproveita as informações de rótulo, resultando no desenvolvimento de perdas específicas para múltiplos rótulos que consideram a sobreposição entre os conjuntos de rótulos das amostras. Um exemplo é a Equação 2(ZAIGRAJEW; ZIEBA,

2022):

$$\mathcal{L}_{ml}^{sup} = \sum_{i \in I} \mathcal{L}_{ml,i}^{sup} = \sum_{i \in I} \frac{-1}{|N(i)|} \sum_{p \in N(i)} s_{i,p} \cdot \log \left(\frac{e^{(z_i \cdot z_p)/\tau}}{\sum_{a \in A(i)} e^{(z_i \cdot z_a)/\tau}} \right), \quad (2)$$

sendo $N(i)$ o conjunto de amostras positivas relativas à imagem atual i e $|N(i)|$ o número de termos de $N(i)$. Os valores de z_i representam as projeções lineares das características da imagem na iteração atual da somatória; z_a representa as projeções lineares das características da amostra positiva de i, denominada a; p é a amostra a ser comparada na iteração, e z_p é à projeção linear da amostra. O termo $s_{i,p}$ indica o peso da perda baseada na similaridade entre os rótulos de i e p. Por fim, τ é o coeficiente de temperatura, que controla a penalização das amostras negativas (ZAIGRAJEW; ZIEBA, 2022).

As amostras são classificadas como positivas ou negativas com base na porcentagem de classes compartilhadas e um limiar estabelecido. Amostras com uma porcentagem de classes compartilhadas acima do limiar são consideradas positivas, enquanto que abaixo do limiar, são tratadas como negativas (ZAIGRAJEW; ZIEBA, 2022). No entanto, esta abordagem baseada em limiar fixo apresenta limitações quando aplicada a *datasets* com alta variabilidade na sobreposição de rótulos, motivando o desenvolvimento de métodos mais sofisticados que ajustem dinamicamente os pesos baseados no grau de similaridade entre amostras.

2.6.2.1 Multi-Label Supervised Contrastive Learning

O *Multi-Label Supervised Contrastive Learning* (MulSupCon) (ZHANG; WU, 2024) é uma extensão do aprendizado contrastivo supervisionado especificamente projetada para problemas de classificação multi-rótulo. Diferentemente de abordagens anteriores que extraem representações em nível de rótulo ou mapeiam rótulos para um espaço de *embedding*, o MulSupCon aborda explicitamente a ambiguidade na determinação de amostras positivas quando há diferentes graus de sobreposição de rótulos entre amostras, conforme identificado pelos autores (ZHANG; WU, 2024).

A principal inovação do MulSupCon reside em sua função de perda contrastiva que ajusta pesos dinamicamente baseados na quantidade de sobreposição que uma amostra compartilha com a âncora. Esta abordagem reconhece que em problemas multi-rótulo, a relação entre amostras não é binária (positiva ou negativa), mas sim um espectro contínuo baseado na similaridade dos conjuntos de rótulos. Amostras com maior sobreposição de rótulos recebem pesos maiores na função de perda, permitindo que o modelo aprenda representações que capturam melhor as correlações entre múltiplos rótulos. Zhang e Wu (ZHANG; WU, 2024) demonstraram através de análise de gradientes que esta abordagem performa melhor sob circunstâncias multi-rótulo, validando o método em *datasets* amplamente utilizados como MS-COCO e NUS-WIDE.

Embora o aprendizado contrastivo tenha demonstrado eficácia em diversos contextos, sua aplicação em problemas de segmentação semântica fracamente supervisionada

multi-rótulo apresenta limitações significativas. Uma das principais limitações reside na necessidade de tamanhos de *batch* substancialmente grandes para garantir diversidade suficiente de amostras negativas e estabilidade no treinamento. Chen *et al.* (CHEN et al., 2020) demonstraram empiricamente que o aprendizado contrastivo se beneficia significativamente de *batch sizes* maiores comparado ao aprendizado supervisionado, sendo que *batches* típicos de 64 ou mais amostras são frequentemente necessários para operação efetiva. Com *batch sizes* reduzidos, o número de amostras negativas disponíveis é limitado, resultando em gradientes instáveis e aprendizado subótimo das representações contrastivas.

Em problemas de WSSS multi-rótulo, a determinação de amostras positivas e negativas é particularmente ambígua, conforme destacado por Zhang e Wu (ZHANG; WU, 2024) no contexto do MulSupCon. Diferentemente de problemas de classificação single-label, onde a relação entre amostras é binária (pertencem à mesma classe ou não), em problemas multi-rótulo amostras podem compartilhar algumas classes mas não outras, criando um espectro contínuo de similaridade. Esta sobreposição parcial de rótulos exibe uma grande ambiguidade na determinação de amostras positivas quando há diferentes graus de sobreposição de rótulos entre amostras, conforme observado por Zhang e Wu (ZHANG; WU, 2024). Zaigrajew e Zieba (ZAIGRAJEW; ZIEBA, 2022) propuseram uma abordagem onde amostras são classificadas como positivas ou negativas com base na porcentagem de classes compartilhadas e um limiar estabelecido, no entanto, a seleção adequada deste limiar permanece um desafio em *datasets* com alta variabilidade de sobreposição de rótulos.

Além disso, classes raras podem não apresentar amostras positivas suficientes dentro de um *batch* para permitir aprendizado efetivo. Em *datasets* com desequilíbrio severo de classes, classes minoritárias podem aparecer em apenas uma ou duas amostras por *batch*, dificultando o estabelecimento de relações contrastivas significativas. Esta limitação é particularmente problemática quando combinada com a necessidade de *batch sizes* grandes, pois mesmo com *batches* maiores, classes raras podem continuar sub-representadas, comprometendo a capacidade do modelo de aprender representações discriminativas para essas classes.

Em resumo, o aprendizado contrastivo representa uma abordagem promissora para aprender representações úteis a partir de dados não rotulados ou parcialmente rotulados. Métodos como o SimCLR (CHEN et al., 2020) estabeleceram a base para aprendizado auto-supervisionado eficaz, demonstrando que a composição de aumentos de dados, transformações não lineares aprendíveis e *batch sizes* maiores são componentes críticos para o aprendizado efetivo de representações contrastivas. Extensões como as perdas multi-rótulo propostas por Zaigrajew e Zieba (ZAIGRAJEW; ZIEBA, 2022) e o MulSupCon (ZHANG; WU, 2024) adaptam esses princípios para contextos onde múltiplas classes podem coexistir, abordando explicitamente a ambiguidade na determinação de amostras

positivas. No entanto, as limitações práticas relacionadas a requisitos de *batch size* substancialmente grandes (conforme evidenciado empiricamente por Chen *et al.* (CHEN et al., 2020)) e a ambiguidade inerente na definição de amostras positivas/negativas em problemas multi-rótulo (conforme destacado por Zhang e Wu (ZHANG; WU, 2024)) motivam o desenvolvimento de estratégias alternativas que incorporem princípios contrastivos de forma mais flexível e eficiente em termos computacionais, especialmente em cenários com restrições de *hardware* ou *datasets* com desequilíbrio severo de classes.

2.6.3 Módulos Auxiliares

O *Pixel Correlation Module* (PCM) é um módulo desenvolvido para refinar os CAMs utilizando informações contextuais, que avalia a similaridade entre as características de cada *pixel* com base na distância cosseno (d_c), permitindo que *pixels* semelhantes influenciem as ativações uns dos outros, conforme indicado na Equação 3 (WANG et al., 2020):

$$d_c(\mathbf{x}_i, \mathbf{x}_j) = \frac{\theta(\mathbf{x}_i)^\top \theta(\mathbf{x}_j)}{\|\theta(\mathbf{x}_i)\| \cdot \|\theta(\mathbf{x}_j)\|}, \tag{3}$$

em que $\theta(\mathbf{x}_i)$ e $\theta(\mathbf{x}_j)$ correspondem às representações $\mathbf{x}_i$ e $\mathbf{x}_j$ de características do *pixels* i e j após a aplicação da transformação θ, que mapeia $\mathbf{x}_i$ e $\mathbf{x}_j$ para um novo espaço de características. O símbolo $\top$ representa operador de transposição matricial. Além disso, o denominador representa o produto das magnitudes dos vetores $\theta(\mathbf{x}_i)$ e $\theta(\mathbf{x}_j)$, utilizado para normalizar o produto interno obtido do cálculo $\theta(\mathbf{x}_i)^T\theta(\mathbf{x}_j)$ (WANG et al., 2020).

O PCM calcula uma média ponderada das ativações dos *pixels* vizinhos, utilizando medida de similaridade como peso, e aplica a função ReLU para ativar apenas as similaridades positivas, suprimindo valores negativos, como representado na Equação 4. A estrutura do PCM é projetada para manter a intensidade de ativação do CAM original, evitando conexões residuais que poderiam alterar essa intensidade e reduzindo a complexidade do modelo ao eliminar funções de incorporação que poderiam levar ao *overfitting*. Como resultado, o PCM melhora a consistência dos CAMs, resultando em ativações mais refinadas (WANG et al., 2020).

No entanto, o custo computacional do PCM apresenta uma limitação significativa. Dado que o módulo calcula a similaridade entre todos os pares de *pixels* na imagem, conforme indicado pela soma $\sum_{\forall j}$ na Equação 4, a complexidade computacional do PCM é quadrática em relação ao número de *pixels* na imagem, ou seja, $O(n^2)$, onde n representa o número de *pixels*. Para uma imagem de resolução $H \times W$, isso resulta em $O(H^2W^2)$ operações, o que pode ser computacionalmente custoso para imagens de alta resolução. Além disso, o cálculo das transformações $\theta(\mathbf{x}_i)$ e $\theta(\mathbf{x}_j)$ para cada par de *pixels* requer operações de produto interno e normalização, aumentando ainda mais o custo computacional e o consumo de memória durante o processamento (WANG et al., 2020). Esta

complexidade quadrática pode limitar a aplicabilidade do PCM em cenários que exigem processamento eficiente ou quando se trabalha com imagens de grande dimensão.

É importante notar que o PCM foi utilizado no modelo SSDB-II (CAI; ABHAYARATNE, 2023), que é um dos modelos do estado da arte comparados neste trabalho. No entanto, o PCM não foi incorporado na arquitetura do modelo proposto. Esta decisão foi motivada pelo fato de que o SegFormer-B0, utilizado como *backbone* no modelo proposto, já possui uma estrutura hierárquica que facilita a extração de CAMs de qualidade através de sua arquitetura de codificador multi-estágio (XIE et al., 2021).

O SegFormer emprega uma arquitetura hierárquica baseada em Mix Transformer (MiT), que mantém representações bidimensionais em diferentes estágios do codificador, preservando informações espaciais em múltiplas escalas (XIE et al., 2021). Esta estrutura hierárquica permite que as CAMs sejam geradas diretamente das camadas de normalização dos estágios finais do codificador, capturando informações contextuais em diferentes níveis de abstração sem necessidade de módulos adicionais de refinamento como o PCM. Além disso, o SegFormer utiliza um mecanismo de *Efficient Self-Attention* que reduz a complexidade computacional de $O(N^2)$ para $O(N/R^2)$ através de redução de sequência, onde N representa o número de *tokens* na sequência e R é o fator de redução, tornando o processamento significativamente mais eficiente em comparação com a complexidade quadrática do PCM (XIE et al., 2021). Esta redução de complexidade é alcançada através da aplicação de operações de atenção sobre sequências reduzidas, que requerem menos memória e computação durante o processamento.

A combinação da estrutura hierárquica do SegFormer com sua *Efficient Self-Attention* permite que o modelo capture dependências globais e locais de forma eficiente, gerando CAMs que já incorporam informações contextuais através dos mecanismos de atenção multi-escala. Desta forma, a adição do PCM, que possui complexidade quadrática $O(H^2W^2)$, não apenas seria redundante, mas também aumentaria significativamente o custo computacional do modelo sem proporcionar ganhos substanciais na qualidade das CAMs geradas. Esta abordagem mantém a eficiência computacional do modelo enquanto preserva a capacidade de gerar CAMs precisas para geração de pseudo-rótulos (XIE et al., 2021).

$$y_i = \frac{1}{C(\mathbf{x}_i)} \sum_{\forall j} \text{ReLU}\left(\frac{\theta(\mathbf{x}_i)^\top \theta(\mathbf{x}_j)}{\|\theta(\mathbf{x}_i)\| \cdot \|\theta(\mathbf{x}_j)\|}\right) \hat{y}_j, \tag{4}$$

sendo que y_i representa a CAM refinada para o pixel i, $C(\mathbf{x}_i)$ é uma função de normalização responsável por calcular a soma das similaridades para o *pixel* i e $\hat{y}_j$ é a CAM para o *pixel* j. A função ReLU é aplicada para garantir que apenas as similaridades positivas sejam ativadas. A estrutura do módulo PCM é representada pela Figura 7.

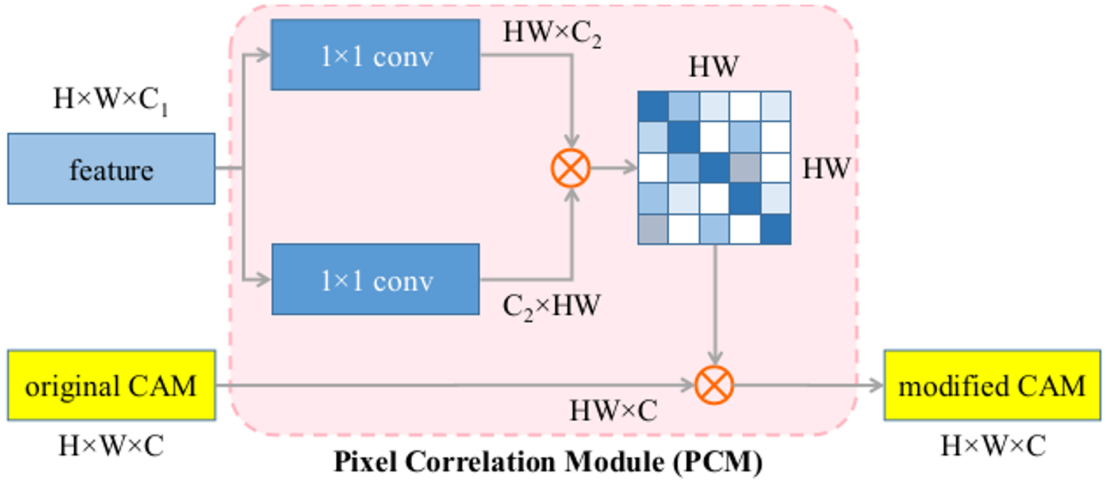

Figura 7 – Estrutura do PCM, retirado de (WANG et al., 2020).

2.6.4 Arquiteturas WSSS do Estado da Arte

2.6.4.1 SEAM

O *Self-supervised Equivariant Attention Mechanism* (SEAM) (WANG et al., 2020) é um método de segmentação semântica fracamente supervisionada que explora a invariância e equivariância das representações de características através de um mecanismo de atenção auto-supervisionado. Diferentemente de abordagens que utilizam arquiteturas de múltiplos ramos para treinamento simultâneo, o SEAM requer uma etapa separada para gerar pseudo-rótulos, que são então utilizados para treinar o segmentador.

O princípio fundamental do SEAM baseia-se na aplicação de transformações espaciais (como rotações, translações e escalas) nas imagens de entrada e na garantia de que as previsões do modelo permaneçam consistentes sob essas transformações. O mecanismo de atenção equivariante força o modelo a produzir mapas de ativação semelhantes para uma imagem original e suas versões transformadas através de uma função de perda de consistência. Esta abordagem melhora a robustez das CAMs geradas, reduzindo a sensibilidade a variações espaciais e melhorando a qualidade dos pseudo-rótulos utilizados na etapa de segmentação.

A arquitetura do SEAM utiliza um *backbone* convolucional (tipicamente ResNet) para extração de características, seguido por módulos de atenção que capturam relações contextuais em diferentes escalas. O processo de treinamento envolve duas etapas principais: (1) geração de CAMs refinadas através do mecanismo de atenção equivariante, e (2) utilização dessas CAMs como pseudo-rótulos para treinar um segmentador separado. Esta abordagem de duas etapas contrasta com métodos como o SSDB-Net, que realizam classificação e segmentação simultaneamente em uma única etapa de treinamento.

No contexto de segmentação de imagens de alimentos, o SEAM reportou um mIoU de 11,49% no *dataset* FoodSeg-103 (WANG et al., 2020), demonstrando a viabilidade da

abordagem para domínios específicos. No entanto, a necessidade de uma etapa separada para geração de pseudo-rótulos pode introduzir ruídos adicionais no processo de treinamento, uma limitação que foi abordada por métodos subsequentes como o SSDB-Net, que alcançou 14,79% de mIoU no mesmo *dataset* através de treinamento simultâneo dos ramos de classificação e segmentação.

2.6.4.2 *Separate and Conquer*

O modelo *Separate and Conquer* (SeCo) é um dos principais modelos de segmentação fracamente supervisionada no estado da arte. Essa arquitetura ganhou destaque por sua abordagem para lidar com o problema de coocorrência de classes, que pode prejudicar a segmentação de cenas mais complexas, onde o modelo tende a confundir a identificação de objetos individuais (YANG et al., 2024).

A abordagem proposta é baseada em duas etapas principais. Na primeira etapa, chamada de decomposição ou separação via redes de atenção, a imagem é dividida em pequenas regiões (*patches*) e cada *patch* recebe um rótulo baseado nas informações extraídas das CAMs. O mecanismo de decomposição utiliza redes de atenção que aprendem a separar representações de objetos que frequentemente coocorrem, permitindo que o modelo distinga entre diferentes classes mesmo quando aparecem juntas no mesmo contexto. Especificamente, o processo de atribuição de rótulos a *patches* baseia-se na análise das ativações das CAMs: *patches* com ativações altas para uma classe específica recebem o rótulo correspondente, enquanto *patches* com ativações ambíguas ou baixas são tratados de forma especial através de mecanismos de correção baseados em similaridade.

A correção baseada em similaridade reduz rótulos ruidosos através da análise de características semânticas entre *patches* vizinhos. *Patches* com características semelhantes tendem a pertencer à mesma classe, permitindo que o modelo corrija rótulos inicialmente atribuídos incorretamente através da comparação de similaridade cosseno entre as representações de características dos *patches*. Este processo iterativo melhora progressivamente a qualidade dos pseudo-rótulos gerados na primeira etapa.

Na segunda etapa, o processo de refinamento das características aprendidas é realizado por meio do aprendizado contrastivo supervisionado e destilação de conhecimento. O aprendizado contrastivo é aplicado de forma a facilitar a distinção entre co-contextos: amostras de *patches* pertencentes à mesma classe são aproximadas no espaço de características, enquanto *patches* de classes diferentes são afastados, mesmo quando essas classes frequentemente coocorrem. Este contraste semântico é fundamental para o sucesso do SeCo, pois permite que o modelo aprenda representações discriminativas que capturam diferenças sutis entre objetos coocorrentes.

A arquitetura completa do SeCo utiliza um *backbone* convolucional (tipicamente ResNet) para extração de características, seguido por módulos de atenção que realizam a decomposição na primeira etapa. O decodificador processa as características refinadas para

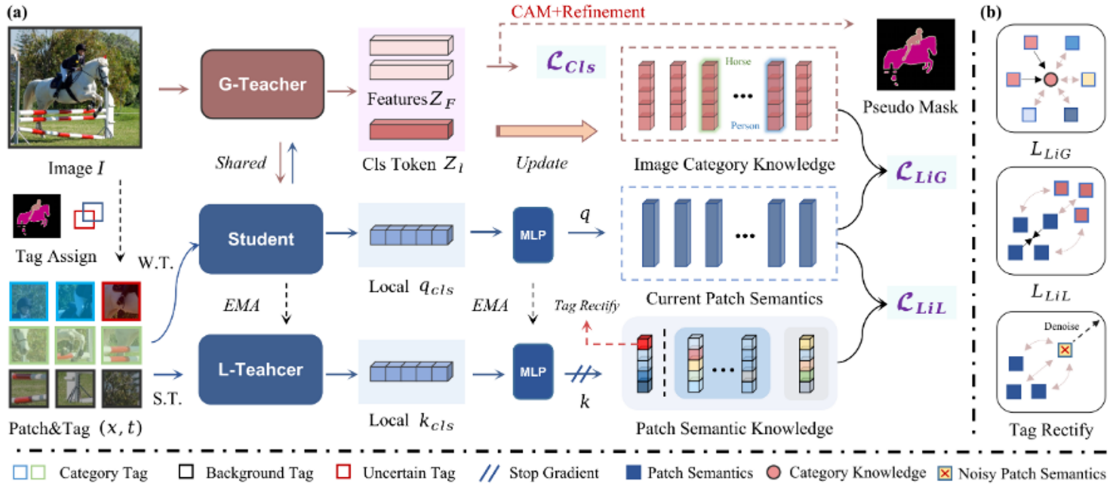

Figura 8 – Estrutura do modelo SeCo, retirado de (YANG et al., 2024).

gerar as segmentações finais. A abordagem de destilação de conhecimento implementa uma estratégia *two-teachers one-student*, onde dois modelos professores (treinados com diferentes inicializações ou configurações) geram pseudo-rótulos que são utilizados para supervisionar o treinamento de um modelo estudante. Esta abordagem reduz o impacto de erros individuais dos professores, melhorando a robustez do modelo final, conforme ilustrado na Figura 8.

O SeCo demonstrou resultados quantitativos significativos, alcançando 74,0% de mIoU no conjunto de validação do Pascal VOC e 73,8% no conjunto de teste (YANG et al., 2024). Em *datasets* mais desafiadores como MS COCO, o modelo alcançou 46,7% de mIoU, demonstrando sua capacidade de lidar com cenas complexas com múltiplos objetos e classes coocorrentes. Estes resultados posicionam o SeCo como um dos métodos mais eficazes para segmentação fracamente supervisionada em cenários com alta complexidade semântica.

2.6.4.3 *From SAM to CAMs*

Esse trabalho teve como objetivo explorar como modelos modernos do tipo *zero-shot* podem ser integrados e utilizados para aprimorar outros modelos baseados em CAM. O modelo de base escolhido foi o *Segment Anything Model* (SAM), que adota uma abordagem de segmentação projetada para ser um modelo generalista, aplicável a qualquer cenário sem a necessidade de treinamento especifico. O SAM foi pré-treinado em *datasets* com bilhões de exemplos (KIRILLOV et al., 2023; KWEON; YOON, 2024).

A Figura 9 ilustra a arquitetura simplificada do SAM, na qual um vetor característica de uma imagem é gerado, seguido pela adição de características auxiliares (como texto, *bounding box*, máscara, ou ponto). Por fim, o modelo produz as máscaras de segmentação (KIRILLOV et al., 2023).

O S2C (KWEON; YOON, 2024) propôs dois módulos importantes para o processo de

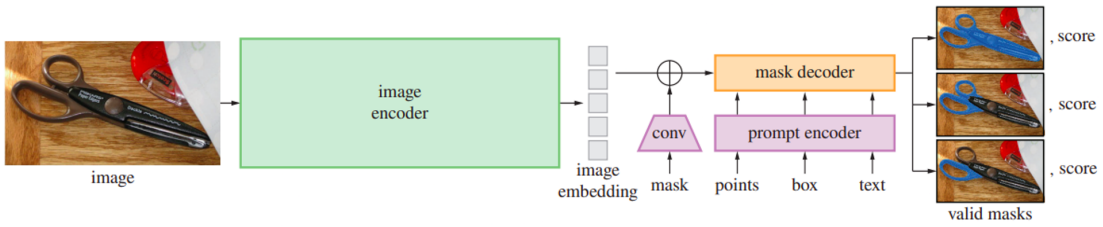

Figura 9 – Estrutura do modelo SAM, retirado de (KIRILLOV et al., 2023).

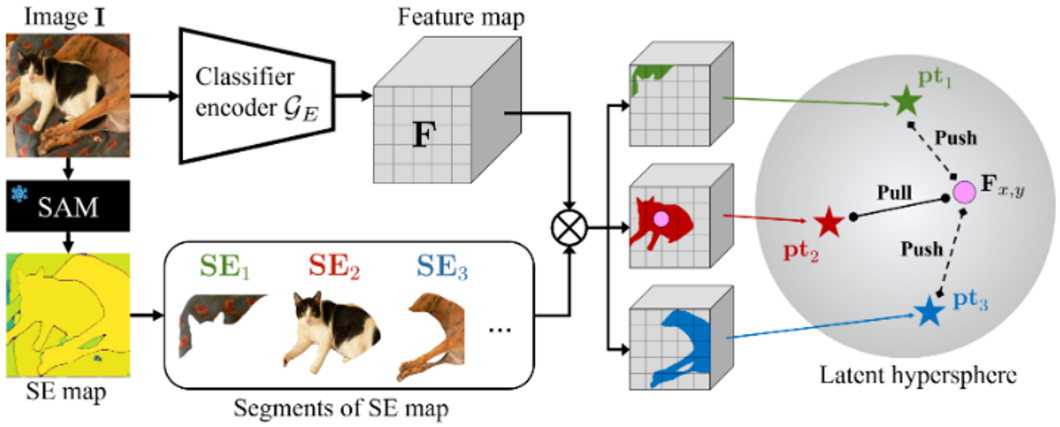

Figura 10 – Estrutura do modelo SSC, retirado de (KWEON; YOON, 2024).

transferência. O primeiro, chamado SSC, utiliza o modelo SAM congelado para extrair áreas de segmentação da imagem e, em seguida, usa essas áreas extraídas para treinar um codificador de forma contrastiva. Esse processo força as características extraídas pelo codificador a se aproximarem das características médias (protótipos) das respectivas áreas segmentadas (KWEON; YOON, 2024), conforme ilustrado na Figura 10.

Por fim, o segundo módulo proposto, o CAMs-based Prompting Module (CPM), utiliza as CAMs geradas previamente pelo classificador como entrada auxiliar para o modelo SAM congelado. O SAM retorna uma máscara para cada classe da CAM, que é agregada com base na confiança da predição do SAM e utiliza para o autotreinamento de forma iterativa (KWEON; YOON, 2024), conforme ilustrado na Figura 11.

2.6.4.4 DuPL: Dual Student with Trustworthy Progressive Learning

O DuPL (Dual Student with Trustworthy Progressive Learning) (WU et al., 2024) é um *framework* de segmentação semântica fracamente supervisionada que emprega uma arquitetura de duplo estudante para mitigar o viés de confirmação (*confirmation bias*) comum em modelos de rede única. O paradigma de duplo estudante treina duas sub-redes siamesas que se supervisionam mutuamente através de pseudo-rótulos cruzados, permitindo que cada estudante aprenda de forma mais robusta sob supervisão fraca.

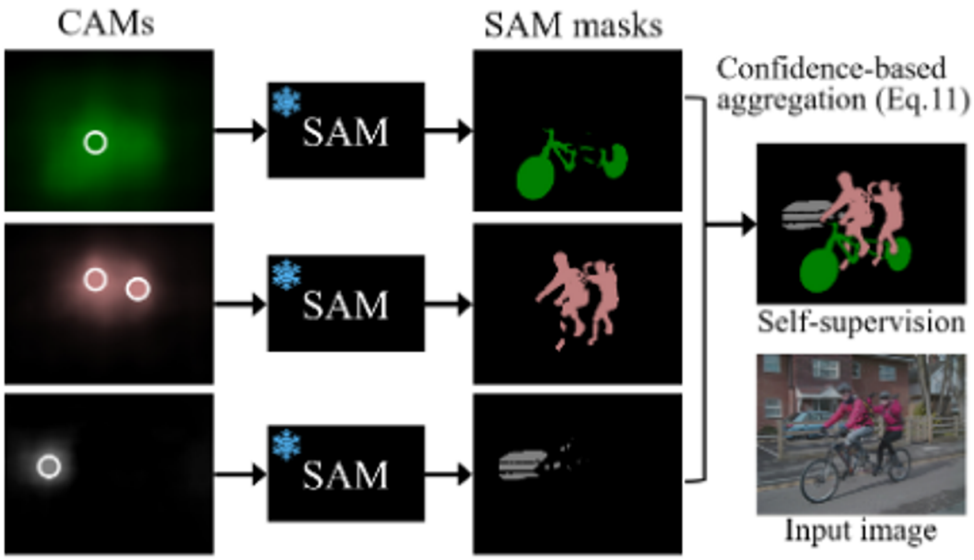

Figura 11 – Estrutura do modelo CPM, retirado de (KWEON; YOON, 2024).

A arquitetura DuPL incorpora vários mecanismos para manter diversidade entre os estudantes e melhorar a confiabilidade dos pseudo-rótulos. O *Dynamic Threshold Adjustment* (DTA) ajusta dinamicamente os limiares de confiança para geração de pseudo-rótulos durante o treinamento, adaptando-se à qualidade crescente das predições. O *Adaptive Noise Filtering* (ANF) identifica e filtra pseudo-rótulos ruidosos baseando-se em métricas de consistência entre os dois estudantes. Além disso, uma função de perda de discrepância (*discrepancy loss*) é empregada para garantir que os estudantes mantenham diversidade suficiente, evitando colapso em soluções idênticas.

O processo de aprendizado progressivo do DuPL envolve atualizações iterativas dos pseudo-rótulos, onde cada estudante gera segmentações que são utilizadas para supervisionar o outro. Esta supervisão cruzada permite que o modelo refine progressivamente suas predições, melhorando a qualidade dos pseudo-rótulos ao longo do treinamento. A regularização de consistência garante que ambos os estudantes concordem em regiões de alta confiança, enquanto a perda de discrepância encoraja diversidade em regiões ambíguas.

Embora o DuPL demonstre resultados robustos em *datasets* acadêmicos, sua dependência de módulos computacionalmente pesados, especialmente o ANF, aumenta significativamente o custo de treinamento e a complexidade de parâmetros. Esta limitação torna o treinamento contínuo ou em dispositivos embarcados impraticável para implantação industrial, motivando o desenvolvimento de arquiteturas de duplo estudante simplificadas que preservem a robustez enquanto reduzem o overhead computacional.

