Uma Investigação sobre Técnicas para o Treinamento de Detectores de Objetos de Um Estágio em Cenários de Escassez de Rótulos

João Victor Dias Sobrinho

Projeto de Graduação apresentado ao Curso de Engenharia Eletrônica e de Computação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Miguel Elias Mitre Campista

Rio de Janeiro

Março de 2026

Uma Investigação sobre Técnicas para o Treinamento de Detectores de Objetos de Um Estágio em Cenários de Escassez de Rótulos

João Victor Dias Sobrinho

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO ELETRÔNICO E DE COMPUTAÇÃO

Autor:

João Victor Dias Sobrinho

Orientador:

Prof. Miguel Elias Mitre Campista, D.Sc.

Examinador:

Prof. Luís Henrique Maciel Kosmalski Costa, Dr.

Examinador:

Prof. Rodrigo de Souza Couto, D.Sc.

Rio de Janeiro

Março de 2026

Declaração de Autoria e de Direitos

Eu, *João Victor Dias Sobrinho* CPF *156.533.767-01*, autor da monografia *Uma Investigação sobre Técnicas para o Treinamento de Detectores de Objetos de Um Estágio em Cenários de Escassez de Rótulos*, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na disciplina de Projeto de Graduação da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.

2. Excetuam-se do item 1. eventuais transcrições de texto, figuras, tabelas, conceitos e idéias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.

3. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.

4. O autor pode, excepcionalmente, encaminhar à Comissão de Projeto de Graduação, a não divulgação do material, por um prazo máximo de 01 (um) ano, improrrogável, a contar da data de defesa, desde que o pedido seja justificado, e solicitado antecipadamente, por escrito, à Congregação da Escola Politécnica.

5. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.

6. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.

7. Por ser verdade, firmo a presente declaração.

João Victor Dias Sobrinho

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Escola Politécnica - Departamento de Eletrônica e de Computação
Centro de Tecnologia, bloco H, sala H-217, Cidade Universitária
Rio de Janeiro - RJ CEP 21949-900

Este exemplar é de propriedade da Universidade Federal do Rio de Janeiro, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es).

AGRADECIMENTO

Agradeço aos meus pais por todo o apoio, incentivo e pelos inúmeros sacrifícios realizados ao longo da minha formação, oferecendo-me oportunidades sem as quais eu não teria chegado até aqui.

Estendo esses agradecimentos aos amigos e colegas que conheci na UFRJ, cuja parceria foi de essencial relevância para a minha trajetória acadêmica, seja nos momentos de estudo e colaboração, seja nos momentos de descontração, que também contribuíram para essa caminhada.

Por fim, agradeço a todo o corpo docente da UFRJ, que, ao compartilhar seus conhecimentos, enriqueceu significativamente a minha formação acadêmica. Destaco, em especial, o professor Miguel, meu orientador, que, desde o início da minha graduação, ofereceu-me constante apoio, orientação e valiosos aconselhamentos.

RESUMO

O treinamento de modelos de detecção de objetos tipicamente pressupõe a disponibilidade de grandes volumes de dados com rótulos de alta qualidade. Em implementações realistas, que podem impor restrições quanto à capacidade de produção desses rótulos, a escassez de dados rotulados torna-se um fator que limita a viabilidade do desenvolvimento desses sistemas. Na literatura, diversos trabalhos exploram o uso de técnicas de detecção de objetos semissupervisionada (*Semi-Supervised Object Detection - SSOD*) como uma alternativa, utilizando dados rotulados e não rotulados simultaneamente através da geração de rótulos pelo próprio modelo durante o treinamento. Apesar dos avanços na área, a aplicação de SSOD em modelos de detecção de um estágio, como os modelos da família YOLO, enfrenta desafios, já que os detectores de um estágio, caracterizados por realizarem a localização e a classificação de objetos em uma única passagem pela rede, são particularmente sensíveis ao uso de rótulos ruidosos ou de baixa qualidade. Este trabalho investiga duas abordagens para o treinamento de um modelo de detecção de objetos de um estágio em contexto de escassez de rótulos: técnicas clássicas de SSOD e o uso de um modelo de fundação multimodal pré-treinado, capaz de processar dados textuais e imagens, como gerador de rótulos. Os experimentos com técnicas clássicas de SSOD resultaram em desempenho inferior ao valor de referência obtido utilizando apenas a parcela rotulada dos dados. Por outro lado, a abordagem baseada no uso de modelos de fundação mostrou-se como uma alternativa promissora, resultando em ganhos de até $14,8\%$ de desempenho em cenários de escassez severa de dados rotulados e viabilizando o treinamento mesmo em situações de ausência completa de rotulação manual.

Palavras-Chave: detecção de objetos, modelos de fundação, aprendizado semissupervisionado.

ABSTRACT

The training of object detection models typically relies on the availability of large amounts of high-quality labeled data. In realistic settings, restrictions on access to annotations often lead to label-scarcity scenarios, which may limit the feasibility of developing such systems. Several recent works explore semi-supervised object detection (SSOD) techniques as an alternative to mitigate these challenges by simultaneously leveraging labeled and unlabeled data. Despite advances, applying SSOD techniques to single-stage object detectors, such as those from the YOLO family, is still challenging, since these detectors, which perform object localization and classification in a single pass through the network, are particularly sensitive to noisy, low-quality labels. This work investigates two approaches to training single-stage object detection models under a label-scarcity scenario: classical SSOD techniques and the use of a pre-trained multi-modal foundation model, capable of dealing with textual and visual information, as a label generator. Results suggest that the SSOD is ill-suited for the analyzed scenario, as all techniques produced performances inferior to the supervised baseline, which consists of training with only the labeled data available. In contrast, the usage of foundation models as label generators has emerged as a promising alternative, achieving performance improvements of up to 14.8% in severe labeled data scarcity conditions and enabling effective training even under the complete absence of labeled data.

Key-words: object detection, foundation models, semi-supervised learning.

SIGLAS

BDD100K – Berkeley DeepDrive 100K

COCO - Common Objects in Context

CR - Consistency Regularization

DSAT - Dynamic Self-Adaptive Threshold

EMA - Exponential Moving Average

FCOS - Fully Convolutional One-Stage Object Detection

FL - Federated Learning

IoU - Intersection over Union

KL - Kullback-Leibler

mAP - Mean Average Precision

MSE - Mean Squared Error

NMS - Non Maximum Suppression

PAC - Probably Approximately Correct

PL - Pseudo-Label

SAM - Segment Anything Model

SSD - Single Stage Detector

SSL - Semi-Supervised Learning

SSOD - Semi-Supervised Object Detection

SSFOD - Semi-Supervised Federated Object Detection

VLM - Visual Language Model

VRU - Vulnerable Road User

YOLO - You Only Look Once

Sumário

1 Introdução **1**

2 Fundamentação Teórica **5**

2.1 Detecção de Objetos 5

2.1.1 Detectores de Dois Estágios 6

2.1.2 Detectores de Um Estágio 7

2.2 Aprendizado semissupervisionado 8

2.2.1 Formalização do Problema 9

2.2.2 Premissas 10

2.2.3 Principais Abordagens 10

3 Trabalhos Relacionados **14**

3.1 Aprendizado semissupervisionado para classificação de imagens 14

3.2 Aprendizado semissupervisionado para detecção de objetos 15

3.3 SSOD em Detectores de Um Estágio 17

3.4 Modelos de Fundação para Geração de Pseudo-Rótulos 18

4 Configuração Experimental e Estratégias para Detecção de Objetos com Poucos Rótulos **20**

4.1 Conjunto de Dados, Métricas e Ambiente Experimental 20

4.2 Detecção de Objetos Semissupervisionada 21

4.2.1 Arquitetura do Modelo 21

4.2.2 Estratégias de Treinamento 22

4.2.3 Configuração Experimental 23

4.3 Geração de Rótulos com Modelos de Fundação 24

4.3.1 Modelo SAM3 24

4.3.2 Pipeline de Geração de Pseudo-Rótulos 25
4.3.3 Configuração Experimental 26

5 Avaliação Experimental de Estratégias para Detecção de Objetos com Poucos Rótulos **30**

5.1 Técnicas de Detecção de Objetos Semissupervisionada 30
5.1.1 Comparação de Desempenho de Detecção Máximo 31
5.1.2 Investigação sobre o Desempenho do S4OD 32
5.1.3 Discussão sobre Técnicas de Detecção de Objetos Semissupervisionada 33
5.2 Geração de Rótulos para Detecção de Objetos com o SAM3 34
5.2.1 Comparação entre Estratégias de Treinamento 35
5.2.2 Impacto do Limiar de Confiança no Desempenho de Treinamento 38
5.2.3 Tempo de Execução de Treinamento 40
5.2.4 Análise por Classe 41
5.2.5 Discussão sobre a Geração de Rótulos para Detecção de Objetos com o SAM3 42

6 Conclusões e Trabalhos Futuros **44**

Bibliografia **48**

Lista de Figuras

2.1 Fluxo simplificado de um detector de objetos de dois estágios. Após a extração de características da imagem, o primeiro estágio gera múltiplas regiões de interesse. No segundo estágio, uma única cabeça de detecção é aplicada a cada região proposta, realizando simultaneamente a classificação dos objetos e o refinamento de suas localizações espaciais de maneira repetida. 6

2.2 Fluxo simplificado do funcionamento de um detector de objetos de um estágio. Nesse tipo de modelo a classificação e a localização de todos os objetos da imagem ocorrem com uma única aplicação do mapa de características na cabeça de detecção. 8

2.3 Diagrama de exemplo de uso de pseudo-rotulação. 12

2.4 Diagrama de exemplo de uso de regularização por consistência. 13

3.1 Exemplo ilustrativo do conceito de âncoras em detecção de objetos. A imagem é dividida em uma grade espacial e múltiplas caixas âncora, em vermelho, são consideradas em cada célula. Durante o treinamento, cada objeto é associado à âncora que tem a maior sobreposição com a caixa de referência, em vermelho. 16

4.1 Ilustração conceitual da composição dos conjuntos de dados utilizados em cada configuração experimental. As caixas representam apenas os tipos de dados presentes em cada configuração, não correspondendo a proporções ou quantidades relativas. Caixas azuis correspondem a dados rotulados manualmente, caixas amarelas a dados pseudo-rotulados e caixas cinza a dados descartados. 28

5.1 Comparação do desempenho máximo (mAP 50-95) alcançado por diferentes técnicas clássicas de SSOD e pelo treinamento supervisionado. 31
5.2 Evolução temporal da métrica mAP 50-95 durante o treinamento supervisionado e com a técnica S4OD. 33
5.3 Comparação de desempenho de detecção entre os cenários supervisionado, híbrido simétrico, híbrido total e pseudo-rotulado ($\tau = 0.5$). O eixo inferior refere-se à proporção de dados rotulados manualmente destinada ao treinamento, enquanto o eixo superior indica a proporção de dados pseudo-rotulados utilizados. 35
5.4 Impacto do limiar de confiança τ no desempenho do modelo para diferentes configurações de rotulação considerando o uso de 1%, 2%, 5% e 10% de dados. 39
5.5 Relação entre desempenho (mAP 50-95) e tempo total de treinamento para diferentes distribuições de dados, considerando $\tau = 0.5$. 41
5.6 Mapa de calor relacionando ganho de desempenho de detecção (AP 50-95) por classe ao incluir dados pseudo-rotulados e para diferentes quantidades de dados rotulados manualmente. Comparação entre caso supervisionado e caso híbrido simétrico. 42

Lista de Tabelas

4.1 Configurações utilizadas nos experimentos com pseudo-rótulos gerados pelo SAM3. 27

5.1 Análise de qualidade dos pseudo-rótulos gerados pelo SAM3 em comparação aos rótulos manuais do conjunto BDD100K ($\tau = 0.5$). 37

5.2 Impacto do limiar de confiança τ na densidade de objetos por imagem e número de imagens sem objetos para rótulos manuais e pseudo-rótulos gerados pelo SAM3. 40

Capítulo 1

Introdução

Entre as diversas classes de aplicações viabilizadas ou aprimoradas pelos avanços recentes em algoritmos de aprendizado de máquina e pela evolução de *hardware* especializado, a visão computacional destaca-se por seu amplo impacto em tarefas complexas. Em particular, a detecção de objetos tem sido extensivamente empregada em aplicações de grande relevância, como a identificação de usuários vulneráveis das vias (*Vulnerable Road Users – VRUs*) em sistemas de transporte inteligentes [1]. O desempenho dessas técnicas apresentou avanços significativos com o advento do aprendizado profundo, possibilitando níveis de acurácia antes inatingíveis [2, 3]. Entretanto, esses métodos dependem fortemente da disponibilidade de grandes volumes de dados rotulados, o que representa um desafio para o projeto e implementação realista desses sistemas em diversos cenários de aplicação [4].

Tarefas como a classificação de imagens e detecção de objetos são comumente implementadas por meio de técnicas de aprendizado supervisionado, que envolvem o treinamento de modelos a partir de anotações humanas que servem de referência ao ajuste dos seus parâmetros. Na ausência de dados rotulados, o modelo perde a capacidade de comparar suas predições com os valores esperados, inviabilizando o processo de aprendizado. No caso específico de detecção de objetos, a aquisição de rótulos de alta qualidade é um processo especialmente custoso. Diferentemente do treinamento para classificação, que associa categorias a imagens inteiras, as anotações utilizadas em detecção de objetos devem definir tanto a categoria quanto a região que contém cada objeto de interesse na imagem, podendo

cada imagem apresentar diversos objetos. Em diferentes contextos realistas, como aplicações embarcadas, sistemas distribuídos ou ambientes com coleta contínua de dados, a geração de rótulos torna-se um fator limitante para a viabilização de treinamento ou atualização de modelos. O aprendizado federado (*Federated Learning – FL*) é um exemplo de cenário em que a escassez de rótulos impõe relevante limitação, especialmente no caso de modelos de visão computacional. Isso ocorre porque, no FL, assume-se que o treinamento é realizado localmente nos dispositivos clientes, transferindo para o usuário final a responsabilidade pela coleta e rotulação dos dados.

