Detecção de objetos em cenários de escassez de rótulos usando pseudo-rótulos gerados pelo SAM3

João V. D. Sobrinho[1], **Miguel E. M. Campista**[1]

[1]Universidade Federal do Rio de Janeiro, GTA/DEL-Poli/PEE-COPPE
Rio de Janeiro – RJ – Brasil

{joaosobrinho,miguel}@gta.ufrj.br

Abstract. *Label scarcity remains a major challenge for training object detection models. Semi-supervised object detection methods attempt to address this issue by exploiting unlabeled data through pseudo-labeling, but they often exhibit instability in single-stage detectors. In this work, we investigate the use of the foundation model Segment Anything Model 3 (SAM3) as an automatic label generator for training object detection models. Experiments show that using SAM3-generated labels can outperform purely supervised training when the same amount of labeled data is used. These results highlight the potential of foundation models to reduce the reliance on manual annotations in object detection.*

Resumo. *A escassez de dados rotulados é um dos principais desafios para o treinamento de modelos de detecção de objetos. Técnicas de detecção semissupervisionada abordam esse problema explorando dados não rotulados por meio de pseudo-rótulos, porém podem apresentar instabilidade em detectores de um estágio. Este trabalho investiga o uso do modelo de fundação Segment Anything Model 3 (SAM3) como gerador automático de rótulos para o treinamento de detectores de objetos. Experimentos indicam que o uso de rótulos gerados pelo SAM3 pode superar o treinamento puramente supervisionado considerando a mesma quantidade de dados rotulados manualmente, evidenciando o potencial de modelos de fundação para reduzir a dependência de anotações humanas.*

1. Introdução

O desenvolvimento de algoritmos para aprendizado profundo promoveu ganhos de desempenho em diversas tarefas, viabilizando a resolução de problemas que, até então, apresentavam complexidade proibitiva [Zou et al. 2023]. A detecção de objetos destaca-se como um problema de visão computacional de particular dificuldade. No caso de classificação, o modelo responsável pela tarefa é treinado para gerar uma única categoria que defina a imagem como um todo. Já para detecção de objetos, o modelo deve não somente categorizar, como localizar todos os objetos de interesse pertencentes à cena avaliada. O treinamento desse tipo de modelo, portanto, envolve o uso de rótulos de produção mais custosa quando comparados aos utilizados em classificação, o que pode reduzir a disponibilidade de dados rotulados e, consequentemente, gerar desafios quanto à implementação do sistema de detecção.

Trabalhos da literatura apontam o uso de técnicas de aprendizado semissupervisionado como uma alternativa para lidar com o problema de escassez de rótulos.

Nesse paradigma, busca-se explorar estruturas presentes nos dados não rotulados como fonte de informação complementar aos dados rotulados disponíveis [Chapelle et al. 2009, Lee 2013]. No contexto da detecção de objetos semissupervisionada (*Semi-Supervised Object Detection - SSOD*), técnicas como a geração automática de rótulos, conhecidos como pseudo-rótulos [Lee 2013], e a regularização por consistência [Sohn et al. 2020] são amplamente adotadas na literatura. Entretanto, a aplicação dessas abordagens em detectores de um estágio, como os modelos da família *You Only Look Once* (YOLO) [Redmon et al. 2016], pode ser desafiadora, uma vez que esses modelos apresentam maior sensibilidade ao uso de pseudo-rótulos de baixa qualidade [Zhang et al. 2022]. Além disso, muitas técnicas de SSOD introduzem procedimentos adicionais de treinamento que aumentam o custo computacional do processo, o que pode limitar a aplicabilidade dessas soluções em cenários restritivos.

Paralelamente às abordagens de SSOD, o surgimento de modelos de fundação multimodais abriu novas possibilidades para a geração automática de anotações. Esses modelos, treinados em larga escala sobre grandes volumes de dados visuais e textuais, são capazes de executar tarefas de visão computacional a partir de instruções textuais, mesmo sem treinamento específico para o domínio da atividade final desejada. Dessa forma, é possível utilizar esse tipo de modelo como um gerador de pseudo-rótulos para o treinamento de detectores especializados, potencialmente mitigando limitações associadas às abordagens tradicionais de SSOD.