2.6.4.5 *Single-Step Dual Branch Network*

Single-Step Dual Branch Network (SSDB-Net) (CAI; ABHAYARATNE, 2023) é um modelo proposto para abordar SS de ingredientes utilizando *datasets* públicos de imagens de alimentos. No estudo de Cai *et al.* (CAI; ABHAYARATNE, 2023), foram propostas duas versões do modelo SSDB, cuja principal diferença é o compartilhamento de pesos entre os *backbones* utilizados. Na versão SSDB I, os dois ramos (classificação e segmentação) compartilham os mesmos pesos do *backbone* ResNet-38, resultando em uma arquitetura mais compacta mas com menor capacidade de especialização. A versão SSDB II utiliza pesos independentes para cada ramo, permitindo que cada *backbone* se especialize na tarefa específica (classificação ou segmentação), resultando em melhor desempenho. Esta diferença é fundamental: pesos independentes permitem que o ramo de classificação aprenda representações otimizadas para reconhecimento de classes em nível de imagem, enquanto o ramo de segmentação aprende representações otimizadas para localização espacial e geração de CAMs precisas.

O modelo SSDB II, ilustrado na Figura 12, possui dois ramos: um para classificação e um outro para segmentação. Cada ramo utiliza um *backbone* ResNet-38 com pesos independentes, pré-treinado no *dataset* Food-101. O ramo de classificação contém um módulo de atenção que captura contexto global da imagem através de mecanismos de atenção espacial e de canal. Este módulo de atenção funciona aplicando transformações que destacam regiões relevantes da imagem para a classificação, permitindo que o modelo foque em características discriminativas mesmo em presença de múltiplos ingredientes. A saída do módulo de atenção é processada por uma camada GAP, seguida por uma cabeça de classificação que produz os *logits* de classificação. A saída deste ramo é usada para suprimir camadas falsamente ativadas das CAMs geradas pelo ramo de segmentação, incorporando informação de contexto global que ajuda a refinar as ativações locais.

A função de perda utilizada no ramo de classificação é chamada *Multi-Label Soft Margin* (MLSM), descrita pela Equação 5(CAI; ABHAYARATNE, 2023):

$$\mathbf{l}_{\text{cla}}(\mathbf{z}, \mathbf{l}) = -\frac{1}{C}\sum_{c=1}^{C}\left[\mathbf{l}_c \log\left(\frac{1}{1+e^{-\mathbf{z}_c}}\right) + (1-\mathbf{l}_c)\log\left(\frac{e^{-\mathbf{z}_c}}{1+e^{-\mathbf{z}_c}}\right)\right], \tag{5}$$

sendo $\mathbf{z}$ os *logits* preditos, $\mathbf{l}$ é o vetor de rótulos binários, C é o número total de classes, $\mathbf{l}_c$ é o rótulo da classe c e $\mathbf{z}_c$é o valor predito para a classe c.

O ramo de segmentação, que é responsável pela geração das CAMs, possui uma cabeça de classificação, uma camada GAP para a obtenção das CAMs, e um módulo PCM para o refinamento das CAMs (CAI; ABHAYARATNE, 2023). A interação entre os dois ramos durante o treinamento ocorre através de supervisão cruzada: o ramo de classificação fornece informação de contexto global que é utilizada para refinar as CAMs geradas pelo ramo de segmentação, enquanto o ramo de segmentação gera CAMs que são diretamente utilizadas como supervisão para treinar o segmentador, evitando a necessidade de uma

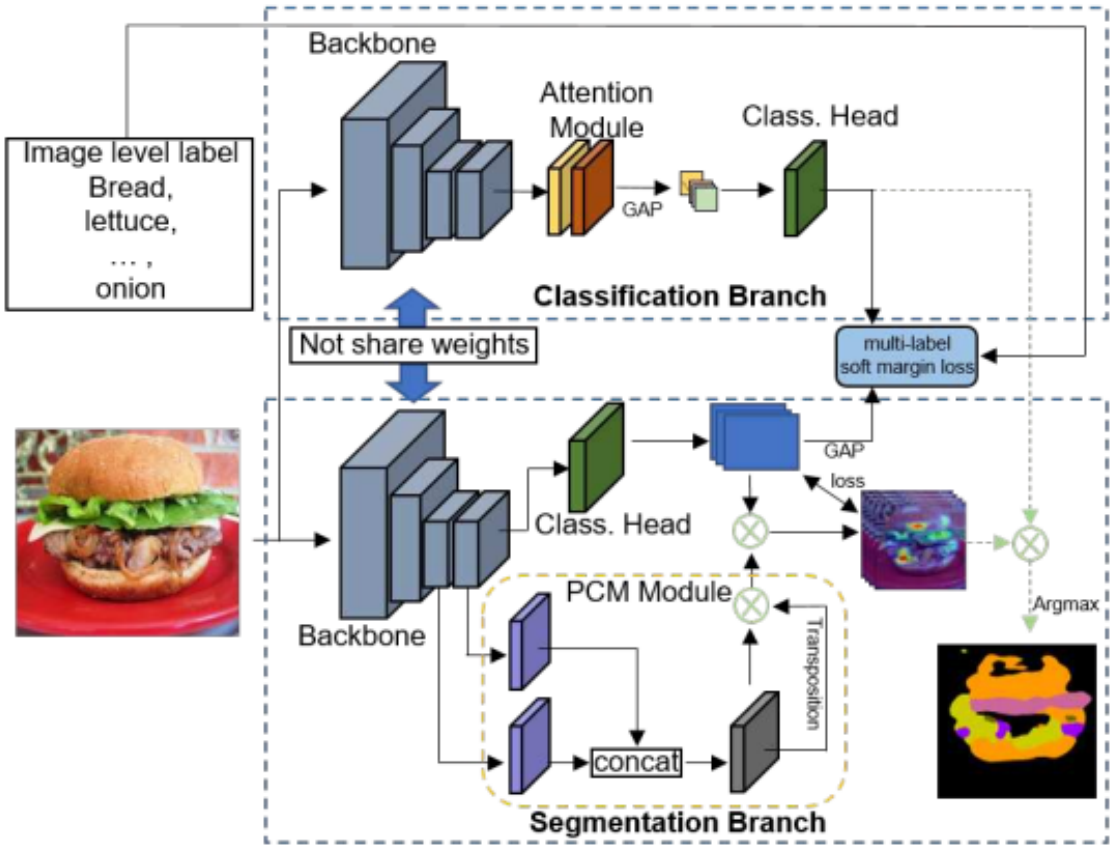

Figura 12 – Arquitetura do modelo SSDB II (CAI; ABHAYARATNE, 2023).

etapa separada de geração de pseudo-rótulos como no SEAM. Esta abordagem de treinamento simultâneo em uma única etapa (*single-step*) é uma das principais vantagens do SSDB-Net sobre métodos tradicionais que requerem múltiplas etapas de treinamento.

O módulo PCM utilizado no ramo de segmentação refina as CAMs através da análise de similaridade entre características de *pixels* vizinhos, melhorando a consistência espacial das ativações. As estratégias de aumento de dados utilizadas incluem espelhamento horizontal, rotações, suavização Gaussiana e *color dithering*, técnicas comuns em problemas de segmentação de imagens que aumentam a robustez do modelo a variações de iluminação, orientação e aparência.

O SSDB-Net II reportou um mIoU de 14,79% no *dataset* FoodSeg-103 (CAI; ABHAYARATNE, 2023), superando significativamente o SEAM, que alcançou 11,49% de mIoU no mesmo *dataset*. Esta melhoria de aproximadamente 3,3 pontos percentuais demonstra a eficácia da abordagem de treinamento simultâneo em uma única etapa, que evita a acumulação de erros que ocorre em métodos de múltiplas etapas como o SEAM, onde pseudo-rótulos gerados na primeira etapa podem conter ruídos que são propagados para a etapa de segmentação.

Este capítulo apresentou o embasamento teórico das principais técnicas e métodos utilizados nesta pesquisa, incluindo Redes Neurais Artificiais, Redes Neurais Convolucionais, Vision Transformers, arquiteturas para Segmentação Semântica, técnicas de aprendizado

fracamente supervisionado baseadas em CAMs e aprendizado contrastivo. Os conceitos discutidos aqui fundamentam a compreensão dos modelos e métodos apresentados nos capítulos subsequentes, fornecendo a base teórica necessária para o desenvolvimento e avaliação do modelo proposto.

Capítulo 3

Revisão bibliográfica

Neste capítulo é apresentado o estado da arte de técnicas propostas para a Segmentação Semântica de ingredientes em imagens de alimentos.

Este capítulo apresenta a revisão bibliográfica do estado da arte relacionado à Segmentação Semântica de ingredientes em imagens de alimentos, focando em trabalhos relacionados ao problema específico investigado. Diferentemente do capítulo anterior, que apresentou os fundamentos teóricos e conceitos gerais das técnicas utilizadas, este capítulo revisa métodos e modelos específicos desenvolvidos para problemas similares, permitindo identificar lacunas e oportunidades para o desenvolvimento do modelo proposto.

3.1 Segmentação Semântica de ingredientes em imagens de alimentos

A segmentação semântica de ingredientes em imagens de alimentos apresenta desafios únicos que a distinguem de problemas de segmentação de objetos genéricos (WU et al., 2021b; CAI; ABHAYARATNE, 2023). Ingredientes alimentares exibem alta variabilidade intra-classe devido a variações em preparação, apresentação e condições de captura (SALVADOR et al., 2017; LAN et al., 2023), enquanto fronteiras semânticas frequentemente são ambíguas quando ingredientes se sobrepõem ou se misturam (CAI; ABHAYARATNE, 2023). Além disso, a segmentação de ingredientes requer granularidade fina para aplicações industriais, onde distinções sutis entre classes visualmente similares são críticas para controle de qualidade. Apesar dos avanços, muitos métodos assumem condições de captura e qualidade de imagem não presentes em ambientes industriais, onde sucesso em *datasets* acadêmicos não garante desempenho adequado.

Esta revisão bibliográfica inicia com uma análise de métodos de SS fracamente supervisionados baseados em CAMs, destacando modelos desenvolvidos para o setor alimentício. Em seguida, são apresentados modelos baseados em CAMs aplicados a imagens gerais. Ainda no contexto de SS fracamente supervisionada de alimentos, é apresentado e discutido um modelo fundamentado em *Multiple Instances Learning* (MIL). Por fim, é abordado um modelo que se compara ao estado da arte em SS totalmente supervisionada de ingredientes, além de um estudo sobre o impacto de diversas técnicas de aumento de dados (*data augmentation*) na tarefa proposta.

3.1.1 Segmentação Semântica fracamente supervisionada

O SSDB-Net (CAI; ABHAYARATNE, 2023) é um modelo de SS fracamente supervisionado voltado para a segmentação de imagens de alimentos. Diferente de modelos tradicionais como o *Self-supervised Equivariant Attention Mechanism* (SEAM) (WANG et al., 2020), que utilizam uma etapa separada para gerar pseudo-rótulos para o treinamento da segmentação, o SSDB-Net explora sua arquitetura de duas ramificações para realizar o treinamento simultâneo nas tarefas de classificação e segmentação. O CAM gerado pelo modelo é usado diretamente por meio da função de perda para treinar o segmentador, evitando possíveis ruídos oriundos do uso de pseudo-rótulos, como ocorre no treinamento do modelo SEAM.

Cada ramo do modelo SSDB-Net II (CAI; ABHAYARATNE, 2023) utiliza um *backbone* Resnet-38 (HE et al., 2016; WU; SHEN; HENGEL, 2019) pré-treinado com pesos independentes. Os *backbones* são re-treinados para a tarefa de classificação usando o *dataset* Food-101 (BOSSARD; GUILLAUMIN; GOOL, 2014), aplicando técnicas de aumento de dados (*data augmentation*) comuns em problemas de segmentação de imagens, como espelhamento, rotação, suavização Gaussiana e *color dithering*. Além do aumento de dados, a arquitetura também conta com um módulo de atenção no ramo de classificação, para utilização do contexto da imagem no treinamento do classificador. O próprio estudo (CAI; ABHAYARATNE, 2023) reportou que esse módulo não melhorou o resultado geral do modelo e constitui complexidade desnecessária no ramo de classificação. O estudo reportou um mIoU médio de 14,79% no *dataset* FoodSeg103 (WU et al., 2021b), superando o modelo SEAM (WANG et al., 2020), que alcançou 11,49% na mesma métrica (CAI; ABHAYARATNE, 2023).Embora sejam métricas baixas em valor absoluto, são esperadas dada a dificuldade do *dataset* (muitas classes e distribuição long-tail). No trabalho de (WANG et al., 2020), o SEAM alcança 64,5% de mIoU no conjunto de validação do Pascal VOC 2012. No entanto, o desempenho relativamente baixo (14,79% mIoU) indica dependência de CAMs limpas e refinadas, enquanto a avaliação foi realizada exclusivamente em *datasets* acadêmicos, não contemplando desafios específicos de imagens industriais de baixa qualidade ou condições de captura adversas.

No campo de SS fracamente supervisionada de imagens de alimentos com modelos ba-

seados em CAMs, o modelo proposto por Wang *et al.* (WANG et al., 2017) adapta a rede VGG-16 (SIMONYAN; ZISSERMAN, 2014), originalmente desenvolvida para aprendizado supervisionado, para distinguir pixels de alimentos em uma imagem num problema de SS fracamente supervisionada. As principais modificações incluem a adição de uma camada convolucional com 1024 canais e a remoção da primeira camada totalmente conectada da rede. A camada de *max pooling* foi substituída pela *Global Max-Average Pooling* (GMAP), que combina *Global Max Pooling* (GMP) e *Global Average Pooling* (GAP)(LIN, 2013). Estas novas camadas permitem ao modelo lidar com imagens de diferentes tamanhos e melhorar a eficácia na extração de características.

A camada GMAP (WANG et al., 2017) atua como uma camada *pooling* generalizada que combina as características dos métodos GMP e GAP (LIN, 2013), permitindo uma melhor representação das regiões dos objetos nas imagens. A arquitetura modificada processa imagens RGB de 224×224 pixels e gera um vetor de características que é usado para classificação.

O modelo de Wang *et al.* (WANG et al., 2017) também incorpora o conceito de *kernel* adaptativo e pode ser estendido para classificação multi-categoria, embora essa extensão não seja o foco principal do estudo. O objetivo é adaptar a rede para operar de forma eficaz em cenários em que anotações completas não estão disponíveis, utilizando apenas informações parciais ou fracamente anotadas.

Para avaliação, foi utilizado um *dataset* proveniente de um estudo de vida livre (em que participantes registram a dieta em condições naturais) do sistema *Technology Assisted Dietary Assessment* (TADA)(ZHU et al., 2010) (sistema de avaliação dietética assistida por tecnologia), composto por 1453 imagens de 56 alimentos comumente consumidos. Os *datasets* Caltech-256 (GRIFFIN et al., 2007), UECFOOD-256 (KAWANO; YANAI, 2014) e Food-101 (BOSSARD; GUILLAUMIN; GOOL, 2014) foram empregados para o treinamento do modelo. Devido à época da publicação, não havia disponibilidade de *datasets* públicos específicos para segmentação de ingredientes de imagens de alimentos. Além disso, o modelo não reporta métricas de eficiência computacional, como consumo de VRAM (memória de vídeo da GPU) ou vazão (*throughput*) de inferência, e não foi avaliado sob condições de imagens de baixa qualidade típicas de ambientes industriais, limitando sua aplicabilidade prática para cenários com restrições de *hardware* ou qualidade de imagem degradada.

Fora do escopo de segmentação de ingredientes de imagens de alimentos, o estudo de Yang *et al.* (YANG et al., 2024) aborda o problema de coocorrência de classes, que frequentemente leva a ativações falsas. A abordagem proposta divide a imagem em *patches* e atribui rótulos em nível de imagem para cada *patch* com base nas informações obtidas pela CAM. Para reduzir a presença de rótulos ruidosos, é aplicada uma correção baseada em similaridade. Além disso, uma arquitetura de destilação de conhecimento (GOU et al., 2021) *two-teachers one-student* foi desenvolvida, inspirada em arquiteturas de destila-

ção de conhecimento, para construir conhecimento multigranular e promover o contraste semântico, facilitando a distinção entre co-contextos, ou seja, distinguindo classes que aparecem normalmente em um mesmo contexto, como exemplo, trilhos e trens.

Na arquitetura SeCo (YANG et al., 2024), cada *teacher* e o *student* são compostos por um codificador com *backbone* ViT-B (DOSOVITSKIY et al., 2021) pré-treinado no ImageNet (WIGHTMAN, 2021), e um decodificador com uma cabeça de segmentação formada por quatro camadas convolucionais 3×3. Os rótulos atribuídos aos *patches* são usados para formar pares positivos ou negativos no aprendizado contrastivo. O estudo de ablação mostrou que a inclusão da perda contrastiva melhora significativamente a revocação e o mIoU.

Os experimentos foram realizados utilizando o *dataset* Pascal VOC 2012 (EVERINGHAM et al., 2012) com aumento de dados (*data augmentation*), alançando 74,0% e 73,8% de mIoU nos conjuntos de validação e teste, respectivamente. O modelo também foi treinado no *dataset* MS COCO (LIN et al., 2014), obtendo 46,7% de mIoU no conjunto de validação. No entanto, o método SeCo requer aprendizado contrastivo explícito, que tipicamente demanda *batch sizes* substancialmente grandes (não reportado no estudo, mas comum em métodos contrastivos) para funcionar efetivamente. Além disso, a avaliação foi realizada exclusivamente em *datasets* de objetos genéricos (Pascal VOC e MS COCO), não contemplando desafios específicos de segmentação de ingredientes alimentares ou condições de imagens industriais.

O modelo S2C (KWEON; YOON, 2024) propõe o uso eficiente do modelo base *Segment Anything Model* (SAM) (KIRILLOV et al., 2023) para segmentação de objetos. O estudo de Kweon *et al.* (KWEON; YOON, 2024) observou que muitos modelos utilizam o SAM como etapa de pós-processamento ou como método para facilitar a inferência *zero-shot*, no entanto, os autores destacam que essa prática pode tornar o *pipeline* de treinamento vulnerável a ruídos presentes nas CAMs.

Para explorar o conhecimento do modelo base, o S2C (KWEON; YOON, 2024) propôs dois métodos: SAM-*Segment Contrasting* (SSC) e CAM-based Prompting Module (CPM). O SSC utiliza o SAM (KIRILLOV et al., 2023) para segmentar a imagem em múltiplos segmentos, gerando um vetor que representa a média das características de cada segmento, o qual atua como uma âncora no aprendizado contrastivo, atraindo características do segmento correspondente e repelindo características não relacionadas.

Por outro lado, o CPM refina as CAMs obtidas durante o treinamento utilizando o SAM (KIRILLOV et al., 2023), empregando os máximos locais de cada CAM como incitadores para o segmentador.

No S2C (KWEON; YOON, 2024), o classificador utiliza uma ResNet-38 (HE et al., 2016; WU; SHEN; HENGEL, 2019) como codificador de características, seguida por uma cabeça de classificação com uma camada convolucional 1×1 para gerar as CAMs. O segmentador, por sua vez, utiliza o DeepLab (CHEN et al., 2017) com ResNet-38 (HE et

al., 2016; WU; SHEN; HENGEL, 2019) como *backbone*, enquanto o SAM (KIRILLOV et al., 2023) emprega o modelo pré-treinado com ViT-H (HE et al., 2022) .

O estudo de ablação demonstrou que a inclusão das perdas SSC e CPM resultou em um aumento de 26,0% no mIoU em relação ao *baseline* no conjunto de validação do Pascal VOC 2012.

Durante o treinamento, foram aplicadas técnicas de aumento de dados (*data augmentation*), como espelhamento, redimensionamento, recorte e *color jittering*. O modelo foi treinado no Pascal VOC 2012 (EVERINGHAM et al., 2012) por seis épocas, totalizando oito horas em duas *Graphics Processing Units* (GPUs) Tesla V100s, obtendo 78,2% e 77,5% de mIoU nos conjuntos de validação e teste, respectivamente. No *dataset* MS COCO (LIN et al., 2014), o modelo alcançou 49,8% de mIoU no conjunto de validação. Apesar do desempenho competitivo, o S2C apresenta limitações significativas de eficiência computacional: requer o modelo SAM pré-treinado com ViT-H, que é computacionalmente pesado, e o tempo de treinamento de oito horas em duas GPUs de alta capacidade indica alto custo computacional. Além disso, o estudo não reporta consumo de VRAM, *throughput* de inferência ou viabilidade com *batch sizes* limitados, métricas críticas para implantação industrial.

O DuPL (WU et al., 2024) é um método de SS fracamente supervisionada que utiliza duas redes estudante (dual student) com supervisão cruzada por pseudo-rótulos, CAM, Ajuste Dinâmico de Limiar (DTA), *Adaptive Noise Filtering* (ANF) e perda de discrepância para mitigar o viés de confirmação em *pipelines* de estágio único. O modelo foi treinado no Pascal VOC 2012 e avaliado no conjunto de validação do Pascal VOC 2012, reportando 74,1% de mIoU. A supervisão é em nível de imagem e o foco é em objetos gerais.

No campo de imagens de alimentos, o estudo (VLACHOPOULOU; SARAFIS; PAPADOPOULOS, 2023) apresentou uma abordagem inovadora para SS de ingredientes, que é baseada em MIL, em contraste com o uso tradicional de CAMs. A técnica de MIL, para SS fracamente supervisionada, envolve a seleção aleatória de um número K de *patches* de uma imagem, agrupados em "*bags*". Cada *patch* é processado pelo *backbone* do modelo ResNet-34 (HE et al., 2016), gerando K vetores de características de dimensão 128.

Um mecanismo de atenção específico do MIL é empregado para identificar a classe-alvo de alimentos na imagem e gerar um mapa de calor que destaca as regiões relevantes da classe. Esse mapa de calor é criado atribuindo pesos aos vetores de características de cada *patch*, refletindo a relevância de cada região da imagem para a classe-alvo. A combinação ponderada dessas características resulta no mapa de calor, que é então utilizado para segmentar a imagem ao aplicar um limiar em seus valores. Dessa forma, além de classificar a imagem, o modelo também fornece segmentações precisas das áreas correspondentes à classe-alvo.

O modelo (VLACHOPOULOU; SARAFIS; PAPADOPOULOS, 2023) foi treinado

utilizando o *dataset* FoodSeg103 (WU et al., 2021b), dividido em conjuntos de treino, validação e teste. Para simplificar o problema, foram consideradas duas meta-classes: "Padaria", que inclui categorias como torta de ovo, biscoito, bolo e pão, e "Carne", composta por bife, porco, frango, pato, salsicha, carne frita e cordeiro. No pré-processamento, as imagens foram redimensionadas para 512×512 pixels. Uma imagem é atribuída a uma classe se pelo menos 7,6% da imagem for ocupada pela categoria correspondente.

Nos resultados, a meta-classe "Padaria" alcançou um índice de Interseção sobre União (IoU) de 47,4%, enquanto "Carne" obteve 53,4%, um mIoU total de 50,4% considerando as duas classes, refletindo o desempenho do modelo na SS para cada meta-classe. No entanto, a simplificação do problema para apenas duas meta-classes representa uma granularidade grosseira que não atende aos requisitos de aplicações industriais, onde é necessário distinguir ingredientes em nível de granularidade fina. Esta limitação torna o modelo inadequado para cenários que requerem distinção precisa entre múltiplas classes de ingredientes visualmente similares.

Apesar dos avanços apresentados pelos métodos de SS fracamente supervisionada, observa-se que a maioria assume condições de captura e qualidade de imagem não presentes em ambientes industriais. Métodos que dependem de aprendizado contrastivo explícito frequentemente requerem *batch sizes* substancialmente grandes, tornando-se inviáveis em cenários com restrições de *hardware*. Além disso, a maioria dos trabalhos não reporta métricas críticas para implantação industrial, como consumo de VRAM e *throughput* de inferência.

3.1.2 Segmentação Semântica supervisionada de ingredientes

Apesar do desempenho superior alcançado por métodos totalmente supervisionados, esses métodos exigem anotação em nível de pixel, que é incompatível com o cenário estudado. A anotação pixel-level de imagens industriais requer especialistas treinados e consome tempo substancialmente maior que anotação em nível de imagem, tipicamente exigindo de 30 minutos a várias horas por imagem dependendo da complexidade. Para um *dataset* industrial com milhares de imagens, o custo e tempo de anotação tornam métodos totalmente supervisionados inviáveis do ponto de vista prático e de escalabilidade. Portanto, embora esses métodos estabeleçam benchmarks importantes, a abordagem fracamente supervisionada é necessária para viabilizar aplicações industriais.

Semelhante ao S2C (KWEON; YOON, 2024), o FoodSAM (LAN et al., 2023) integra o SAM (KIRILLOV et al., 2023) para a SS supervisionada de ingredientes, formando um *pipeline* com três modelos principais: um segmentador semântico M_s, o SAM M_a e o detector de objetos M_d. O *pipeline* adota estratégias como correspondência de máscara-categoria, mesclagem, e seleção prévia baseada em *prompts*. Além da SS de alimentos, o artigo aborda outros desafios, como a segmentação de instância (HAFIZ; BHAT, 2020) e a segmentação panóptica (KIRILLOV et al., 2019), que combina características das

duas abordagens anteriores. No processo, o SAM gera máscaras de segmentos, o segmentador semântico atribui categorias a essas máscaras, e o detector identifica classes de não-alimentos para o *background*. Integrando esses componentes por meio de uma estratégia de mesclagem, o modelo produz resultados tanto em nível de instância quanto panóptico.

Foram utilizados no estudo (LAN et al., 2023) os *datasets* FoodSeg103 (WU et al., 2021b) e UECFoodPix Complete (OKAMOTO; YANAI, 2021) para o treinamento. O modelo SAM pré-treinado na versão ViT-H (HE et al., 2022) foi usado, com o UniDet (ZHOU; KOLTUN; KRÄHENBÜHL, 2022) como detector de objetos, o SETR (ZHENG et al., 2021) com ViT-16/B (DOSOVITSKIY et al., 2021) e *Multi-Level feature Aggregation* (MLA) (ZHENG et al., 2021) como decodificador para SS no treinamento do FoodSeg103 (WU et al., 2021b), e o DeepLabv3+ (CHEN et al., 2018) como segmentador semântico no treinamento do UECFoodPix Complete.

No artigo (LAN et al., 2023), o modelo foi avaliado na tarefa de SS usando os *datasets* públicos FoodSeg103 (WU et al., 2021b) e UECFoodPix Complete (OKAMOTO; YANAI, 2021), alcançando mIoUs de 46,42% e 66,14%, respectivamente, estabelecendo o estado da arte para esses *datasets*. Observa-se também que os resultados da linha de base de cada experimento apresentaram valores significativos como métrica absoluta, com o SETR (ZHENG et al., 2021) alcançando 45,1% de mIoU no FoodSeg103 (WU et al., 2021b) e DeepLabV3+ (CHEN et al., 2018) alcançando 65,61% de mIoU no UECFoodPix Complete (OKAMOTO; YANAI, 2021).

O estudo de Takayuki *et al.* (TAKAYUKI et al., 2022) investigou como diferentes tipos de aumento de dados (*data augmentation*) afetam a segmentação supervisionada de imagens de alimentos, bem como o impacto dos tipos de imagens escolhidas para o *dataset*. Foram empregadas técnicas de corte, espelhamento, rotação, *color jittering* e composição de imagens. Esta última técnica consiste em extrair partes de alimentos, adicioná-la a outra imagem e suavizar as bordas, sendo semelhante ao método *copy-paste*. O estudo utilizou um *dataset* próprio com imagens de ingredientes recém-preparados, tanto individuais quanto misturados, contemplando classes como cenoura, repolho, brotos, carne de porco, pimentão verde, cebola e cogumelo shimeji. O modelo utilizado foi o U-Net (RONNEBERGER; FISCHER; BROX, 2015), treinado com 10.000 imagens, durante 30 épocas, com *batch-size* de 8 e *learning rate* de 0,0010. A acurácia alcançada foi de 86,6% ao utilizar composição, corte e manipulação de cores, enquanto que, no cenário de imagens com ingredientes individuais sem composição, a acurácia foi de 48,3% (Tabela 2).