Diante desse cenário, o aprendizado semissupervisionado (*Semi-Supervised Learning – SSL*) surge como uma alternativa promissora para reduzir a dependência de dados rotulados. Essas técnicas exploram simultaneamente dados rotulados e não rotulados durante o treinamento, buscando extrair informação adicional a partir da estrutura inerente aos dados não anotados, utilizando o conhecimento obtido na parcela rotulada como um guia para a aprendizagem. No contexto da detecção de objetos, essa abordagem é conhecida como detecção de objetos semissupervisionada (*Semi-Supervised Object Detection – SSOD*), e apresenta desafios adicionais quando comparada à classificação, devido à necessidade de estimar não somente a classe, mas também a localização espacial dos objetos.

No contexto da detecção de objetos semissupervisionada, os modelos de detecção utilizados geralmente pertencem a duas classes principais: detectores de dois estágios e detectores de um estágio. Nos detectores de dois estágios, o processo de detecção envolve dois procedimentos distintos. Primeiro, é feita uma etapa de identificação de regiões de interesse, onde o modelo detecta possíveis regiões candidatas a conterem objetos. A partir dessas regiões, o detector inicia uma segunda fase, em que o modelo refina o posicionamento das caixas delimitadoras e atribui uma classe ao objeto localizado. Os detectores de um estágio, por outro lado, não exigem uma etapa explícita de geração de regiões de interesse, produzindo as caixas e classes para todos os objetos da imagem de uma só vez. Essa diferença promove menores tempos de inferência aos detectores de um estágio quando comparados aos de dois estágios, mas esse ganho de velocidade tipicamente acompanha uma redução de

desempenho de detecção.

Apesar do relevante avanço das técnicas de SSOD, diversos trabalhos reportam dificuldades na sua aplicação a modelos de detecção de um estágio, como os da família YOLO [5]. Esses modelos, amplamente utilizados em aplicações práticas devido à sua baixa latência e eficiência computacional, mostram-se particularmente sensíveis a erros nos rótulos utilizados no treinamento, o que pode comprometer o processo de treinamento quando técnicas semissupervisionadas clássica são empregadas [6], uma vez que comumente envolvem a geração automática desses rótulos, porém com possível introdução de erros. Em paralelo às abordagens baseadas em SSOD, surgiram os modelos de fundação multimodais, capazes de integrar conhecimento semântico aprendido em larga escala com a capacidade de executar tarefas de visão computacional. Modelos desse tipo, como o Segment Anything Model 3 (SAM3) [7], permitem a geração de anotações automáticas a partir de instruções em linguagem natural, mesmo na ausência de treinamento específico para o conjunto de dados alvo. Esse avanço abre a possibilidade de utilizar tais modelos como geradores de pseudo-rótulos de maior qualidade, potencialmente mitigando algumas das limitações observadas nas abordagens tradicionais de SSOD.

Neste contexto, este trabalho investiga estratégias para o treinamento de modelos de detecção de objetos de um estágio em cenários de escassez de dados rotulados. São analisadas tanto técnicas de detecção semissupervisionada clássicas da literatura, como pseudo-rotulação e regularização por consistência, quanto o uso de modelos de fundação, em particular o SAM3, como geradores automáticos de pseudo-rótulos. A avaliação das abordagens tem enfoque em modelos da família YOLO, considerando sua relevância em aplicações com restrições de latência e recursos computacionais. Para treinamentos conduzidos, utilizou-se o conjunto de dados BDD100K [8], que é composto por cem mil imagens de câmeras frontais de veículos em ambientes variados, contendo anotações de dez classes. Resultados obtidos com técnicas de SSOD sugerem que essas abordagens não apresentam desempenho suficiente para superar o valor de referência. Por outro lado, o uso do modelo SAM3 como gerador de pseudo-rótulos demonstrou ser uma alternativa promissora tanto para viabilizar o treinamento dos modelos avaliados em cenários de escassez de dados rotulados

manualmente, quanto em casos de completa ausência desses rótulos.

Este trabalho é dividido em seis capítulos. O Capítulo 2 introduz os conceitos relevantes à compreensão do trabalho, como detecção de objetos e aprendizado semissupervisionado. Já o Capítulo 3 apresenta trabalhos relacionados. O Capítulo 4 descreve os experimentos e a metodologia utilizada no desenvolvimento deste trabalho. O Capítulo 5 apresenta e discute os resultados obtidos. Por fim, o Capítulo 6 apresenta uma conclusão acerca do trabalho realizado, apontando possíveis direções futuras.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos necessários para a compreensão do problema de detecção de objetos em cenários de escassez de dados rotulados. Inicialmente, são introduzidos conceitos básicos referentes aos modelos de detecção de objetos, sendo apontadas diferenças entre essa tarefa e a de classificação. Em seguida, são discutidos dois dos principais tipos de modelos de detecção, descrevendo e comparando detectores de um e dois estágios. Adicionalmente, o conceito de aprendizado semissupervisionado é introduzido, acompanhado de sua formalização, premissas e principais abordagens.

2.1 Detecção de Objetos

A detecção de objetos é uma tarefa fundamental da visão computacional que consiste em identificar e localizar objetos de interesse em uma imagem. Diferentemente da classificação, em que uma única classe é atribuída à imagem como um todo, a detecção de objetos envolve a predição simultânea das classes dos objetos presentes na cena e de suas respectivas posições espaciais, tipicamente representadas por caixas delimitadoras (*bounding boxes*). O requisito adicional de localização espacial torna a tarefa de detecção mais complexa do que a classificação.

Com o avanço dos métodos baseados em aprendizado profundo, dois tipos de detectores de objetos passaram a predominar na literatura: os detectores de dois estágios e os detectores de um estágio [9]. Essas abordagens diferem na forma como

organizam o processo de detecção, em particular na relação entre a localização dos objetos e a sua classificação.

2.1.1 Detectores de Dois Estágios

Os detectores de dois estágios foram algumas das primeiras abordagens para a detecção de objetos baseadas em aprendizado profundo [9]. Em seu trabalho seminal, Girshick et al. [3] propuseram a utilização de redes neurais profundas para classificar regiões de interesse extraídas da imagem em uma etapa anterior, separando o processo de detecção de objetos entre um momento de geração de regiões candidatas e outro em que essas regiões são classificadas e suas posições são refinadas. Essa formulação influenciou o paradigma de detecção em dois estágios, em que a precisão da detecção é otimizada por meio do refinamento sucessivo de propostas de regiões de interesse [10, 11].

Ao identificar inicialmente um conjunto de regiões da imagem com potencial de conter objetos, a abordagem de detecção em dois estágios reduz o espaço de busca. Dessa forma, a tarefa executada na segunda etapa é simplificada. Por consequência, após os procedimentos de refinamento e classificação das regiões, o desempenho de detecção desse tipo de modelo é tipicamente superior quando comparado ao dos modelos de um estágio [9]. A Figura 2.1 ilustra de forma simplificada o funcionamento desses detectores.

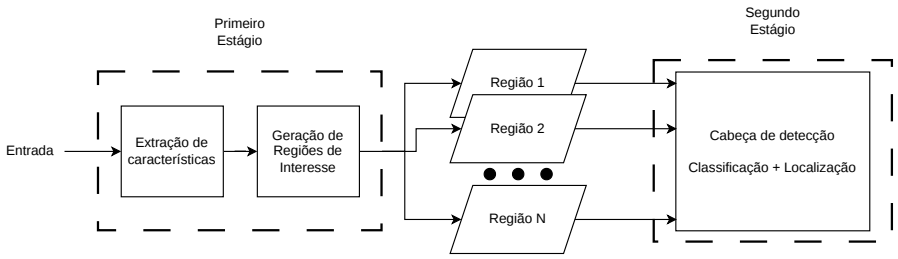

Figura 2.1: Fluxo simplificado de um detector de objetos de dois estágios. Após a extração de características da imagem, o primeiro estágio gera múltiplas regiões de interesse. No segundo estágio, uma única cabeça de detecção é aplicada a cada região proposta, realizando simultaneamente a classificação dos objetos e o refinamento de suas localizações espaciais de maneira repetida.

Entretanto, o ganho de desempenho observado com o uso desse modelo é acompanhado de desafios. Ao realizar a execução sequencial dessas etapas, esse tipo de abordagem implica maior custo computacional e maior latência, uma vez que um número significativo de regiões candidatas deve ser avaliado individualmente e isso pode indicar múltiplas execuções da parte da rede neural referente à cabeça de detecção. Como consequência, embora apresentem elevado desempenho em termos de precisão, detectores de dois estágios tendem a ser menos adequados para aplicações em tempo real ou em cenários com restrições de recursos computacionais, como sistemas embarcados e aplicações veiculares.

2.1.2 Detectores de Um Estágio

Os detectores de um estágio surgiram como alternativa tanto às abordagens de visão computacional clássicas [12, 13] quanto aos modelos de dois estágios [5]. Nessa abordagem, a predição das classes e das caixas delimitadoras é realizada diretamente a partir da imagem de entrada, em uma única etapa, sem a geração explícita de regiões candidatas intermediárias.

O trabalho que inaugura essa classe de detectores com o modelo *You Only Look Once (YOLO)* reformula a tarefa de detecção como um problema de regressão, diferindo das abordagens baseadas em classificação de regiões. Nessas abordagens, regiões candidatas da imagem são inicialmente geradas e posteriormente classificadas individualmente por um modelo de reconhecimento visual. Em contraste, os modelos de um estágio realizam simultaneamente a classificação dos objetos e a estimação de suas localizações espaciais em uma aplicação das características da imagem na cabeça de detecção. A Figura 2.2 apresenta uma visão simplificada do funcionamento desse tipo de modelo, destacando a geração simultânea de classes e posições, sem a necessidade de executar a cabeça de detecção em múltiplas iterações, utilizando diretamente as características extraídas da imagem.

Ao permitir que todo o processo de detecção seja realizado sem o uso de regiões de interesse, os detectores de um estágio apresentam ganhos de desempenho em termos de latência e de redução de custos computacionais e energéticos. Tais características

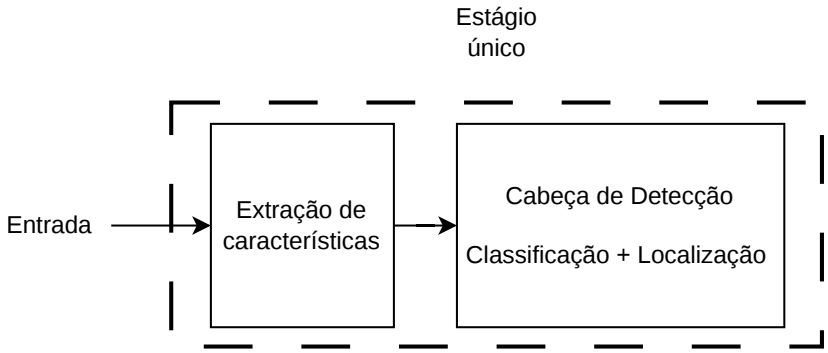

Figura 2.2: Fluxo simplificado do funcionamento de um detector de objetos de um estágio. Nesse tipo de modelo a classificação e a localização de todos os objetos da imagem ocorrem com uma única aplicação do mapa de características na cabeça de detecção.

tornam esses modelos particularmente atraentes para aplicações em que há restrições de recursos ou há demanda de execução em tempo real, como ocorre em cenários veiculares, com dispositivos embarcados e móveis.

Entretanto, as reduções observadas no custo computacional acompanham perdas em termos de qualidade de detecção, especialmente para objetos pequenos [9, 5]. Como a classificação e a localização são realizadas de forma acoplada, erros nos rótulos podem impactar diretamente o processo de otimização, podendo degradar significativamente o desempenho do modelo. Essa sensibilidade torna os detectores de um estágio particularmente desafiadores em cenários de escassez de dados rotulados ou na presença de pseudo-rótulos ruidosos, o que se encaixa no contexto investigado no presente trabalho e que será discutido na seção seguinte.

2.2 Aprendizado semissupervisionado

O paradigma de aprendizado de máquina semissupervisionado (*Semi-Supervised Learning - SSL*) é constituído das estratégias de treinamento utilizadas no cenário em que se tem à disposição tanto dados rotulados quanto não rotulados. Essa configuração é particularmente interessante, porque melhor se aproxima de cenários reais. Para muitos casos de uso, a captura dos dados é trivial, porém a sua rotulação pode

ser desafiadora quando o volume de dados é muito grande, exige conhecimento especialista ou não se pode terceirizar a tarefa devido a preocupações com privacidade.

2.2.1 Formalização do Problema

A análise teórica de problemas de aprendizado supervisionado é frequentemente realizada no modelo de aprendizagem provavelmente aproximadamente correta (*Probably Approximately Correct - PAC*). Nesse *framework*, assume-se que os dados são amostrados de uma distribuição desconhecida $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$. No contexto do aprendizado semissupervisionado, considera-se um conjunto de dados rotulados e um conjunto de dados não rotulados. Define-se $\mathcal{X}$ como o espaço de amostras e $\mathcal{Y}$ como o espaço de rótulos. A distribuição de geração dos dados é $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$, e $\mathcal{D}_{\mathcal{X}}$ é a distribuição marginal sobre as amostras. O conjunto de amostras rotuladas é $\mathcal{S}_{rot} = (x_i, y_i)_{i=1}^{n_{rot}}$, com $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, e $(x_i, y_i) \sim \mathcal{D}$. O conjunto não rotulado é $\mathcal{S}_n = (x_j)_{j=1}^{n_n}$, em que $x_j \in \mathcal{X}$, $x_j \sim \mathcal{D}_{\mathcal{X}}$. Tipicamente, assume-se que o número de amostras rotuladas é muito menor que o de não rotuladas, $n_{rot} \ll n_n$. Dado um espaço de hipóteses $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, o objetivo de um algoritmo de SSL L é usar ambos os conjuntos, $\mathcal{S}_{rot}$ e $\mathcal{S}_n$, para produzir uma hipótese, ou função de predição, $h \in \mathcal{H}$ que generalize melhor do que uma hipótese treinada usando apenas $\mathcal{S}_{rot}$.