Neste contexto, este trabalho investiga o uso do modelo de fundação multimodal *Segment Anything Model 3* (SAM3) [Carion et al. 2025] como gerador automático de rótulos para o treinamento de detectores de objetos de um estágio em cenários de escassez de dados rotulados. Particularmente, este trabalho avalia a qualidade dos pseudo-rótulos produzidos pelo modelo e o impacto de sua utilização no treinamento de um detector da família YOLO em diferentes configurações de disponibilidade de dados rotulados. Os resultados obtidos sugerem que o uso de rótulos gerados pelo SAM3 pode promover ganhos de desempenho ao detector seja em combinação a uma pequena parcela de dados rotulados manualmente, ou na ausência completa de anotações produzidas por humanos.

O restante deste trabalho é organizado da seguinte forma. A Seção 2 apresenta trabalhos relacionados. A Seção 3 descreve os experimentos e a metodologia experimental adotada neste trabalho, definindo o conjunto de dados, os modelos utilizados e as configurações de rotulação avaliadas. A Seção 4 apresenta e discute os resultados obtidos. Por fim, a Seção 5 apresenta as conclusões alcançadas, apontando possíveis direções futuras.

2. Trabalhos Relacionados

2.1. Detecção de Objetos Semissupervisionada

A detecção de objetos semissupervisionada (*Semi-Supervised Object Detection - SSOD*) busca explorar dados não rotulados para reduzir a dependência de grandes conjuntos de dados anotados manualmente. A maioria das abordagens propostas na literatura baseia-se na geração de pseudo-rótulos a partir de um modelo previamente treinado, frequentemente combinada com técnicas de regularização por consistência para explorar a estrutura dos dados não rotulados [Lee 2013, Sohn et al. 2020].

Apesar dos avanços recentes, trabalhos da literatura apontam que a aplicação de técnicas de detecção de objetos semissupervisionada a detectores de um estágio pode ser mais desafiadora quando comparada a detectores de dois estágios [Zhang et al. 2022]. Esses trabalhos observam que a qualidade dos pseudo-rótulos tem impacto significativo no processo de treinamento, podendo afetar a estabilidade ou limitar os ganhos de desempenho obtidos nesses modelos.

Existem, entretanto, propostas direcionadas a detectores de um estágio. O método S4OD [Zhang et al. 2022] introduz mecanismos para controlar a qualidade dos pseudo-rótulos durante o treinamento. Contudo, essa abordagem requer procedimentos adicionais de validação para estimar dinamicamente limiares de confiança, o que aumenta a complexidade e o custo computacional do processo de treinamento.

2.2. Modelos de Fundação em Visão Computacional

O recente surgimento de modelos de fundação treinados em larga escala abriu novas possibilidades para diversas tarefas de visão computacional. Esses modelos são treinados em grandes volumes de dados visuais ou multimodais e apresentam forte capacidade de generalização para diferentes tarefas, mesmo sem treinamento específico no domínio alvo [Bommasani et al. 2022]. Exemplos desse tipo de modelo incluem modelos de visão–linguagem *(Vision Language Model - VLM)*, como o SAM3 [Carion et al. 2025].

Diversos trabalhos têm explorado o uso direto de modelos de fundação para tarefas de visão computacional. Figueiredo e Melo [Figueiredo and Melo 2025] investigam o uso de diferentes VLMs para classificação de imagens *zero-shot*, avaliando o impacto de estratégias de *prompting* no desempenho dos modelos. De forma semelhante, Roy et al. [Roy et al. 2025] utilizam modelos de fundação para detectar objetos fora de contexto (*out-of-context*) em imagens. Nessas abordagens, o próprio modelo de fundação é utilizado diretamente para realizar a tarefa desejada, não envolvendo o treinamento adicional de modelos especializados.