Referência	**Modelo**	***Dataset* de treinamento**	***Dataset* de avaliação de SS**	**Principais Técnicas**	**Supervisão**	**Foco de Segmentação**	**Resultados**
(CAI; ABHAYARATNE, 2023)	SSDB-II	Food-101	val FoodSeg103	CAM, Atenção, PCM, RNC	nível de imagem	Ingredientes	14,79% mIoU
(WANG et al., 2020)	SEAM	Pascal VOC 2012	val FoodSeg103	CAM, Atenção equivariante, PCM	nível de imagem	Ingredientes	11,49% mIoU
(WANG et al., 2020)	SEAM	Pascal VOC 2012	val Pascal VOC 2012	CAM, Atenção equivariante, PCM	nível de imagem	Objetos Gerais	64,5% mIoU
(YANG et al., 2024)	SeCo	Pascal VOC 2012	val Pascal VOC 2012	CAM, Aprendizado contrastivo, Destilação de conhecimento	nível de imagem	Objetos Gerais	74,00% mIoU
(YANG et al., 2024)	SeCo	MS COCO	val MS COCO	CAM, Aprendizado contrastivo, Destilação de conhecimento	nível de imagem	Objetos Gerais	46,70% mIoU
(KWEON; YOON, 2024)	S2C	Pascal VOC 2012	val Pascal VOC 2012	CAM, Aprendizado contrastivo, SAM prompting	nível de imagem	Objetos Gerais	78,20% mIoU
(KWEON; YOON, 2024)	S2C	Pascal VOC 2012	test Pascal VOC 2012	CAM, Aprendizado contrastivo, SAM prompting	nível de imagem	Objetos Gerais	77,50% mIoU
(KWEON; YOON, 2024)	S2C	MS COCO	val MS COCO	CAM, Aprendizado contrastivo, SAM prompting	nível de imagem	Objetos Gerais	49,80% mIoU
(WU et al., 2024)	DuPL	Pascal VOC 2012	val Pascal VOC 2012	Dual student, CAM, DTA, ANF, perda de discrepância	nível de imagem	Objetos Gerais	74,1% mIoU
(VLACHOPOULOU; SARAFIS; PAPADOPOULOS, 2023)	-	FoodSeg103	val FoodSeg103	MIL, Atenção	nível de imagem	Carne e Padaria	50,40% mIoU
(WANG et al., 2017)	-	Caltech-256, UECFOOD-256, Food-101	-	CAM	nível de imagem	Ingredientes	Grafico de precisão e revocação
(LAN et al., 2023)	FoodSam	UECFoodPix Complete	val UECFoodPix Complete	SAM	nível de pixel	Alimentos	66,14% mIoU
(LAN et al., 2023)	DeepLabV3+	UECFoodPix Complete	val UECFoodPix Complete	-	nível de pixel	Alimentos	65,61% mIoU
(LAN et al., 2023)	FoodSam	FoodSeg103	val FoodSeg103	SAM	nível de pixel	Ingredientes	46,42% mIoU
(ZHENG et al., 2021)	SETR	FoodSeg103	val FoodSeg103	-	nível de pixel	Ingredientes	45,10% mIoU
(TAKAYUKI et al., 2022)	U-Net	Dataset proprietário	val dataset proprietário	Aumento de dados com composição	nível de pixel	Ingredientes	86,60% Acurácia
(TAKAYUKI et al., 2022)	U-Net	Dataset proprietário	val dataset proprietário	Aumento de dados sem composição	nível de pixel	Ingredientes	48,30% Acurácia

Tabela 2 – Resumo de trabalhos relacionados: modelos, *datasets*, técnicas e resultados. Onde *val* significa o conjunto de validação do *dataset* e *test* indica o conjunto de teste do *dataset*.

3.2 Lacunas Identificadas no Estado da Arte

A análise dos trabalhos relacionados revela várias lacunas significativas no estado da arte que justificam o desenvolvimento do modelo proposto. Estas lacunas são particularmente relevantes para aplicações industriais de segmentação semântica de ingredientes em imagens de montagem de lanches *fast-food*.

3.2.1 Eficiência Computacional

A maioria dos trabalhos revisados não reporta consumo de VRAM, *throughput* de inferência ou tempo de treinamento por época. Essas métricas são críticas para implantação industrial, onde restrições de *hardware* são comuns. Modelos existentes do estado da arte, como o DuPL e o S2C, apresentam limitações significativas em termos de eficiência computacional. O DuPL, por exemplo, utiliza módulos computacionalmente pesados, especialmente o *Adaptive Noise Filtering* (ANF), que aumentam substancialmente o consumo de memória e o tempo de processamento. Métodos que dependem de componentes complexos como refinamento iterativo de pseudo-rótulos com múltiplas etapas de processamento ou módulos de atenção pesados tornam-se menos adequados para implantação em ambientes industriais com restrições de recursos computacionais. Esta limitação é crítica para aplicações que requerem processamento em tempo real ou implantação em dispositivos embarcados.

3.2.2 Adequação para Imagens Industriais

A maioria dos modelos do estado da arte foi desenvolvida e avaliada em *datasets* acadêmicos, como Pascal VOC e MS COCO, que apresentam características visuais distintas das imagens capturadas em ambientes industriais. Imagens de montagem de lanches capturadas por câmeras CFTV apresentam desafios específicos, incluindo baixa resolução, condições de iluminação variáveis, oclusões frequentes e artefatos de compressão de vídeo, conforme discutido no Capítulo 1. Métodos avaliados exclusivamente em imagens de alta qualidade podem degradar significativamente quando aplicados a imagens de CFTV industriais. Estes desafios não são adequadamente abordados por modelos treinados exclusivamente em *datasets* acadêmicos, que tipicamente apresentam imagens de alta qualidade e condições controladas.

3.2.3 Granularidade de Classes

Os *datasets* públicos de segmentação de alimentos, como FoodSeg-103 e UECFoodPix, segmentam ingredientes em nível grosso, agrupando classes visualmente similares ou funcionalmente relacionadas. Granularidade fina não é detalhe, é requisito industrial para controle de qualidade. Para aplicações industriais de controle de qualidade, é necessário

distinguir ingredientes em nível de granularidade fina. Por exemplo, enquanto *datasets* públicos podem segmentar genericamente "queijo", o contexto industrial requer distinção entre "*cheddar*", "queijo" e "*cream cheese*", cada um com características visuais e funcionais distintas. O *dataset* VSS utilizado neste trabalho apresenta 17 classes específicas que refletem esta necessidade de granularidade fina, conforme detalhado no Capítulo 4.

3.2.4 Requisitos de Batch Size para Aprendizado Contrastivo

Nenhum dos trabalhos revisados que utilizam aprendizado contrastivo explícito reporta viabilidade com *batch sizes* menores que 32. Métodos de aprendizado contrastivo explícito, como o MulSupCon, requerem tamanhos de *batch* substancialmente grandes (tipicamente 64 ou mais) para funcionar efetivamente. Esta limitação é particularmente problemática em cenários com restrições de *hardware*: modelos baseados em Transformers consomem quantidades significativas de memória, o que limita o *batch size* praticável e torna *batch sizes* grandes impraticáveis em ambientes com recursos limitados. Esta restrição de *batch size* tem implicações diretas na viabilidade de métodos que dependem de aprendizado contrastivo explícito, conforme discutido em detalhes no Capítulo 2, motivando a necessidade de estratégias alternativas que não dependam de *batch sizes* grandes para funcionar efetivamente.

3.2.5 Estabilidade e Robustez

A maioria dos trabalhos reporta apenas o melhor resultado ou média sem desvio padrão. Ausência de validação cruzada ou análise de variabilidade entre *folds* indica que robustez não foi adequadamente avaliada na maioria dos estudos revisados. Modelos existentes frequentemente apresentam alta variabilidade entre diferentes divisões de dados ou *folds* de validação cruzada, indicando sensibilidade a mudanças na distribuição de dados. Esta variabilidade é problemática para aplicações industriais, onde desempenho previsível e robustez a mudanças de dados são essenciais para implantação confiável. A ausência de análises estatísticas robustas, como testes de significância ou intervalos de confiança, dificulta a avaliação da reprodutibilidade e confiabilidade dos métodos propostos em diferentes condições operacionais.

Este capítulo apresentou o estado da arte de técnicas propostas para a Segmentação Semântica de ingredientes em imagens de alimentos, incluindo métodos fracamente supervisionados baseados em CAMs, abordagens baseadas em Multiple Instances Learning, e modelos totalmente supervisionados. A análise crítica da revisão bibliográfica revelou limitações específicas dos métodos existentes em relação a eficiência computacional, adequação para imagens industriais, granularidade de classes, requisitos de *batch size* e estabilidade entre *folds*. Essas lacunas identificadas justificam o desenvolvimento de uma

abordagem adaptada ao problema específico de segmentação de ingredientes em imagens de montagem de lanches *fast-food*, que será detalhada no próximo capítulo.

Capítulo 4

Materiais e Métodos

Neste capitulo são apresentadas as bases de dados e a metodologia utilizadas nesta pesquisa, destacando a integração das técnicas propostas e sua aplicação para resolver o problema de pesquisa, destacando os procedimentos para alcançar os resultados obtidos.

4.1 Base de dados

As próximas seções apresentam uma descrição das bases de imagens que foram empregadas para o desenvolvimento dos modelos de aprendizado profundo.

4.1.1 *Visio Sandwich Segmentation*

O *dataset Visio Sandwich Segmentation* (VSS) é proprietário e não está publicamente disponível. Foi construído a partir de imagens capturadas durante o processo de montagem de sanduíches submarinos em estabelecimentos de uma rede de *fast-food*. A coleta foi realizada utilizando câmeras de circuito fechado de televisão (CFTV) modelo Hikvision, configuradas com distância focal de 2,8 mm e posicionadas a aproximadamente 50 cm da pista de montagem. Esta configuração foi escolhida para garantir uma cobertura adequada da área de trabalho, permitindo a captura detalhada dos sanduíches durante todas as etapas de preparação, enquanto mantém uma distância segura que não interfere nas operações dos funcionários.

O processo de seleção de frames foi automatizado através de um algoritmo desenvolvido pela empresa Visio.ai, que combina técnicas de detecção e rastreamento de objetos com análise de estabilidade temporal. O sistema identifica momentos de pausa no movimento dos sanduíches (instantes em que o objeto permanece estacionário para receber ingredientes adicionais) através da análise de sinais de posicionamento (coordenadas do

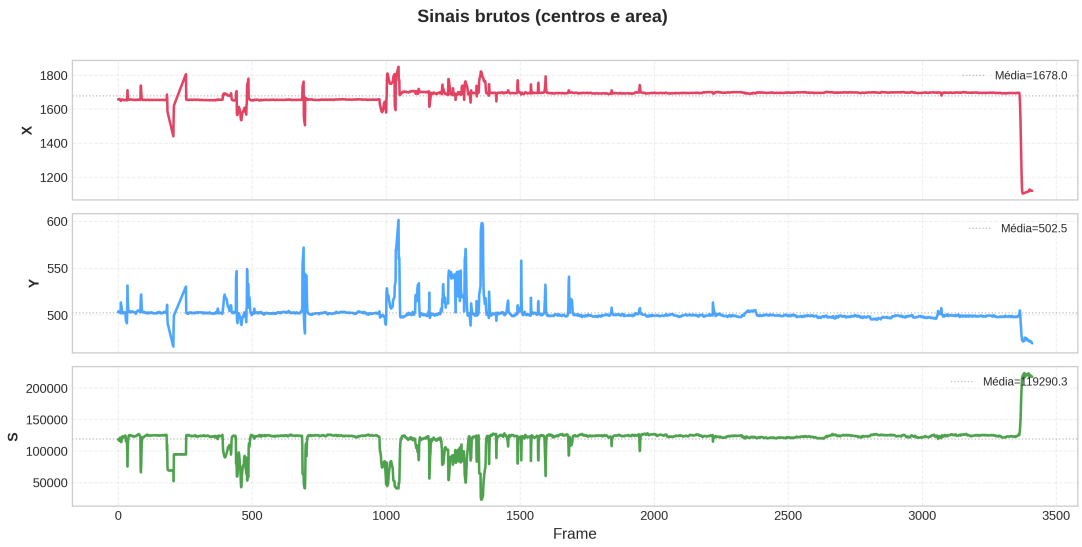

Figura 13 – Sinais brutos de posicionamento (coordenadas X e Y do centroide) e área (S) do *bounding box* ao longo da sequência de vídeo. Cada sinal é apresentado em subplot separado devido à diferença de escala entre os valores.

centroide) e área do *bounding box* ao longo da sequência de vídeo, conforme ilustrado na Figura 13.

Os sinais brutos são suavizados utilizando um filtro gaussiano unidimensional (passa-baixa), cujo parâmetro de suavização é determinado automaticamente por uma busca exaustiva que testa diferentes combinações de parâmetros (sigma: 1–11, threshold: 0,1–3,5), selecionando a configuração que maximiza o número de intervalos estáveis identificados, limitado a menos de 10 intervalos para evitar fragmentação excessiva. Embora o filtro passa-baixa atenue sinais de alta frequência, como o objetivo é identificar momentos de pausa (eventos de baixa frequência), o filtro é adequado ao propósito. O número de 10 intervalos foi definido com base na quantidade média típica de ingredientes por lanche (aproximadamente 10).

A Figura 14 apresenta os resultados desta busca, evidenciando a seleção automática dos parâmetros ótimos. Para cada intervalo estável detectado, o frame com maior área de *bounding box* é selecionado, priorizando instantes em que o sanduíche está mais próximo da câmera e, consequentemente, apresenta maior resolução.

A Figura 15 mostra a máscara binária resultante, indicando quais frames pertencem a intervalos estáveis. Os frames candidatos são então submetidos a filtros de qualidade que descartam imagens com dimensões inferiores a 32×32 pixels, baixa nitidez (medida através do operador Canny com threshold de 8), razão de aspecto inadequada (largura/altura $<$ 0,5) ou oclusão significativa por mãos de operadores (IoU $>$ 0,3 com detecções de pessoas). A Figura 16 apresenta exemplos de imagens rejeitadas por estes critérios, contrastando com as imagens selecionadas mostradas na Figura 17.

Finalmente, duplicatas são removidas através de comparação de *embeddings* extraídos de uma rede ResNet50 pré-treinada, utilizando similaridade cosseno com threshold de

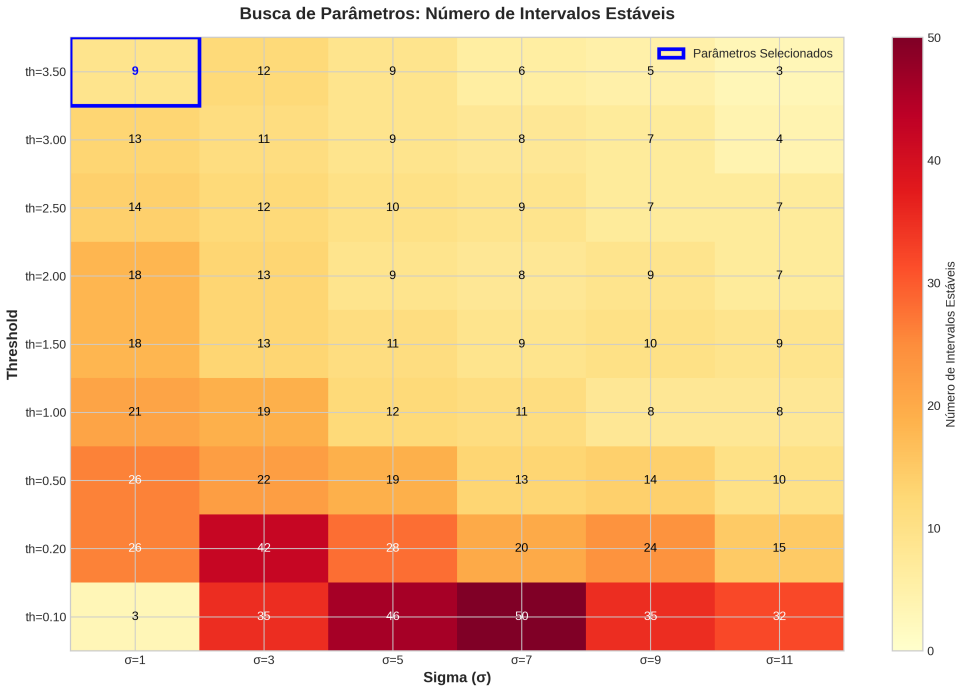

Figura 14 – Heatmap da busca exaustiva de parâmetros mostrando o número de intervalos estáveis encontrados para cada combinação de sigma e threshold. A célula destacada em azul indica os parâmetros selecionados automaticamente.

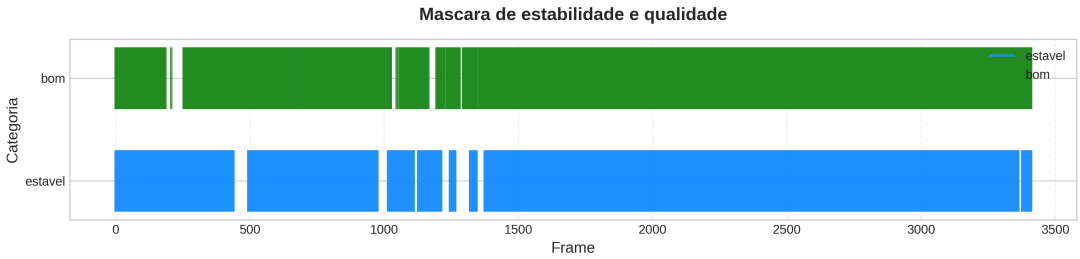

Figura 15 – Máscara binária indicando frames pertencentes a intervalos estáveis (verde) e frames de qualidade adequada (azul). Os intervalos estáveis correspondem a momentos de pausa no movimento dos sanduíches.

(a) Imagens borradas (b) Dimensões insuficientes (c) Frames instáveis

Figura 16 – Exemplos de imagens rejeitadas pelos filtros de qualidade aplicados durante a coleta. Estes exemplos ilustram os critérios de exclusão: baixa nitidez, dimensões inferiores a 32×32 pixels, e frames fora dos intervalos estáveis identificados.

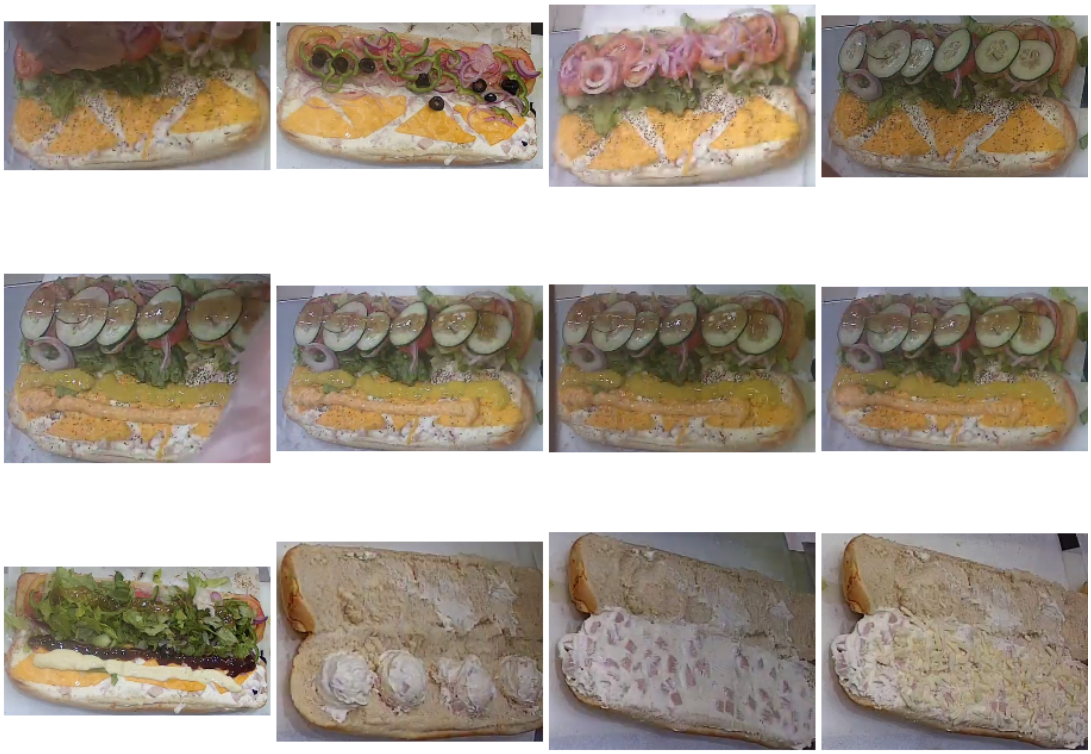

Figura 17 – Amostra das imagens finais selecionadas após todas as etapas de filtragem e deduplicação. As imagens apresentam boa qualidade, nitidez adequada e representam momentos estáveis da montagem dos sanduíches.

0,9. A maior parte das duplicatas é removida automaticamente por esse processo; um conjunto restante foi removido manualmente após inspeção do *dataset*.

A coleta ocorreu em diversos estabelecimentos da rede, em diferentes locais e horários, resultando em condições operacionais variáveis que impactam diretamente as características das imagens coletadas. A ausência de controle uniforme sobre condições de iluminação, decorrente da captura em ambientes reais de produção, introduz variações significativas no brilho médio das imagens (61,26–222,06), refletindo diferentes períodos do dia, configurações de iluminação artificial e posicionamento relativo das fontes de luz. Adicionalmente, devido ao ambiente de coleta, algumas imagens podem apresentar as mãos dos operadores durante a montagem, o que constitui uma característica inerente ao processo de produção e não uma limitação do sistema de captura.

A distância fixa das câmeras (50 cm) e a qualidade dos equipamentos utilizados resultam em variações consideráveis na resolução e proporção das imagens finais, conforme apresentado na Tabela 3. As imagens coletadas apresentam alturas variando entre 105–549 pixels, larguras entre 71–748 pixels, e razões de aspecto entre 0,5–3,03, refletindo tanto variações na distância do sanduíche em relação à câmera quanto diferenças no posicionamento e orientação dos objetos durante a montagem. A quantidade de pixels varia de 8.236 a 300.884, evidenciando a diversidade de escalas presentes no *dataset*. Esta

Métricas	Valor mínimo	Valor máximo
Altura (px)	105	549
Largura (px)	71	748
Proporção (*aspect ratio*) (adimensional)	0,5	3,03
Brilho médio (escala 0–255)	61,26	222,06
Quantidade de pixels (px)	8236	300884

Tabela 3 – Características das imagens do *dataset Visio Sandwich Segmentation*.

variabilidade, embora represente um desafio para algoritmos de segmentação, é intencional e reflete a natureza realista do ambiente de produção, garantindo que o *dataset* seja representativo das condições operacionais encontradas em estabelecimentos da rede.

O *dataset* abrange um total de 17 classes, incluindo ingredientes e elementos contextuais presentes nos sanduíches. As categorias de ingredientes incluem:

a) Vegetais: alface, tomate, cebola, pimentão, pepino, picles e azeitona;

b) Laticínios e molhos: *cheddar*, queijo, *cream cheese* e molho;

c) Carnes: agregadas na meta-classe Carne, incluindo frango e outras proteínas, exceto bacon, que é tratado como classe separada devido à sua aparência, custo e textura distintas;

d) Carnes processadas: salame e presunto agrupados na meta-classe Processados;

e) Pão: representando a base estrutural dos sanduíches;

f) Mão/pessoa: capturando casos onde ocorre interferência humana, útil para avaliar a robustez do modelo contra oclusões e ruído externo.

O *dataset* contém um total de 17.658 imagens coletadas de vídeos de preparação de sanduíches capturados por câmeras CFTV. Dessas imagens, 17.282 possuem anotações em nível de imagem, enquanto 2.214 imagens receberam anotações em nível de pixel para suportar avaliação supervisionada e validação do treinamento fracamente supervisionado. A distribuição das anotações entre as classes está detalhada na Tabela 4. Uma comparação do VSS com outros *datasets* de segmentação de alimentos está apresentada na Tabela 6

O processo de anotação envolveu três anotadores trabalhando simultaneamente sob diretrizes padronizadas de rotulagem. Cada anotação de nível de pixel foi subsequentemente revisada por um anotador diferente para garantir consistência de rótulos e precisão de limites. Para anotações em nível de imagem, foi estabelecido um canal de comunicação compartilhado para esclarecer casos ambíguos envolvendo ingredientes visualmente similares. Além disso, uma amostra aleatória compreendendo aproximadamente 10% das imagens foi revisada manualmente, com ênfase em classes sub-representadas onde o ruído de anotação poderia ter maior impacto no treinamento do modelo. A anotação de 2.214 imagens para segmentação em nível de pixel exigiu aproximadamente 1.305 horas (35 minutos por imagem), enquanto as 17.282 anotações em nível de imagem levaram aproximadamente 192 horas (40 segundos por imagem), representando uma redução de

Classe	Treino (Image-Level)	Validação (Pixel-Level)	Total de Imagens
background	15473	2185	17658
Picles	545	96	641
Pao	15184	2179	17363
Cebola	5151	829	5980
Queijo	14192	2006	16198
Carne	11506	1724	13230
Cheddar	5934	942	6876
Azeitona	3242	529	3771
Pimentao	2311	421	2732
Molho	3922	632	4554
Alface	10163	1602	11765
Mao_Pessoa	5247	725	5972
Cream–cheese	3028	406	3434
Tomate	7706	1290	8996
Processados	1253	175	1428
Bacon	758	97	855
Pepino	3027	620	3647

Tabela 4 – Estatísticas do Dataset VSS

aproximadamente 85% no tempo de anotação manual. Os valores de IoU e acurácia das anotações estão detalhados na Tabela 5.

4.2 Arquitetura do Modelo

A metodologia proposta visa realizar SS fracamente supervisionada para inspeção visual em nível de ingrediente na montagem de sanduíches *fast-food*. O *framework* aborda os desafios inerentes de imagens industriais de baixa qualidade, anotações limitadas em nível de pixel e alta variabilidade intra-classe típica de ambientes de produção reais. A abordagem combina preparação eficiente de dados, geração de pseudo-rótulos e uma arquitetura de treinamento de duplo estudante construída sobre um *backbone* SegFormer-B0. O *pipeline* proposto foi projetado para alcançar robustez contra desequilíbrio de classes, oclusão e degradação de imagem, mantendo eficiência computacional adequada para implantação industrial.

4.2.1 Backbone e Cabeças de Segmentação

O SegFormer-B0 serve como extrator de características devido ao seu codificador baseado em Transformers hierárquico e decodificador *Multilayer Perceptron* (MLP) com menor complexidade computacional. Cada rede estudante é composta por um par codificador-decodificador que produz mapas de probabilidade de classe por pixel. Representações intermediárias são usadas para supervisão cruzada entre estudantes.

Classe	**IoU (%)**	**Acurácia (%)**
Bg	64,95	82,87
Picles	84,21	4,96
Pão	38,09	58,85
Cebola	65,81	67,85
Queijo	38,36	65,92
Carne	37,85	35,60
Cheddar	45,98	9,36
Azeitona	76,31	32,92
Pimentão	57,26	19,06
Molho	13,36	42,95
Alface	51,91	55,07
Mão/Pessoa	61,30	27,47
Cream-cheese	52,94	9,81
Tomate	62,65	72,57
Processados	83,31	45,51
Bacon	79,43	50,42
Pepino	57,58	75,11
Valor médio	57,13	44,49

Tabela 5 – Valores de IoU (%) e Acurácia (%) por classe obtidos em anotações em nível de pixel do conjunto de validação VSS.

Característica	**UECFoodPix Complete**	**FoodSeg-103**	**MyFood**	**VSS**
Total de imagens anotadas em nível de pixel	10000	7118	1250	2214
Número de classes	103	104	9	16
Tipo de anotação	tipos de pratos	ingredientes	ingredientes	ingredientes

Tabela 6 – Comparação de datasets de segmentação de alimentos. Os datasets FoodSeg-103 (WU et al., 2021b) e Food-101 (BOSSARD; GUILLAUMIN; GOOL, 2014) foram utilizados apenas para experimentos acadêmicos de troca de domínio, não sendo utilizados para pré-treino do modelo final.

A arquitetura de duplo estudante, inspirada em DuPL (WU et al., 2024), adapta o paradigma de duplo estudante para *backbones* baseados em Transformers para imagens de alimentos, permitindo supervisão cruzada robusta sob rótulos fracos e melhorando a confiabilidade da segmentação em cenários industriais. Conforme ilustrado na Figura 18, cada rede estudante processa a mesma imagem de entrada e gera mapas de ativação de classe (CAMs), que são refinados progressivamente através de atualizações iterativas de pseudo-rótulos. Este projeto melhora a generalização, mitiga o sobreajuste a supervisão ruidosa e suporta requisitos de inspeção em tempo real.

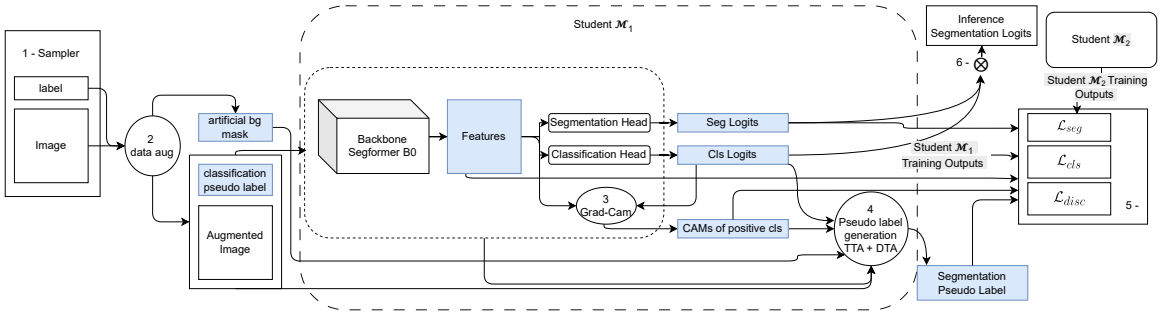

Figura 18 – Visão geral do *pipeline* do modelo proposto. A Etapa 1 aplica um amostrador aleatório ponderado para balancear amostras de treinamento. Na Etapa 2, pré-processamento, aumento e transformações informadas por domínio geram pseudo-rótulos e máscaras de região. O modelo realiza conjuntamente classificação e segmentação usando Grad-CAM, com pseudo-rótulos refinados através de aumento no tempo de teste e limiarização dinâmica. Durante a inferência, máscaras de segmentação são filtradas por saídas de classificação para suprimir classes inativas. Fonte: Próprio autor.

4.2.2 Geração de Mapas de Ativação de Classe

Os CAMs são gerados a partir da camada de normalização (LayerNorm) do quarto estágio do codificador de cada rede e refinados através de aumento no tempo de teste (TTA), que já contempla múltiplas escalas e transformações geométricas. Os mapas de ativação são normalizados e fundidos usando estratégias de ponderação adaptativa para formar pseudo-rótulos preliminares que guiam iterações subsequentes de treinamento.

4.3 Geração de Pseudo-Rótulos

Para permitir refinamento progressivo, o estágio de geração de pseudo-rótulos incorpora múltiplos mecanismos: aumento no tempo de teste (TTA) e ajuste dinâmico de limiar (DTA). Durante cada iteração, pixels com alta confiança de ativação são tratados como pseudo-rótulos confiáveis, enquanto regiões incertas são refinadas via propagação de rótulos suaves.