Entretanto, a simples existência de dados não rotulados ($\mathcal{S}_n$) não garante que eles serão úteis. Os *frameworks* teóricos clássicos para aprendizado supervisionado, como o modelo PAC, não fornecem garantias formais para o aprendizado semissupervisionado. No modelo PAC clássico assume-se independência entre a distribuição de dados $\mathcal{D}_{\mathcal{X}}$ e a função-alvo h^*, sendo a única crença prévia que h^* pertence a uma classe $\mathcal{H}$. Mesmo que a distribuição $\mathcal{D}_{\mathcal{X}}$ fosse totalmente conhecida, qualquer função em $\mathcal{H}$ ainda seria uma candidata válida. Nesse cenário, os dados não rotulados não contribuem para restringir o espaço de hipóteses, pois a estrutura da distribuição não impõe restrições sobre qual é a função-alvo correta.

Para que o SSL seja teoricamente viável, é necessária uma extensão do modelo PAC. Balcan & Blum [14] propuseram um modelo PAC aumentado que introduz uma noção de compatibilidade (χ). Essa noção formaliza a crença fundamental do SSL:

a de que a função-alvo h^* não é arbitrária, mas sim compatível com a estrutura da distribuição de dados $\mathcal{D}_\mathcal{X}$. Em termos intuitivos, isso significa assumir que a organização dos dados no espaço de entrada contém informação relevante sobre os rótulos. Os dados não rotulados $\mathcal{S}_n$ tornam-se então valiosos, pois permitem ao algoritmo estimar propriedades da distribuição $\mathcal{D}_\mathcal{X}$ e avaliar a compatibilidade $\chi(h, \mathcal{D}_\mathcal{X})$ de uma hipótese h. Diferentes trabalhos da literatura expressam essa compatibilidade por meio de premissas estruturais sobre os dados, como as hipóteses de suavidade, agrupamento e variedade, discutidas na subseção seguinte.

2.2.2 Premissas

Entretanto, não é óbvio que a introdução de dados não rotulados vá gerar ganhos de desempenho para o modelo. Em seu trabalho seminal, Chapelle et al. [15] sugerem três premissas básicas para o funcionamento do SSL, resumidas a seguir:

- **Suavidade:** Dados semelhantes devem gerar predições semelhantes.
- **Agrupamento:** Dados que fazem parte de um mesmo agrupamento provavelmente são da mesma classe.
- **Variedade:** Os dados de alta dimensionalidade podem ser representados por uma variedade de baixa dimensionalidade.

2.2.3 Principais Abordagens

2.2.3.1 Pseudo-Rotulação

Cunhado por Lee [16], as abordagens baseadas no uso de Pseudo-Rótulos (*Pseudo-Labels - PL*) utilizam um modelo treinado nos dados rotulados para inferir rótulos de maneira automática sobre as amostras não rotuladas. Esse processo é comumente realizado de duas principais formas. Uma estratégia é a de gerar os pseudo-rótulos de maneira concomitante ao treinamento nos dados rotulados, de forma que é possível construir um novo conjunto com todos os dados e realizar o treinamento novamente a cada iteração. Outra forma de utilizar a técnica de pseudo-rotulação é conhecida como *self-training*, *self-learning* ou auto-aprendizado e se diferencia da anterior por ser caracterizada por duas etapas de treinamento distintas. Nesse caso, primeiro é

feito um treinamento completo com os dados rotulados. Em seguida, são gerados os pseudo-rótulos para todos os dados não rotulados disponíveis, que são então combinados aos rotulados e, todos juntos, são utilizados para uma segunda etapa de treinamento supervisionado.

Independentemente do processo que seja escolhido para a implementação dessa abordagem, o uso de pseudo-rótulos apresenta algumas limitações. Caso o treinamento com os dados rotulados não torne o modelo treinado capaz de gerar rótulos de boa qualidade, diz-se que o modelo gerou rótulos ruidosos. Rótulos ruidosos correspondem a anotações incorretas ou imprecisas associadas aos dados de treinamento, podendo envolver erros de classificação do objeto ou imprecisões na sua localização espacial. Uma estratégia tipicamente implementada para lidar com esse problema é a filtragem de pseudo-rótulos pela confiança atribuída à predição, em que define-se um limiar mínimo de confiança τ. Dessa forma, são aceitos como pseudo-rótulos válidos apenas aqueles gerados com valores de confiança superiores a τ. O valor de confiança corresponde à probabilidade estimada pelo próprio modelo para a predição realizada. Nos detectores da família YOLO [5], por exemplo, esse valor é tipicamente calculado a partir da combinação entre a probabilidade da classe prevista e a confiança de presença de objeto na região analisada.

A Figura 2.3 ilustra de forma simplificada o fluxo de uso de pseudo-rotulação. Inicialmente, um modelo treinado com dados rotulados é utilizado para gerar predições sobre amostras não rotuladas. Essas predições são então utilizadas como pseudo-rótulos, podendo ser filtradas com base em um limiar de confiança τ. As amostras não rotuladas acompanhadas dos pseudo-rótulos aceitos são então incorporadas ao processo de treinamento, podendo ser utilizadas de forma simultânea com os dados rotulados.

2.2.3.2 Regularização por consistência

A regularização por consistência (*Consistency Regularization - CR*) baseia-se fundamentalmente na premissa de suavidade, segundo a qual exemplos próximos no espaço de entrada devem possuir rótulos semelhantes. Ao forçar o modelo a produzir predições estáveis sob pequenas perturbações nos dados não rotulados, essa

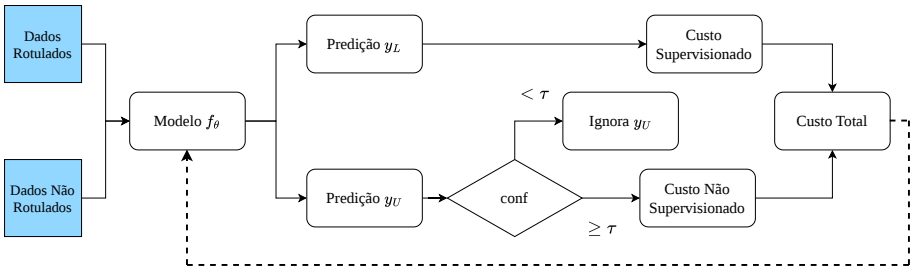

Figura 2.3: Diagrama de exemplo de uso de pseudo-rotulação.

técnica explora a estrutura local da distribuição não rotulada. No caso de dados de imagem, essas perturbações podem corresponder a transformações como rotações leves, recortes, inversões horizontais ou variações de brilho e contraste.

A imposição de semelhança entre as predições é feita por meio da adição de um termo de regularização à função de custo do treinamento. Esse termo é implementado por alguma métrica de distância $d(h(x_1), h(x_2))$, em que $d(\cdot)$ é a função de distância e $h(x_i)$ são as diferentes predições do modelo para entradas afetadas por perturbações. Métricas como a divergência de Kullback-Leibler (KL) ou o erro médio quadrático (*Mean Squared Error - MSE*) são comumente usadas como medida de distância.

A Figura 2.4 descreve um fluxo de execução simples da regularização por consistência. Nesse exemplo, duas perturbações distintas α e β são aplicadas à mesma amostra x, gerando duas versões modificadas da entrada. O modelo produz predições para ambas as versões, e a função de regularização penaliza diferenças entre essas predições, incentivando o modelo a manter respostas consistentes diante de pequenas variações na entrada.

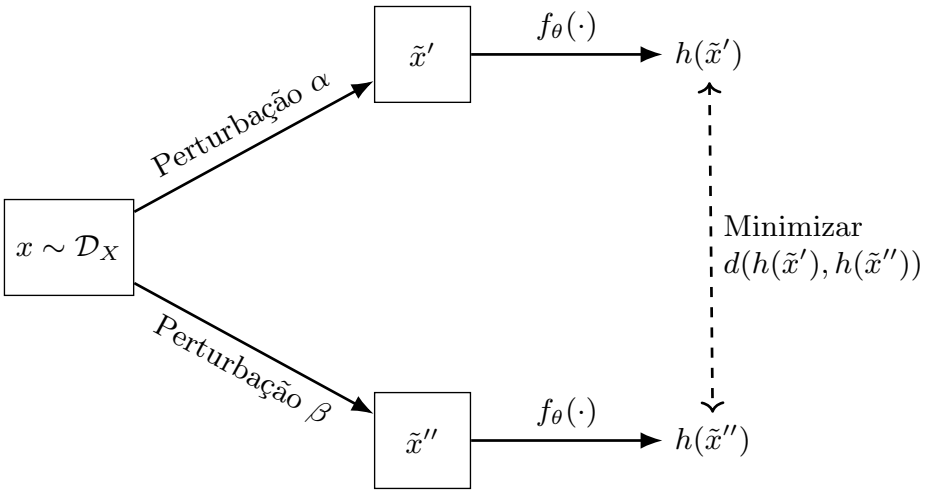

Figura 2.4: Diagrama de exemplo de uso de regularização por consistência.

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta os principais trabalhos relacionados ao problema da detecção de objetos semissupervisionada. Inicialmente, são discutidos métodos clássicos de aprendizado semissupervisionado desenvolvidos no contexto de classificação de imagens, que servem de base para abordagens posteriores. Em seguida, são apresentados os principais paradigmas de SSOD. Adicionalmente, são discutidas limitações específicas desses métodos quando aplicados a detectores de um estágio e os trabalhos que buscam mitigar essas limitações. Por fim, são apresentados trabalhos recentes que exploram modelos de fundação como alternativas para a geração de rótulos.

3.1 Aprendizado semissupervisionado para classificação de imagens

O estudo sobre SSL é um tema amplamente investigado na literatura científica. O uso de pseudo-rótulos, proposto por Lee et al. [16], é uma das técnicas mais relevantes desse tema. Apesar da ampla adoção dessa técnica, o seu uso pode gerar problemas como instabilidade de treinamento e produção de rótulos ruidosos para o treinamento.

Outra abordagem de destaque na literatura, e que é utilizada para mitigar o problema de rótulos ruidosos, é a regularização por consistência [17, 18], que utiliza a hipótese de suavidade do SSL para sugerir que entradas levemente perturbadas

devem produzir predições consistentes no espaço de predições. Métodos como o Π-Model [17] e o Mean Teacher [19] impõem essa consistência ao penalizar diferenças entre predições obtidas a partir de diferentes versões da mesma amostra, geralmente utilizando perturbações estocásticas ou modelos professores baseados em médias móveis dos parâmetros do aluno.

3.2 Aprendizado semissupervisionado para detecção de objetos

As primeiras tentativas de aplicar aprendizado semissupervisionado à detecção de objetos surgiram como extensões diretas de métodos utilizados em tarefas de classificação de imagens. Diferentemente da classificação, no entanto, a detecção de objetos envolve a produção de predições fundamentalmente mais complexas, simultaneamente executando a tarefa de classificação e definindo a posição do objeto classificado na imagem. Dessa forma, o processo de detecção passa a admitir erros não apenas na classificações, como também na localização dos objetos na cena. No contexto de SSOD, mesmo que o modelo identifique corretamente a qual categoria o objeto identificado pertence, pequenos erros no posicionamento desse objeto podem levar à rejeição incorreta dessa predição devido a um baixo valor de interseção sobre a união (*Intersection over Union - IoU*). Esse efeito torna os modelos de detecção mais sensíveis ao ruído de rotulação do que os modelos de classificação.

Sohn et al. [20] propuseram uma das primeiras abordagens para SSOD, combinando pseudo-rotulação com regularização por consistência. Embora o método apresente ganhos relevantes em relação ao treinamento supervisionado, os autores destacam que a estabilidade do treinamento depende da qualidade inicial dos pseudo-rótulos gerados.

Jeong et al. [21] empregam uma técnica de regularização por interpolação ao contexto de SSOD, realizando testes em modelos SSD [22]. Apesar de realizar testes em detectores de um estágio, a proposta não é projetada especificamente para esse tipo de modelo, além de envolver o uso de âncoras de detecção.

Âncoras de detecção consistem em caixas pré-definidas posicionadas em cada célula da grade da imagem e utilizadas como ponto de partida para a estimação das caixas delimitadoras finais. Durante o treinamento, cada objeto presente na imagem é associado à âncora que maximiza a sobreposição com a caixa de referência, permitindo ao modelo aprender ajustes de tamanho e posicionamento que aproximem a âncora da localização correta do objeto. A Figura 3.1 ilustra esse conceito, mostrando múltiplas âncoras com diferentes escalas e proporções associadas a uma mesma região da imagem.

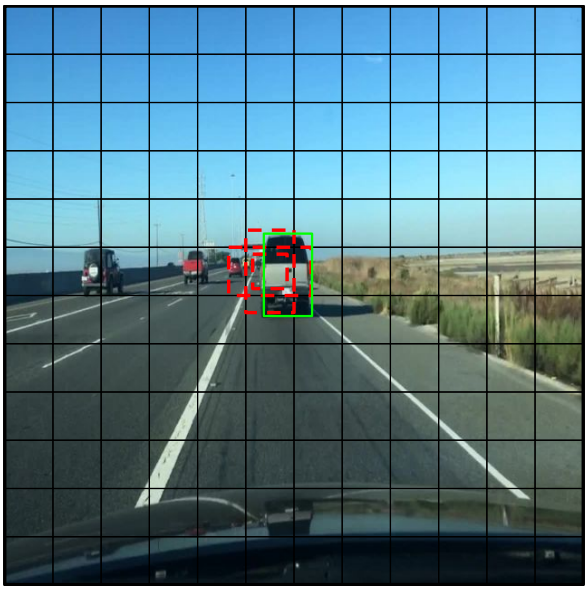

Figura 3.1: Exemplo ilustrativo do conceito de âncoras em detecção de objetos. A imagem é dividida em uma grade espacial e múltiplas caixas âncora, em vermelho, são consideradas em cada célula. Durante o treinamento, cada objeto é associado à âncora que tem a maior sobreposição com a caixa de referência, em vermelho.