Em contraste com essas abordagens, modelos de fundação também têm sido explorados como ferramentas para geração automática de anotações. Bhaskar et al. [Bhaskar et al. 2025] utilizam um modelo de fundação como gerador de pseudo-rótulos para treinar detectores YOLOv5, combinando essa estratégia com co-treinamento para reduzir o impacto de rótulos de baixa qualidade. Entretanto, existe um número reduzido de estudos que investigam o uso de modelos de fundação como geradores de pseudo-rótulos para o treinamento de detectores de um estágio modernos em cenários de escassez de dados rotulados.

Nesse contexto, a principal contribuição deste trabalho consiste na investigação do uso do SAM3 como gerador de pseudo-rótulos para o treinamento de um detector YOLO11 em diferentes cenários de escassez de rótulos. Diferentemente de abordagens de SSOD, nas quais os pseudo-rótulos são gerados pelo próprio detector durante o treinamento, a estratégia avaliada neste trabalho utiliza um modelo de fundação pré-treinado como responsável pela geração das anotações que serão utilizadas no treinamento do modelo alvo.

3. Geração de Rótulos com Modelos de Fundação

3.1. Segment Anything Model 3

O *Segment Anything Model 3* (SAM3) [Carion et al. 2025] é um modelo de fundação treinado em larga escala para segmentação de imagens. O modelo utiliza uma arquitetura baseada em *transformers* capaz de produzir máscaras de segmentação a partir de instruções semânticas fornecidas na forma de *prompts* textuais. Apesar de ser projetado para segmentação, as máscaras geradas pelo modelo podem ser convertidas em caixas delimitadoras, permitindo sua aplicação em tarefas de detecção de objetos. A escolha do SAM3 neste trabalho foi motivada pelo seu lançamento recente e pelo número ainda reduzido de estudos investigando sua utilização como gerador de pseudo-rótulos para treinamento de detectores de objetos.

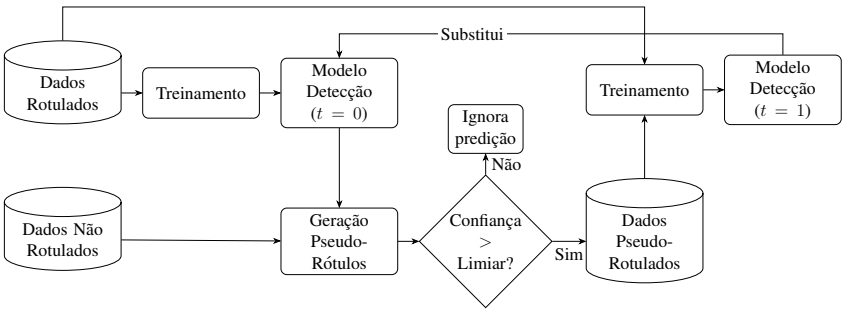
Neste trabalho, o SAM3 é utilizado como gerador automático de pseudo-rótulos. Diferentemente das abordagens tradicionais de SSOD, nas quais o próprio detector produz as anotações utilizadas durante o seu treinamento, os rótulos são gerados por um modelo externo pré-treinado. Esse desacoplamento elimina o ciclo de realimentação responsável pelo fenômeno de viés de confirmação observado em métodos baseados em pseudo-rotulação, em que pseudo-rótulos de baixa qualidade degradam o desempenho do modelo ao serem utilizados no treinamento. Além disso, a geração de rótulos pode ser realizada de forma *offline*, antes do treinamento do detector, evitando a introdução de etapas adicionais e potencialmente custosas durante o processo de treinamento.