4.3.1 Aumento no Tempo de Teste (TTA)

O processo de geração de pseudo-rótulos, ilustrado pela Figura 19,

começa com a extração de Mapas de Ativação de Classe (CAMs) dos dois modelos estudante usando Grad-CAM. Para aumentar a robustez dos pseudo-rótulos, foi implementada uma estratégia de Aumento no Tempo de Teste que aplica seis transformações geométricas distintas:

a) espelhamento horizontal;

b) espelhamento vertical;

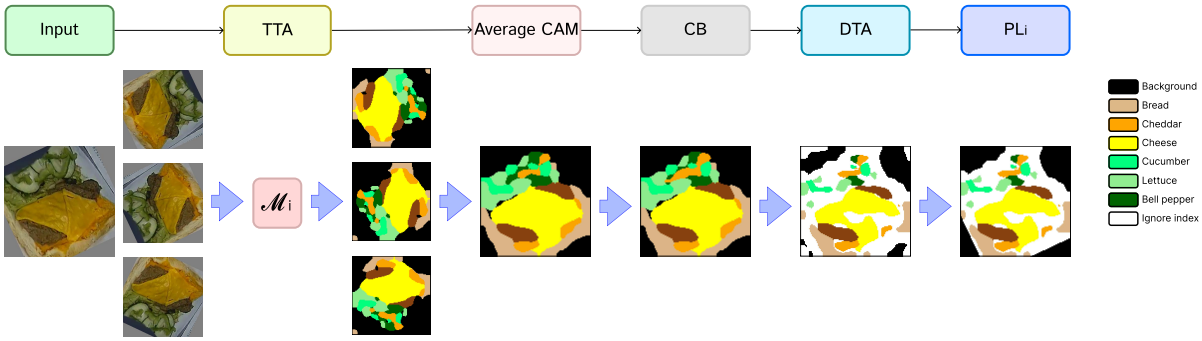

Figura 19 – *Pipeline* passo a passo de geração de pseudo-rótulos mostrando saídas intermediárias para classes de ingredientes alimentares. A combinação de TTA e DTA produz um sinal de pseudo-supervisão estável e reduz o viés de confirmação.

c) escalas de 0,8x e 1,2x;

d) rotações de 90° e 270°.

Cada transformação é processada pelo modelo estudante correspondente, produzindo CAMs que são então revertidos para sua orientação original. Os CAMs resultantes são combinados através de uma média ponderada, onde a imagem original tem peso de 0,4, e as transformações têm pesos empiricamente escolhidos decrescentes (0,15; 0,15; 0,1; 0,1; 0,05; 0,05), priorizando a predição original enquanto incorpora a diversidade das transformações.

Após combinar os CAMs, um algoritmo de balanceamento de classes (CB) garante que cada classe presente na imagem tenha pelo menos uma fração mínima (3%) de pixels no pseudo-rótulo final. Este balanceamento é realizado através de ajustes iterativos de pesos de classe, multiplicando por um fator de 1,05 as classes que não atendem ao limiar mínimo, até que todas as classes satisfaçam o critério ou o número máximo de iterações seja atingido.

4.3.2 Ajuste Dinâmico de Limiar (DTA)

Após as etapas de fusão TTA e balanceamento de classes, um mecanismo de limiarização adaptativa baseado em confiança de predição é aplicado para filtrar ativações de baixa certeza, conforme ilustrado na Figura 19. Este processo, inspirado em DuPL (WU et al., 2024), visa gerar pseudo-rótulos mais confiáveis distinguindo entre regiões certas, incertas e inválidas.

Pixels com confiança abaixo de um limiar baixo (τ_l) são atribuídos à classe de fundo (0). Aqueles com confiança entre τ_l e um limiar alto (τ_h) são considerados incertos e recebem um índice de ignorar (255), excluindo-os de contribuir para a perda de segmentação. Isso previne que o modelo aprenda a partir de predições ruidosas e de baixa confiança.

O limiar alto (τ_h) é atualizado dinamicamente ao longo do treinamento usando um agendamento cosseno que decai de $0,70$ para $0,55$ ao longo das épocas. Esta estratégia

permite que o modelo seja inicialmente conservador em sua geração de pseudo-rótulos, mitigando viés de confirmação nos estágios iniciais. À medida que o treinamento progride e o modelo estabiliza, torna-se progressivamente mais permissivo, encorajando a exploração e incorporação de regiões inicialmente incertas. Isso resulta em refinamento mais suave de pseudo-rótulos e melhora na consistência da segmentação.

Além disso, uma máscara de fundo artificial é aplicada a regiões geradas por operações de aumento de dados (por exemplo, rotação, recorte, ofuscação). Pixels dentro desta máscara são forçados à classe de fundo, garantindo que o modelo não tente segmentar áreas inválidas criadas artificialmente.

Os pseudo-rótulos finais refinados são usados para supervisionar as cabeças de segmentação via perda CrossEntropy. A combinação de TTA e esta limiarização dinâmica fornece um sinal de pseudo-supervisão mais estável e robusto, reduzindo efetivamente o viés de confirmação, uma limitação comum em abordagens WSSS de rede única.

4.4 Funções de Perda e Estratégia de Treinamento

O objetivo total de treinamento combina múltiplos componentes de perda projetados para equilibrar precisão de localização e consistência inter-classe:

a) Perda de Segmentação ($\mathcal{L}_{seg}$): penaliza má classificação no nível de pixel usando supervisão cruzada.

b) Perda de Discrepância ($\mathcal{L}_{disc}$): promove diversidade maximizando desacordo entre representações de características intermediárias.

c) Perda de Classificação ($\mathcal{L}_{cls}$): penaliza má classificação no nível de classificação.

4.4.1 Perda de Classificação

A perda de classificação emprega MultiLabel Soft Margin Loss para classificação multi-rótulo:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \log\left(1 + \exp(-y_{ic} \cdot f_{ic})\right), \tag{6}$$

em que N é o tamanho do *batch*, C é o número de classes, $y_{ic} \in \{0, 1\}$ são os rótulos binários, e f_{ic} são os *logits* de predição.

4.4.2 Perda de Segmentação

A perda de segmentação utiliza supervisão cruzada com pseudo-rótulos gerados pelo modelo estudante par:

$$\begin{aligned} \mathcal{L}_{\text{seg}} = {} & CE\left(\text{Logits}_{\text{seg1}}, \text{PseudoRótulos}_2\right) \\ & + CE\left(\text{Logits}_{\text{seg2}}, \text{PseudoRótulos}_1\right), \end{aligned} \tag{7}$$

em que $CE(\cdot, \cdot)$ denota a perda de entropia cruzada.

4.4.3 Perda de Discrepância

No contexto deste trabalho, as restrições computacionais impostas pelo *hardware* disponível limitaram o tamanho de *batch* a 12 durante o treinamento, conforme detalhado na Seção 4.6. Este tamanho de *batch* é substancialmente inferior ao requerido por métodos de aprendizado contrastivo online como o MulSupCon (ZHANG; WU, 2024), que tipicamente demandam *batch sizes* de 64 ou mais para operação efetiva, conforme evidenciado empiricamente por Chen *et al.* (CHEN et al., 2020) que demonstraram que o aprendizado contrastivo se beneficia significativamente de *batch sizes* maiores. Consequentemente, embora o paradigma de duplo estudante também opere no espaço de representações para garantir representações distintas entre os estudantes, a implementação de aprendizado contrastivo explícito através de métodos como MulSupCon não se mostrou viável sob essas restrições computacionais, uma vez que a limitação de amostras negativas disponíveis em *batches* pequenos comprometeria a estabilidade do treinamento e a qualidade das representações aprendidas.

Esta limitação motivou a adoção de estratégias alternativas. A perda de discrepância foi selecionada como mecanismo que induz representações distintas entre as redes sem a necessidade de *batch sizes* grandes. Esta função de perda promove diversidade de características entre as duas redes estudante através da minimização de sua similaridade cosseno, reduzindo o viés de confirmação e diversificando as representações (*features*) entre as redes; isso permite que o modelo explore o espaço de representações de forma eficiente mesmo sob restrições de memória. A abordagem é complementada por mecanismos de supervisão cruzada entre estudantes, que em conjunto garantem que os estudantes mantenham diversidade suficiente enquanto convergem em regiões de alta confiança, mitigando assim o viés de confirmação e fomentando representações complementares.

A perda de discrepância é definida como:

$$\mathcal{L}_{\text{disc}} = \frac{1}{B} \sum_{b=1}^{B} \left[1 + \text{cos\_sim}\left(F_1^b, F_2^b \right) \right], \tag{8}$$

em que B é o tamanho do *batch*, F_1^b e F_2^b representam os mapas de características achatados das primeira e segunda redes estudante para a b-ésima imagem no *batch*, e $\text{cos\_sim}(\cdot, \cdot)$ denota a função de similaridade cosseno:

$$\text{cos\_sim}(F_1, F_2) = \frac{F_1 \cdot F_2}{\|F_1\| \|F_2\|}. \tag{9}$$

Como a similaridade cosseno varia de -1 (perfeitamente dissimilar) a 1 (perfeitamente similar), minimizar $\mathcal{L}_{\text{disc}}$ maximiza diretamente a dissimilaridade de características entre as duas redes. Esta formulação garante que os estudantes desenvolvam representações de características diversas enquanto evita soluções triviais onde a perda poderia conduzir as características em direção à anti-correlação.

4.4.4 Perda Total e Curriculum Learning

A função de perda composta final é definida como:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{disc}} + \lambda_3 \mathcal{L}_{\text{cls}}, \tag{10}$$

em que λ_1, λ_2 e λ_3 são fatores de ponderação determinados empiricamente.

O treinamento prossegue em uma estratégia de aprendizado curricular com ponderação adaptativa de perda baseada no progresso do treinamento. A perda de discrepância ($\mathcal{L}_{disc}$), definida na Equação 8, é aplicada apenas quando o progresso do treinamento está abaixo de 50%, pois a imposição de diversidade em estágios iniciais ajuda a estabelecer representações de características distintas, enquanto estágios posteriores se beneficiam da convergência em direção a predições consistentes. A perda de classificação recebe ponderação mais alta (20,0) durante estágios iniciais para estabelecer representações de características robustas antes de introduzir complexidade de segmentação, então reduz para 1,0 em estágios posteriores para permitir que o refinamento de segmentação domine o processo de otimização.

O sistema de perda multi-estágio implementa ativação progressiva de diferentes componentes de perda conforme mostrado na Tabela 9: alta ponderação inicial em $\mathcal{L}_{cls}$ (20,0) garante aprendizado robusto de características a partir dos rótulos fracos antes de confiar na tarefa de segmentação. A redução subsequente permite que o refinamento de segmentação domine, com uma ponderação menor (0,3) para prevenir que pseudo-rótulos ruidosos desestabilizem os estágios posteriores do treinamento.

4.5 Amostragem e Aumento de Dados

Uma estratégia de amostragem balanceada por classe é empregada para mitigar os efeitos de severo desequilíbrio de classes. Classes minoritárias (por exemplo, Picles e Bacon) são super-amostradas através de repetição controlada e recorte sintético. O aumento de dados inclui:

a) transformações geométricas: rotação, espelhamento e escalonamento;

b) ajustes fotométricos: brilho, contraste e saturação;

c) simulação de oclusão;

d) crops de domínio: recortes centrais focados na região do lanche, permitindo ao modelo aprender características específicas de cada ingrediente. Por exemplo, ao desabilitar temporariamente a classe “pão” durante o treinamento com crops centrais, o modelo aprende a identificar melhor as bordas e características do pão em contextos variados.

A estratégia de crops de domínio e a inclusão de amostras negativas (imagens sem determinadas classes) são particularmente importantes para o aprendizado, pois forçam o

Hiperparâmetro	Valor
Épocas	20
Tamanho do *batch*	12
Tamanho da imagem	448×448 pixels
Amostras por época	3000
Número de classes	17
Semente aleatória	42

Tabela 7 – Hiperparâmetros gerais de treinamento.

Parâmetro	*Backbone*	Cabeças
Taxa de aprendizado	$4,1 \times 10^{-5}$	$2,6 \times 10^{-2}$
Decaimento de peso	$4,3 \times 10^{-5}$	$1,7 \times 10^{-5}$
Parâmetros do otimizador AdamW: $\beta_1 = 0,9$ e $\beta_2 = 0,999$		

Tabela 8 – Parâmetros do otimizador AdamW para diferentes grupos de parâmetros.

modelo a distinguir entre presença e ausência de ingredientes, reduzindo falsos positivos e melhorando a robustez da classificação. Estas técnicas melhoram a robustez do modelo sob condições do mundo real.

A normalização é realizada no intervalo [0, 1], convertendo os valores de *pixels* do intervalo [0, 255] para [0, 1], e normalização baseada em média e desvio padrão, ajustando os pixels da imagem para que sigam médias e desvios padrão definidos. No caso, a maioria dos modelos utiliza os valores do *dataset ImageNet* (DENG et al., 2009), que possui médias de [0,485; 0,456; 0,406], e desvio padrão de [0,229; 0,224; 0,225], respectivamente, para os canais RGB.

4.6 Detalhes de Implementação

O treinamento foi conduzido em uma GPU NVIDIA GeForce RTX 5090 usando o *framework* PyTorch. Os principais hiperparâmetros estão detalhados nas Tabelas 7 e 8.

Um agendador do tipo *Cosine Annealing Warm Restarts* foi empregado para decaimento da taxa de aprendizado ao longo do treinamento. O *backbone* SegFormer-B0 foi pré-treinado no ImageNet (WIGHTMAN, 2021) e ajustado para a tarefa específica. O sistema de perda multi-estágio usa pesos adaptativos que variam ao longo do treinamento, implementando uma estratégia de aprendizado curricular que ativa diferentes componentes de perda em estágios específicos.

Os hiperparâmetros apresentados neste trabalho foram sistematicamente otimizados usando otimização bayesiana através do Optuna, empregando o amostrador Tree-structured Parzen Estimator (TPE) para exploração eficiente do espaço de hiperparâmetros. Uma busca abrangente foi conduzida sobre 14 hiperparâmetros críticos do modelo proposto, incluindo taxas de aprendizado para diferentes componentes do modelo (*backbone*, cabeças),

parâmetros de decaimento de peso e configurações de agendador de treinamento.

Estágio de Treinamento	**Componentes de Perda**	**Pesos**
Estágio 1 (0-30%)	$\mathcal{L}_{cls}$	20,0
	$\mathcal{L}_{disc}$	1,0
Estágio 2 (30-50%)	$\mathcal{L}_{cls}$	20,0
	$\mathcal{L}_{seg}$	10,0
	$\mathcal{L}_{disc}$	1,0
Estágio 3 (50-100%)	$\mathcal{L}_{cls}$	1,0
	$\mathcal{L}_{seg}$	0,3

Tabela 9 – Sistema de Perda Multi-Estágio com Aprendizado Curricular

4.7 Experimento de Troca de Domínio

Além do experimento principal realizado no *dataset* VSS, foi conduzido um experimento de troca de domínio utilizando os *datasets* públicos Food-101 (BOSSARD; GUILLAUMIN; GOOL, 2014) e FoodSeg-103 (WU et al., 2021b). Este experimento teve finalidade puramente acadêmica e não foi utilizado para pré-treino do modelo final apresentado neste trabalho. O objetivo foi avaliar a capacidade de transferência de conhecimento entre domínios de alimentos através de um protocolo de treinamento no FoodSeg-103 utilizando pesos pré-treinados no ImageNet (WIGHTMAN, 2021).

Este experimento seguiu um protocolo similar ao utilizado por Cai *et al.* (CAI; ABHAYARATNE, 2023) no modelo SSDB-II, que reportou um mIoU de 14,79% no FoodSeg-103 utilizando pré-treino no *dataset* Food-101. No entanto, diferentemente do SSDB-II, que utiliza pré-treino no Food-101, este experimento investigou especificamente a transferência de conhecimento a partir de pesos pré-treinados no ImageNet (WIGHTMAN, 2021), permitindo uma análise do impacto do pré-treino em domínios genéricos versus domínios específicos de alimentos.

Uma observação interessante do estudo de Cai *et al.* (CAI; ABHAYARATNE, 2023) é o desempenho notavelmente alto da classe *seaweed* no FoodSeg-103, que alcançou 44,59% de IoU (5º lugar) apesar de possuir apenas 7 amostras no conjunto de treinamento, conforme demonstrado nas Tabelas 10 e 11. Este resultado é particularmente surpreendente quando comparado com outras classes do top 10 que possuem centenas de amostras. Uma hipótese plausível é que os autores tenham realizado mapeamento direto desta classe do Food-101 para o FoodSeg-103 durante o pré-treino, permitindo transferência de conhecimento que compensa a escassez de dados de treinamento.

O experimento utilizou a divisão padrão do FoodSeg-103, com 4.983 imagens para treinamento e 2.135 para validação, conforme estabelecido pelos criadores do *dataset* (WU et al., 2021b). Diferentemente do experimento principal no VSS, que utiliza validação cru-

zada com quatro *folds*, este experimento seguiu o protocolo padrão do *dataset* utilizando apenas a divisão treino/validação, garantindo comparabilidade com outros trabalhos que utilizam o mesmo conjunto de dados.

Classe	**Número de amostras (treino)**
background	4.982
bread	991
carrot	881
chicken duck	848
sauce	818
tomato	790
potato	785
steak	728
broccoli	704
ice cream	636
seaweed	7

Tabela 10 – Distribuição de amostras por classe no conjunto de treinamento do FoodSeg-103. As 10 classes com maior número de amostras são apresentadas, juntamente com a classe *seaweed* para comparação.

Ranking	**Classe**	**IoU SSDB-II (%)**	**Número de amostras (treino)**
1	background	64,50	4.982
2	corn	51,81	343
3	green beans	51,20	152
4	broccoli	50,94	704
5	seaweed	44,59	7
6	noodles	44,37	187
7	strawberry	43,98	388
8	rice	43,40	464
9	asparagus	40,90	187
10	French beans	37,86	253

Tabela 11 – Comparação entre as 10 classes com melhor desempenho (maior IoU) no SSDB-II (CAI; ABHAYARATNE, 2023) e suas respectivas quantidades de amostras no conjunto de treinamento do FoodSeg-103. A classe *seaweed* é destacada por apresentar alto desempenho (44,59% IoU, 5º lugar) apesar de possuir apenas 7 amostras, sugerindo possível transferência de conhecimento do Food-101 através de mapeamento direto durante o pré-treino combinado.

Os resultados quantitativos deste experimento (mIoU do modelo proposto e comparação com SEAM e SSDB) são apresentados no Capítulo 5, na Seção 5.10.

4.8 Métricas de Avaliação

Para a avaliação do modelo proposto foram utilizadas as seguintes métricas principais:

4.8.1 Intersection over Union (IoU) e mean IoU (mIoU)

O IoU mede a sobreposição entre a predição e o valor verdadeiro para cada classe c:

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{11}$$

, em que TP (True Positive) representa pixels corretamente preditos como classe c, FP (False Positive) representa pixels incorretamente preditos como classe c, e FN (False Negative) representa pixels da classe c não detectados.

O mIoU é a média do IoU sobre todas as classes:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c, \tag{12}$$

em que C é o número total de classes (incluindo *background*).

4.8.2 Precisão e Revocação

A precisão mede a proporção de pixels corretamente preditos dentre todos os pixels preditos como classe c:

$$\text{Precisão}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}. \tag{13}$$

A revocação mede a capacidade do modelo de identificar todos os pixels de uma classe c:

$$\text{Revocação}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}. \tag{14}$$

4.8.3 F1-Score

O F1-Score é a média harmônica entre precisão e revocação:

$$\text{F1}_c = 2 \cdot \frac{\text{Precisão}_c \cdot \text{Revocação}_c}{\text{Precisão}_c + \text{Revocação}_c}. \tag{15}$$

Estas métricas são calculadas por classe e também em média para avaliar o desempenho geral do modelo, conforme detalhado no Capítulo 5.

Este capítulo apresentou as bases de dados e a metodologia utilizadas nesta pesquisa, destacando a integração das técnicas propostas e sua aplicação para resolver o problema de pesquisa. Foram descritos o *dataset* Visio Sandwich Segmentation, os procedimentos de preparação de dados, a arquitetura do modelo baseada em SegFormer-B0 com paradigma de duplo estudante, as estratégias de geração de pseudo-rótulos, as funções de perda e a estratégia de treinamento. Os detalhes de implementação e as métricas de avaliação apresentados aqui fornecem a base metodológica para a compreensão dos resultados experimentais apresentados no próximo capítulo.

Capítulo 5

Metodologia Experimental e Resultados

Neste capítulo são apresentados os resultados experimentais obtidos com o modelo proposto, incluindo comparações com métodos do estado da arte, estudos de ablação, análise de simplificação de componentes, análise de sensibilidade de hiperparâmetros, análises quantitativas e qualitativas, avaliação de eficiência computacional e custo-desempenho, e resultados no dataset *FoodSeg103.*

5.1 Configuração Experimental

A avaliação experimental apresentada neste capítulo visa demonstrar a adequação do modelo proposto para aplicações industriais, considerando não apenas métricas de desempenho absolutas, mas também estabilidade, eficiência computacional e viabilidade de implantação. Diferentemente de avaliações que buscam maximizar métricas isoladas em cenários controlados, os experimentos aqui conduzidos refletem restrições práticas de ambientes industriais, onde robustez e previsibilidade operacional são tão importantes quanto precisão. Esta abordagem justifica a escolha de métricas como mIoU e IoU por classe, que capturam diferentes aspectos do desempenho de segmentação, e o emprego de validação cruzada com quatro *folds*, priorizando robustez e generalização sobre desempenho pontual em um único *fold*.

Os experimentos foram conduzidos utilizando uma GPU NVIDIA GeForce RTX 5090 com o *framework* PyTorch. A estratégia de validação cruzada foi empregada com quatro *folds*, garantindo uma avaliação robusta do desempenho do modelo. Os hiperparâmetros utilizados foram sistematicamente otimizados através de otimização bayesiana usando Optuna, conforme detalhado no Capítulo 4.

O modelo proposto foi comparado com métodos do estado da arte, incluindo DuPL (WU et al., 2024), que representa uma abordagem recente e de alto desempenho em WSSS de duplo estudante, fornecendo um *benchmark* forte para a simplificação focada em aplicações industriais proposta neste trabalho. Além disso, foi realizada uma comparação com o SegFormer-B0 supervisionado para estabelecer uma referência de desempenho superior.

5.2 Comparação com Métodos do Estado da Arte

A comparação com métodos do estado da arte visa avaliar o posicionamento do modelo proposto em relação a abordagens recentes e de alto desempenho, bem como estabelecer uma referência de desempenho superior através de métodos totalmente supervisionados. Nesta análise, é fundamental observar não apenas os valores médios de mIoU, mas também a variabilidade entre *folds* de validação cruzada, que indica robustez e previsibilidade operacional. Além disso, deve-se considerar que métodos comparados podem apresentar custos ocultos em termos de consumo de memória (VRAM), requisitos de tamanho de *batch* para operação efetiva, e custos de anotação, aspectos críticos para aplicações industriais que frequentemente não são capturados por métricas de desempenho isoladas.

A Tabela 12 apresenta o desempenho médio do modelo fracamente supervisionado proposto, DuPL, SEAM e SegFormer-B0 supervisionado através de quatro *folds*. Enquanto DuPL representa uma abordagem recente e de alto desempenho em WSSS de duplo estudante, SEAM serve como uma linha de base pioneira para avaliar o impacto da evolução arquitetural de métodos baseados em CNN clássicos para designs modernos baseados em Transformers no domínio de alimentos (WANG et al., 2020; CAI; ABHAYARATNE, 2023; WU et al., 2021b). Embora não maximize mIoU absoluto, o modelo proposto apresenta maior estabilidade entre *folds*, característica crítica em cenários industriais. Os resultados revelam que o método proposto supera DuPL em 2,7 pontos percentuais e supera significativamente SEAM, que alcançou um CAM mIoU de $28,3 \pm 3,8\%$ (WU et al., 2024; WANG et al., 2020). Quando treinado sob as mesmas condições que os outros modelos, SEAM falhou em aprender características de classificação efetivas, permanecendo abaixo de 50% de precisão média (mean Average Precision, mAP). Obter resultados estáveis foi viável apenas após pré-treinar um *backbone* ResNet-38 (WU; SHEN; HENGEL, 2019) por 200 épocas especificamente no *dataset* VSS na tarefa de classificação e usar esses pesos para inicialização. Esta lacuna de desempenho é esperada, pois SEAM é um método WSSS pioneiro utilizando um *backbone* CNN clássico que converge significativamente mais lento que a arquitetura baseada em Transformers SegFormer (XIE et al., 2021) e carece da modelagem de contexto global avançada encontrada em arquiteturas baseadas em Transformers (WANG et al., 2020; WU; SHEN; HENGEL, 2019). O método proposto mantém um desvio padrão pequeno e consistente entre *folds*, enquanto DuPL e SEAM exibem maior variabilidade, permitindo implantação mais confiável em ambientes industriais (WU et

al., 2024; WANG et al., 2020).

Além disso, o método proposto apresenta desempenho competitivo em comparação com sua contraparte totalmente supervisionada, alcançando 89,2% do mIoU do modelo supervisionado (43,9% versus 49,2% mIoU). Esta proximidade de desempenho, combinada com a redução significativa de custo de anotação discutida na Seção 5.9, demonstra a eficácia da abordagem fracamente supervisionada proposta. É importante destacar que, embora o DuPL apresente CAM IoU superior, o modelo proposto é comparável ao DuPL em segmentação final, indicando que a qualidade dos pseudo-rótulos gerados é adequada para treinar um módulo de segmentação efetivo, mesmo partindo de CAMs com desempenho inferior.

Modelo	**CAM IoU (média)**	**Segmentação IoU (média)**
Métodos WSSS		
SEAM	$28,3 \pm 3,8$	—
DuPL	$43,0 \pm 5,8$	$41,2 \pm 6,8$
Modelo proposto	$34,4 \pm 1,4$	$\mathbf{43,9 \pm 0,6}$
Métodos totalmente supervisionados		
SegFormer-B0	—	$49,2 \pm 0,5$

Tabela 12 – Comparação de mIoU médio obtido por CAMs e segmentação completa. Valores são percentuais. Fonte: Próprio autor.

Para avaliar significância estatística, foram conduzidos um teste t de duas amostras e um teste de postos sinalizados de Wilcoxon. Conforme a Tabela 13, os p-valores obtidos (0,4834 no teste t e 0,6250 no Wilcoxon) estão acima do limiar usual de significância (0,05), de modo que os testes não rejeitam a hipótese nula de igualdade entre o modelo proposto e o DuPL. Dado o pequeno tamanho da amostra ($n = 4$ *folds*), o poder estatístico é limitado: com tão poucas observações, tanto o teste paramétrico quanto o não paramétrico têm baixa capacidade de detectar diferenças reais entre as médias (ou postos), mesmo que existissem. Em cenários com mais *folds* ou repetições, amostras maiores confeririam maior poder para detectar diferenças menores com significância estatística; aqui, a conclusão permanece conservadora: os resultados são interpretados como *consistentes com* desempenho comparável, sem afirmar equivalência estatística forte.

Teste	**Estatística**	**p-valor**
Teste t de duas amostras	-0,7963	0,4834
Wilcoxon postos sinalizados	3,0000	0,6250

Tabela 13 – Comparação estatística do modelo proposto versus DuPL através dos *folds*. Fonte: Próprio autor.

5.3 Estudos de Ablação

A Tabela 14 apresenta um estudo de ablação dos componentes do modelo, permitindo identificar quais elementos são essenciais para o funcionamento do método e quais contribuem de forma incremental. Partindo de uma linha de base fracamente supervisionada que utiliza apenas $\mathcal{L}_{cls}$ para classificação, cada adição produz um aumento monotônico tanto no CAM IoU quanto no IoU de segmentação, resultando em ganhos de 10,3 e 42,4 pontos percentuais sobre a linha de base, respectivamente.

A introdução de $\mathcal{L}_{seg}$ representa um componente essencial, pois adiciona supervisão direta para a tarefa de segmentação, permitindo que o modelo aprenda representações espaciais além da classificação. A remoção deste componente impacta negativamente tanto a qualidade dos CAMs quanto o desempenho de segmentação, indicando dependência crítica do modelo em supervisão explícita para segmentação. Os resultados mostram que esta adição não apenas melhora a qualidade dos CAMs, mas também leva a um desempenho de segmentação que supera os próprios CAMs, demonstrando que o módulo de segmentação aprende representações complementares às ativações de classificação.

A perda de discrepância ($\mathcal{L}_{disc}$) contribui para a estabilidade do treinamento, reduzindo a variância entre os *folds* de 2,3% para 1,3% em CAM IoU. Embora o ganho absoluto seja incremental (0,4 pontos percentuais em CAM IoU), sua remoção mostra dependência de mecanismos de regularização para garantir robustez em arquiteturas de duplo estudante, onde a diversidade entre redes estudante deve ser balanceada com convergência em regiões de alta confiança.

O aumento no tempo de teste (TTA) produz o maior salto em desempenho de segmentação (de 30,4% para 39,7%), indicando que este componente é essencial para a qualidade dos pseudo-rótulos gerados. A remoção de TTA impacta negativamente a geração de pseudo-rótulos, pois o modelo perde a capacidade de impor consenso entre múltiplas visões da mesma imagem, resultando em pseudo-rótulos mais ruidosos e menos confiáveis. Este componente aproxima um *ensemble* sobre perturbações geométricas e fotométricas, reduzindo a variância de predição e impondo consenso de limites entre visões.

O Amostrador Ponderado contribui para estabilização do treinamento e melhoria de predições de classes raras, lidando com desequilíbrio de classes no cenário multi-rótulo. Embora o ganho seja incremental em termos de mIoU médio, sua remoção impacta negativamente classes minoritárias, conforme evidenciado na análise por classe (Seção 5.7), onde classes raras apresentam desempenho inferior sem este componente.

Em conjunto, esses resultados indicam que cada módulo aborda um modo de falha distinto, resultando em melhorias complementares e um modelo mais estável, com menor variância sob mudanças de distribuição.