O uso de âncoras introduz maior complexidade de configuração, pois o desempenho final do detector pode depender do ajuste adequado de hiperparâmetros adicionais, como o número, escalas e proporções das âncoras utilizadas [23].

Trabalhos subsequentes passaram a investigar arquiteturas baseadas em professor-aluno como forma de mitigar o impacto de rótulos ruidosos. Nessa linha, Zhou et al. [24] propuseram uma combinação de técnicas de aprendizado ativo com o uso de um modelo professor atualizado de forma progressiva para gerar pseudo-rótulos

mais estáveis ao longo do treinamento.

3.3 SSOD em Detectores de Um Estágio

Apesar dos avanços obtidos com abordagens baseadas em arquiteturas professor-aluno e técnicas baseadas em CR, a maior parte da literatura em SSOD concentra-se em detectores de dois estágios [6]. Esses modelos apresentam maior robustez a ruídos de rótulos, o que facilita a utilização de pseudo-rótulos durante o treinamento quando comparados com modelos de um estágio.

Zhang et al. [25] propuseram o método S4OD, uma das primeiras abordagens explicitamente projetadas para o treinamento semissupervisionado de detectores de um estágio. Diferentemente de métodos originalmente desenvolvidos para detectores de dois estágios, o S4OD busca lidar com a maior sensibilidade desses modelos ao ruído de pseudo-rótulos.

O método introduz dois principais componentes. O primeiro é o *Dynamic Self-Adaptive Threshold (DSAT)*, que faz um controle dinâmico do limiar de confiança utilizado na decisão de aceitar ou não predições como pseudo-rótulos válidos durante o treinamento. O limiar é escolhido com base em uma busca do valor que maximize o *F1-Score* obtido no conjunto de validação, tentando encontrar de limiar que permitam uma geração de pseudo-rótulos com equilíbrio de quantidade e qualidade.

O segundo componente é o *NMS-UNC*, uma variação do algoritmo de filtragem de predições *Non Maximum Suppression (NMS)*, que incorpora ao processo de seleção predições uma estimativa da incerteza posicional das caixas geradas pelo modelo. Os autores argumentam que, quando as predições densas geradas por detectores de um estágio apresentam menor espalhamento espacial, a predição final tende a ser mais precisa. Baseando-se nessa ideia, utiliza-se o desvio padrão das coordenadas das predições redundantes geradas pelo detector como uma estimativa da incerteza na regressão das caixas delimitadoras, que funciona como um critério adicional de filtragem complementar ao limiar de confiança.

Adicionalmente, Kim et al. [6] investigaram o cenário de detecção de objetos semissupervisionada federada *(Semi-Supervised Federated Object Detection - SSFOD)*, em que a exigência de privacidade dos clientes impõe um cenário de escassez de rótulos. No caso considerado pelos autores, apenas o servidor possui dados rotulados, enquanto os clientes possuem apenas dados não rotulados. Para lidar com essa configuração, os autores propõem uma estratégia baseada em dois estágios. Inicialmente, é feito um ajuste fino dos parâmetros do *backbone* do detector no conjunto de dados do servidor. Em seguida, os clientes realizam treinamento local utilizando seus dados não rotulados com base em pseudo-rótulos gerados por um modelo professor obtido por média móvel exponencial (*Exponential Moving Average - EMA*) dos parâmetros do modelo. Adicionalmente, os autores utilizam uma regularização de ortogonalidade para mitigar o problema de heterogeneidade entre clientes, tornando o sistema mais robusto a dados de clientes não independentes e identicamente distribuídos (não-IID).

3.4 Modelos de Fundação para Geração de Pseudo-Rótulos

Recentemente, o surgimento de modelos de fundação multimodais, que são treinados em larga escala, com grandes volumes de dados, sendo capazes de serem adaptados para diversos domínio e e aplicações, abriu novas possibilidades para a geração automática de rótulos. Modelos como o Segment Anything Model (SAM) [26] são capazes de realizar tarefas de visão computacional, como segmentação semântica e detecção de objetos, mesmo sem treinamento específico no domínio alvo. Isso é possível devido à natureza multimodal desse tipo de modelo. Ou seja, a sua capacidade de processar e interpretar diferentes tipos de dados, como imagens, áudio ou texto.

Wang et al. [27] propuseram adaptar a estrutura do SAM para adequá-la à tarefa de detecção de objetos fracamente supervisionada. Nessa proposta, os autores definem componentes adicionais à arquitetura do modelo, como geradores adaptativos de *prompts* e adaptadores de domínio. Entretanto, a abordagem sugerida exige

que o modelo seja executado durante a inferência, inviabilizando a implementação desse tipo de sistema em dispositivos com restrições de recursos computacionais e energéticos.

Bhaskar et al. [28] utilizam um modelo de fundação como gerador de rótulos para o treinamento de modelos de detecção de objeto, YOLOv5. Para mitigar o impacto do ruído dos rótulos produzidos pelo modelo, os autores combinam essa geração a um esquema de co-treinamento para realizar filtragem e eliminação de rótulos ruidosos. Apesar da proposta apresentar relevante ganho de desempenho, a análise realizada pelos autores não envolve modelos mais modernos como o YOLOv11 e o modelo de fundação SAM3.

Neste trabalho, considera-se o problema do treinamento de detectores de objetos de um estágio em cenários de escassez de dados rotulados, cenário recorrente na literatura de SSOD. O recorte adotado, com foco em detectores de um estágio da família YOLO, é relativamente pouco explorado na literatura, já que a maioria dos trabalhos de SSOD considera arquiteturas de dois estágios. Este estudo busca investigar experimentalmente a viabilidade de diferentes estratégias nesse contexto, incluindo técnicas clássicas de SSOD e o uso de modelos de fundação como geradores de pseudo-rótulos. As análises realizadas utilizam detectores modernos da família YOLO como modelos base e consideram o uso do modelo de fundação multimodal SAM3 para geração automática de rótulos, o que ainda é pouco explorado em trabalhos voltados ao treinamento de detectores de um estágio.

Capítulo 4

Configuração Experimental e Estratégias para Detecção de Objetos com Poucos Rótulos

Este trabalho investiga possíveis soluções para o desafio de realizar o treinamento de um modelo de detecção de um estágio em cenários de escassez de dados rotulados. Essa investigação foi conduzida por meio de duas principais frentes. A primeira foi a avaliação da aplicabilidade de técnicas clássicas de SSL e SSOD ao contexto de interesse. Em seguida, foi feita uma exploração quanto ao uso de modelos de fundação multimodais como geradores automáticos de rótulos para detecção de objetos.

4.1 Conjunto de Dados, Métricas e Ambiente Experimental

Todos os experimentos realizados neste trabalho utilizam o conjunto de dados BDD100K [8], comumente utilizado em pesquisas de visão computacional voltadas ao contexto veicular [6, 29, 30]. O conjunto é composto por cem mil imagens capturadas em cenários urbanos, acompanhadas de anotações para dez classes de objetos. Esse conjunto é disponibilizado pelos autores com uma divisão dos dados de 70% para treino, 20% para teste e 10% para validação, que são mantidos fixos para todos os experimentos realizados no presente trabalho. O contexto de aquisição dos dados torna o BDD100K particularmente relevante para aplicações como de-

tecção de usuários vulneráveis das vias (*Vulnerable Road Users – VRUs*), condução autônoma e sistemas avançados de assistência ao motorista. Adicionalmente, o conjunto apresenta diversas variações de condições ambientais, como diferentes iluminações, condições climáticas e densidades de tráfego, o que o torna um cenário desafiador e aproxima os experimentos de um contexto realista.

O desempenho dos modelos em termos de qualidade de detecção é avaliado por meio da métrica *Mean Average Precision* (mAP), considerando o intervalo de interseção sobre união *(Intersection over Union - IoU)* de 50% até 95%, que corresponde a um dos padrões adotados na literatura de detecção de objetos [31]. Essa métrica avalia simultaneamente a precisão da classificação e a qualidade da localização espacial das caixas delimitadoras.

Os experimentos foram executados em uma máquina equipada com 15 GB de memória RAM e GPU *NVIDIA RTX A4000*, utilizando o sistema operacional *Debian 12*. O treinamento dos modelos foi realizado utilizando a biblioteca de aprendizado profundo PyTorch.

4.2 Detecção de Objetos Semissupervisionada

4.2.1 Arquitetura do Modelo

Nos experimentos baseados em técnicas clássicas de SSL e SSOD, foi utilizado um modelo da família YOLOv8. Essa arquitetura foi escolhida devido à disponibilidade de uma implementação em PyTorch com acesso direto ao código-fonte do modelo [32], o que permite realizar modificações no fluxo de treinamento. Esse controle sobre o modelo é necessário para a implementação das estratégias investigadas, uma vez que técnicas como PL e CR envolvem procedimentos de difícil implementação na distribuição oficial do modelo, como a realização de múltiplas inferências por etapa de treinamento e a aplicação de perturbações aos dados de entrada.

4.2.2 Estratégias de Treinamento

Foram avaliadas diferentes estratégias de SSOD da literatura. As estratégias investigadas são descritas a seguir.

Supervisionado: O treinamento supervisionado é utilizado como referência. Nessa abordagem, apenas os dados rotulados manualmente são utilizados durante o treinamento, e as amostras consideradas como não rotuladas são ignoradas. Essa configuração estabelece uma base para a análise dos outros métodos, permitindo avaliar se o aproveitamento de dados não rotulados resulta em ganhos de desempenho.

Auto-Aprendizado: Nesta estratégia, o treinamento é dividido em duas etapas. Inicialmente, o modelo é treinado de forma supervisionada utilizando apenas os dados rotulados. Em seguida, esse modelo é utilizado para gerar pseudo-rótulos para o conjunto de dados não rotulados, que são então incorporados ao conjunto de treinamento em uma segunda etapa supervisionada.

Aluno-Professor: Na estratégia aluno-professor, os dados rotulados e não rotulados são utilizados simultaneamente durante todo o treinamento. Essa configuração segue o uma estrátégia de arquiteturas *teacher–student* amplamente utilizadas em aprendizado semissupervisionado. Em particular, o modelo professor é obtido a partir da média móvel exponencial (*Exponential Moving Average - EMA*) dos pesos do aluno, o que é inspirado na técnica *MeanTeacher* [19]. Nessa estratégia, o professor é responsável por gerar pseudo-rótulos para cada lote de dados, que são atualizados ao longo do treinamento acompanhando a evolução do modelo. As predições do professor são utilizadas como referência para os dados não rotulados do aluno, sendo aplicadas perturbações diferentes às entradas de cada modelo. O treinamento do aluno combina a perda supervisionada calculada sobre os dados rotulados com um termo de regularização por consistência, que incentiva o modelo a produzir predições semelhantes às geradas pelo professor para as amostras não rotuladas.

S4OD: Neste trabalho, o método S4OD é utilizado como uma das estratégias de SSOD avaliadas, sendo implementadas as duas técnicas descritas no artigo original.

Essa estratégia foi escolhida devido ao trabalho original propor, especificamente, o tratamento dos desafios de realizar SSOD com modelos de um estágio.

4.2.3 Configuração Experimental

Os experimentos de SSOD foram realizados utilizando uma divisão fixa entre dados rotulados e não rotulados, sendo 25% do conjunto de treinamento tratado como rotulado e os 75% restantes como não rotulados. Essa proporção foi repetida em todos os experimentos de SSOD.

O treinamento foi realizado por um total de 400 épocas em todos os experimentos. Esse valor foi definido de forma empírica a partir de experimentos preliminares, nos quais se observou que o treinamento supervisionado de referência atingia um regime de desempenho relativamente estável antes desse limite. A adoção de um número elevado de épocas teve como objetivo evitar que diferenças entre as estratégias avaliadas fossem influenciadas por interrupções prematuras do treinamento.

Com exceção dos experimentos puramente supervisionados, todos os treinamentos partem de um período inicial de aquecimento, em que apenas os dados rotulados são utilizados. Esse procedimento tem como objetivo reduzir o impacto de pseudo-rótulos de baixa qualidade nas fases iniciais do treinamento [25, 20, 33]. No caso da estratégia de auto-aprendizado, as primeiras 200 épocas são tratadas como aquecimento, correspondendo à etapa supervisionada inicial. Nas estratégias aluno-professor e S4OD, foi adotado um aquecimento mais curto, de 25 épocas, escolhido de forma empírica como compromisso entre estabilidade inicial e custo computacional.

Nos experimentos baseados na estratégia aluno-professor, o modelo professor é obtido a partir de uma EMA dos pesos do modelo aluno. A regularização por consistência é implementada por meio da aplicação de perturbações às amostras não rotuladas, sendo utilizadas aumentações fracas para o professor e aumentações mais fortes para o aluno. Como estratégia de mitigação do ruído de rótulos, a utilização de pseudo-rótulos é controlada por um limiar de confiança, cujo valor é ajustado dinamicamente ao longo do treinamento.

Além das estratégias descritas, os experimentos foram realizados com congelamento parcial das camadas do modelo. Nessa técnica de regularização, os pesos do *backbone* são mantidos fixos durante parte do treinamento, sendo atualizados apenas os parâmetros da cabeça de detecção. Ao impedir a atualização de uma parcela significativa dos parâmetros do modelo, o congelamento reduz o espaço efetivo de hipóteses que pode ser explorado durante o treinamento, funcionando como um mecanismo de regularização. No contexto deste trabalho, essa estratégia busca limitar a propagação de ruído proveniente de pseudo-rótulos de baixa qualidade e aumentar a estabilidade do treinamento.