A Figura 1 apresenta uma comparação entre o fluxo tradicional de pseudo-rotulação utilizado em SSOD e a abordagem adotada neste trabalho. No caso da Figura 1(a), o próprio detector produz rótulos para o seu treinamento, formando um ciclo de realimentação que pode reforçar erros do modelo. Já na Figura 1(b), esse fluxo é quebrado em dois momentos $t = 0$, em que o SAM3 gera os pseudo-rótulos, e $t = 1$ em que esses dados recém rotulados são utilizados no treinamento. Essa separação permite não somente a eliminação da realimentação do erro, como também a possibilidade de operação *offline*, não incorrendo custo adicional em tempo de treinamento.

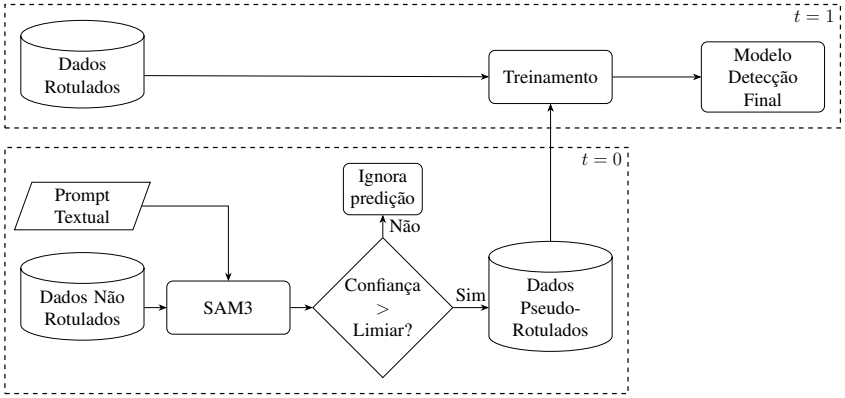
A estratégia avaliada neste trabalho pode ser interpretada como uma forma de transferência indireta de conhecimento, em que um modelo de fundação de grande porte é utilizado para gerar anotações destinadas ao treinamento de um detector mais leve e especializado. Ferramentas como a biblioteca *autodistill* [Roboflow 2023] implementam essa ideia, fornecendo acesso a destilação de modelos de fundação para diferentes arquiteturas de destino. Entretanto, devido ao lançamento recente do SAM3, a biblioteca ainda não oferecia suporte a esse modelo durante a realização deste trabalho. Por esse motivo, foi utilizado um fluxo simplificado para geração automática de pseudo-rótulos diretamente pelo SAM3.

3.2. Geração de Pseudo-Rótulos

A geração de pseudo-rótulos é realizada em uma etapa prévia ao treinamento do detector. Todas as imagens do conjunto de treinamento são processadas pelo SAM3, que produz segmentações a partir de *prompts* textuais correspondentes às classes do conjunto de dados. Neste trabalho, foram utilizadas como *prompts* as traduções diretas do nome de cada classe do conjunto utilizado para o idioma português. Essa escolha busca reduzir a influência de vieses associados ao processo de produção de *prompts* mais complexos, além



(a) Pseudo-rotulação tradicional.



(b) Geração de pseudo-rótulos com SAM3 avaliada neste trabalho.

Figura 1. Comparação entre o fluxo tradicional de geração de pseudo-rótulos e o método adotado neste trabalho.

de aproximar a avaliação de um cenário mais automatizado de geração de rótulos, sem necessidade de ajustes humanos possivelmente dependentes de conhecimento especialista. As máscaras resultantes são convertidas em caixas delimitadoras e filtradas por meio de um limiar de confiança τ. As anotações aceitas são então convertidas para o formato COCO [Lin et al. 2014], utilizado pelo detector YOLO durante o treinamento.