Configuração	**CAM IoU (%)**	**Segmentação IoU (%)**
Linha de base ($\mathcal{L}_{cls}$ apenas)	$24,1 \pm 1,9$	–
+ $\mathcal{L}_{seg}$	$28,1 \pm 2,3$	$29,7 \pm 3,1$
+ $\mathcal{L}_{disc}$	$28,5 \pm 1,3$	$30,4 \pm 2,1$
+ Aumento no tempo de teste	$29,8 \pm 2,2$	$39,7 \pm 1,6$
+ Aumento de dados de domínio	$33,5 \pm 2,7$	$42,5 \pm 1,6$
+ Amostrador ponderado	$\mathbf{34,4 \pm 1,4}$	$\mathbf{43,9 \pm 0,6}$

Tabela 14 – Estudo de ablação dos componentes do modelo. Resultados são reportados como média ± desvio padrão através dos *folds*.

5.3.1 Análise de Simplificação de Componentes

Para avaliar a complexidade do *pipeline* e a contribuição de simplificações, foram testadas três variantes: (1) Aumento no Tempo de Teste (TTA) simples (3 transformações em vez de 7); (2) sem balanceamento de classes (CB); (3) limiar dinâmico com decaimento linear em vez de cosseno (Ajuste Dinâmico de Limiar, DTA).

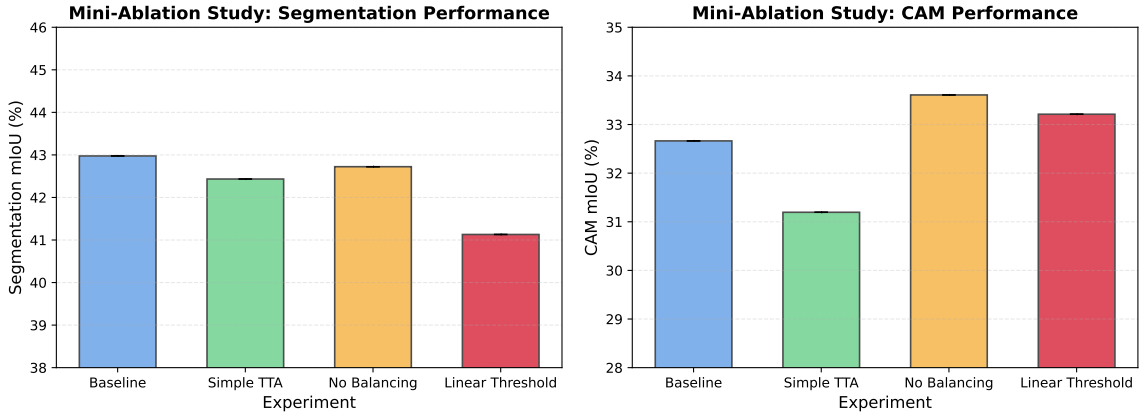

Figura 20 – Comparação da configuração baseline com variantes simplificadas. Esquerda: mIoU de segmentação. Direita: mIoU de CAM.

A configuração baseline (completa) alcançou 43,0% de mIoU de segmentação e 32,7% de mIoU de CAM, estabelecendo a referência para comparação. A variante TTA simples alcançou 42,4% de mIoU de segmentação (−0,6 p.p.); a redução de custo (cerca de 57% menos forward passes) pode justificar essa simplificação em cenários com restrição de recursos. A variante sem balanceamento alcançou 42,7% (−0,3 p.p.); o impacto é pequeno quando o amostrador ponderado já está presente. A variante com limiar linear alcançou 41,1% (−1,9 p.p.), representando a maior queda; o agendamento cosseno deve ser mantido. Em conjunto, os resultados indicam hierarquia de importância (cosseno, TTA, balanceamento) e que o método permite simplificações seletivas conforme restrições de implantação.

5.4 Análise de Sensibilidade de Hiperparâmetros

Foi realizada uma análise de sensibilidade de quatro hiperparâmetros críticos (τ_l, τ_{h0}, τ_{hT}, λ_{dis}) em quatro *folds*, para avaliar robustez e adequação a implantação industrial.

5.4.0.1 Limiar baixo (τ_l)

O limiar τ_l controla a confiança mínima para um pixel ser atribuído a uma classe de primeiro plano em vez de fundo. Valores testados: 0,01; 0,05; 0,1; 0,3. O melhor desempenho foi obtido com $\tau_l = 0{,}1$ (44,75% mIoU ± 0,5%); o pior com $\tau_l = 0{,}01$ (43,5%). O valor padrão 0,05 alcançou 44,25%.

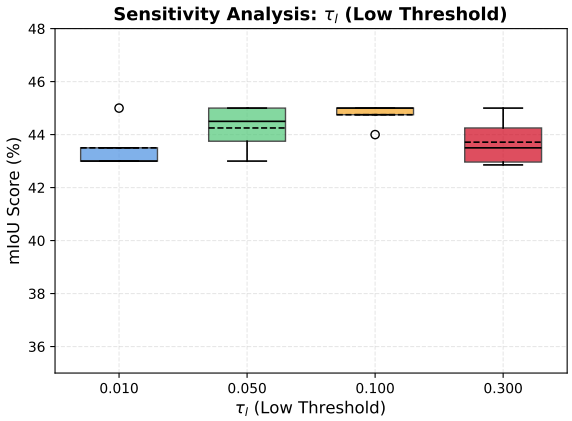

Figura 21 – Sensibilidade ao parâmetro τ_l (limiar baixo). Distribuição do mIoU de segmentação nos *folds*.

5.4.0.2 Limiar alto inicial (τ_{h0})

O parâmetro τ_{h0} define o ponto inicial do agendamento de Ajuste Dinâmico de Limiar (DTA). Valores testados: 0,5; 0,6; 0,7; 0,9. O melhor desempenho foi obtido com $\tau_{h0} = 0{,}7$ (45,0% mIoU ± 0,82%).

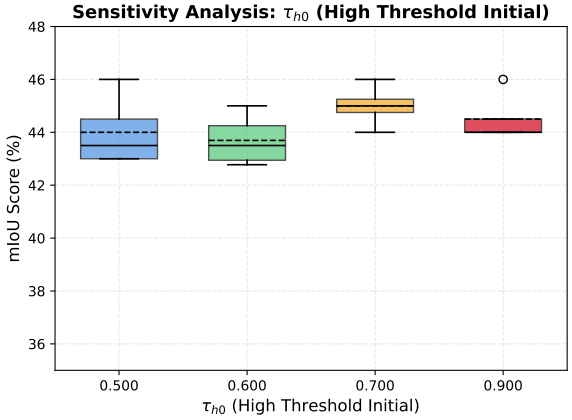

Figura 22 – Sensibilidade ao parâmetro τ_{h0} (limiar alto inicial).

5.4.0.3 Limiar alto final (τ_{hT})

O parâmetro τ_{hT} define o ponto final do agendamento DTA. Valores testados: 0,35; 0,4; 0,55; 0,7. O melhor desempenho foi obtido com $\tau_{hT} = 0{,}4$ (44,33% mIoU ± 1,53%). O valor $\tau_{hT} = 0{,}55$ apresentou o pior desempenho (36,71%, dp 16,78%); esse valor mostrou-se instável entre *folds*.

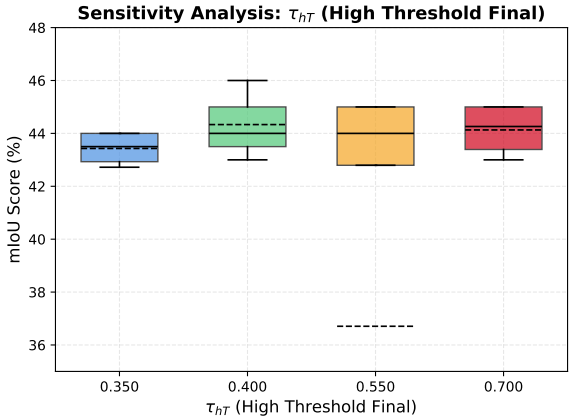

Figura 23 – Sensibilidade ao parâmetro τ_{hT} (limiar alto final).

5.4.0.4 Peso da perda de discrepância (λ_{dis})

O parâmetro λ_{dis} controla a força da regularização de diversidade entre as redes estudante. Valores testados: 0; 0,1; 0,25; 0,5. O melhor desempenho foi obtido com $\lambda_{\text{dis}} = 0$ (44,62% mIoU ± 0,48%); o valor padrão 0,1 ficou próximo (44,52%). A sensibilidade a esse parâmetro foi a mais baixa entre os testados.

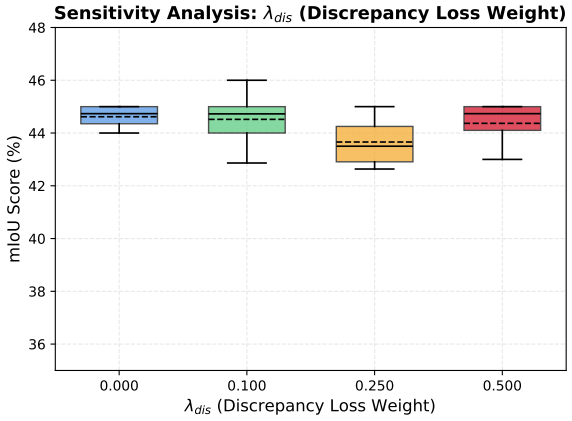

Figura 24 – Sensibilidade ao peso da perda de discrepância λ_{dis}.

Em conjunto, a análise de sensibilidade indica sensibilidade baixa a moderada na maioria dos parâmetros; τ_{hT} é o mais sensível. O método mostra-se robusto para implantação com pouca re-otimização.

5.5 Eficiência Computacional

Em aplicações industriais, a eficiência computacional frequentemente importa mais que pequenos ganhos marginais em métricas de desempenho, pois determina viabilidade de implantação, custos operacionais e escalabilidade. Modelos que requerem alto consumo de VRAM implicam custos significativos em *hardware* especializado, limitando a capacidade de implantação em dispositivos acessíveis ou processamento paralelo de múltiplas instâncias. Da mesma forma, *throughput* de inferência determina diretamente a capacidade de processamento em tempo real, essencial para aplicações de controle de qualidade onde cada frame deve ser processado dentro de janelas temporais restritas. Neste contexto, um modelo que oferece desempenho comparável com menor custo computacional representa uma solução superior para ambientes industriais, mesmo que não maximize métricas absolutas em cenários controlados.

No cenário industrial, esta aplicação depende de alto *throughput* sob restrições de recursos. Para quantificar isso, foi realizado um *benchmark* de *throughput* de inferência e pico de VRAM em uma GPU NVIDIA GeForce RTX 5090 através de tamanhos de *batch* $\{1, 2, 4, 8\}$.

A Tabela 15 mostra que o modelo proposto usa aproximadamente 33% menos VRAM e apresenta *throughput* (vazão) de 1,5 a 5 vezes superior ao DuPL, consistentemente entre qualquer tamanho de *batch* escolhido. Em termos práticos, a redução de VRAM de 848,5 MB para 185,9 MB (*batch*=1) permite implantação em *hardware* mais acessível ou processamento paralelo de múltiplas instâncias, reduzindo custos de infraestrutura. O *throughput* superior, alcançando mais de 1.100 imagens por segundo com *batch*=8, garante processamento em tempo real mesmo sob condições de alta carga, essencial para aplicações de inspeção contínua em linhas de produção. Esses ganhos são atribuídos ao projeto arquitetural, que emprega um *backbone* menor e um conjunto de componentes leves, resultando em menos custo computacional mantendo desempenho comparável ao DuPL, conforme demonstrado na Seção 5.2. Enquanto SEAM (WANG et al., 2020) mostra um *throughput* inicial ligeiramente superior ao DuPL com tamanho de *batch* igual a 1 (80,7 versus 59,7 imagens/s), ele falha em escalar efetivamente com *batches* maiores. Especificamente, o *throughput* de SEAM estabiliza em aproximadamente 83-85 imagens/s, enquanto o modelo proposto alcança um pico de 1.110,9 imagens/s com tamanho de *batch* igual a 8. Isso indica que o *design* com menor custo computacional baseado em SegFormer apresenta *throughput* mais de 13 vezes superior ao SEAM em cenários de alto *throughput*. Além disso, SEAM exibe o maior consumo de VRAM com tamanho de *batch* igual a 8 (2.273,5 MB), que é quase o dobro da pegada de memória do modelo proposto (1.191,2 MB). O gargalo de desempenho observado em SEAM pode ser atribuído principalmente à sua dependência de atenção equivariante auto-supervisionada e ao Módulo de Correlação de Pixel (PCM). Notavelmente, o PCM utiliza um mecanismo de auto-atenção para refinar CAMs calculando similaridade inter-pixel através de todo o mapa de características

(WANG et al., 2020), resultando em complexidade computacional quadrática relativa às dimensões espaciais. Esta complexidade inerente introduz sobrecarga significativa e limita escalabilidade durante a inferência. Em contraste, a arquitetura de duplo estudante simplificada proposta alcança um equilíbrio superior entre velocidade e eficiência de memória, tornando-a o candidato mais adequado para implantação industrial em tempo real.

Modelo	**Tamanho *batch***	***Batch*/s**	**Imagens/s**	**VRAM (MB)**
DuPL	1	59,7	59,7	848,5
	2	80,4	160,7	981,2
	4	86,4	345,5	1.247,6
	8	94,7	757,2	1.781,2
Proposto	1	293,7	293,7	185,9
	2	307,1	614,2	327,9
	4	225,6	902,2	620,1
	8	138,9	1.110,9	1.191,2
SEAM	1	80,7	80,7	690,2
	2	42,8	85,6	891,5
	4	21,4	85,6	1.325,9
	8	10,4	83,3	2.273,5

Tabela 15 – *Throughput* de inferência e consumo de VRAM para DuPL, SEAM e o modelo proposto. Fonte: Próprio autor.

5.6 Análise Qualitativa

A Figura 25 mostra resultados representativos de segmentação para o modelo proposto, DuPL e SegFormer supervisionado. Nesta análise qualitativa, é fundamental observar o *trade-off* fundamental entre completude de objetos e nitidez de bordas, que reflete decisões de *design* orientadas para aplicações industriais. Observe que o modelo proposto prioriza detecção de presença/ausência e localização aproximada de ingredientes, características mais relevantes para inspeção visual em linhas de produção do que precisão absoluta de contornos.

Nota-se que o modelo proposto produz segmentação coerente, alcançando maior completude de objetos que o DuPL, enquanto classes pequenas ou translúcidas (por exemplo, picles, cebolas) permanecem desafiadoras. Este comportamento reflete explicitamente o *trade-off* entre nitidez de bordas e completude de região discutido na Seção 6.3, com o método proposto favorecendo detecção de presença/ausência relevante para aplicações industriais, onde identificar corretamente quais ingredientes estão presentes é frequentemente mais valioso que segmentação precisa de limites. Esta escolha de *design* é intencional e adequada para o contexto de aplicação, conforme justificado nas limitações assumidas (Seção 6.2).

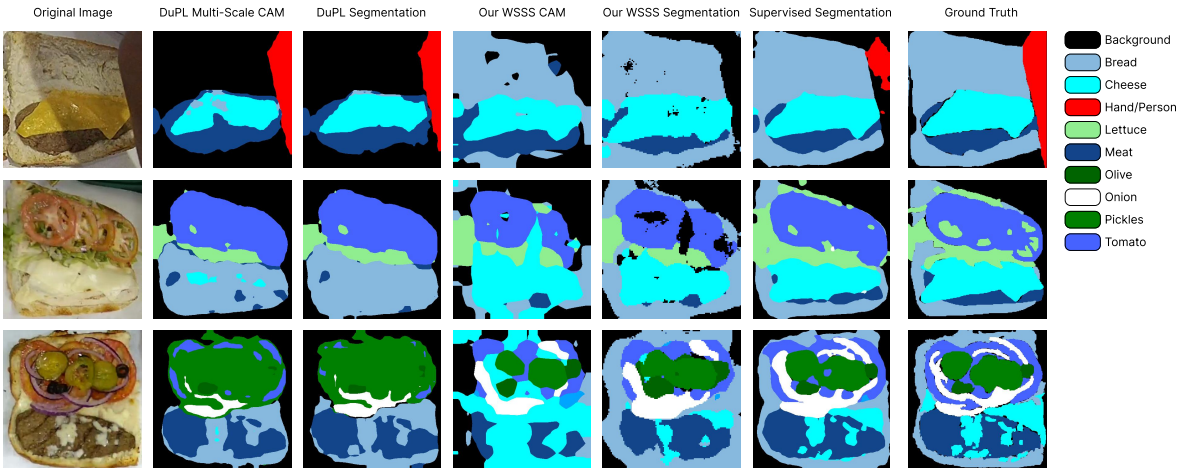

Figura 25 – Comparação qualitativa de saídas de segmentação. O modelo proposto exibe maior completude de objetos, enquanto o DuPL preserva limites mais nítidos. O SegFormer supervisionado fornece uma referência de desempenho superior.

Conforme mostrado na primeira imagem da Figura 25, ambos os modelos fracamente supervisionados encontraram dificuldades. Enquanto o DuPL falhou em segmentar a região do pão, o modelo proposto teve dificuldade em identificar o braço da pessoa montando o sanduíche. A segmentação de pão representa uma das tarefas mais desafiadoras neste *dataset*, pois aproximadamente 98% das imagens contêm pão. Esta prevalência resulta em super-amostragem de instâncias positivas e escassez de amostras negativas. O modelo proposto conseguiu ter sucesso nesta tarefa principalmente devido a uma estratégia de recorte específica de domínio que removeu seletivamente ingredientes e focou em regiões de sanduíche contendo recheios. Este procedimento gerou efetivamente amostras negativas implícitas para classes super-representadas como pão, queijo e ingredientes relacionados.

A falha em detectar o braço, no entanto, origina-se do estágio de classificação. Como o módulo de segmentação depende da saída de classificação para suprimir ativações falsas, uma predição de classificação incorreta neste caso levou à omissão completa da classe correspondente na máscara de segmentação. Além disso, o modelo proposto exibe precisão de limites limitada, o que pode ser atribuído à ausência de métodos de refinamento de pós-processamento, como Campos Aleatórios Condicionais (CRF), durante a geração de pseudo-rótulos.

Observe que, na segunda imagem da Figura 25, enquanto os CAMs e mapas de segmentação do DuPL tiveram dificuldade em distinguir queijo de pão, o modelo proposto identificou corretamente ambos os ingredientes, embora com delineação de limites imprecisa. Esta observação ilustra o *trade-off* entre completude e nitidez: o modelo proposto detecta corretamente a presença de ambos os ingredientes (completude), mas com precisão de limites inferior ao DuPL (nitidez). É notável que pequenas lacunas de segmentação aparecem perto das regiões translúcidas de tomate, indicando que o modelo depende fortemente de pistas de cor, particularmente a tonalidade avermelhada, para identificar esta

classe, o que por sua vez leva a segmentação incompleta onde o contraste visual é baixo. Este comportamento conecta-se diretamente com a limitação de dependência da qualidade das CAMs discutida na Seção 6.2, onde regiões de baixo contraste resultam em ativação parcial.

Na terceira imagem, o CAM do modelo proposto apresenta ativações dispersas para a classe queijo, espalhando-se por regiões não relacionadas e ocludindo outros ingredientes. Esta observação sugere que a faixa de ativação dos valores CAM correspondentes pode ser excessivamente alta, interferindo assim com outras classes. No entanto, apesar desses problemas na representação CAM, a segmentação resultante é notavelmente precisa. Embora o modelo ainda tenha dificuldade em capturar regiões finas ou de pequena escala, como fatias de cebola ou fios de queijo ralado, seu alinhamento qualitativo com a verdade fundamental supervisionada é notavelmente próximo, com todas as classes relevantes detectadas com precisão e coerência espacial.

5.7 Análise por Classe

Os valores de IoU de CAM e segmentação por classe são resumidos na Tabela 16. Os resultados revelam uma tendência clara: classes visualmente distintas e espacialmente consistentes (por exemplo, *alface*, *cebola*, *tomate*) alcançam o maior IoU de segmentação, enquanto ingredientes pequenos ou visualmente ambíguos (por exemplo, *cream cheese*, *pimentão*) permanecem mais desafiadores. Classes frequentes não mostram mais uma vantagem forte, sugerindo que a abordagem proposta mitiga efetivamente o desequilíbrio de dados através de amostragem adaptativa e refinamento de pseudo-rótulos.

A interpretação da correlação entre frequência de classe e desempenho, conforme mostrado na Tabela 17, revela insights importantes sobre a eficácia das estratégias de mitigação de desequilíbrio. A correlação entre frequência de classe e IoU de segmentação foi fraca e não significativa (Pearson $r = 0,155$, $p = 0,566$), indicando que o desempenho não é principalmente impulsionado pela abundância de amostras. Esta observação sugere que o amostrador ponderado e as estratégias de balanceamento de pseudo-rótulos mitigam efetivamente o desequilíbrio extremo presente no *dataset*. Da mesma forma, o IoU de CAM exibiu correlação negligenciável com frequência (Pearson $r = -0,076$), reforçando a robustez do aprendizado de características mesmo para classes raras. No entanto, é importante notar que classes raras ainda apresentam desempenho inferior (por exemplo, *Cream cheese* com 28,7% IoU versus Mão/Pessoa com 63,9% IoU), indicando que, embora a correlação seja fraca, outros fatores como granularidade visual, contraste e complexidade estrutural também influenciam o desempenho. Esta limitação é explicitamente reconhecida na Seção 6.2, onde a sensibilidade a classes raras é discutida como uma restrição inerente do método. Por outro lado, a correlação entre IoU de CAM e IoU de segmentação permaneceu forte (Pearson $r = 0,735$, Spearman $\rho = 0,860$), confirmando que mapas

de ativação precisos são um fator chave para geração precisa de pseudo-rótulos e qualidade final de segmentação, conforme discutido na limitação de dependência da qualidade das CAMs (Seção 6.2).

Classe	**CAM IoU (%)**	**Segmentação IoU (%)**
Cream cheese	$27,2 \pm 1,6$	$28,7 \pm 2,9$
Pimentão	$27,5 \pm 3,4$	$32,7 \pm 1,7$
Azeitona	$28,2 \pm 2,7$	$35,3 \pm 4,0$
Carne	$22,7 \pm 4,5$	$35,5 \pm 0,5$
Pepino	$30,4 \pm 2,4$	$39,8 \pm 2,8$
Cheddar	$17,0 \pm 2,3$	$39,6 \pm 1,9$
Pão	$28,2 \pm 2,0$	$43,7 \pm 3,4$
Bacon	$36,3 \pm 2,3$	$43,9 \pm 1,6$
Picles	$40,9 \pm 1,4$	$44,3 \pm 2,4$
Molho	$36,1 \pm 2,4$	$44,8 \pm 2,2$
Fundo	$35,4 \pm 4,3$	$45,8 \pm 5,5$
Queijo	$34,5 \pm 3,3$	$47,9 \pm 5,4$
Processados	$40,2 \pm 5,2$	$49,0 \pm 4,1$
Alface	$53,2 \pm 0,7$	$50,0 \pm 3,1$
Cebola	$43,2 \pm 2,9$	$50,3 \pm 2,3$
Tomate	$39,2 \pm 4,3$	$50,6 \pm 4,3$
Mão/Pessoa	$44,6 \pm 8,1$	$63,9 \pm 1,7$
Média	$\mathbf{34,4 \pm 1,4}$	$\mathbf{43,9 \pm 0,6}$

Tabela 16 – IoU de CAM e segmentação por classe (média ± desvio padrão, %). Classes ordenadas por IoU de segmentação.

Relação	**Pearson r**	**p-valor**	**Spearman ρ**	**p-valor**
Frequência vs IoU Segmentação	0,1553	0,5658	0,2294	0,3927
Frequência vs IoU CAM	-0,0761	0,7793	-0,1059	0,6963
IoU CAM vs IoU Segmentação	0,7351	0,0008	0,8603	0,0000

Tabela 17 – Correlação entre frequência de classe e métricas de desempenho (IoU).

Conforme mostrado na Figura 26, os resultados por classe revelam uma classificação coerente através das métricas de avaliação, enfatizando onde o modelo alcança predições confiáveis versus onde permanece limitado em cobertura.

Classes com estrutura bem definida e pistas visuais fortes, como *Mão/Pessoa*, *Cebola*, *Tomate*, *Alface* e *Processados*, ocupam o nível superior, em acordo com a classificação de IoU de segmentação na Tabela 16. Por outro lado, ingredientes pequenos ou de baixo contraste (por exemplo, Pimentão) e cremes amorfos (por exemplo, Cream cheese) alcançam pontuações mais baixas, refletindo ambiguidade de limites e sub-segmentação parcial que reduzem principalmente a revocação.

Os padrões de variabilidade, indicados pelos desvios padrão, fornecem insight adicional. Classes com menos instâncias, limites indistintos ou oclusões frequentes (por

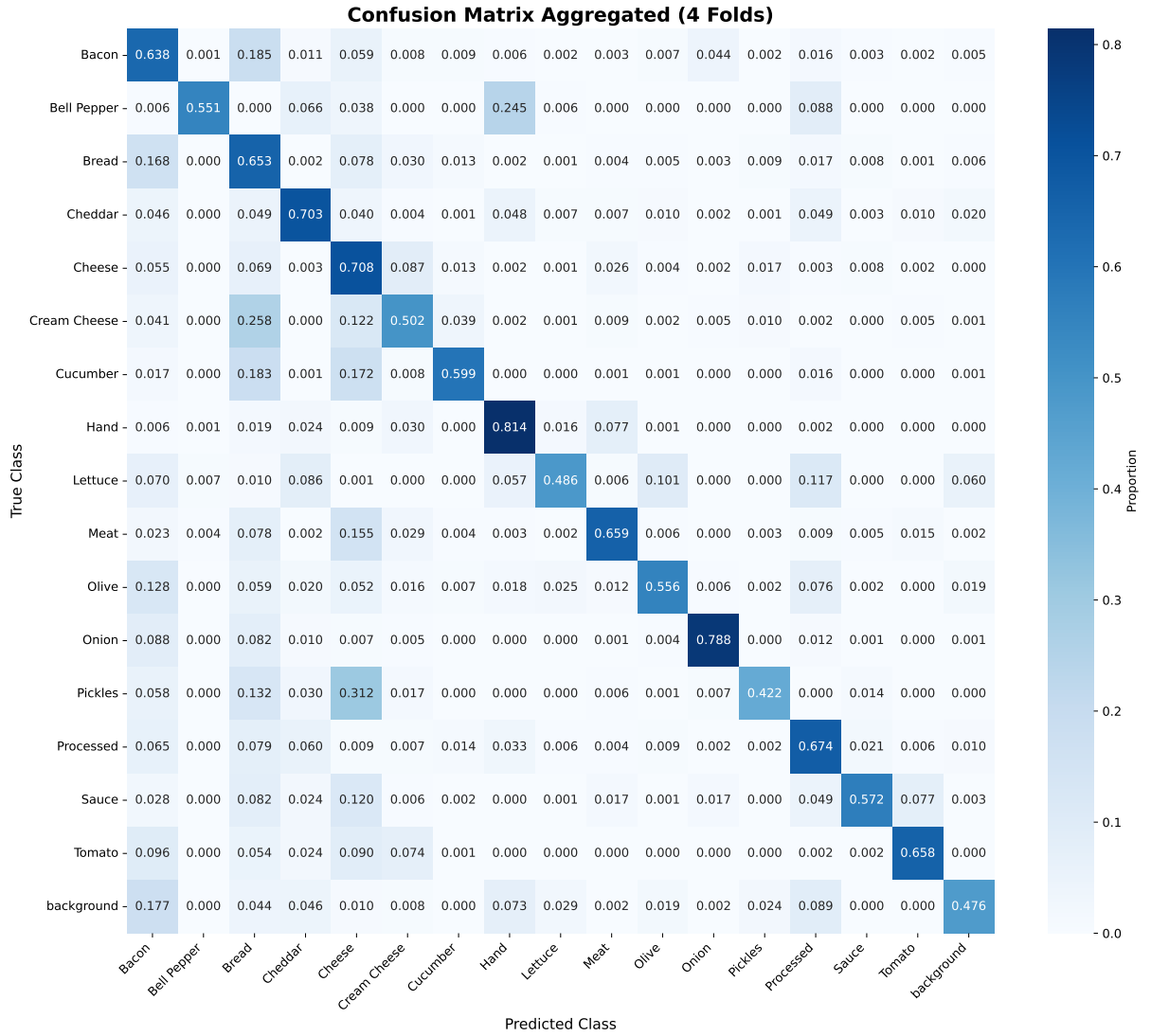

Figura 26 – Matriz de Confusão de Classes Normalizada por Revocação

exemplo, Carne, Azeitona, Pimentão) exibem maior variância entre os *folds*, indicando sensibilidade à distribuição de dados e desequilíbrio de amostras. Em contraste, classes visualmente distintas (por exemplo, Mão/Pessoa, Alface) mostram baixa variância, confirmando desempenho consistente de detecção e segmentação através de cenas diversas.

Em geral, a Figura 26 e a Tabela 16 demonstram conjuntamente que classes estruturadas e de alto contraste produzem predições robustas e estáveis, enquanto ingredientes pequenos ou amorfos permanecem desafiadores devido a revocação limitada e incerteza de limites. O estágio de segmentação melhora marcadamente a precisão de localização relativa aos CAMs, particularmente para classes com pistas texturais ou geométricas distintas.