4.3 Geração de Rótulos com Modelos de Fundação

A segunda linha experimental deste trabalho investiga o uso de modelos de fundação como fonte alternativa de rótulos para o treinamento de detectores de objetos em cenários de escassez de dados rotulados. Diferentemente das abordagens clássicas de SSOD, que dependem da geração iterativa de pseudo-rótulos a partir do próprio detector, essa metodologia avalia a utilização de um modelo externo, pré-treinado em larga escala, como gerador direto de anotações.

4.3.1 Modelo SAM3

O *Segment Anything Model 3* (*SAM3*) é a versão mais recente de um modelo de fundação multimodal voltado para tarefas de segmentação visual generalista. Seu funcionamento baseia-se no uso de arquiteturas do tipo *transformer*, inspiradas no paradigma de detectores baseados em atenção introduzido pelo modelo DETR [34]. Nesse modelo, a imagem de entrada e os *prompts* textuais são inicialmente codificados em representações vetoriais, que são então combinadas em uma etapa de fusão multimodal. Essas representações são posteriormente processadas por um decodificador que produz máscaras de segmentação correspondentes aos objetos descritos pelos *prompts*. Embora o modelo seja voltado principalmente para a geração de máscaras de segmentação, essas máscaras podem ser convertidas em caixas delimitadoras, permitindo sua utilização em outras tarefas de visão computacional, como a detecção de objetos.

Nessa abordagem, o *Segment Anything Model 3* [7] (SAM3) é empregado como gerador automático de rótulos. O SAM3 é utilizado apenas na etapa de rotulação, não sendo ajustado nem adaptado ao conjunto de dados alvo. Essa estratégia baseia-se na hipótese de que, ao aproveitar o conhecimento semântico adquirido durante seu pré-treinamento, seja possível gerar anotações de boa qualidade. Essa característica do modelo permite avaliar a viabilidade do uso do modelo em cenários de restrição severa ou até mesmo na ausência completa de dados rotulados, utilizando apenas rótulos gerados por um modelo de fundação. Diferentemente das técnicas clássicas de detecção de objetos semissupervisionadas, o processo de rotulação não é influenciado pelo detector nem envolve realimentação durante o treinamento, evitando problemas de instabilidade associados à propagação de erros e ao viés de confirmação.

Essa abordagem pode ser interpretada como uma forma de transferência ou destilação indireta de conhecimento. Nesse cenário, toma-se proveito de um modelo grande e lento, mas detentor de vasta capacidade de processamento, para realizar o treinamento de um modelo leve e especializado, adequado ao funcionamento em aplicações que demandam inferência em tempo real. A biblioteca *autodistill* [35] implementa essa ideia, fornecendo acesso facilitado à destilação de modelos de fundação para diferentes arquiteturas de destino. Entretanto, devido ao lançamento recente do SAM3, a biblioteca ainda não havia disponibilizado suporte a esse modelo durante a realização deste trabalho, o que levou à adoção de um pipeline simplificado de geração de rótulos com o modelo.

4.3.2 Pipeline de Geração de Pseudo-Rótulos

Foi utilizado um pipeline para a geração automática de pseudo-rótulos a partir do SAM3, no qual todas as imagens do conjunto de treinamento, ou seja, 70% das imagens do conjunto BDD100k [8], são processadas em etapa anterior ao treinamento do detector. Para cada imagem, o SAM3 produz um conjunto de detecções a partir de instruções semânticas, que são filtradas com base em um limiar de confiança τ. As anotações aceitas são então convertidas para o formato *COCO* [31], compatível com o treinamento supervisionado do YOLO.

O processo de geração de pseudo-rótulos baseia-se na definição de uma ontologia, que estabelece a correspondência entre os rótulos de interesse e suas descrições semânticas . Como o SAM3 é um modelo multimodal, essa ontologia é utilizada na forma de *prompts* textuais, orientando a geração das predições para cada imagem. A ontologia adotada neste trabalho foi construída a partir das classes do conjunto de dados BDD100K, com cada *prompt* fornecendo uma descrição simples e direta da classe correspondente. Como o nome das classes do conjunto de dados utilizado já é descritivo, optou-se por adotar uma tradução dos nomes para o idioma português, como o *prompt* "Pessoa" para a classe "*person*".

4.3.3 Configuração Experimental

Nos experimentos baseados em pseudo-rótulos gerados pelo SAM3, foi adotado o modelo YOLOv11 utilizando a implementação oficial da Ultralytics. Essa escolha ocorreu devido à possibilidade de uso do modelo sem a necessidade de implementar modificações na estrutura do detector ou nos fluxos de carregamento de dados e treinamento.

Os experimentos foram organizados de acordo com diferentes distribuições de rótulos, sendo variada a proporção de dados rotulados manualmente, pseudo-rotulados e dados não utilizados no treinamento. Essas configurações incluem cenários puramente supervisionados, cenários híbridos e cenários em que o treinamento é realizado apenas com dados pseudo-rotulados. Essas configurações são apresentadas na Tabela 4.1 e são posteriormente analisadas. Cada configuração define a proporção de dados rotulados manualmente, pseudo-rotulados e não utilizados durante o treinamento.

Em todos os casos, realizou-se os mesmos experimentos para valores de limiar de confiança $\tau \in \{0.5, 0.7, 0.9\}$, permitindo analisar a robustez da abordagem em relação à quantidade e à qualidade dos pseudo-rótulos utilizados. Em todas as configurações, o número de épocas de treinamento foi fixado em 200 para todos os experimentos dessa metodologia, uma vez que foi possível observar ganhos de desempenho sem a necessidade de treinamentos prolongados. As configurações foram

Tabela 4.1: Configurações utilizadas nos experimentos com pseudo-rótulos gerados pelo SAM3.

Modalidade	**Rotulados** (%)	**Pseudo-rotulados** (%)	**Descartados** (%)
Supervisionado	1	N/A	99
Supervisionado	2	N/A	98
Supervisionado	5	N/A	95
Supervisionado	10	N/A	90
Híbrido Simétrico	1	1	98
Híbrido Simétrico	2	2	96
Híbrido Simétrico	5	5	90
Híbrido Simétrico	10	10	80
Híbrido Total	1	99	0
Híbrido Total	2	98	0
Híbrido Total	5	95	0
Híbrido Total	10	90	0
Pseudo-rotulado	0	1	99
Pseudo-rotulado	0	2	98
Pseudo-rotulado	0	5	95
Pseudo-rotulado	0	10	90

agrupadas em 4 modalidades, ilustradas na Figura 4.1 e apresentadas na sequência.

Supervisionado: No caso supervisionado, o treinamento do modelo é realizado exclusivamente com dados rotulados manualmente. O restante dos dados, considerados não rotulados, são descartados. Este cenário serve como um valor de referência, modelando o melhor resultado que se pode obter na ausência de qualquer técnica que objetive mitigar o problema de escassez de dados rotulados.

Híbrido Simétrico: O caso híbrido simétrico considera o uso de proporções iguais de dados com rótulos de origem humana e produzidos pelo SAM3, sendo os dados restantes descartados. Esta configuração permite avaliar cenários em que a escassez de rótulos manuais não é tão severa, porém existem restrições quanto ao número de dados coletados. A utilização de níveis iguais de rótulos manuais e pseudo-

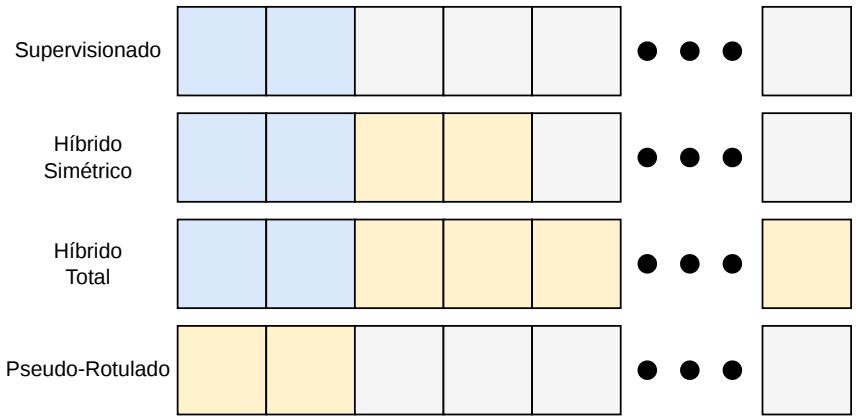

Figura 4.1: Ilustração conceitual da composição dos conjuntos de dados utilizados em cada configuração experimental. As caixas representam apenas os tipos de dados presentes em cada configuração, não correspondendo a proporções ou quantidades relativas. Caixas azuis correspondem a dados rotulados manualmente, caixas amarelas a dados pseudo-rotulados e caixas cinza a dados descartados.

rótulos permite avaliar o impacto do uso das anotações artificiais de modo que elas não dominem o treinamento, evitando o possível mascaramento das contribuições geradas pelos dados rotulados manualmente.

Híbrido Total: A modalidade Híbrido Total envolve o uso de uma parcela dos dados com rótulos de origem humana e todo o restante dos dados acompanhados de pseudo-rótulos, não havendo descarte. Esta configuração modela um cenário mais realista, em que existe escassez de rótulos e abundância de dados, assumindo custo baixo para coleta de dados e alto para produção de rótulos. Permite avaliar se grandes volumes de dados pseudo-rotulados compensam a disponibilidade limitada de dados rotulados, assim como se o desequilíbrio das fontes de rotulação pode induzir vieses ao modelo treinado.

Pseudo-rotulado: Nesta modalidade, todos os dados utilizados no treinamento acompanham rótulos gerados pelo SAM3. Essa configuração avalia um cenário extremo, em que há completa ausência de recursos de rotulação, sendo o desafio treinar um modelo para tarefas supervisionadas mesmo na ausência de anotações. Permite avaliar de forma isolada a influência de dados pseudo-rotuladas no desempenho do

modelo treinado.

Capítulo 5

Avaliação Experimental de Estratégias para Detecção de Objetos com Poucos Rótulos

Este capítulo apresenta os experimentos e as análises desenvolvidos com o fim de avaliar a aplicabilidade de diferentes técnicas para realizar o treinamento de modelos de detecção de objetos em cenários de escassez de rótulos. Inicialmente, são analisados os resultados obtidos a partir dos experimentos utilizando técnicas clássicas de SSOD. Em seguida, são analisados e discutidos os resultados de experimentos sobre o uso de modelos de fundação para a geração automática de rótulos, sendo avaliados aspectos como desempenho de detecção e custo computacional de treinamento.

5.1 Técnicas de Detecção de Objetos Semissupervisionada

Esta seção apresenta os resultados obtidos com a aplicação de técnicas clássicas de SSOD em cenários de restrição na disponibilidade de dados rotulados. Essa análise tem como objetivo verificar a viabilidade do uso dessas técnicas como alternativa ao treinamento puramente supervisionado, considerando o cenário de escassez de rótulos e o uso de modelos de detecção de um estágio.

5.1.1 Comparação de Desempenho de Detecção Máximo

A Figura 5.1 apresenta o valor máximo da métrica mAP 50-95 observada ao longo de todo o treinamento, independentemente da época em que esse valor foi alcançado, para cada uma das abordagens avaliadas, incluindo o valor de referência supervisionado.

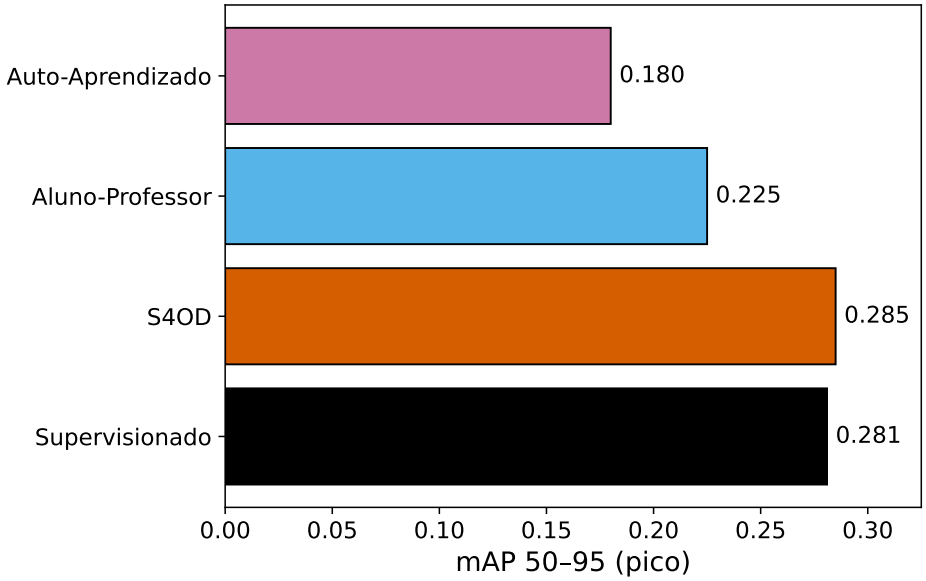

Figura 5.1: Comparação do desempenho máximo (mAP 50-95) alcançado por diferentes técnicas clássicas de SSOD e pelo treinamento supervisionado.

Nesta análise inicial, nota-se que apenas o método S4OD apresentou desempenho superior ao gerado pelo treinamento puramente supervisionado. Nos outros casos, como nas estratégias baseadas em pseudo-rotulação e regularização por consistência, o desempenho máximo obtido é inferior ao valor de referência supervisionado, mesmo com a inclusão de dados não rotulados no treinamento.

Com exceção do método S4OD, esses resultados indicam que, no contexto avaliado, a utilização dos dados não rotulados por meio de técnicas clássicas de SSOD não é suficiente para melhorar o desempenho de detectores de um estágio, sendo inclusive prejudicial ao treinamento. Tal comportamento reforça observações da literatura, que sugerem que esses modelos são particularmente sensíveis à qualidade dos rótulos utilizados durante o treinamento [6]. Observa-se ainda que, entre as

técnicas avaliadas, a abordagem de auto-aprendizado apresentou o pior desempenho, sugerindo que estratégias que realizam a geração de pseudo-rótulos de forma simultânea ao treinamento tendem a provocar menor degradação do desempenho.