4. Avaliação do SAM3 como gerador de rótulos

4.1. Conjunto de Dados, Métricas e Ambiente Experimental

Os experimentos utilizam o conjunto de dados BDD100K [Yu et al. 2020], amplamente utilizado em pesquisas de visão computacional voltadas a cenários de condução urbana. O conjunto contém 100 mil imagens anotadas com dez classes de objetos, disponibilizadas com uma divisão de 70% para treinamento, 10% para validação e 20% para teste. Optou-se por utilizar esse conjunto de dados para avaliar a estratégia proposta em um cenário complexo e mais próximo de aplicações reais de detecção de objetos urbanos. O desempenho dos modelos é avaliado pela métrica *mean Average Precision* (mAP) no intervalo de *Intersection over Union* (IoU) de 0,5 a 0,95, adotada neste trabalho por corresponder ao padrão utilizado em benchmarks de detecção de objetos [Lin et al. 2014] e trabalhos de SSOD [Sohn et al. 2020], permitindo avaliar de forma simultânea a capacidade de localização e classificação de modelos de detecção de objetos. Os experimentos foram executados em uma GPU NVIDIA RTX A4000.

Os experimentos utilizam a implementação oficial do detector YOLO11

nano [Jocher and Qiu 2024]. Diferentes casos de treinamento são avaliados variando a proporção de dados rotulados manualmente e pseudo-rotulados, buscando explorar diferentes cenários de escassez de rótulos. Além disso, os experimentos consideram diferentes valores para o limiar de confiança utilizado na filtragem dos pseudo-rótulos, $\tau \in \{0{,}5; 0{,}7; 0{,}9\}$. Cada configuração experimental foi executada 3 vezes, com diferentes inicializações aleatórias, exceto no caso híbrido total devido ao elevado custo associado. A partir dos resultados obtidos, são calculados valores de média e intervalo de confiança para as métricas avaliadas, com o objetivo de mitigar o impacto de variações estocásticas na análise comparativa entre as configurações.

As diferentes estratégias de rotulação avaliadas são apresentadas na Tabela 1 e descritas em seguida.

Tabela 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

Supervisionado: treinamento utilizando apenas dados rotulados manualmente, servindo como caso de referência para comparação com os outros cenários avaliados.

Híbrido Simétrico: uso de quantidades iguais de dados rotulados manualmente e pseudo-rotulados, sendo os dados restantes descartados. Permite avaliar o impacto dos pseudo-rótulos em condições de igualdade entre os volumes de dados rotulados manualmente e pseudo-rotulados.

Híbrido Total: uso de uma pequena fração de dados rotulados manualmente e pseudo-rótulos para todo o restante do conjunto, não havendo descarte de dados. Representa cenários com escassez de rótulos, porém com abundância de dados sem anotações.

Pseudo-rotulado: treinamento realizado utilizando apenas anotações geradas pelo SAM3, representando um caso extremo de escassez de rótulos produzidos por humanos.

4.2. Comparação entre Estratégias de Treinamento

Com o objetivo de contextualizar os valores de desempenho de detecção observados nos cenários avaliados neste trabalho, a Tabela 2 apresenta resultados de referência para o

treinamento puramente supervisionado utilizando volumes crescentes de dados rotulados manualmente, variando de 1% até 100% do conjunto de treinamento.

Tabela 2. Desempenho do treinamento supervisionado considerando diferentes proporções de dados rotulados manualmente.

Dados Rotulados (%)	**1**	**2**	**5**	**10**	**25**	**50**	**75**	**100**
mAP 50–95	0,1153	0,1386	0,1664	0,1996	0,2285	0,2455	0,2549	0,2587

Os resultados observados indicam um aumento progressivo do desempenho à medida que o volume de dados utilizado é aumentado, como esperado. Nota-se também que o modelo treinado com o conjunto de dados completo obteve um mAP de $0{,}2587$, representando uma referência de limite superior de desempenho para as outras estratégias avaliadas neste trabalho.