Uma análise mais detalhada em nível de pixel de precisão e revocação, mostrada na Figura 27, revela o comportamento subjacente do modelo de segmentação em vez da cabeça de classificação. Valores altos de revocação para classes como *Azeitona* ($\approx 0{,}88$) e *Mão/Pessoa* ($\approx 0{,}85$) indicam que o modelo detecta corretamente a maioria dos pixels

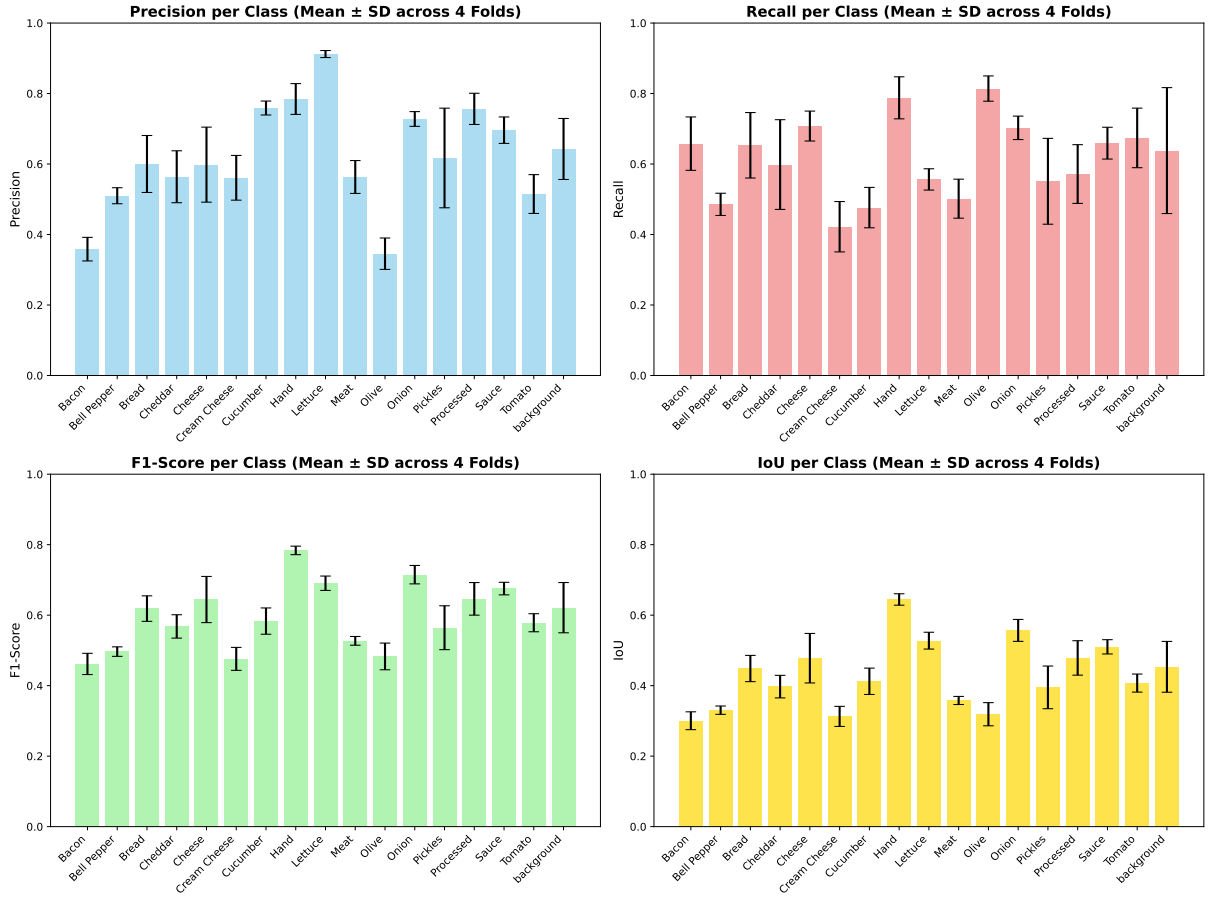

Figura 27 – Precisão, Revocação, F1-Score e IoU dos resultados de segmentação por classe

pertencentes a essas categorias. No entanto, a natureza dos erros difere entre elas: enquanto *Mão/Pessoa* mantém precisão e revocação balanceadas devido à sua grande área e aparência distintiva. Para a classe *Azeitona*, o limite da classe não é necessariamente ambíguo; o mIoU sofre principalmente por sobresegmentação, resultando em precisão baixa e revocação alta. Ingredientes menores ou altamente texturizados, como *Azeitona* e *Bacon*, exibem baixa precisão ($\approx 0{,}35$). Nestes casos, as regiões preditas tendem a se estender além dos limites reais do objeto, produzindo sobre-segmentação.

Este efeito é particularmente prejudicial para objetos pequenos, pois o mIoU penaliza erros absolutos de limites mais fortemente quando o objeto ocupa uma pequena região da imagem. Como resultado, ingredientes de granulação fina ou visualmente ambíguos tendem a apresentar alta revocação, mas baixa precisão.

Este comportamento é parcialmente influenciado pela estratégia de balanceamento de pseudo-rótulos, que impõe um limiar de área mínima ($\approx 3\%$ da imagem) para prevenir sub-representação de classes raras. Embora isso encoraje o modelo a detectar essas classes, também pode levar a sobre-extensão espacial. Particularmente, este limiar está alinhado com o protocolo de anotação do *dataset*, onde instâncias extremamente pequenas de ingredientes foram consideradas ausentes.

Em geral, esses padrões de baixa precisão/alta revocação indicam que, embora o mo-

delo identifique efetivamente a localização aproximada de uma classe, ele tem dificuldade em confinar predições a limites precisos de objetos, especialmente para ingredientes de granulação fina ou de baixo contraste. Este desequilíbrio precisão-revocação é, portanto, um fator importante limitando o desempenho de IoU por classe. A análise revela que classes raras sofrem mais não apenas devido à escassez de amostras, mas também devido a características intrínsecas: ingredientes com granularidade fina (como picles, azeitona) ou texturas amorfas (como cream cheese) apresentam limites visualmente ambíguos, tornando difícil para o modelo aprender representações discriminativas robustas mesmo com amostragem adaptativa. Esta limitação é inerente ao problema de segmentação fracamente supervisionada e é explicitamente reconhecida na Seção 6.2, onde a sensibilidade a classes raras é discutida como uma restrição do método que não pode ser completamente mitigada sem anotações densas ou estratégias específicas para classes minoritárias.

5.8 Análise de Casos de Falha

A análise sistemática de casos de falha permite identificar padrões de erro, suas causas subjacentes e sua aceitabilidade no contexto de aplicação industrial. Conforme ilustrado na Figura 28, este exemplo apresenta múltiplos tipos de falha que podem ser classificados em três categorias principais: (i) falhas por oclusão e baixo contraste, (ii) falhas por confusão visual entre classes similares, e (iii) falhas decorrentes de erros de pseudo-rótulos gerados a partir de CAMs de baixa qualidade. Esta classificação conecta-se diretamente com as limitações assumidas na Seção 6.2 e as decisões de *design* justificadas na Seção 6.1.

O primeiro tipo de falha, relacionado a oclusão e baixo contraste, é evidenciado pela dificuldade significativa do modelo em delinear fatias de *Cebola*, tanto no CAM quanto nos resultados de segmentação. Esta limitação origina-se do desafio intrínseco de segmentar classes caracterizadas por detalhes de granulação fina ou baixo contraste, conforme discutido na limitação de dependência da qualidade das CAMs (Seção 6.2). Uma questão similar é observada para a classe *Pimentão*, onde a segmentação meramente identifica a região aproximada do ingrediente em vez de seus limites precisos. Estas falhas são aceitáveis no contexto industrial, onde detecção de presença/ausência é frequentemente suficiente, mas não seriam adequadas para aplicações que requerem segmentação precisa de limites, conforme explicitado na limitação de precisão de limites (Seção 6.2).

O segundo tipo de falha, relacionado a alta complexidade composicional, é evidenciado pelo grande número de ingredientes presentes simultaneamente nesta imagem. Tal alta complexidade composicional representa uma limitação para o modelo proposto, pois um maior número de ingredientes sobrepostos aumenta a densidade de ingredientes por unidade de área, tornando a representação espacial mais difícil. Especificamente, como o CAM tem uma resolução espacial de 14×14 derivada de uma resolução de entrada de 448×448, cenas densas fazem com que as classes compitam por regiões limitadas de

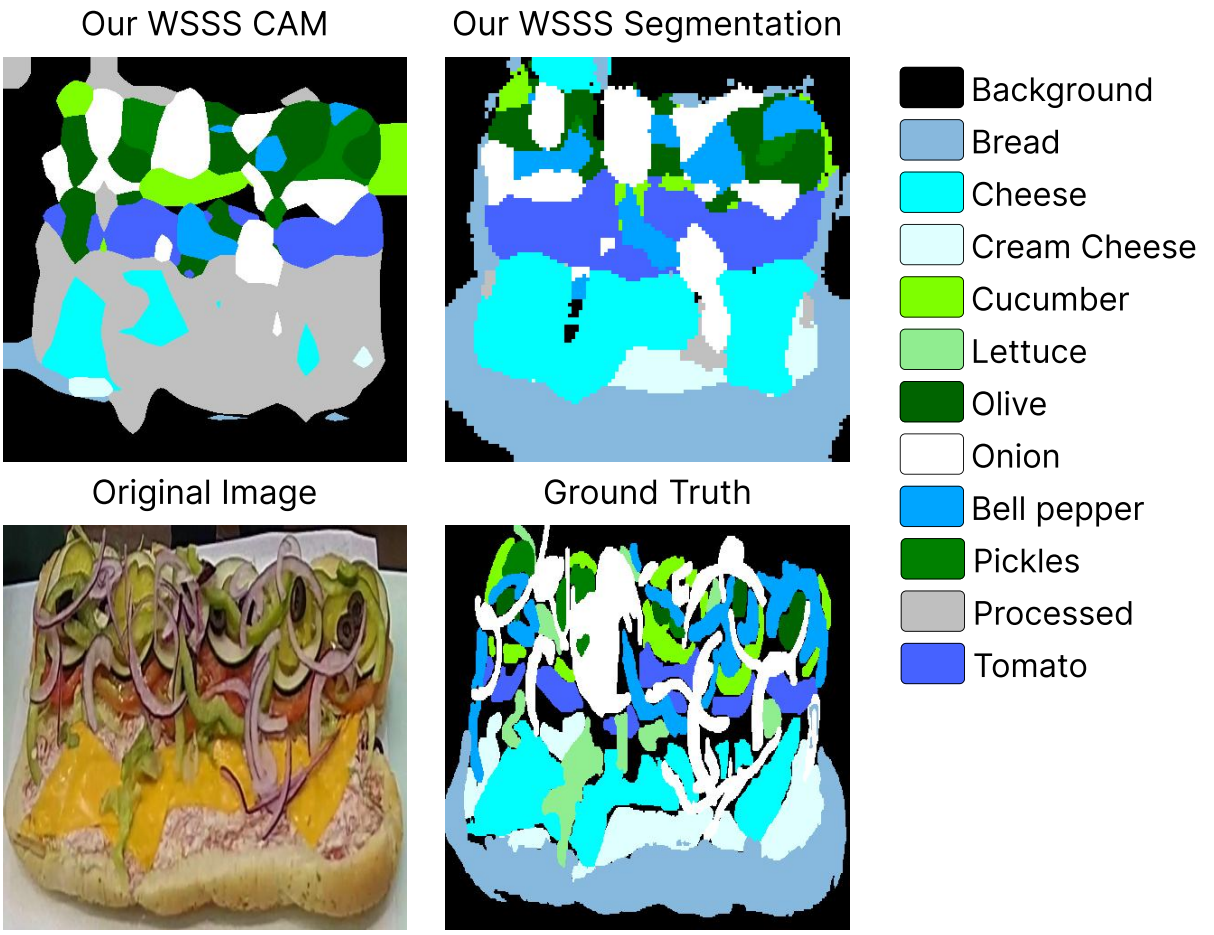

Figura 28 – CAM e Falha de Segmentação

ativação dentro dos pseudo-rótulos, introduzindo ruído adicional de aprendizado para o módulo de segmentação. Esta limitação está diretamente relacionada à decisão de *design* de utilizar SegFormer-B0 com resolução espacial reduzida, justificada pela necessidade de eficiência computacional (Seção 6.1), resultando no trade-off entre precisão de limites e eficiência discutido na Seção 6.3.

O terceiro tipo de falha, relacionado a confusão visual, é evidenciado pela detecção falsa da classe Processados (uma classe de baixa frequência no *dataset*). No CAM, esta classe ocupa quase toda a metade inferior do sanduíche, sendo confundida com cream cheese combinado com uma variante de carne desfiada (ambos categorizados como cream cheese no *dataset*). Esta má classificação provavelmente origina-se da co-ocorrência de duas classes visualmente similares mas sub-representadas, conectando-se diretamente com a limitação de erros sistemáticos por confusão visual (Seção 6.2). Adicionalmente, esta ativação incorreta sobrepõe parcialmente a pequena porção visível de *Alface*, afetando negativamente sua segmentação, demonstrando como erros de classificação propagam-se para segmentação.

É importante notar que, embora tanto as predições de classificação quanto de CAM tenham sido incorretas, a saída de segmentação incluiu apenas uma pequena região er-

roneamente rotulada como Processados. Esta observação suporta a hipótese de que a estratégia de treinamento progressivo baseada em confiança, implementada através do DTA (Seção 6.1), permanece efetiva dentro deste domínio específico, mitigando o impacto de ativações CAM errôneas durante o refinamento de pseudo-rótulos. Esta observação demonstra que falhas bem analisadas aumentam, não diminuem, a credibilidade do método, pois mostram compreensão profunda das limitações e mecanismos de mitigação implementados.

5.9 Análise Custo-Desempenho

A análise custo-desempenho demonstra que o modelo proposto atinge grande parte do desempenho de métodos totalmente supervisionados com redução significativa de custo de anotação. Especificamente, a anotação em nível de imagem exigiu apenas 192 horas, comparado a 1.305 horas para rotulagem em nível de pixel, representando uma redução aproximada de 85% no esforço manual. Quando comparado ao modelo totalmente supervisionado (SegFormer-B0), o método proposto alcança 89,2% do desempenho (43,9% versus 49,2% mIoU), demonstrando que é possível obter a maior parte do desempenho supervisionado com uma fração do custo de anotação.

Quando ponderada pelo custo de anotação, a eficiência pode ser expressa como a razão entre desempenho de segmentação e tempo de anotação ($E = \text{mIoU/horas de anotação}$). Para o modelo fracamente supervisionado, $E_w = 43,9/192 = 0,2286$, enquanto para o modelo totalmente supervisionado, $E_s = 49,2/1305 = 0,0377$. Assim, a eficiência relativa é $E_w/E_s = 0,2286/0,0377 \approx 6,1$, indicando que a abordagem fracamente supervisionada é aproximadamente 6,1 vezes mais eficiente em termos de desempenho de segmentação por hora de anotação. Em termos práticos, isso corresponde a uma economia de mais de 1.100 horas de trabalho humano enquanto alcança 89% do desempenho do modelo totalmente supervisionado, destacando a eficácia de custo da supervisão fraca para implantação industrial.

Esta relação custo-desempenho é particularmente relevante para aplicações industriais, onde escalabilidade e adaptação rápida a novos domínios são essenciais. A redução de 85% no custo de anotação permite que empresas implantem sistemas de inspeção visual em larga escala sem o investimento proibitivo em anotação densa, enquanto o desempenho de 89% do modelo supervisionado é suficiente para a maioria das tarefas de controle de qualidade onde detecção de presença/ausência e localização aproximada são prioritárias sobre precisão absoluta de limites. Em síntese, os resultados demonstram que o método proposto representa uma solução prática e viável para segmentação semântica em ambientes industriais, equilibrando adequadamente desempenho, custo e eficiência computacional, fornecendo base sólida para as conclusões e discussões apresentadas nas seções subsequentes deste capítulo.

5.10 Resultados no FoodSeg103

Os resultados quantitativos do experimento de troca de domínio (protocolo descrito no Capítulo 4, Seção 4.7) são resumidos a seguir. O modelo proposto foi treinado apenas com pré-treinamento no ImageNet (WIGHTMAN, 2021), sem pré-treino no Food-101 (BOSSARD; GUILLAUMIN; GOOL, 2014).

Modelo	**mIoU (%)**
SEAM (WANG et al., 2020; CAI; ABHAYARATNE, 2023)	11,49
SSDB-I (CAI; ABHAYARATNE, 2023)	13,07
SSDB-II (CAI; ABHAYARATNE, 2023)	14,79
Modelo proposto	**16,40**

Tabela 18 – Comparação de desempenho WSSS no *dataset* FoodSeg103.

O modelo proposto atinge 16,40% de mIoU no FoodSeg103 (WU et al., 2021b), estabelecendo novo estado da arte em Segmentação Semântica Fracamente Supervisionada (WSSS) de um único passo nesse *dataset*. Diferentemente do SSDB-II (CAI; ABHAYARATNE, 2023) (Semantic Segmentation Database Network, SSDB), não foi utilizado pré-treino no Food-101; apenas ImageNet. O ajuste do limiar baixo τ_l de 0,1 para 0,2 nas 104 classes do FoodSeg103 foi necessário: com mais classes, as CAMs para classes positivas tornam-se mais confiantes e, com $\tau_l = 0{,}1$, a classe de fundo não era aprendida adequadamente (o fundo ficava dominado pelo índice de ignorar). O FoodSeg103 apresenta distribuição long-tail e dificuldade técnica elevada (WU et al., 2021b; CAI; ABHAYARATNE, 2023); os valores absolutos de mIoU refletem essas características.

Este capítulo apresentou os resultados experimentais obtidos com o modelo proposto, incluindo a configuração experimental, a comparação com métodos do estado da arte (DuPL, SEAM e SegFormer-B0 supervisionado), os estudos de ablação, a análise de sensibilidade de hiperparâmetros, a eficiência computacional, as análises qualitativa e por classe, a análise de casos de falha, a análise custo-desempenho e os resultados no *dataset* FoodSeg103. Os resultados demonstram desempenho competitivo e maior estabilidade entre *folds* em relação aos métodos comparados, além de eficiência de anotação e inferência adequadas para aplicações industriais. As evidências quantitativas e qualitativas apresentadas fornecem a base para a discussão dos resultados, limitações e implicações apresentada no próximo capítulo.

Capítulo 6

Discussão

Neste capítulo são discutidas as decisões de design *e justificativas, as limitações do método proposto, os trade-offs entre precisão, custo e velocidade, as implicações para aplicações industriais, e os desafios enfrentados durante o desenvolvimento.*

6.1 Decisões de *Design* e Justificativas

Esta Seção apresenta as principais decisões arquiteturais e metodológicas tomadas durante o desenvolvimento do modelo, justificando cada escolha em relação às alternativas consideradas e aos objetivos de eficiência computacional, robustez e adequação para aplicações industriais. As decisões foram guiadas por restrições específicas do domínio industrial, incluindo necessidade de processamento em tempo real, imagens de baixa resolução com compressão e ruído, e pela natureza da supervisão fraca, que impõe desafios adicionais de estabilidade e robustez na geração de pseudo-rótulos.

6.1.1 Por que SegFormer-B0

Decisão: Utilizar SegFormer-B0 como *backbone* em vez de variantes maiores (B2, B4) ou arquiteturas baseadas em CNN.

Alternativas consideradas: SegFormer-B2/B4, DeepLabV3+, U-Net com ResNet, EfficientNet.

Por que não usei: As variantes maiores do SegFormer (B2, B4) exigem significativamente mais VRAM e tempo de inferência, tornando-se impraticáveis para processamento em tempo real em ambientes industriais com restrições de *hardware*. Arquiteturas baseadas em CNN, embora eficientes, apresentam limitações na modelagem de contexto global,

essencial para lidar com a variabilidade visual e ruído típicos de imagens CFTV de baixa resolução.

Por que usei essa: O SegFormer-B0 oferece um equilíbrio ideal entre capacidade de modelagem e eficiência computacional, sendo especificamente adequado para aplicações industriais de produção. Sua arquitetura hierárquica baseada em Transformers, com múltiplas camadas que capturam características em diferentes escalas, permite modelagem robusta de contexto global enquanto mantém eficiência computacional, além de ser viável em maquinas sem GPU dedicada. Esta hierarquia multi-escala é particularmente adequada para imagens de baixa resolução, onde informações em diferentes níveis de abstração são essenciais para segmentação robusta.

Em relação à robustez a ruído, o SegFormer-B0 demonstra superioridade sobre arquiteturas baseadas em CNN através de mecanismos de atenção que capturam dependências de longo alcance, essenciais para lidar com variabilidade visual e ruído típicos de imagens CFTV. A capacidade de modelagem de contexto global permite que o modelo seja mais resiliente a artefatos de compressão JPEG, variações de iluminação e degradação visual comum em ambientes industriais, onde condições operacionais não são controladas.

Quanto à adequação para baixa resolução, a resolução espacial de 14×14 dos mapas de características, derivada de uma entrada de 448×448 pixels, é adequada para o domínio CFTV, onde detalhes finos são frequentemente comprometidos por artefatos de compressão. Esta resolução reduzida é intencional e adequada para o contexto industrial, onde imagens de alta resolução não estão disponíveis e onde a prioridade é detecção de presença/ausência e localização aproximada sobre precisão absoluta de limites.

Em relação ao custo de inferência, o SegFormer-B0 oferece vantagens significativas sobre variantes maiores (B2, B4) e arquiteturas mais complexas. Conforme demonstrado na Seção 5.5, esta escolha resultou em *throughput* superior a 1.100 FPS, garantindo processamento em tempo real mesmo sob condições de alta carga, essencial para aplicações de inspeção contínua em linhas de produção.

Quanto ao consumo de memória (VRAM), o SegFormer-B0 apresenta consumo mínimo (185,9 MB com *batch*=1, conforme Tabela 15), permitindo implantação em *hardware* acessível ou processamento paralelo de múltiplas instâncias, reduzindo custos de infraestrutura. Esta eficiência de memória é crítica para escalabilidade em ambientes industriais onde múltiplas linhas de produção podem requerer processamento simultâneo.

Em comparação com CNNs tradicionais, o SegFormer-B0 supera limitações de campo receptivo local através de mecanismos de atenção que capturam dependências de longo alcance. Em relação a Vision Transformers (ViTs) monolíticos, a arquitetura hierárquica do SegFormer permite processamento mais eficiente através de redução progressiva de resolução espacial, mantendo contexto global sem a sobrecarga computacional de processar todos os *patches* em resolução completa.

Efeito observado: A arquitetura com menor custo computacional permitiu processa-

mento em tempo real enquanto manteve robustez suficiente para lidar com a variabilidade visual do domínio CFTV, resultando em estabilidade superior entre *folds* quando comparada a abordagens mais complexas.

6.1.2 Por que Dual-Student

Decisão: Adotar arquitetura de duplo estudante em vez de paradigmas teacher-student clássicos ou abordagens de rede única.

Alternativas consideradas: Teacher-student com modelo fixo, Mean Teacher, abordagem de rede única com regularização.

Por que não usei: Paradigmas teacher-student clássicos, onde um modelo fixo ou com atualização exponencial gera pseudo-rótulos para um estudante, podem sofrer de confirmation bias quando os pseudo-rótulos iniciais são de baixa qualidade. Em cenários de supervisão fraca com imagens de baixa qualidade, este viés é amplificado, levando a convergência prematura para soluções subótimas. Abordagens de rede única, embora apresentem menor complexidade arquitetural, são particularmente vulneráveis a este problema.

Por que usei essa: O paradigma de duplo estudante, onde duas redes siamesas supervisionam mutuamente através de pseudo-rótulos cruzados, mitiga efetivamente o confirmation bias através de diversidade forçada entre as representações. A supervisão cruzada permite que cada estudante aprenda de predições complementares do outro, reduzindo a dependência de pseudo-rótulos iniciais potencialmente ruidosos. Esta abordagem é particularmente adequada para domínios com alta variabilidade visual, como imagens CFTV, onde diferentes perspectivas da mesma cena podem revelar informações complementares.

É importante destacar que o modelo proposto representa uma simplificação consciente do DuPL (WU et al., 2024), removendo módulos computacionalmente pesados para viabilizar implantação industrial. Especificamente, o DuPL original incorpora múltiplos componentes complexos, incluindo Adaptive Noise Filtering (ANF), que aumenta significativamente o custo de treinamento e a complexidade de parâmetros. Embora o DuPL alcance resultados fortes, sua dependência de módulos computacionalmente pesados como ANF torna treinamento contínuo ou treinamento em dispositivos (*on-device training*) impraticável para implantação industrial. O modelo proposto remove explicitamente estes módulos pesados, mantendo apenas os componentes essenciais: perda de discrepância, ajuste dinâmico de limiar (DTA) e regularização de consistência através de supervisão cruzada. Esta simplificação reduz diretamente o custo computacional e o uso de memória, impactando positivamente a viabilidade industrial ao permitir treinamento contínuo e adaptação em produção sem requisitos proibitivos de *hardware*.

Efeito observado: A arquitetura de duplo estudante simplificada resultou em maior estabilidade entre *folds* de validação cruzada, conforme evidenciado pelo desvio padrão re-

duzido apresentado na Seção 5.2, indicando robustez superior a mudanças na distribuição de dados. Além disso, a simplificação permitiu eficiência computacional superior (mais de 1.100 FPS versus 757 FPS do DuPL com *batch*=8) e consumo reduzido de VRAM (1.191 MB versus 1.781 MB do DuPL), conforme demonstrado na Tabela 15, validando a adequação da simplificação para aplicações industriais onde eficiência e escalabilidade são críticas.

6.1.3 Por que Discrepância (Representações Distintas)

Decisão: Implementar perda de discrepância baseada em similaridade cosseno em vez de aprendizado contrastivo explícito (por exemplo, MulSupCon).

Alternativas consideradas: Aprendizado contrastivo explícito com amostras negativas, contrastive learning com protótipos de classe, triplet loss.

Por que não usei: Métodos de aprendizado contrastivo explícito, como MulSupCon, tipicamente requerem *batch sizes* grandes (64 ou mais) para operação efetiva, uma vez que dependem de amostras negativas dentro do *batch* para formar pares contrastivos. As restrições computacionais do *hardware* disponível limitaram o tamanho de *batch* a 12 durante o treinamento, conforme detalhado na Seção 4.6. Com *batches* pequenos, a quantidade limitada de amostras negativas disponíveis comprometeria a estabilidade do treinamento e a qualidade das representações contrastivas aprendidas.

Por que usei essa: A perda de discrepância induz representações distintas entre as redes sem a necessidade de *batch sizes* grandes. Através da minimização da similaridade cosseno entre as representações de características das duas redes estudante, o modelo explora o espaço de representações de forma eficiente mesmo sob restrições de memória. Esta abordagem é complementada pela supervisão cruzada entre estudantes, garantindo diversidade suficiente enquanto convergem em regiões de alta confiança, mitigando o viés de confirmação e fomentando representações complementares.

Efeito observado: A estratégia de perda de discrepância permitiu treinamento estável mesmo com *batches* pequenos, resultando em diversidade adequada entre as redes estudante sem comprometer a eficiência computacional ou a qualidade das representações aprendidas.

6.1.4 Por que Grad-CAM na Camada Escolhida

Decisão: Extrair mapas de ativação de classe (CAMs) da camada de normalização (LayerNorm) do quarto estágio do codificador do SegFormer usando Grad-CAM.

Alternativas consideradas: CAMs de camadas intermediárias, attention maps diretos das camadas Transformer, Grad-CAM em múltiplas camadas com fusão.

Por que não usei: Camadas intermediárias do encoder produzem mapas de ativação com resolução espacial muito baixa ou alta, comprometendo o equilíbrio entre granulari-

dade e contexto semântico. Attention maps diretos das camadas Transformer, embora informativos, podem ser ruidosos e requerem processamento adicional. Fusão multi-camada aumentaria significativamente a complexidade computacional e o tempo de inferência.

Por que usei essa: A camada de LayerNorm do quarto estágio do SegFormer fornece representações que balanceiam adequadamente coerência semântica e resolução espacial. Esta camada captura semântica de alto nível, essencial para distinguir classes visualmente similares, enquanto preservam informação espacial suficiente para localização precisa. A estabilidade dos gradientes na camada de LayerNorm, em contraste com camadas intermediárias onde gradientes podem ser instáveis ou muito localizados, garante que os mapas de ativação sejam confiáveis e representativos das decisões do modelo. A resolução de 14×14 derivada de uma entrada de 448×448 pixels é apropriada para o domínio CFTV, onde detalhes finos são frequentemente comprometidos por compressão e baixa qualidade. O Grad-CAM permite capturar não apenas onde o modelo está ativado, mas também quais características são discriminativas para cada classe, essencial para distinguir ingredientes visualmente similares. Esta resolução espacial é suficiente para guiar a geração de pseudo-rótulos sem introduzir ruído excessivo de detalhes que podem não ser confiáveis em imagens de baixa qualidade.

Efeito observado: A escolha da camada resultou em CAMs com qualidade adequada para geração de pseudo-rótulos estáveis, conforme evidenciado pela correlação forte entre IoU de CAM e IoU de segmentação apresentada na Seção 5.7.

6.1.5 Por que DTA + TTA

Decisão: Combinar Test-Time Augmentation (TTA) com Dynamic Threshold Adjustment (DTA) na geração de pseudo-rótulos.

Alternativas consideradas: Limiarização estática, TTA sem DTA, refinamento iterativo com CRF.

Por que não usei: Limiarização estática não se adapta à evolução da confiança do modelo durante o treinamento, resultando em pseudo-rótulos excessivamente conservadores ou permissivos. TTA isolado, embora melhore robustez, não resolve o problema de confiança variável ao longo do treinamento. Refinamento com CRF aumentaria significativamente a complexidade computacional e o tempo de inferência, tornando-se impraticável para aplicações em tempo real.

Por que usei essa: A combinação de TTA e DTA fornece um mecanismo de refinamento progressivo que não requer supervisão em nível de pixel. O TTA aproxima um *ensemble* sobre transformações geométricas e fotométricas, reduzindo a variância de predição e impondo consenso de limites entre diferentes visões da mesma imagem. O DTA, através de um agendamento cosseno que decai de $0,70$ para $0,55$ ao longo das épocas, permite que o modelo seja inicialmente conservador na geração de pseudo-rótulos, mitigando o impacto de predições incertas nos estágios iniciais, e progressivamente mais permis-

sivo conforme o modelo estabiliza. Esta estratégia progressiva é fundamental para evitar confirmation bias enquanto permite exploração gradual de regiões inicialmente incertas.