Já no caso específico do método S4OD, embora o valor máximo de mAP observado seja similar ao supervisionado, com uma diferença de aproximadamente 0.04, esse resultado não é suficiente para confirmar um ganho de desempenho. Para avaliar se esse comportamento corresponde a uma melhoria real ou apenas fenômenos transitórios ao longo do treinamento, é necessária uma análise da evolução temporal do método, apresentada a seguir.

5.1.2 Investigação sobre o Desempenho do S4OD

Para verificar se o valor de pico de desempenho observado com o uso do método S4OD corresponde a um ganho que se sustenta ao longo do treinamento, foi realizado um exame da evolução temporal da métrica mAP no conjunto de validação. Os resultados obtidos com o S4OD são comparados aos produzidos pelo treinamento puramente supervisionado. A Figura 5.2 apresenta a evolução do desempenho ao longo das épocas de treinamento, bem como a evolução do limiar de confiança utilizado para a seleção de pseudo-rótulos no método S4OD, representado pela linha pontilhada no gráfico.

Observa-se que, embora a técnica S4OD apresente picos de desempenho superiores ao regime supervisionado, a evolução do seu desempenho de detecção é instável ao longo das épocas. Conforme observado na Figura 5.2, após o período inicial de aquecimento ocorre uma queda abrupta do limiar de confiança utilizado para a seleção dessas anotações. Esse limiar é ajustado dinamicamente pelo mecanismo DSAT (*Dynamic Self-Adaptive Threshold*), que define o valor do limiar com base no desempenho do modelo no conjunto de validação, selecionando o valor que maximiza o *F1-score* das predições utilizadas como pseudo-rótulos. No cenário observado, esse processo leva o limiar rapidamente ao seu valor mínimo, permitindo a inclusão de um grande volume de predições de baixa confiança como pseudo-rótulos válidos. Como consequência, o modelo pode reforçar os seus próprios erros ao longo das épocas,

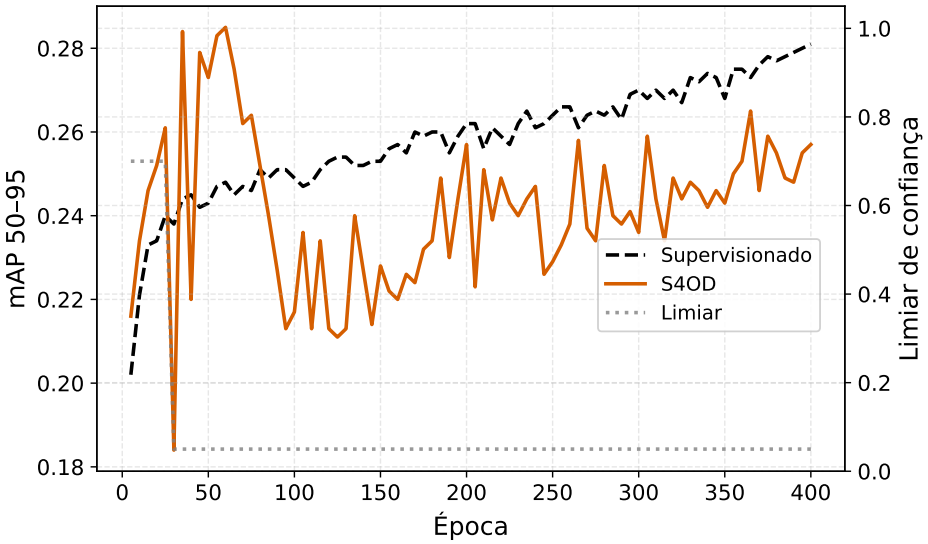

Figura 5.2: Evolução temporal da métrica mAP 50-95 durante o treinamento supervisionado e com a técnica S4OD.

caracterizando um ciclo de viés de confirmação negativo que leva à degradação do desempenho.

5.1.3 Discussão sobre Técnicas de Detecção de Objetos Semissupervisionada

Os resultados obtidos apontam para limitações das técnicas clássicas de SSOD quando aplicadas a detectores de um estágio. Mesmo utilizando técnicas de CR, arquitetura de aluno e professor com EMA e filtragem de confiança, os modelos não foram capazes de igualar o desempenho obtido ao realizar o treinamento do modelo com apenas a parcela rotulada dos dados.

Mesmo uma técnica projetada para mitigar os desafios de realizar SSOD em detectores de um estágio, como o S4OD, apresentou comportamento instável no contexto avaliado. Cabe ressaltar, entretanto, que o trabalho original propõe o uso de um período de aquecimento maior, além de conjuntos de dados mais simples e arquiteturas distintas das utilizadas neste trabalho, como detectores FCOS e RetinaNet.

O uso das configurações originais do método implicaria na execução de treinamentos em escala significativamente maior do que a adotada neste trabalho. No artigo

original, os autores utilizam múltiplas GPUs, períodos extensos de aquecimento e um número elevado de iterações de treinamento. No cenário experimental considerado neste estudo, essas configurações foram reduzidas devido às limitações de recursos computacionais disponíveis. Além disso, os experimentos foram conduzidos em um conjunto de dados de maior complexidade e utilizando uma arquitetura de detector mais recente, o que também limita o escopo desta análise.

Dessa forma, os resultados obtidos indicam que, sob as condições experimentais adotadas, incluindo a escala de treinamento, a complexidade do conjunto de dados e a arquitetura do detector, a abordagem S4OD não se mostrou suficiente para viabilizar o treinamento de detectores de um estágio em cenários de escassez de rótulos. Essas limitações motivam a investigação de alternativas que reduzam a dependência do treinamento do modelo com pseudo-rótulos ruidosos, como o uso de modelos de fundação para geração de rótulos de forma desacoplada ao treinamento do modelo de detecção, explorado na seção seguinte.

5.2 Geração de Rótulos para Detecção de Objetos com o SAM3

O uso de modelos de fundação como geradores de pseudo-rótulos para o treinamento de modelos de detecção de um estágio em cenários de escassez de rótulos é investigado nesta seção. Objetiva-se, portanto, avaliar se essa abordagem é uma alternativa viável ao uso das técnicas clássicas de SSOD no contexto descrito. Com o fim de verificar tal viabilidade, foram feitos experimentos com essa técnica baseando-se em diferentes eixos de análise. Inicialmente, é verificado o desempenho dessa abordagem considerando quatro esquemas de rotulação: supervisionado, híbrido simétrico, híbrido total e pseudo-rótulos. Em uma etapa seguinte, avalia-se o impacto da variação do limiar de confiança adotado durante a filtragem dos pseudo-rótulos no desempenho final do modelo de detecção. Adicionalmente, é analisada a relação entre custo computacional do treinamento e o desempenho de detecção para cada configuração estudada. Por fim, é feita uma análise exploratória do impacto da adição de pseudo-rótulos no treinamento do modelo em termos das classes usadas

no conjunto de dados.

5.2.1 Comparação entre Estratégias de Treinamento

Para avaliar o impacto da adição de pseudo-rótulos ao treinamento do modelo em cenários de escassez de rótulos, foram avaliadas diferentes estratégias, cujos resultados são apresentados na Figura 5.3. A figura utiliza dois eixos horizontais, sendo o eixo inferior correspondente à proporção de dados rotulados manualmente (aplicável aos cenários supervisionado e híbridos), e o eixo superior referente à proporção de dados pseudo-rotulados (aplicável apenas ao cenário pseudo-rotulado). Em todos os cenários, são comparados casos em que cada parcela dos dados disponíveis é utilizada, sendo os valores avaliados 1%, 2%, 5% e 10%, com o excedente descartado em todos os casos exceto o híbrido total. Os resultados apresentados correspondem ao experimento conduzido com limiar de confiança fixado em $\tau = 0.5$.

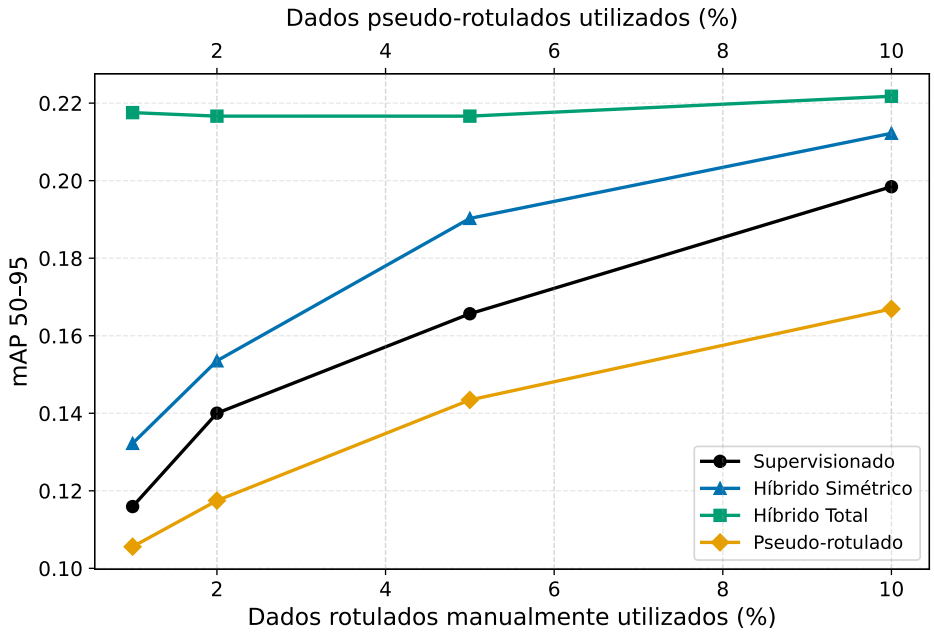

Figura 5.3: Comparação de desempenho de detecção entre os cenários supervisionado, híbrido simétrico, híbrido total e pseudo-rotulado ($\tau = 0.5$). O eixo inferior refere-se à proporção de dados rotulados manualmente destinada ao treinamento, enquanto o eixo superior indica a proporção de dados pseudo-rotulados utilizados.

Observa-se que, para todos os percentuais analisados de dados rotulados manualmente, o desempenho obtido no cenário híbrido é superior ao caso supervisionado, apresentando ganhos de até 11% no caso de treino com 5% de dados rotulados manualmente. Diferente dos resultados obtidos com as técnicas clássicas de SSOD, a introdução de maiores volumes de dados pseudo-rotulados não resulta na degradação do desempenho do modelo treinado, o que indica que os pseudo-rótulos produzidos apresentam qualidade suficiente para contribuir com informação complementar à extraída dos dados rotulados manualmente.

Nota-se também que, no cenário híbrido total, o desempenho do modelo permanece relativamente estável, mesmo com variações na quantidade de dados rotulados manualmente. Esse comportamento sugere que o uso de proporções tão grandes de pseudo-rótulos reduz significativamente a dependência do treinamento em relação aos rótulos de origem humana, dominando o treinamento do modelo com as informações ruidosas dos pseudo-rótulos. Observa-se ainda que o caso híbrido total apresentou desempenho superior para todos os volumes de dados rotulados manualmente analisados. Entretanto, a diminuição da diferença de desempenho entre as abordagens à medida em que o volume de dados rotulados manualmente aumenta, indica uma saturação do modelo com relação à informação adicionada pelos dados pseudo-rotulados.

Retornando à Figura 5.3, a curva referente ao cenário pseudo-rotulado apresenta os resultados obtidos quando o treinamento é realizado exclusivamente com anotações automáticas, variando-se a fração de dados utilizada com base no eixo superior do gráfico. Os resultados indicam uma tendência de aumento do desempenho à medida que mais pseudo-rótulos são utilizados no treinamento. Esse comportamento demonstra que os pseudo-rótulos gerados pelo SAM3 contribuem de forma suficiente para conduzir o aprendizado do modelo de detecção, mesmo em cenários de completa ausência de dados com rótulos de origem humana.

Entretanto, os valores absolutos observados são inferiores aos obtidos nos cenários híbrido simétrico e supervisionado. Esse resultado sugere que os pseudo-rótulos introduzem ruído de rótulos ao treinamento, mesmo que inferior ao observado nos

casos de SSOD clássicos e que não inviabiliza completamente o treinamento no caso avaliado. Uma possível origem para essa degradação de qualidade dos pseudo-rótulos quando comparados aos rótulos de origem humana é oriunda da natureza imperfeita dos modelos de fundação, que não são capazes de reproduzir perfeitamente a rotulação humana. Para avaliar a hipótese de introdução de ruído ao treinamento pelos pseudo-rótulos, foi feita uma análise comparativa entre essas anotações e as disponibilizadas no conjunto de dados utilizado. A Tabela 5.1 apresenta métricas relevantes para a avaliação em termos da qualidade de localização, capacidade de percepção da existência de objetos, precisão de detecção e a quantidade de pseudo-rótulos que não correspondem a nenhum objeto no conjunto de dados.

Tabela 5.1: Análise de qualidade dos pseudo-rótulos gerados pelo SAM3 em comparação aos rótulos manuais do conjunto BDD100K ($\tau = 0.5$).

Métrica	**Valor**
IoU médio	0.33
IoU ≥ 0.5 (%)	38.39
Precisão (%)	46.02
Falsos positivos (%)	53.98

Os resultados apresentados na Tabela 5.1 indicam um valor de 33% de sobreposição média entre as anotações manuais e geradas pelo SAM3, ilustrado pela métrica de interseção sobre união (*Intersection over Union - IoU*), dando indícios de que não há alinhamento perfeito entre as caixas delimitadoras. Adicionalmente, nota-se que menos de 40% dos pseudo-rótulos apresentam sobreposição maior que 50% com as caixas de referência, contribuindo para a observação da limitação de precisão espacial dos objetos nas imagens. Outros resultados desta análise apontam para uma precisão de detecção inferior ao valor de 50% e uma taxa de falsos positivos superior à metade dos pseudo-rótulos, sugerindo relevante produção de detecções incorretas, que não correspondem aos objetos observados nos dados de referência. Tais métricas são consistentes com a hipótese de geração de ruído de rótulos devido à limitação do modelo gerador das anotações automáticas.