A Figura 2 apresenta a comparação entre as estratégias supervisionada, híbrida simétrica, híbrida total e pseudo-rotulada considerando $\tau = 0{,}5$. Para todos os percentuais de dados rotulados manualmente avaliados, a configuração híbrida simétrica apresenta desempenho superior ao treinamento puramente supervisionado, com ganhos de até 11% quando apenas 5% dos dados possuem rótulos produzidos por humanos. Nesse cenário, a configuração híbrida simétrica atinge $73{,}9\%$ do desempenho obtido pelo treinamento supervisionado utilizando 100% dos dados rotulados manualmente, enquanto o cenário híbrido total alcança $83{,}7\%$ desse valor utilizando apenas 5% de anotações humanas. Adicionalmente, observa-se que os desempenhos são consistentes entre diferentes execuções dos experimentos, não gerando sobreposição entre os intervalos de confiança observados, o que reforça a robustez dos resultados.

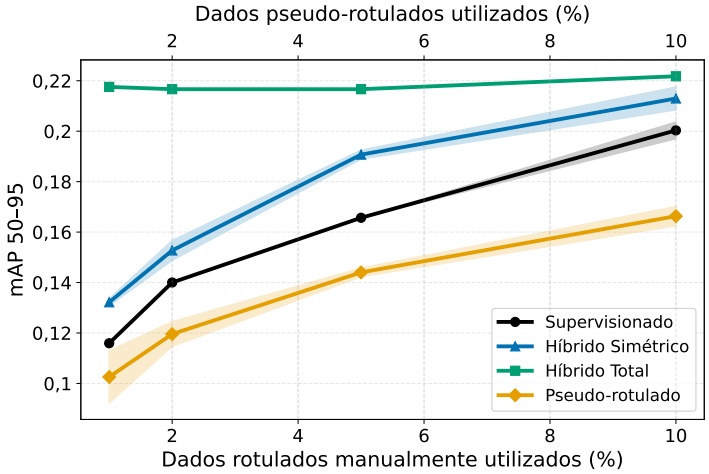


Figura 2. Comparação de desempenho de detecção entre os cenários supervisionado, híbrido simétrico, híbrido total e pseudo-rotulado ($\tau = 0{,}5$). O eixo inferior refere-se à proporção de dados rotulados manualmente destinada ao treinamento, enquanto o eixo superior apresenta a proporção de dados pseudo-rotulados utilizados e aplica-se apenas a curva do caso de treinamento exclusivamente pseudo-rotulado.

Observa-se também que o cenário híbrido total mantém desempenho relativamente estável, mesmo com variações na quantidade de dados rotulados manualmente,

indicando que grandes volumes de pseudo-rótulos podem reduzir a dependência do treinamento em relação a rótulos humanos. Por outro lado, o treinamento exclusivamente pseudo-rotulado apresenta desempenho inferior aos cenários híbridos, embora ainda seja capaz de promover aprendizado ao detector mesmo na ausência completa de anotações manuais. Esse resultado aponta para uma possível introdução de informações imprecisas ao processo de treinamento com o uso de pseudo-rótulos, possivelmente limitando a contribuição desses rótulos ao aprendizado do modelo. Para investigar a qualidade das anotações automáticas, foi realizada uma comparação entre os pseudo-rótulos gerados pelo SAM3 e os rótulos do conjunto BDD100K, apresentada na Tabela 3.

Tabela 3. 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 $\geq$ 0,5 (%)	38,39
Precisão (%)	46,02
Falsos positivos (%)	53,98

Os resultados apresentados na Tabela 3 indicam que os rótulos gerados pelo SAM3 apresentam alinhamento limitado com os rótulos do BDD100K, com IoU médio de $0{,}33$ e menos de 40% das caixas apresentando sobreposição superior a $0{,}5$. A precisão de detecção inferior ao valor de 50% e elevada taxa de falsos positivos apontam para 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. Esses resultados também são consistentes com limitações apontadas por Carion et al. [Carion et al. 2025]

A Figura 3 apresenta exemplos que ilustram esse cenário, comparando as anotações manuais com pseudo-rótulos produzidos pelo SAM3 e predições geradas por um modelo YOLO treinado na configuração híbrida simétrica. Para facilitar a visualização, apenas as anotações e predições referentes à classe "*car*" foram sinalizadas na imagem. É possível observar que o SAM3 produz pseudo-rótulos ruidosos, gerando detecções incorretas e omitindo objetos presentes na cena, o que se agrava à medida que o valor do limiar de confiança aumenta. Esse comportamento é consistente com os resultados apresentados na Tabela 3 ao ilustrar casos de baixa precisão e falsos positivos.