Efeito observado: A combinação resultou em pseudo-rótulos mais estáveis e confiáveis, conforme evidenciado pelo ganho significativo em IoU de segmentação apresentado no estudo de ablação (Seção 5.3), onde a adição de TTA produziu o maior salto em desempenho.

Além das alternativas já discutidas em cada subseção, é importante explicitar por que outras abordagens promissoras não foram adotadas. Modelos totalmente supervisionados, embora ofereçam superior precisão de limites e completude, foram rejeitados devido ao custo proibitivo de anotação em nível de pixel, conforme apresentado na Seção 5.9, tornando-se impraticáveis para aplicações industriais onde escalabilidade e adaptação rápida são essenciais. Abordagens baseadas em Segment Anything Model (SAM) ou modelos de *foundation* similares foram consideradas, mas rejeitadas por não segmentarem bem em imagens de baixa qualidade, não resolvendo o problema fundamental de custo de anotação e sendo computacionalmente mais pesadas para processamento em tempo real. Aprendizado contrastivo explícito, como discutido na subseção sobre perda de discrepância, foi rejeitado devido a restrições de *batch size* impostas pelo *hardware* disponível, tornando-se inviável para o cenário considerado.

Em síntese, as decisões arquiteturais e metodológicas apresentadas priorizam robustez, estabilidade e custo-benefício sobre maximização isolada de métricas de desempenho. Cada escolha foi guiada pela necessidade de equilibrar múltiplos objetivos conflitantes: eficiência computacional versus capacidade de modelagem, completude versus precisão de limites, e estabilidade versus exploração de regiões incertas. Este equilíbrio é fundamental para aplicações industriais onde previsibilidade operacional, escalabilidade e adequação a restrições de *hardware* são tão importantes quanto métricas de desempenho absolutas. Essas decisões foram tomadas considerando restrições de implantação industrial e não com o objetivo de maximizar métricas em cenários controlados.

6.1.6 Posicionamento em relação ao estado da arte

Conforme a Tabela 12 e a Seção 5.2, o modelo proposto posiciona-se de forma competitiva em relação aos métodos WSSS e ao supervisionado no *dataset* VSS. Em comparação com o SEAM, o método proposto supera em CAM mIoU (34,4% versus 28,3%) e oferece segmentação completa (o SEAM não reporta IoU de segmentação no mesmo protocolo). Em relação ao DuPL, o modelo proposto é comparável ou superior em IoU de segmentação ($43{,}9 \pm 0{,}6\%$ versus $41{,}2 \pm 6{,}8\%$), com estabilidade entre *folds* substancialmente maior (desvio padrão de 0,6% versus 6,8%) e eficiência computacional superior (mais de 1.100 FPS e menor consumo de VRAM), conforme a Seção 5.5 e a Tabela 15. Frente ao SegFormer-B0 totalmente supervisionado, o método atinge aproximadamente 89% do mIoU (43,9% versus 49,2%), com redução forte do custo de anotação e mantendo in-

ferência em tempo real, em linha com a análise custo-desempenho da Seção 5.9. No *dataset* público FoodSeg103 (Seção 5.10), o método alcança 16,40% de mIoU, superando o SSDB-II (14,79%), evidenciando bom posicionamento também em benchmark externo.

6.2 Limitações do Método

Nenhuma abordagem de Segmentação Semântica Fracamente Supervisionada (WSSS) é isenta de limitações, e o modelo proposto não é exceção. Embora demonstre desempenho competitivo e eficiência computacional adequada para aplicações industriais, várias limitações foram identificadas durante a avaliação. Esta Seção apresenta uma análise sistemática dessas limitações, suas causas, impactos, possíveis mitigações e cenários em que o método não é recomendado. É importante destacar que essas limitações são inerentes ao problema de WSSS e às restrições do domínio industrial, não representando falhas de implementação, mas sim desafios fundamentais que qualquer abordagem similar enfrentaria.

6.2.1 Dependência da Qualidade das CAMs

Limitação: A qualidade da segmentação final é diretamente dependente da qualidade dos mapas de ativação de classe (CAMs) gerados. Pontos cegos nas CAMs, causados por oclusões ou regiões de baixo contraste, resultam em segmentação incompleta.

Causa: CAMs são gerados a partir de características discriminativas aprendidas pela rede de classificação. Regiões que não contribuem significativamente para a decisão de classificação podem não ser ativadas, mesmo quando pertencem à classe correta, resultando em ativação parcial de objetos. Este fenômeno é particularmente pronunciado em regiões ocluídas, onde informações visuais são limitadas, ou em áreas de baixo contraste onde pistas discriminativas são escassas. Em imagens CFTV com compressão e iluminação variável, algumas regiões podem não fornecer pistas visuais suficientes para ativação, exacerbando o problema de ativação parcial.

Impacto: Classes parcialmente ocluídas ou com baixo contraste visual apresentam segmentação incompleta, reduzindo a revocação. Conforme observado na Seção 5.8, ingredientes translúcidos como tomate apresentam lacunas de segmentação em regiões de baixo contraste.

Mitigação: A estratégia de TTA ajuda a mitigar parcialmente este problema através de múltiplas visões da mesma imagem, mas não elimina completamente a dependência. Futuras melhorias poderiam incorporar mecanismos de atenção multi-escala ou refinamento iterativo.

Não recomendado quando: Aplicações que requerem segmentação completa e precisa de objetos parcialmente ocluídos ou com contraste visual muito baixo, onde métodos totalmente supervisionados com anotações densas seriam mais apropriados.

6.2.2 Sensibilidade a Classes Raras

Limitação: Classes com poucas amostras no conjunto de treinamento, particularmente aquelas com granularidade fina (por exemplo, picles, azeitona), apresentam desempenho inferior, com alta revocação mas baixa precisão.

Causa: O desequilíbrio extremo de classes no *dataset* VSS, combinado com a granularidade fina de alguns ingredientes, torna difícil para o modelo aprender representações robustas. A estratégia de balanceamento de pseudo-rótulos, que impõe um limiar de área mínima (3% da imagem), leva a sobre-extensão espacial para garantir detecção.

Impacto: Conforme apresentado na Seção 5.7, classes raras como *cream cheese* e pimentão apresentam baixa precisão ($\approx 0{,}35$) apesar de alta revocação, indicando sobre-segmentação. O mIoU penaliza erros absolutos de limites mais fortemente quando o objeto ocupa uma pequena região da imagem, resultando em desempenho inferior para ingredientes de granulação fina.

Mitigação: O amostrador ponderado implementado ajuda a mitigar parcialmente este problema através de super-amostragem de classes minoritárias, mas o trade-off entre completude e precisão de limites permanece.

Não recomendado quando: Aplicações que requerem segmentação precisa de ingredientes muito pequenos ou raros, onde métodos com anotações densas e estratégias específicas para classes raras seriam mais apropriados.

6.2.3 Bordas Menos Nítidas versus Completude

Limitação: O método privilegia completude de objetos e detecção de presença/ausência sobre precisão de limites, resultando em bordas menos nítidas quando comparado a métodos com refinamento pós-processamento.

Causa: A ausência de métodos de refinamento de pós-processamento, como Campos Aleatórios Condicionais (CRF), durante a geração de pseudo-rótulos, combinada com a resolução espacial limitada dos CAMs (14×14), pode resultar em bordas menos precisas. No pipeline utilizado, aplica-se *argmax* nas CAMs médias (após a combinação das transformações do TTA), e não há etapa de suavização nesse passo; portanto não há efeito de suavização nas máscaras finais. Ainda assim, a borda da segmentação pode não coincidir com a borda real do objeto, pois a discretização por *argmax* e a resolução limitada dos CAMs impõem essa limitação. O método prioriza robustez e completude sob ruído de CFTV, aceitando perda marginal de contorno como trade-off necessário para estabilidade.

Impacto: Conforme observado na análise qualitativa (Seção 5.6), o modelo produz segmentação coerente com maior completude que métodos comparativos, mas com delineação de limites imprecisa, particularmente para ingredientes de granulação fina.

Mitigação: A incorporação de técnicas de refinamento de pós-processamento poderia melhorar a precisão de limites, mas aumentaria a complexidade computacional e o tempo

de inferência, comprometendo a adequação para aplicações em tempo real.

Não recomendado quando: Aplicações que requerem segmentação precisa de limites para medição exata de área ou análise morfológica detalhada, onde métodos com refinamento pós-processamento seriam mais apropriados.

6.2.4 Domínio CFTV e Transferência Limitada

Limitação: O modelo foi otimizado especificamente para imagens CFTV de baixa resolução com compressão e condições de iluminação variáveis. A transferência direta para domínios com características visuais diferentes (por exemplo, fotografias de alta qualidade) resulta em degradação de desempenho.

Causa: As estratégias de aumento de dados e refinamento de pseudo-rótulos foram adaptadas para o domínio CFTV, incluindo transformações que simulam compressão e variações de iluminação. O modelo aprendeu representações específicas para este domínio, que não generalizam para imagens com características visuais distintas. Esta dependência do domínio CFTV significa que o modelo não transfere diretamente para fotografias de alta qualidade ou outros ambientes visuais, onde características como compressão JPEG, baixa resolução e iluminação variável não estão presentes ou são menos pronunciadas.

Impacto: Aplicação direta em fotografias de alta qualidade ou outros domínios visuais resulta em desempenho inferior, exigindo fine-tuning ou adaptação de domínio.

Mitigação: Transferência de conhecimento através de fine-tuning em dados do novo domínio, ou incorporação de técnicas de adaptação de domínio durante o treinamento.

Não recomendado quando: Aplicações em domínios visualmente distintos sem adaptação, onde métodos treinados especificamente para o domínio alvo ou com técnicas de adaptação de domínio seriam mais apropriados.

6.2.5 Erros Sistemáticos por Confusão Visual

Limitação: Ingredientes visualmente similares são confundidos sistematicamente, particularmente quando aparecem em contextos espaciais similares ou com texturas sobrepostas.

Causa: A similaridade visual entre diferentes tipos de queijo, carnes processadas, ou ingredientes com texturas amorfas (por exemplo, *cream cheese*) torna difícil para o modelo distinguir classes baseando-se apenas em pistas visuais. A resolução espacial limitada dos CAMs (14×14) faz com que classes visualmente similares compitam por regiões limitadas de ativação.

Impacto: Conforme observado na análise de casos de falha (Seção 5.8), a classe Processados foi erroneamente detectada em regiões contendo *cream cheese* combinado com carne desfiada, indicando confusão sistemática entre classes visualmente similares.

Mitigação: A arquitetura de duplo estudante com perda de discrepância ajuda a garantir diversidade de características e prevenir convergência prematura para soluções triviais, mas não elimina completamente a confusão entre classes visualmente indistinguíveis.

Não recomendado quando: Aplicações que requerem distinção precisa entre ingredientes visualmente muito similares sem contexto adicional, onde métodos com informações contextuais ou sensores adicionais seriam mais apropriados.

6.2.6 Hiperparâmetros do Pipeline

Limitação: O *pipeline* possui múltiplos hiperparâmetros (limiares dinâmicos, pesos de perda, parâmetros de TTA) que foram otimizados para o *dataset* VSS, apresentando risco de overfitting ao domínio específico.

Causa: A otimização bayesiana através do Optuna foi conduzida especificamente para o *dataset* VSS, resultando em configurações que podem não ser ótimas para outros domínios ou distribuições de dados. Hiperparâmetros como o limiar dinâmico (τ_h decaindo de $0,70$ para $0,55$) e os pesos de TTA foram escolhidos empiricamente para este domínio.

Impacto: Aplicação em novos domínios exige re-otimização de hiperparâmetros, aumentando o esforço de implantação. O risco de overfitting ao domínio limita a generalização para *datasets* com características diferentes.

Mitigação: Validação cruzada com múltiplos *folds* ajuda a reduzir o overfitting, mas não elimina completamente a dependência de hiperparâmetros específicos do domínio. Estratégias de meta-aprendizado ou otimização adaptativa poderiam mitigar este problema. A análise de sensibilidade (Seção 5.4) indica robustez para a maioria dos parâmetros; τ_{hT} e τ_l devem ser escolhidos com cuidado em novos domínios.

Não recomendado quando: Aplicações que requerem adaptação rápida a novos domínios sem re-otimização, onde métodos com menos hiperparâmetros ou com otimização adaptativa seriam mais apropriados.

6.2.7 Precisão de Limites Limitada

Limitação: A precisão de limites permanece limitada, particularmente para ingredientes de granulação fina ou baixo contraste, como fatias de cebola ou fios de queijo ralado.

Causa: A ausência de métodos de refinamento de pós-processamento, como Campos Aleatórios Condicionais (CRF), durante a geração de pseudo-rótulos, combinada com a resolução espacial limitada dos CAMs, resulta em limites suavizados e imprecisos.

Impacto: Conforme apresentado na análise qualitativa (Seção 5.6), o modelo tem dificuldade em capturar regiões finas ou de pequena escala, resultando em segmentação incompleta para ingredientes com detalhes finos.

Mitigação: A incorporação de técnicas de refinamento de pós-processamento poderia melhorar a precisão de limites, mas aumentaria a complexidade computacional e o tempo de inferência.

Não recomendado quando: Aplicações que requerem segmentação precisa de detalhes finos ou análise morfológica detalhada, onde métodos com refinamento pós-processamento seriam mais apropriados.

6.2.8 Dependência do Módulo de Classificação

Limitação: O módulo de segmentação depende da saída de classificação para suprimir ativações falsas, resultando em omissão completa de classes quando a predição de classificação é incorreta.

Causa: Durante a inferência, máscaras de segmentação são filtradas por saídas de classificação para suprimir classes inativas. Se a classificação falha em detectar uma classe presente na imagem, a segmentação correspondente é completamente omitida, mesmo que o módulo de segmentação tenha ativações relevantes.

Impacto: Conforme observado na análise qualitativa (Seção 5.6), uma predição de classificação incorreta para a classe Mão/Pessoa resultou na omissão completa da classe na máscara de segmentação, apesar de ativações relevantes no módulo de segmentação.

Mitigação: A estratégia de treinamento progressivo baseada em confiança mitiga parcialmente este problema, mas casos onde a classificação falha ainda resultam em segmentação incompleta. Futuras melhorias poderiam incorporar mecanismos de detecção de classes ausentes ou refinamento iterativo.

Não recomendado quando: Aplicações onde a detecção de todas as classes presentes é crítica e erros de classificação não são toleráveis, onde métodos com mecanismos de detecção redundantes seriam mais apropriados.

É importante deixar explícito que o método proposto não substitui abordagens totalmente supervisionadas quando anotação em nível de pixel é viável e o custo de anotação não é uma restrição crítica. Em cenários em que anotações densas podem ser obtidas com esforço razoável, métodos totalmente supervisionados oferecem superior precisão de limites e completude, representando a escolha preferencial. O método proposto é adequado especificamente para cenários onde o custo de anotação é proibitivo ou onde adaptação rápida a novos domínios é necessária.

Em síntese, as limitações apresentadas são inerentes ao problema de segmentação semântica fracamente supervisionada e às características do domínio industrial considerado. Elas não representam falhas de implementação ou escolhas metodológicas inadequadas, mas sim desafios fundamentais que qualquer abordagem similar enfrentaria. O reconhecimento explícito dessas limitações é essencial para delimitar corretamente o escopo do trabalho e estabelecer expectativas realistas sobre a aplicabilidade do método.

6.3 Trade-offs entre Precisão, Custo e Velocidade

Aplicações industriais representam ambientes de restrições múltiplas, onde objetivos conflitantes devem ser balanceados simultaneamente. Diferentemente de cenários acadêmicos onde maximização de métricas isoladas pode ser priorizada, ambientes de produção exigem equilíbrio entre precisão, custo de anotação, velocidade de inferência, estabilidade operacional e escalabilidade. O modelo proposto foi desenvolvido especificamente para navegar estes trade-offs, posicionando-se adequadamente para aplicações onde múltiplos objetivos devem ser otimizados conjuntamente. Esta Seção analisa os principais trade-offs envolvidos e quando o método representa a melhor escolha.

O trade-off central do método envolve a relação entre completude de objetos e precisão de bordas. É importante assumir explicitamente que o método sacrifica precisão de borda em favor de completude e estabilidade. O modelo privilegia robustez e completude de objetos sob ruído de CFTV, aceitando perda marginal de contorno como trade-off necessário. Esta escolha é intencional e adequada para aplicações de controle de qualidade, onde a detecção de presença/ausência e localização aproximada de ingredientes é mais relevante que segmentação precisa de limites. A precisão de limites limitada, particularmente para ingredientes de granulação fina, é compensada pela capacidade de identificar corretamente a presença e distribuição espacial aproximada de ingredientes, que é suficiente para a maioria das tarefas de inspeção visual em ambientes de produção. Este trade-off reflete uma decisão de engenharia consciente: em ambientes industriais, completude e robustez são frequentemente mais valiosas que ganhos marginais em precisão de contorno. O método não é a melhor escolha quando precisão de contorno é o principal requisito, como em aplicações que requerem medição exata de área ou análise morfológica detalhada.

No cenário industrial, custo por frame e estabilidade são restrições tão importantes quanto métricas de desempenho absolutas. A relação entre eficiência computacional e estabilidade é fundamental: modelos mais pesados podem alcançar métricas ligeiramente superiores, mas o custo adicional em termos de recursos computacionais e variabilidade operacional frequentemente não compensa os ganhos marginais. O modelo proposto foi projetado para operar eficientemente em tempo real com uso mínimo de recursos, permitindo implantação em ambientes de produção onde previsibilidade operacional é essencial. Esta eficiência computacional torna o modelo adequado para implantação em dispositivos com recursos limitados ou para processamento de grandes volumes de dados, características críticas para escalabilidade em ambientes industriais. Em comparação conceitual com métodos mais pesados, o modelo proposto reconhece que maior mIoU absoluto não necessariamente se traduz em melhor adequação para aplicações industriais, onde estabilidade, eficiência e robustez são frequentemente mais valiosas que ganhos marginais em precisão.

É importante deixar explícito que o modelo proposto não busca maximizar mIoU absoluto, mas sim otimizar a relação custo-desempenho considerando múltiplos objetivos simultaneamente. Esta filosofia de *design* reconhece que em aplicações industriais, ganhos

marginais em precisão frequentemente não justificam aumentos proporcionais em custo computacional, complexidade de implantação ou variabilidade operacional.

O método proposto é mais adequado para cenários em que (i) custo de anotação é uma consideração crítica, (ii) aplicações requerem processamento em tempo real com recursos limitados, (iii) detecção de presença/ausência e localização aproximada são suficientes para as tarefas operacionais, e (iv) robustez e estabilidade sob condições adversas (ruído, compressão, iluminação variável) são prioritárias; e não deve ser priorizado quando (i) precisão de contorno é essencial para medição exata de área ou análise morfológica, (ii) aplicações requerem distinção precisa entre ingredientes visualmente muito similares sem contexto adicional, (iii) domínios visualmente distintos sem adaptação, ou (iv) detecção completa de todas as classes presentes é crítica e erros de classificação não são toleráveis. Nestes casos, métodos totalmente supervisionados com anotações densas, refinamento pós-processamento, ou técnicas específicas para os desafios identificados seriam mais apropriados. O método prioriza completude e estabilidade sobre precisão de borda, sendo adequado para monitoramento contínuo e inspeção sistemática, mas não para aplicações onde precisão de contorno é o requisito principal.

6.4 Implicações para Aplicações Industriais

A segmentação semântica desempenha um papel fundamental no controle de qualidade industrial, permitindo inspeção automatizada e objetiva de produtos em linhas de produção. No contexto específico de montagem de sanduíches *fast-food*, a segmentação por ingrediente fornece informações quantitativas sobre composição e distribuição, suportando decisões operacionais críticas para garantia de qualidade, padronização e eficiência. Os resultados obtidos demonstram a viabilidade de *frameworks* WSSS leves e eficientes para inspeção visual em larga escala, preenchendo a lacuna entre modelos acadêmicos e os requisitos rigorosos de custo, velocidade e robustez de implantação industrial em tempo real. Esta Seção discute as implicações práticas do método para aplicações de controle de qualidade em ambientes de produção.

No contexto de controle de qualidade, a segmentação por ingrediente permite estimar (i) presença/ausência de ingredientes específicos, (ii) cobertura aproximada por classe através de estimativa de área, e (iii) distribuição espacial dos insumos na montagem do sanduíche. Essas medidas podem suportar rotinas de auditoria e treinamento, permitindo identificação sistemática de desvios de padrão e fornecendo *feedback* objetivo para operadores. A capacidade de detectar presença/ausência é particularmente valiosa para garantir que ingredientes obrigatórios estejam presentes e que restrições dietéticas sejam respeitadas, enquanto a estimativa de cobertura permite avaliação aproximada de quantidade e distribuição.

As decisões operacionais suportadas incluem (i) treinamento de operadores através

de *feedback* visual objetivo sobre distribuição de ingredientes, (ii) auditoria de qualidade através de análise sistemática de conformidade com padrões estabelecidos, e (iii) padronização de processos através de identificação de desvios e ajustes em tempo real. A estabilidade do modelo, evidenciada pelo pequeno desvio padrão entre *folds* (0,6% para segmentação IoU, conforme apresentado na Seção 5.2), é crucial para aplicações industriais onde desempenho previsível e robustez a mudanças de dados permitem implantação mais confiável.

Os requisitos de implantação são favoráveis para ambientes industriais: (i) latência de inferência permite processamento em tempo real, (ii) consumo mínimo de VRAM permite implantação em *hardware* acessível, e (iii) estabilidade entre *folds* indica robustez a mudanças na distribuição de dados, essencial para ambientes de produção onde condições podem variar. A eficácia do método em mitigar desequilíbrio de classes através de amostragem adaptativa e refinamento de pseudo-rótulos é particularmente relevante para aplicações industriais, onde distribuições de dados podem variar ao longo do tempo ou entre diferentes linhas de produção.

A viabilidade de implantação industrial (*industrial deployability*) do método proposto é determinada por múltiplos fatores além de métricas de desempenho isoladas. Em relação ao custo de anotação, o método reduz significativamente o investimento inicial necessário para implantação, permitindo que empresas escalem sistemas de inspeção visual sem o custo proibitivo de anotação densa. Conforme demonstrado na Seção 5.9, a abordagem fracamente supervisionada é aproximadamente 6,1 vezes mais eficiente em termos de mIoU por hora de anotação, representando economia de mais de 1.100 horas de trabalho humano enquanto alcança 89% do desempenho do modelo totalmente supervisionado. Esta eficiência de anotação é crítica para adaptação rápida a novos domínios e escalabilidade em ambientes industriais onde múltiplos produtos ou linhas de produção podem requerer segmentação.

Quanto à estabilidade entre *folds*, o método proposto apresenta desvio padrão substancialmente menor ($0,6\%$ versus $6,8\%$ do DuPL) entre partições de validação cruzada, indicando robustez superior a mudanças na distribuição de dados. Esta estabilidade é fundamental para aplicações industriais onde desempenho previsível e consistência operacional são essenciais para implantação confiável, permitindo que o sistema mantenha desempenho consistente mesmo quando condições operacionais variam.

Em relação ao consumo de memória (VRAM), o método proposto apresenta consumo mínimo (185,9 MB com *batch*=1, conforme Tabela 15), permitindo implantação em *hardware* acessível ou processamento paralelo de múltiplas instâncias. Esta eficiência de memória reduz custos de infraestrutura e permite escalabilidade em ambientes industriais onde múltiplas linhas de produção podem requerer processamento simultâneo.

Quanto à viabilidade de treinamento contínuo (*continuous training*), a simplificação consciente do DuPL, removendo módulos computacionalmente pesados como Adaptive

Noise Filtering (ANF), torna o método adequado para adaptação em produção. A arquitetura com menor custo computacional permite que o modelo seja retreinado ou ajustado incrementalmente conforme novas condições operacionais surgem, sem requisitos proibitivos de *hardware*. Esta capacidade de adaptação contínua é essencial para ambientes industriais onde condições podem variar ao longo do tempo (mudanças sazonais, novos produtos, variações de iluminação) e onde métodos que requerem retreinamento completo seriam impraticáveis.

Embora desenvolvido especificamente para montagem de sanduíches *fast-food*, o método proposto pode ser generalizado para outros cenários de inspeção visual baseados em CCTV com características similares. A abordagem é adequada para aplicações de inspeção visual industrial onde imagens de baixa resolução, dados ruidosos e desequilíbrio de classes são características comuns. Especificamente, o método pode ser adaptado para: (i) inspeção visual industrial em linhas de produção onde múltiplos componentes devem ser identificados e segmentados; (ii) controle de qualidade baseado em CCTV onde condições operacionais não são controladas e imagens apresentam compressão e variações de iluminação; e (iii) ambientes ruidosos e de baixa resolução onde métodos totalmente supervisionados seriam proibitivos devido ao custo de anotação. A generalização para estes cenários requer adaptação das estratégias de aumento de dados e refinamento de pseudo-rótulos para características específicas do novo domínio, mas a arquitetura fundamental e os princípios metodológicos permanecem aplicáveis. Esta generalização demonstra a robustez do método e sua adequação para uma ampla gama de aplicações industriais além do contexto específico de *fast-food*.

É importante destacar que a abordagem proposta é adequada especificamente para monitoramento contínuo e inspeção sistemática em linhas de produção, não para inspeção pontual de alta precisão onde métodos totalmente supervisionados com refinamento pós-processamento seriam mais apropriados. O método é projetado para fornecer *feedback* contínuo e consistente sobre qualidade e conformidade, permitindo identificação proativa de desvios e ajustes em tempo real, características essenciais para ambientes de produção onde volume e velocidade são críticos.

Os riscos operacionais envolvidos devem ser considerados cuidadosamente, pois diferentes tipos de erro têm consequências distintas em ambientes de produção. Falsos positivos, onde o sistema detecta incorretamente a presença de ingredientes ausentes, podem resultar em desperdício através de rejeição incorreta de produtos que na verdade estão conformes. Este tipo de erro impacta diretamente a eficiência operacional e os custos de produção. Por outro lado, falsos negativos, onde o sistema falha em detectar ingredientes presentes, podem resultar em produtos não conformes chegando aos clientes, gerando reclamações e problemas de qualidade que podem comprometer a reputação da marca. O método privilegia completude sobre precisão de limites, o que ajuda a reduzir falsos negativos, mas pode aumentar falsos positivos em alguns casos. A escolha entre

estes trade-offs deve ser feita com base nas prioridades operacionais específicas de cada aplicação, considerando o custo relativo de cada tipo de erro no contexto operacional particular.

É fundamental deixar explícito que o sistema proposto apoia a tomada de decisão humana, não substitui a supervisão humana em processos críticos. O método fornece informações quantitativas e objetivas sobre composição e distribuição de ingredientes, permitindo que operadores e supervisores tomem decisões informadas com base em dados objetivos. Em cenários onde decisões críticas de qualidade ou segurança estão envolvidas, a supervisão humana permanece essencial, com o sistema atuando como ferramenta de apoio que aumenta a eficiência e consistência da inspeção, mas não elimina a necessidade de julgamento humano em casos ambíguos ou críticos.

Em síntese, os resultados demonstram a viabilidade do método proposto para implantação em ambientes reais e não controlados, onde condições operacionais variam e restrições de recursos são presentes. A combinação de eficiência computacional, estabilidade operacional e adequação para supervisão fraca torna o método uma solução prática e escalável para inspeção visual em larga escala, preenchendo uma lacuna importante entre modelos acadêmicos e necessidades práticas de produção industrial.

6.5 Desafios Enfrentados e Abordagens

O desenvolvimento do método proposto envolveu desafios práticos e metodológicos significativos, refletindo a complexidade inerente de aplicar segmentação semântica fracamente supervisionada em ambientes industriais reais. Estes desafios surgiram tanto das características específicas do domínio (imagens CFTV de baixa qualidade, desequilíbrio de classes, oclusões) quanto das limitações da supervisão fraca (pseudo-rótulos ruidosos, confirmation bias, instabilidade). Esta Seção apresenta uma análise sistemática desses desafios, as abordagens tentadas, o que funcionou, o que não funcionou, e as lições aprendidas que podem informar futuras pesquisas e aplicações.

6.5.1 Baixa Resolução e Compressão

Desafio: As imagens CFTV apresentam baixa resolução (tipicamente 448×448 pixels ou menores) combinada com artefatos de compressão JPEG e condições de iluminação variáveis, apresentando desafios significativos para a geração de pseudo-rótulos confiáveis.

O que tentou: Inicialmente, tentou-se aplicar técnicas de super-resolução para aumentar a resolução das imagens antes do processamento. Também foram testadas estratégias de aumento de dados que simulavam compressão e variações de iluminação.

O que funcionou: A combinação de TTA e DTA foi essencial para estabilizar o processo de geração de pseudo-rótulos sob essas condições adversas. O TTA, através

de múltiplas transformações geométricas e fotométricas, permite que o modelo aprenda representações robustas que são invariantes a variações de iluminação e compressão. O DTA, através de limiarização adaptativa baseada em confiança, permite que o modelo seja inicialmente conservador e progressivamente mais permissivo, mitigando o impacto de ruído visual nos estágios iniciais do treinamento.

O que não funcionou: Técnicas de super-resolução aumentaram significativamente o tempo de processamento sem melhorias proporcionais em desempenho, tornando-se impraticáveis para aplicações em tempo real. Estratégias de pré-processamento agressivo para remover artefatos de compressão também não melhoraram o desempenho, sugerindo que o modelo aprende a lidar com esses artefatos diretamente.

Lição aprendida: Em vez de tentar remover ou corrigir artefatos de baixa qualidade, é mais eficaz adaptar o modelo para lidar com essas condições diretamente através de estratégias de treinamento robustas e aumento de dados que simulam as condições adversas.