Apesar dos resultados absolutos observados serem inferiores aos alcançados nos cenários supervisionado e híbrido simétrico, a viabilização do treinamento em cenários de escassez completa de dados rotulados manualmente apresenta relevância prática em aplicações nas quais a produção de rótulos é inviável. Nessas situações, a técnica avaliada pode atuar em etapas iniciais do aprendizado, sendo possível uma alternativa às abordagens de solução ao problema de *cold start*, comumente relevante em contextos como aprendizado ativo [36]. Entretanto, o desempenho do modelo treinado por esse tipo de abordagem depende da qualidade dos pseudo-rótulos utilizados no treinamento. Dessa forma, a seleção adequada dos critérios de filtragem de pseudo-rótulos é uma etapa importante para o desenvolvimento dos sistemas avaliados na presente análise.

5.2.2 Impacto do Limiar de Confiança no Desempenho de Treinamento

O limiar de confiança τ utilizado na filtragem dos pseudo-rótulos define a quantidade e a diversidade das anotações utilizadas no treinamento. Valores mais baixos de τ tendem a incluir um maior número de pseudo-rótulos, potencialmente mais ruidosos, enquanto valores mais elevados levam à priorização de predições com maior confiança em detrimento da quantidade de dados disponíveis para aprendizado.

A Figura 5.4 apresenta o impacto da variação do limiar de confiança sobre o desempenho de detecção para os regimes supervisionado, híbrido simétrico e de uso exclusivo de pseudo-rótulos. Para garantir uma avaliação abrangente, a análise foi desdobrada em quatro cenários distintos, variando a proporção de dados base utilizada (1%, 2%, 5% e 10%). Nas configurações híbrida simétrica e supervisionada, esse percentual refere-se aos dados rotulados manualmente, enquanto no cenário pseudo-rotulado, refere-se ao volume de anotações geradas pelo SAM3.

Observa-se que o impacto da variação do limiar de confiança no regime puramente pseudo-rotulado difere de acordo com o volume de dados utilizado. Para as proporções menores (1% e 2%), o desempenho final do modelo apresenta alterações muito discretas. Contudo, nos cenários com maior volume de dados (5% e 10%), a

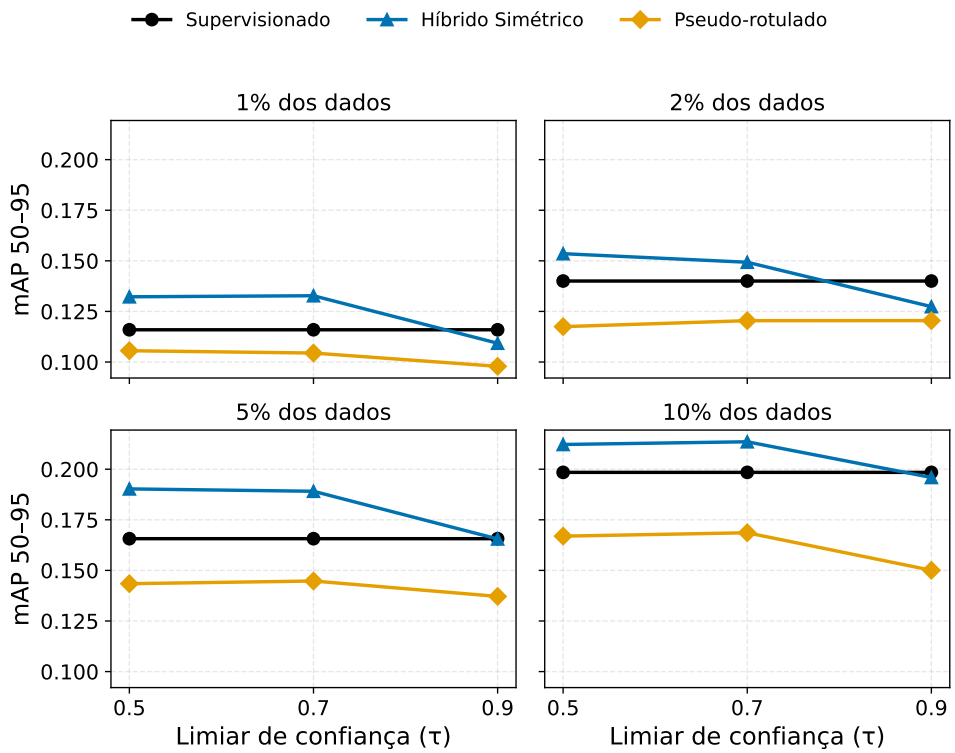

Figura 5.4: Impacto do limiar de confiança τ no desempenho do modelo para diferentes configurações de rotulação considerando o uso de 1%, 2%, 5% e 10% de dados.

drástica redução de anotações imposta por $\tau = 0.9$ passa a prejudicar o aprendizado, resultando em quedas visíveis de desempenho. O rigor desse limiar descarta um volume de dados muito expressivo, conforme apresentado na Tabela 5.2: ao variar τ de 0.7 para 0.9, observa-se um crescimento de 14.33% na quantidade de imagens sem objetos e uma redução de 5.93% na média de objetos por imagem. Essa escassez extrema demonstra que priorizar exclusivamente pseudo-rótulos de alta confiança não promove ganhos de desempenho que superem o benefício de ter um maior volume de dados, mesmo que apresentem algum nível de ruído.

No cenário híbrido simétrico, observa-se uma sensibilidade ainda mais acentuada à escolha do limiar de confiança. Para os valores iniciais ($\tau = 0.5$ e $\tau = 0.7$), a contribuição dos pseudo-rótulos ao processo de aprendizado é positiva, mantendo o desempenho da configuração superior ao caso puramente supervisionado. Entretanto, ao adotar o valor restritivo de $\tau = 0.9$, o modelo híbrido sofre uma brusca degradação ao seu desempenho em todos os volumes avaliados, decaindo a ponto de

Tabela 5.2: Impacto do limiar de confiança τ na densidade de objetos por imagem e número de imagens sem objetos para rótulos manuais e pseudo-rótulos gerados pelo SAM3.

Fonte dos rótulos	τ	**Objetos por imagem**	**Imagens sem objetos (%)**
Rótulos manuais	N/A	18.38	0.20
Pseudo-rótulos	0.5	15.35	0.72
Pseudo-rótulos	0.7	9.28	2.33
Pseudo-rótulos	0.9	3.35	16.66

igualar-se ou tornar-se inferior ao do seu respectivo valor obtido no caso supevisionado.

Apesar das diferenças observadas no comportamento das configurações avaliadas, os resultados apresentados na Tabela 5.2 indicam que o aumento do limiar de confiança reduz a quantidade de anotações utilizadas no treinamento para todos os casos, possivelmente resultando em menor custo computacional associado ao processo de treinamento. Com o fim de verificar se existe relação entre quantidade de rótulos e o custo computacional, são conduzidos experimentos referentes à relação de custo e ao desempenho da abordagem estudada a seguir.

5.2.3 Tempo de Execução de Treinamento

Para complementar a análise baseada exclusivamente na métrica de desempenho em diferentes configurações de distribuição de dados, a Figura 5.5 apresenta a relação entre desempenho de detecção e o tempo total de treinamento, considerando diferentes distribuições de dados.

Em todos os casos avaliados, observa-se um aumento gradual do desempenho à medida que a fração de dados rotulados manualmente cresce, acompanhado por um aumento do tempo de treinamento. Configurações que utilizam quantidades semelhantes de dados, sejam eles rotulados manualmente ou pseudo-rotulados, apresentam custos igualmente semelhantes, sugerindo baixa sensibilidade do custo computacional ao cenário de rotulação adotado. Esses resultados indicam que a parcela dominante do custo envolvido no treinamento dos modelos de detecção não está re-

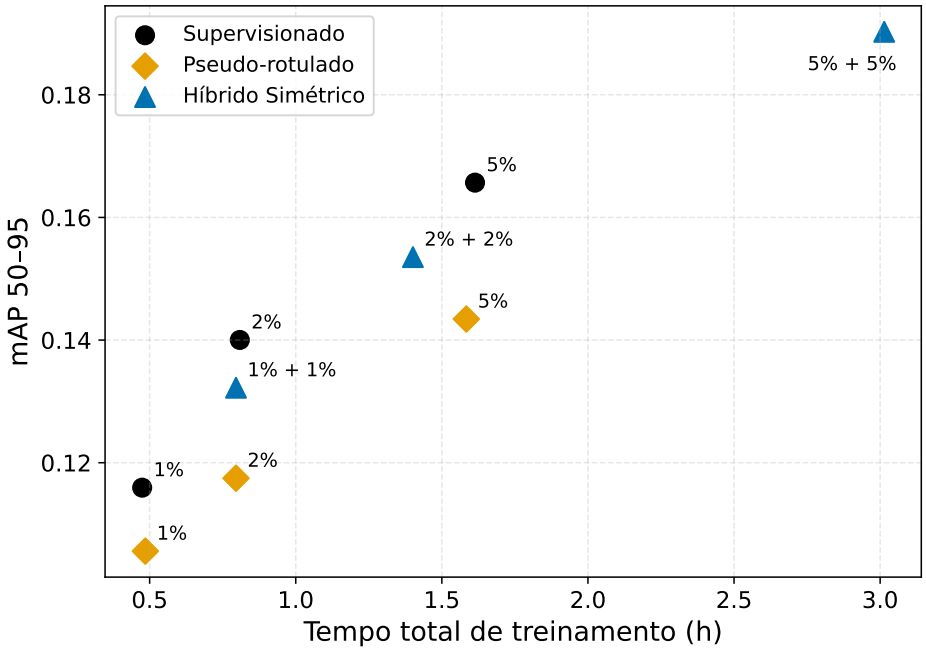

Figura 5.5: Relação entre desempenho (mAP 50-95) e tempo total de treinamento para diferentes distribuições de dados, considerando $\tau = 0.5$.

lacionada com o procedimento de comparação de predições com os rótulos durante o cálculo da função de custo.

5.2.4 Análise por Classe

Embora as avaliações conduzidas neste trabalho apontem para ganhos de desempenho em termos de mAP com a introdução de dados pseudo-rotulados pelo modelo SAM3 em diferentes cenários, o uso de métricas agregadas podem mascarar impactos assimétricos dessa técnica entre as classes do conjunto de dados. Características do processo de geração de pseudo-rótulos, como vieses relacionados à definição da ontologia de rotulação ou ao conjunto de dados utilizado no treinamento do modelo de fundação, podem produzir distribuições de qualidade de rótulos não uniformes [7].

Este trabalho analisa a distribuição de desempenho entre as classes do conjunto de dados, considerando o impacto do uso do SAM3 na geração de pseudo-rótulos. A Figura 5.6 mostra a diferença de precisão média (*Average Precision - AP*) por classe entre o treinamento híbrido e o treinamento puramente supervisionado, considerando diferentes proporções de dados rotulados e utilizando um limiar de confiança fixo

$\tau = 0.5$.

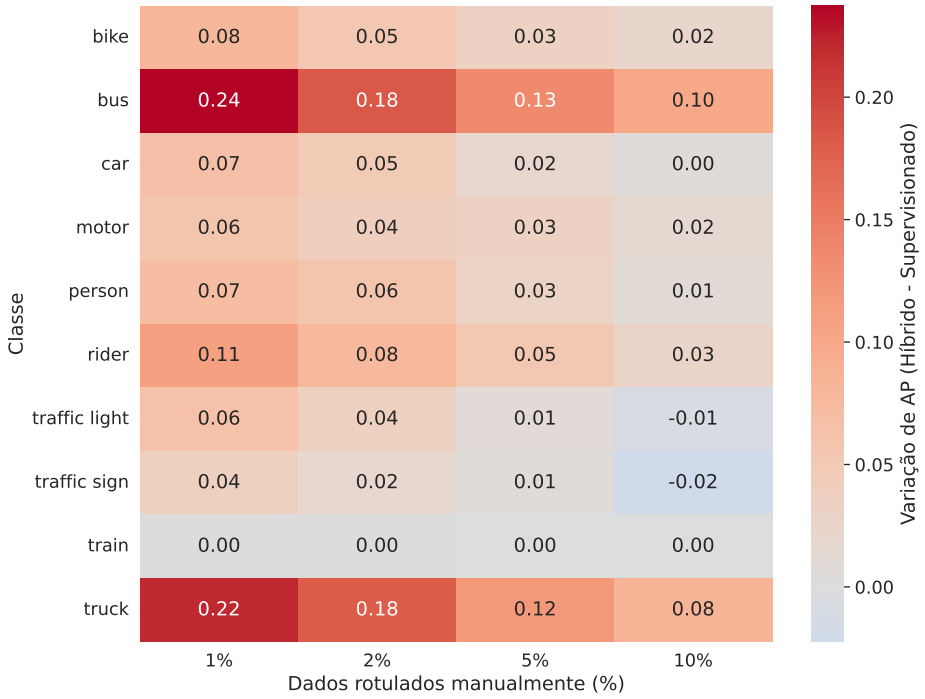

Figura 5.6: Mapa de calor relacionando ganho de desempenho de detecção (AP 50-95) por classe ao incluir dados pseudo-rotulados e para diferentes quantidades de dados rotulados manualmente. Comparação entre caso supervisionado e caso híbrido simétrico.

A análise por classe apresentada na Figura 5.6 demonstra que os ganhos de desempenho provenientes da inclusão de pseudo-rótulos no treinamento não se distribuem de maneira uniforme entre as categorias avaliadas. As classes *bus* e *truck* apresentam os maiores ganhos de desempenho. Em contraste, classes como *traffic sign*, *traffic light* e *car* obtiveram as menores variações, sendo observado valores constantes para a categoria *train*.