6.5.2 Oclusão por Mão/Ingredientes

Desafio: Durante a montagem de sanduíches, as mãos dos operadores frequentemente ocluem ingredientes, e ingredientes podem ser parcialmente cobertos por outros, resultando em segmentação incompleta.

O que tentou: Inicialmente, tentou-se filtrar imagens com oclusão significativa durante a seleção de frames. Também foram testadas estratégias de detecção de mãos para mascarar regiões ocluídas.

O que funcionou: A inclusão da classe Mão/Pessoa no conjunto de classes permitiu que o modelo aprenda a identificar e segmentar regiões ocluídas, fornecendo informações úteis sobre quando a segmentação pode ser incompleta. A estratégia de TTA, através de múltiplas visões da mesma imagem, ajuda a mitigar parcialmente o problema de oclusão através de consenso entre diferentes transformações.

O que não funcionou: Filtrar imagens com oclusão reduziu significativamente o tamanho do *dataset* disponível, comprometendo a capacidade do modelo de aprender a lidar com oclusões. Estratégias de mascaramento de regiões ocluídas também não melhoraram o desempenho, sugerindo que é mais eficaz incluir oclusões como parte do domínio de aprendizado.

Lição aprendida: Em vez de tentar evitar ou mascarar oclusões, é mais eficaz incluí-las como parte do domínio de aprendizado, permitindo que o modelo desenvolva estratégias para lidar com essas condições naturalmente.

6.5.3 Desbalanceamento Extremo de Classes

Desafio: O *dataset* VSS apresenta desequilíbrio severo de classes, com classes frequentes como Pão (98%) e classes raras como Picles (3%), exigindo estratégias específicas para garantir aprendizado adequado de todas as classes.

O que tentou: Inicialmente, tentou-se usar pesos de classe na função de perda para balancear a contribuição de cada classe. Também foram testadas estratégias de sub-amostragem de classes frequentes.

O que funcionou: A implementação de um amostrador ponderado que super-amostra classes minoritárias através de repetição controlada e recorte sintético foi fundamental para garantir que todas as classes fossem adequadamente representadas durante o treinamento. A estratégia de balanceamento de pseudo-rótulos, que impõe um limiar de área mínima (3% da imagem) para prevenir sub-representação de classes raras, também foi essencial para garantir detecção adequada de classes minoritárias.

O que não funcionou: Pesos de classe na função de perda não foram suficientes para lidar com o desequilíbrio extremo, resultando em convergência prematura para classes frequentes. Sub-amostragem de classes frequentes reduziu significativamente o tamanho efetivo do *dataset*, comprometendo a capacidade do modelo de aprender representações robustas.

Lição aprendida: Estratégias de amostragem adaptativa que super-amostram classes minoritárias são mais eficazes que ajustes na função de perda para lidar com desequilíbrio extremo, permitindo que o modelo aprenda representações robustas para todas as classes sem comprometer o aprendizado de classes frequentes.

6.5.4 Dificuldade de Pseudo-label no Início do Treino

Desafio: Nos estágios iniciais do treinamento, os pseudo-rótulos gerados são de baixa qualidade devido à incerteza do modelo, resultando em aprendizado ruidoso e possível confirmation bias.

O que tentou: Inicialmente, tentou-se usar apenas rótulos de classificação nos estágios iniciais, introduzindo pseudo-rótulos gradualmente. Também foram testadas estratégias de thresholding mais conservador nos estágios iniciais.

O que funcionou: A estratégia de DTA, através de um agendamento cosseno que decai de $0,70$ para $0,55$ ao longo das épocas, permite que o modelo seja inicialmente conservador na geração de pseudo-rótulos, mitigando o impacto de predições incertas nos estágios iniciais. A estratégia de aprendizado curricular, através de alta ponderação inicial da perda de classificação (20,0) que reduz para 1,0 nos estágios posteriores, também foi essencial para estabelecer uma base sólida antes de confiar em pseudo-rótulos.

O que não funcionou: Usar apenas rótulos de classificação nos estágios iniciais retardou significativamente a convergência, resultando em desempenho inferior. Thresholding

excessivamente conservador também não melhorou o desempenho, sugerindo que é mais eficaz usar uma estratégia progressiva que permite exploração gradual.

Lição aprendida: Uma estratégia progressiva que começa conservadora e gradualmente se torna mais permissiva é mais eficaz que estratégias binárias (usar ou não usar pseudo-rótulos), permitindo que o modelo explore gradualmente regiões inicialmente incertas enquanto mantém estabilidade.

As lições aprendidas durante o desenvolvimento do método fornecem insights valiosos para futuras pesquisas em segmentação semântica fracamente supervisionada para aplicações industriais. A importância de adaptar o modelo para lidar diretamente com condições adversas, em vez de tentar removê-las ou corrigi-las, sugere que estratégias de treinamento robustas são mais eficazes que pré-processamento agressivo. Esta lição está diretamente conectada às decisões metodológicas apresentadas no Capítulo 4, onde estratégias de aumento de dados que simulam condições adversas foram priorizadas sobre técnicas de pré-processamento para remoção de artefatos. A eficácia de amostragem adaptativa sobre ajustes na função de perda para lidar com desequilíbrio extremo indica que estratégias de nível de dados podem ser mais eficazes que estratégias de nível de otimização, justificando a escolha do amostrador ponderado descrito na Seção 4.5 do Capítulo 4. A superioridade de componentes leves e bem integrados sobre regularização excessiva para alcançar estabilidade sugere que robustez arquitetural é mais valiosa que penalização de complexidade, fundamentando a escolha do SegFormer-B0 e da arquitetura de duplo estudante simplificada apresentadas na Seção 4.2 do Capítulo 4. Finalmente, a eficácia de estratégias progressivas sobre abordagens binárias para lidar com pseudo-rótulos ruidosos indica que adaptabilidade temporal é essencial para supervisão fraca, validando a implementação do DTA e do aprendizado curricular descritos na Seção 4.4 do Capítulo 4.

Em síntese, os desafios enfrentados durante o desenvolvimento refletem a complexidade do domínio industrial real, onde condições operacionais variam, restrições de recursos são presentes, e supervisão densa é proibitiva. A abordagem de enfrentar estes desafios diretamente, em vez de tentar evitá-los ou simplificá-los, resultou em um método robusto e adequado para aplicações práticas. Estes desafios e as soluções desenvolvidas para enfrentá-los fornecem uma base sólida para futuras pesquisas em segmentação semântica fracamente supervisionada para ambientes industriais.

Este capítulo apresentou os resultados experimentais obtidos com o modelo proposto, incluindo comparações com métodos do estado da arte, estudos de ablação, análises quantitativas e qualitativas, e avaliação de eficiência computacional e custo-desempenho. Além disso, foram discutidas as decisões de design e justificativas, as limitações do método, os trade-offs entre precisão, custo e velocidade, as implicações para aplicações industriais, e os desafios enfrentados durante o desenvolvimento. Os resultados demonstram que o modelo proposto apresenta desempenho comparável ao DuPL enquanto mantém maior estabilidade e eficiência computacional, alcançando 89% do desempenho do modelo to-

talmente supervisionado com uma redução de 85% no custo de anotação. As análises apresentadas fornecem insights importantes sobre quando o método é a melhor escolha e quando alternativas seriam mais apropriadas, fornecendo base sólida para as conclusões apresentadas no próximo capítulo.

Capítulo 7

Conclusões

Neste capítulo são sintetizadas as conclusões sobre o trabalho realizado, destacando as contribuições principais, os resultados alcançados e as limitações reconhecidas.

7.1 Síntese dos Resultados

Os principais resultados quantitativos, detalhados no Capítulo 5, destacam-se da seguinte forma. Em comparação com métodos WSSS no *dataset* VSS, o modelo proposto atinge 43,9% de mIoU de segmentação (DuPL 41,2%; SEAM reporta apenas CAM mIoU de 28,3%) e apresenta maior estabilidade entre *folds* (desvio padrão de 0,6% versus 6,8% do DuPL). Frente ao SegFormer-B0 totalmente supervisionado, o método alcança 89,2% do mIoU (43,9% versus 49,2%), com redução de custo de anotação superior a 85% e eficiência de mais de 6× em mIoU por hora de anotação. A eficiência computacional permite inferência a mais de 1.100 FPS, consumo de VRAM de aproximadamente 186 MB (*batch*=1) e viabilidade de execução em CPU, adequando o método a ambientes industriais com restrições de *hardware*.

Este trabalho apresentou um *framework* eficiente de segmentação semântica fracamente supervisionada para inspeção visual em nível de ingrediente na montagem de sanduíches *fast-food*. A integração de componentes leves, incluindo Aumento no Tempo de Teste, aumento de dados informado por domínio e um amostrador multi-rótulo melhorado, em um *pipeline* de duplo estudante coeso, permitiu que o método alcançasse um equilíbrio superior de desempenho e praticidade.

Os objetivos específicos propostos foram atendidos da seguinte forma. O objetivo 1 (avaliar quantitativamente os principais modelos supervisionados e fracamente supervisionados aplicados na segmentação de ingredientes em imagens de *fast-food*) foi atendido por meio da revisão bibliográfica do Capítulo 3, onde foram observados e relatados os pontos

fracos e fortes de cada abordagem, e da avaliação quantitativa comparativa apresentada no Capítulo 5. O objetivo 2 (organizar uma base de imagens para o desenvolvimento do modelo de Segmentação Semântica) foi atendido por meio da organização do *dataset Visio Sandwich Segmentation* (VSS), detalhado no Capítulo 4, contendo 17.658 imagens com anotações em nível de imagem e pixel. O objetivo 3 (investigar estratégias específicas de aumento de dados (*data augmentation*) para o problema em estudo) foi atendido por meio da investigação e implementação de técnicas de aumento de dados informadas por domínio, incluindo transformações geométricas, ajustes fotométricos e simulação de oclusão, apresentadas no Capítulo 4. O objetivo 4 (propor um modelo de *Deep Learning* para Segmentação Semântica de imagens de *fast-food*) foi atendido por meio da proposta de uma arquitetura de duplo estudante baseada em SegFormer-B0, detalhada no Capítulo 4. O objetivo 5 (entender a influência do tamanho e balanceamento dos conjuntos de imagens de treinamento no desempenho do modelo proposto) foi atendido por meio de experimentos de ablação apresentados no Capítulo 5, que investigaram o impacto de estratégias de balanceamento, incluindo o amostrador ponderado para lidar com desequilíbrio de classes. O objetivo 6 (avaliar o desempenho de segmentação do modelo desenvolvido usando métricas quantitativas, como mIoU, precisão, revocação e F1-Score) foi atendido por meio da avaliação exaustiva apresentada no Capítulo 5, utilizando as métricas quantitativas especificadas.

O comportamento observado indica que a abordagem proposta apresenta tendência consistente de desempenho sob diferentes partições dos dados, mantendo robustez frente a variações na distribuição de treinamento. A baixa variabilidade entre *folds* de validação cruzada demonstra que o modelo não é sensível a mudanças na composição do conjunto de dados, característica essencial para aplicações industriais onde a distribuição pode variar ao longo do tempo ou entre linhas de produção. O método exibe adequação ao cenário industrial através de segmentações mais completas em casos desafiadores como ingredientes sobrepostos, embora permaneça limitado em detalhes finos e ingredientes translúcidos.

A viabilidade prática do *framework* é evidenciada por um custo-benefício favorável, alcançando desempenho próximo ao de modelos totalmente supervisionados enquanto reduz substancialmente os custos de anotação. A eficiência computacional observada permite inferência em tempo real, adequando o método para implantação em ambientes industriais com restrições de *hardware*. Além disso, o modelo permite inferência em CPU, o que reduz os requisitos de *hardware* e torna viável a implantação em produção mesmo em máquinas sem GPU, facilitando o uso em ambiente industrial. A combinação de eficiência de anotação, desempenho competitivo e eficiência computacional estabelece a viabilidade da supervisão fraca para aplicações industriais de controle de qualidade.

7.2 Contribuições Principais

As principais contribuições deste trabalho são:

1. **Propor um *framework* WSSS para controle de qualidade em *fast-food*:** Introdução de um *framework* de segmentação semântica fracamente supervisionada adaptado para controle de qualidade de sanduíches *fast-food*, enfatizando eficiência de anotação e desempenho em tempo real. O método alcança 89,2% do desempenho do modelo totalmente supervisionado (43,9% versus 49,2% mIoU) enquanto reduz o custo de anotação em mais de 85%, demonstrando viabilidade prática para aplicações industriais onde custo e escalabilidade são críticos.

2. **Propor arquitetura de duplo estudante simplificada com estratégia de aprendizado curricular:** Arquitetura de duplo estudante baseada em SegFormer-B0 que reduz sobrecarga computacional e uso de memória em relação a métodos do estado da arte, especificamente através da simplificação consciente do DuPL removendo módulos computacionalmente pesados como Adaptive Noise Filtering (ANF). A arquitetura é combinada com estratégia de aprendizado curricular que utiliza pesos adaptativos de perda para ativar progressivamente diferentes componentes de perda em estágios específicos do treinamento, melhorando estabilidade e desempenho final. Esta simplificação resulta em eficiência computacional superior (mais de 1.100 FPS versus 757 FPS do DuPL com *batch*=8) e consumo reduzido de VRAM (1.191 MB versus 1.781 MB do DuPL), tornando o método adequado para treinamento contínuo e adaptação em produção.

3. **Fornecer *benchmark* abrangente de WSSS baseado em Transformers:** Fornecimento de um *benchmark* abrangente de WSSS baseado em Transformers em um *dataset* industrial de alimentos desafiador e de baixa resolução (VSS), demonstrando a adequação do *backbone* SegFormer e dos componentes propostos neste domínio. O método supera DuPL em 2,7 pontos percentuais em segmentação mIoU ($43,9 \pm 0,6\%$ versus $41,2 \pm 6,8\%$) e supera significativamente SEAM ($28,3 \pm 3,8\%$ CAM mIoU), demonstrando superioridade de arquiteturas baseadas em Transformers sobre métodos clássicos baseados em CNN no domínio de alimentos. A estabilidade superior entre *folds* (desvio padrão de $0,6\%$ versus $6,8\%$ do DuPL) indica robustez superior a mudanças na distribuição de dados, essencial para aplicações industriais. Adicionalmente, o modelo foi validado no *dataset* público FoodSeg103, alcançando 16,40% de mIoU e superando SSDB-II (14,79%), conforme Seção 5.10.

4. **Realizar análise abrangente de custo-desempenho:** Realização de uma análise abrangente de custo-desempenho que quantifica que o método proposto alcança 89% do desempenho totalmente supervisionado enquanto é mais de 6 vezes mais eficiente em termos de mIoU por hora de anotação ($E_w = 0,2286$ versus $E_s =$

$0,0377$, razão de $6,1\times$), e opera a mais de 1.100 FPS. Esta análise demonstra a viabilidade econômica da supervisão fraca para aplicações industriais, estabelecendo métricas de eficiência para comparação com métodos totalmente supervisionados e quantificando a economia de mais de 1.100 horas de trabalho humano enquanto mantém desempenho competitivo.

7.3 Limitações Reconhecidas

O modelo proposto apresenta limitações inerentes ao problema de segmentação semântica fracamente supervisionada em imagens industriais de alimentos. Estas limitações não decorrem de falhas de implementação, mas das características intrínsecas do domínio e das restrições da supervisão fraca.

Limitações técnicas: O modelo não resolve a segmentação de classes raras e ingredientes com texturas translúcidas ou suaves, ou com área pequena na imagem. A precisão de limites é limitada para ingredientes de granulação fina ou baixo contraste, como fatias de cebola ou fios de queijo ralado. O modelo não é adequado quando a precisão de contorno é crítica para a aplicação.

Limitações metodológicas: O modelo não é adequado para cenários com alta complexidade composicional, onde múltiplos ingredientes sobrepostos aparecem simultaneamente, devido à resolução espacial limitada dos CAMs gerados a partir de supervisão em nível de imagem. A dependência do módulo de segmentação na saída de classificação para suprimir ativações falsas leva à omissão completa de classes quando a predição de classificação é incorreta, limitando a robustez em casos de classificação errônea.

Limitações de dados: O *dataset* utilizado é proprietário e não está publicamente disponível, limitando a reprodutibilidade direta. Foi fornecida uma descrição detalhada de sua composição, protocolo de anotação e distribuição de classes para garantir transparência e facilitar replicação aproximada.

7.4 Impacto e Aplicabilidade

O método proposto é mais adequado para cenários em que (i) custo de anotação é uma consideração crítica, (ii) aplicações requerem processamento em tempo real com recursos limitados, (iii) detecção de presença/ausência e localização aproximada são suficientes para as tarefas operacionais, e (iv) robustez e estabilidade sob condições adversas (ruído, compressão, iluminação variável) são prioritárias; e não deve ser priorizado quando (i) precisão de contorno é essencial para medição exata de área ou análise morfológica, (ii) aplicações requerem distinção precisa entre ingredientes visualmente muito similares sem contexto adicional, (iii) domínios visualmente distintos sem adaptação, ou (iv) detecção completa de todas as classes presentes é crítica e erros de classificação não são toleráveis.

O *framework* proposto foi desenvolvido para apoiar sistemas de controle de qualidade em ambientes industriais, não para substituir inspeção humana. O sistema fornece segmentações que auxiliam operadores na identificação de problemas de montagem, padronização de processos e documentação de qualidade, mas a decisão final permanece com o operador humano.

O método é adequado para sistemas de monitoramento contínuo em linhas de produção *fast-food*, onde custos de anotação são uma consideração importante e a precisão de contorno não é crítica. A estabilidade do modelo e sua robustez a mudanças de dados são particularmente valiosas para aplicações onde a distribuição de dados varia ao longo do tempo ou entre diferentes linhas de produção, permitindo que o sistema mantenha desempenho consistente sem necessidade de retreinamento frequente. O método é adequado para auditoria de qualidade e padronização de processos, onde segmentações aproximadas são suficientes para identificar problemas de montagem.

O método não é adequado para inspeção de alta precisão onde a exatidão de contorno é crítica, nem para ambientes controlados onde anotação pixel-level é viável e métodos totalmente supervisionados podem ser aplicados. O método não substitui inspeção humana em casos onde ingredientes translúcidos, classes raras ou alta complexidade composicional são frequentes.

Os resultados obtidos podem ser generalizados para outros cenários de inspeção visual baseados em CCTV com características similares (baixa resolução, desequilíbrio de classes, custos de anotação elevados), desde que as limitações identificadas sejam adequadas ao contexto de aplicação. As técnicas desenvolvidas podem ser adaptadas para outros domínios industriais onde imagens de baixa resolução, desequilíbrio de classes e custos de anotação são considerações importantes.

7.5 Trabalhos Futuros

As limitações identificadas neste trabalho indicam direções naturais para pesquisa futura:

a) **Refinamento de precisão de limites:** A limitação de precisão de limites, especialmente para ingredientes de granulação fina ou baixo contraste, pode ser abordada através de técnicas leves de refinamento espacial, como propagação de afinidade de pixel local ou consistência entre ativações CAM e saídas da cabeça de segmentação.

b) **Redução de dependência de classificação:** A dependência do módulo de segmentação na saída de classificação, que leva à omissão completa de classes quando a classificação é incorreta, pode ser reduzida através de arquiteturas que permitam segmentação mesmo quando a classificação é incerta ou incorreta.

c) **Treinamento contínuo e adaptação em produção:** O desenvolvimento de estratégias para treinamento contínuo (*continuous training*) e adaptação incremental

em ambientes de produção representa uma direção importante para pesquisa futura. Embora a simplificação consciente do DuPL tenha tornado o método mais adequado para adaptação em produção ao remover módulos computacionalmente pesados, estratégias específicas para atualização incremental do modelo conforme novas condições operacionais surgem (mudanças sazonais, novos produtos, variações de iluminação) podem melhorar ainda mais a viabilidade de implantação industrial. Isto inclui desenvolvimento de mecanismos para detecção de *drift* de dados, estratégias de amostragem adaptativa para seleção de exemplos para retreinamento, e protocolos para atualização incremental que minimizem interrupção operacional.

A integração dessas estratégias pode melhorar a robustez, eficiência e generalização do modelo, abordando diretamente as limitações técnicas e metodológicas identificadas neste trabalho.

Referências

AHN, J.; KWAK, S. Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2018. p. 4981–4990.

BOSSARD, L.; GUILLAUMIN, M.; GOOL, L. V. Food-101 – mining discriminative components with random forests. In: **European Conference on Computer Vision**. [S.l.: s.n.], 2014. p. 446–461.

CAI, Q.; ABHAYARATNE, C. Ssdb-net: A single-step dual branch network for weakly supervised semantic segmentation of food images. In: IEEE. **2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)**. [S.l.], 2023. p. 1–6.

CARVALHO, M. R. d.; SACILOTTI, A.; FERRARI, R. J. Weakly supervised semantic segmentation for fast-food quality control: A dual-student approach with segformer. **IEEE Access**, 2025. Em análise após 1ª revisão.

CHEN, L.-C. et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 40, n. 4, p. 834–848, 2017.

_____. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: FERRARI, V. et al. (Ed.). **Computer Vision – ECCV 2018**. Cham: Springer International Publishing, 2018. p. 833–851. ISBN 978-3-030-01234-2.

CHEN, T. et al. A simple framework for contrastive learning of visual representations. In: **Proceedings of the 37th International Conference on Machine Learning**. [S.l.]: JMLR.org, 2020. (ICML'20).

CORDTS, M. et al. The cityscapes dataset for semantic urban scene understanding. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 3213–3223.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2009. p. 248–255.

DOSOVITSKIY, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: **International Conference on Learning Representations**. [s.n.], 2021. Disponível em: <https://openreview.net/forum?id=YicbFdNTTy>.

EVERINGHAM, M. et al. **The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results**. 2012. Http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

FENG, D. et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 22, n. 3, p. 1341–1360, 2020.

FREITAS, C. N. C.; CORDEIRO, F. R.; MACARIO, V. Myfood: A food segmentation and classification system to aid nutritional monitoring. In: **2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2020. p. 234–239.

GONCALVES, J. P. et al. Deep learning architectures for semantic segmentation and automatic estimation of severity of foliar symptoms caused by diseases or pests. **Biosystems engineering**, Elsevier, v. 210, p. 129–142, 2021.

GOU, J. et al. Knowledge distillation: A survey. **International Journal of Computer Vision**, Springer, v. 129, n. 6, p. 1789–1819, 2021.

GRIFFIN, G. et al. **Caltech-256 object category dataset**. [S.l.], 2007.

HAFIZ, A. M.; BHAT, G. M. A survey on instance segmentation: state of the art. **International journal of multimedia information retrieval**, Springer, v. 9, n. 3, p. 171–189, 2020.

HAO, S.; ZHOU, Y.; GUO, Y. A brief survey on semantic segmentation with deep learning. **Neurocomputing**, Elsevier, v. 406, p. 302–321, 2020.

HE, K. et al. Masked autoencoders are scalable vision learners. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 16000–16009.

_____. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

JAISWAL, A. et al. A survey on contrastive self-supervised learning. **Technologies**, v. 9, n. 1, 2021. ISSN 2227-7080. Disponível em: <https://www.mdpi.com/2227-7080/9/1/2>.

JUNG, H.; OH, Y. Towards better explanations of class activation mapping. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2021. p. 1336–1344.

K., D. **GradCAM – Enhancing Neural Network Interpretability in the Realm of Explainable AI**. 2023. Blog post. Accessed: 2024-09-03. Disponível em: <https://learnopencv.com/intro-to-gradcam/>.

KAWANO, Y.; YANAI, K. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: **Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)**. [S.l.: s.n.], 2014.

KIRILLOV, A. et al. Panoptic segmentation. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2019. p. 9404–9413.

_____. Segment anything. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2023. p. 4015–4026.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017.

KWEON, H.; YOON, K.-J. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2024. p. 19499–19509.

LAN, X. et al. Foodsam: Any food segmentation. **IEEE Transactions on Multimedia**, IEEE, 2023.

LATEEF, F.; RUICHEK, Y. Survey on semantic segmentation using deep learning techniques. **Neurocomputing**, Elsevier, v. 338, p. 321–348, 2019.

LECUN, Y. et al. Handwritten digit recognition with a back-propagation network. **Advances in neural information processing systems**, v. 2, 1989.

_____. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LI, X. et al. Weakly Supervised Food Image Segmentation via Multi-Scale Feature Fusion. In: **Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2022. p. 123–132.

LIN, M. Network in network. **arXiv preprint arXiv:1312.4400**, 2013.

LIN, T.-Y. et al. Microsoft coco: Common objects in context. In: SPRINGER. **Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13**. [S.l.], 2014. p. 740–755.

MATSUDA, Y.; HOASHI, H.; YANAI, K. Recognition of multiple-food images by detecting candidate regions. In: IEEE. **2012 IEEE international conference on multimedia and expo**. [S.l.], 2012. p. 25–30.

MEYERS, A. et al. Im2calories: towards an automated mobile vision food diary. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2015. p. 1233–1241.

MILIOTO, A.; LOTTES, P.; STACHNISS, C. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In: IEEE. **2018 IEEE international conference on robotics and automation (ICRA)**. [S.l.], 2018. p. 2229–2235.

OKAMOTO, K.; YANAI, K. Uec-foodpix complete: A large-scale food image segmentation dataset. In: SPRINGER. **Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V**. [S.l.], 2021. p. 647–659.

OUASSIT, Y. et al. A brief survey on weakly supervised semantic segmentation. **International Journal of Online & Biomedical Engineering**, v. 18, n. 10, 2022.

RAJU, V. B.; IMTIAZ, M. H.; SAZONOV, E. Food Image Segmentation Using Multi-Modal Imaging Sensors with Color and Thermal Data. **Sensors**, v. 23, n. 2, p. 560, 2023.

RHEUDE, T. et al. Leveraging cam algorithms for explaining medical semantic segmentation. **Machine Learning for Biomedical Imaging**, v. 2, p. 2089–2102, 2024. Disponível em: <https://melba-journal.org/2024:023>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18**. [S.l.], 2015. p. 234–241.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986.

SALVADOR, A. et al. Learning cross-modal embeddings for cooking recipes and food images. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 3020–3028.

SCHUHMANN, C. et al. Laion-5b: An open large-scale dataset for training next generation image-text models. **arXiv preprint arXiv:2210.08402**, 2022.

SELVARAJU, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In: . [S.l.]: Springer, 2020. v. 128, p. 336–359.

SHARMA, U.; ARTACHO, B.; SAVAKIS, A. Gourmetnet: Food segmentation using multi-scale waterfall features with spatial and channel attention. **Sensors**, MDPI, v. 21, n. 22, p. 7504, 2021.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, Springer, v. 6, n. 1, p. 1–48, 2019.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

TAKAYUKI, Y. et al. Training data augmentation for semantic segmentation of food images using deep learning. In: . [S.l.: s.n.], 2022. v. 27, p. 339–342.

TORREY, L.; SHAVLIK, J. Transfer learning. In: **Handbook of research on machine learning applications and trends: algorithms, methods, and techniques**. [S.l.]: IGI global, 2010. p. 242–264.

TOUVRON, H. et al. Training data-efficient image transformers & distillation through attention. In: PMLR. **International conference on machine learning**. [S.l.], 2021. p. 10347–10357.

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VLACHOPOULOU, V.; SARAFIS, I.; PAPADOPOULOS, A. Food image classification and segmentation with attention-based multiple instance learning. In: IEEE. **2023 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP) 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP 2023)**. [S.l.], 2023. p. 1–5.

WANG, Y. et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2020. p. 12275–12284.

_____. Weakly supervised food image segmentation using class activation maps. In: **2017 IEEE International Conference on Image Processing (ICIP)**. [S.l.: s.n.], 2017. p. 1277–1281.

WIGHTMAN, R. **PyTorch Image Models**. 2021. <https://github.com/huggingface/pytorch-image-models>. Accessed: 2024-01-15.

WU, H. et al. Cvt: Introducing convolutions to vision transformers. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2021. p. 22–31.

WU, X. et al. A large-scale benchmark for food image segmentation. In: **Proceedings of ACM international conference on Multimedia**. [S.l.: s.n.], 2021.

WU, Y. et al. DuPL: Dual Student with Trustworthy Progressive Learning for Robust Weakly Supervised Semantic Segmentation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2024. p. 3534–3543.

WU, Z.; SHEN, C.; HENGEL, A. V. D. Wider or deeper: Revisiting the resnet model for visual recognition. **Pattern recognition**, Elsevier, v. 90, p. 119–133, 2019.

XIE, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In: **Advances in Neural Information Processing Systems (NeurIPS 2021)**. [S.l.: s.n.], 2021. p. 12077–12090.

YANG, Z. et al. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2024. p. 3606–3615.

YEGNANARAYANA, B. **Artificial neural networks**. [S.l.]: PHI Learning Pvt. Ltd., 2009.

ZAIGRAJEW, V.; ZIEBA, M. Contrastive learning for multi-label classification. In: **36th Conference on Neural Information Processing Systems (NeurIPS 2022).** [S.l.: s.n.], 2022.

ZHAI, X. et al. Scaling vision transformers. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 12104–12113.

ZHANG, P.; WU, M. Multi-label supervised contrastive learning. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 38, n. 15, p. 16786–16793, Mar. 2024. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/29619>.

ZHANG, Y. et al. Transcam: Transformer attention-based class activation mapping for weakly supervised semantic segmentation. In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. [S.l.: s.n.], 2022. p. 1070–1079.

ZHENG, S. et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2021. p. 6881–6890.

ZHOU, B. et al. Learning deep features for discriminative localization. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 2921–2929.

ZHOU, X.; KOLTUN, V.; KRÄHENBÜHL, P. Simple multi-dataset detection. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 7571–7580.

ZHU, F. et al. The use of mobile devices in aiding dietary assessment and evaluation. **IEEE journal of selected topics in signal processing**, IEEE, v. 4, n. 4, p. 756–766, 2010.

ZHU, L. et al. Deep learning and machine vision for food processing: A survey. **Current Research in Food Science**, v. 4, n. 3, p. 233–249, 2021.