5.2.5 Discussão sobre a Geração de Rótulos para Detecção de Objetos com o SAM3

Os resultados apresentados neste capítulo demonstram que o uso do SAM3 como gerador de pseudo-rótulos é uma estratégia viável para mitigar os efeitos da escassez

de dados rotulados em tarefas de detecção de objetos. A abordagem mostrou-se particularmente vantajosa quando combinada com uma fração de dados rotulados manualmente, proporcionando ganhos de desempenho de até 14.9% no caso híbrido simétrico e 11.8% no caso híbrido total.

Os experimentos referentes à configuração de treinamento com uso exclusivo de dados pseudo-rotulados indicam que, apesar de apresentar desempenho inferior aos correspondentes supervisionado e híbridos, o modelo treinado apenas com pseudo-rótulos é capaz de realizar aprendizado, mesmo na completa ausência de dados rotulados por humanos. Dessa forma, essa abordagem é promissora como uma etapa inicial de treinamento, possivelmente atuando como alternativa para mitigar o desafio de *cold start*, que é um tema relevante em técnicas como as de aprendizado ativo.

A análise da sensibilidade ao limiar de confiança sugere que a abordagem baseada no uso exclusivo de pseudo-rótulos apresenta baixa sensibilidade à escolha desse hiperparâmetro. Por outro lado, observou-se queda relevante de desempenho na abordagem híbrida simétrica, especialmente para valores elevados de τ.

Adicionalmente, a análise referente ao custo computacional de treinamento em cada configuração de rotulação indica que não há relação relevante entre a estratégia adotada e a duração do treinamento. Observa-se, porém, que há predominante influência do volume de dados utilizados no tempo de treinamento, mesmo que esses dados não apresentem rótulos. Essa característica reforça a baixa influência do tempo de processamento dos rótulos no tempo total de treinamento.

Por fim, a comparação entre o desempenho do caso híbrido simétrico e supervisionado para cada classe do conjunto de dados demonstra que os ganhos proporcionados pelo uso dos pseudo-rótulos gerados pelo SAM3 não são observados de maneira uniforme entre as classes, o que pode ser pode ser mascarado em análises baseadas em métricas agregadas como o mAP.

Capítulo 6

Conclusões e Trabalhos Futuros

Este trabalho investigou a viabilidade de diferentes abordagens para o treinamento de modelos de detecção de objetos de um estágio em cenários de escassez de dados rotulados. Essa investigação foi conduzida por duas frentes. Na primeira, técnicas clássicas da literatura de detecção de objetos semissupervisionada (*Semi Supervised Object Detection - SSOD*), incluindo estratégias baseadas no uso de pseudo-rótulos, regularização por consistência e o método S4OD. Na segunda frente, foi avaliado o uso de modelos de fundação, como o *Segment Anything Model 3* (*SAM3*), para gerar pseudo-rótulos destinados ao treinamento do modelo de detecção.

Os resultados referentes às técnicas tradicionais de SSOD demonstram que, no contexto avaliado, essas abordagens apresentaram desempenho inferior quando comparadas ao treinamento puramente supervisionado, corroborando as limitações já apontadas na literatura. Foi possível também observar relevante instabilidade durante o treinamento dos detectores de um estágio, uma vez que apresentam alta sensibilidade ao ruído de rótulos. Mesmo considerando o efeito regulador das técnicas aplicadas, as suas aplicações não foram capazes de mitigar a instabilidade de desempenho gerada pela baixa qualidade de localização desses modelos associada ao ciclo de viés de confirmação negativo gerado pelo treinamento em pseudo-rótulos de baixa qualidade.

Como alternativa ao uso das técnicas clássicas de SSOD, este trabalho explorou o uso do SAM3 como gerador de pseudo-rótulos. Em contraste com as abordagens avaliadas anteriormente, essa estratégia não depende do próprio detector para

construir os seus rótulos de treinamento, eliminando o problema do ciclo de viés de confirmação. Adicionalmente, ao tomar proveito do conhecimento semântico acumulado nestes modelos durante o pré-treinamento, é possível iniciar a geração de rótulos sem qualquer ajuste do modelo, nem a presença de dados rotulados manualmente, possibilitando a aplicabilidade dessa abordagem em cenários de completa ausência de dados rotulados por humanos.

Foram realizados experimentos construídos a partir do treinamento de detectores com dados rotulados pelo modelo, variando diferentes combinações de rótulos reais e artificiais, assim como o hiperparâmetro de limiar de confiança τ para controlar o processo de seleção de pseudo-rótulos. As diferentes estratégias de rotulação são resumidas em quatro tipos básicos: supervisionado, híbrido simétrico, híbrido total e pseudo-rotulado. Foi feita uma comparação direta entre os casos supervisionado e híbrido simétrico, resultando em vantagem para o caso híbrido quando considerada a mesma quantidade de dados rotulados manualmente em ambos os casos, o que demonstra a capacidade da abordagem estudada de tornar possível a utilização de dados inicialmente não rotulados.

Experimentos adicionais avaliaram o uso da abordagem de pseudo-rotulação com o SAM3 para realizar treinamento no cenário de ausência completa de dados rotulados manualmente. Nesta análise, foi possível observar que, apesar de apresentar desempenho inferior aos casos que envolvem o uso de dados rotulados manualmente, o modelo é capaz de aprender mesmo na ausência de rotulação de origem humana. Esse resultado sugere o uso de SAM3 como uma alternativa promissora para iniciar o treinamento de modelos em cenários desafiadores de aprendizado, possivelmente atuando como uma alternativa para mitigar o problema de *cold start*, presente em técnicas como a de aprendizado ativo.

A variação do hiperparâmetro responsável por controlar o limiar da filtragem de pseudo-rótulos baseada na confiança do SAM3 promoveu comportamentos diferentes entre as estratégias de rotulação. O desempenho do modelo treinado no cenário pseudo-rotulado mostrou baixa sensibilidade à variação de τ, não apresentando variação relevante, mesmo no esquema mais rigoroso de filtragem. Já o caso híbrido

simétrico demonstrou maior sensibilidade, saindo de um patamar superior ao valor de referência supervisionado para um valor inferior à medida em que se aumenta o valor de τ. Esse resultado indica que, no cenário avaliado, a redução da quantidade de pseudo-rótulos tem maior influência no desempenho de detecção do modelo treinado do que possíveis ganhos correspondentes ao uso de anotações geradas com maior confiança.

Com relação ao custo computacional associado ao treinamento do modelo, experimentos comparativos indicaram que não há influência significativa entre a estratégia de rotulação e o tempo total de treinamento, sendo predominante o efeito da quantidade de dados utilizados no ajuste do modelo. Dessa forma, se considerados modelos treinados em cenários supervisionados e pseudo-rotulados, porém com a mesma quantidade de dados nos dois casos, o tempo total de treinamento dos modelos é similar.

Por fim, a análise de desempenho por categoria apontou que os ganhos observados com a inclusão de dados pseudo-rotulados no treinamento dos modelos não é distribuído de forma uniforme entre as categorias do conjunto de dados. Classes como *truck* e *bus* apresentaram a maior variação, enquanto *train* teve a menor, mantendo-se constante. Apesar dos resultados positivos obtidos considerando métricas agregadas, a não uniformidade do ganho de desempenho entre classes pode ser um limitante à aplicabilidade do SAM3 como gerador de pseudo-rótulos em aplicações que demandem capacidade de detecção consistente entre as classes.

Dentre as possíveis direções de continuação deste trabalho, destaca-se a expansão da abordagem de geração de pseudo-rótulos com modelos de fundação no contexto de aprendizado federado. Nesse paradigma de aprendizado, a premissa de manutenção da privacidade dos usuários impede que os dados de treinamento do cliente sejam transmitidos ao servidor, o que torna o cliente responsável pela rotulação dos seus dados, possivelmente desestimulando sua participação no treinamento. O uso de modelos de fundação para a geração de pseudo-rótulos é uma alternativa flexível para a viabilização de implementações realistas de treinamento federado, uma vez que reduz a dependência de dados rotulados manualmente. Adicionalmente, o uso de

modelos de fundação pode auxiliar no desenvolvimento de técnicas de aprendizado ativo para reduzir a carga de rotulação do usuário.

Referências Bibliográficas

[1] SILVA, R. M., AZEVEDO, G. F., BERTO, M. V., *et al.*, "Vulnerable Road User Detection and Safety Enhancement: A Comprehensive Survey", *arXiv preprint arXiv:2405.19202*, , 2024.

[2] KRIZHEVSKY, A., SUTSKEVER, I., HINTON, G. E., "ImageNet Classification with Deep Convolutional Neural Networks". In: Pereira, F., Burges, C., Bottou, L., *et al.* (eds.), *Advances in Neural Information Processing Systems*, v. 25, 2012.

[3] GIRSHICK, R., DONAHUE, J., DARRELL, T., *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation", 2014.

[4] CHEN, Y., MANCINI, M., ZHU, X., *et al.*, "Semi-Supervised and Unsupervised Deep Visual Learning: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, n. 3, pp. 1327–1347, 2024.

[5] REDMON, J., DIVVALA, S., GIRSHICK, R., *et al.*, "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

[6] KIM, T., LIN, E., LEE, J., *et al.*, "Navigating Data Heterogeneity in Federated Learning: A Semi-Supervised Federated Object Detection". In: Oh, A., Naumann, T., Globerson, A., *et al.* (eds.), *Advances in Neural Information Processing Systems*, v. 36, pp. 2074–2096, 2023.

[7] CARION, N., GUSTAFSON, L., HU, Y.-T., *et al.*, "SAM 3: Segment Anything with Concepts", 2025.

[8] YU, F., CHEN, H., WANG, X., *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

[9] ZOU, Z., CHEN, K., SHI, Z., *et al.*, "Object Detection in 20 Years: A Survey", *Proceedings of the IEEE*, v. 111, n. 3, pp. 257–276, 2023.

[10] GIRSHICK, R., "Fast R-CNN". In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[11] REN, S., HE, K., GIRSHICK, R., *et al.*, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: Cortes, C., Lawrence, N., Lee, D., *et al.* (eds.), *Advances in Neural Information Processing Systems*, v. 28, 2015.

[12] DALAL, N., TRIGGS, B., "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, v. 1, pp. 886–893 vol. 1, 2005.

[13] FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D., *et al.*, "Object Detection with Discriminatively Trained Part-Based Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 9, pp. 1627–1645, 2010.

[14] BALCAN, M.-F., BLUM, A., "A discriminative model for semi-supervised learning", *J. ACM*, v. 57, n. 3, Mar. 2010.

[15] CHAPELLE, O., SCHOLKOPF, B., ZIEN, EDS., A., "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]", *IEEE Transactions on Neural Networks*, v. 20, n. 3, pp. 542–542, 2009.

[16] LEE, D.-H., "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: *Workshop on Challenges in Representation Learning, ICML*, 2013.

[17] LAINE, S., AILA, T., "Temporal Ensembling for Semi-Supervised Learning", 2017.

[18] MIYATO, T., MAEDA, S.-I., KOYAMA, M., *et al.*, "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 41, n. 8, pp. 1979–1993, 2019.

[19] TARVAINEN, A., VALPOLA, H., "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results", *Advances in neural information processing systems*, v. 30, 2017.

[20] SOHN, K., ZHANG, Z., LI, C.-L., *et al.*, "A Simple Semi-Supervised Learning Framework for Object Detection", 2020.

[21] JEONG, J., VERMA, V., HYUN, M., *et al.*, "Interpolation-Based Semi-Supervised Learning for Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11602–11611, June 2021.

[22] LIU, W., ANGUELOV, D., ERHAN, D., *et al.*, *SSD: Single Shot MultiBox Detector*, Springer International Publishing, p. 21–37, 2016.

[23] TIAN, Z., SHEN, C., CHEN, H., *et al.*, "FCOS: Fully Convolutional One-Stage Object Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] ZHOU, Q., YU, C., WANG, Z., *et al.*, "Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4081–4090, June 2021.

[25] ZHANG, Y., YAO, X., LIU, C., *et al.*, "S4OD: Semi-Supervised learning for Single-Stage Object Detection", 2022.

[26] KIRILLOV, A., MINTUN, E., RAVI, N., *et al.*, "Segment Anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

[27] WANG, H., JIA, T., WANG, Q., *et al.*, "WS-SAM: Generalizing SAM to Weakly Supervised Object Detection With Category Label", *IEEE Transactions on Image Processing*, v. 34, pp. 4052–4066, 2025.

[28] BHASKAR, U., BHATTACHARYA, R., PATEL, A., *et al.*, "Robust Object Detection with Pseudo Labels from VLMs using Per-Object Co-teaching", 2025.

[29] GAO, R., WANG, L., "MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9901–9910, October 2023.

[30] GUPTA, H., KOTLYAR, O., ANDREASSON, H., *et al.*, "Robust Object Detection in Challenging Weather Conditions". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7523–7532, January 2024.

[31] LIN, T.-Y., MAIRE, M., BELONGIE, S., *et al.*, "Microsoft coco: Common objects in context". In: *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755, Springer, 2014.

[32] YUNUSOV, J., "YOLOv8-pt", `https://github.com/jahongir7174/YOLOv8-pt`, 2023, GitHub repository.

[33] LIU, Y.-C., MA, C.-Y., HE, Z., *et al.*, "Unbiased Teacher for Semi-Supervised Object Detection", 2021.

[34] CARION, N., MASSA, F., SYNNAEVE, G., *et al.*, "End-to-End Object Detection with Transformers", 2020.

[35] Roboflow, "autodistill", `https://github.com/autodistill/autodistill`, Jan. 2023, Version 0.1.0, MIT license.

[36] JIN, Q., YUAN, M., LI, S., *et al.*, "Cold-start active learning for image classification", *Information Sciences*, v. 616, pp. 16–36, 2022.