Apesar da baixa qualidade dos pseudo-rótulos gerados pelo SAM3, as predições geradas pelo YOLO treinado em configuração híbrida simétrica, utilizando 5% de dados com anotações manuais, demonstram capacidade de capturar objetos não identificados pelo modelo de referência supervisionado. Como ilustrado na Figura 3, um veículo de menor escala na cena é ignorado pelo modelo puramente supervisionado, mas é detectado pelo modelo híbrido. Adicionalmente, o modelo puramente supervisionado identifica incorretamente uma caminhonete, que seria da classe equivalente "*truck*", como um carro, enquanto o modelo híbrido não comete o mesmo erro.

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

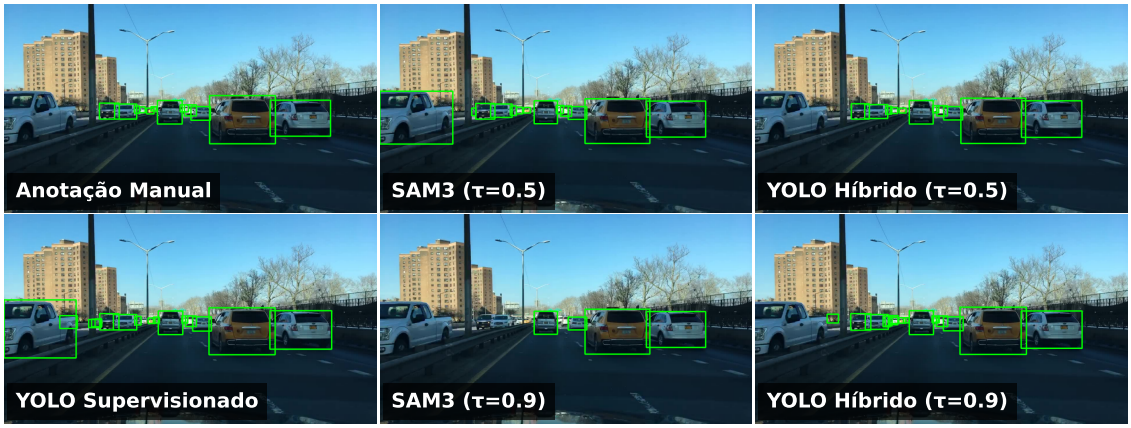


Figura 3. Comparação qualitativa entre anotações manuais referentes apenas à classe carro, produzidas pelo SAM3 e predições do modelo YOLO treinado na configuração híbrida simétrica utilizando 5% de dados rotulados manualmente. Predições consideram filtragem por confiança com limiar $\tau = 0{,}5$ e $\tau = 0{,}9$.

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 [Jin et al. 2022]. 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.

4.3. 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 4 apresenta o impacto da variação do limiar de confiança sobre o ganho relativo de desempenho das configurações híbrida simétrica e pseudo-rotulada em relação ao caso supervisionado. A análise ilustra os quatro volumes de dados utilizados neste trabalho, 1%, 2%, 5% e 10% do conjunto de treino. 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. O alargamento dos intervalos de confiança associado ao aumento de τ sugere que o critério de seleção de pseudo-rótulos mais rigoroso introduz instabilidade ao sistema, o que é corroborado pela Tabela 4. 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. Devido à escassez de anotações por imagem promovida pelo procedimento de filtragem, a informação utilizada para guiar o aprendizado do modelo torna-se limitada, aumentando o impacto de fatores estocásticos como a inicialização dos parâmetros do modelo treinado

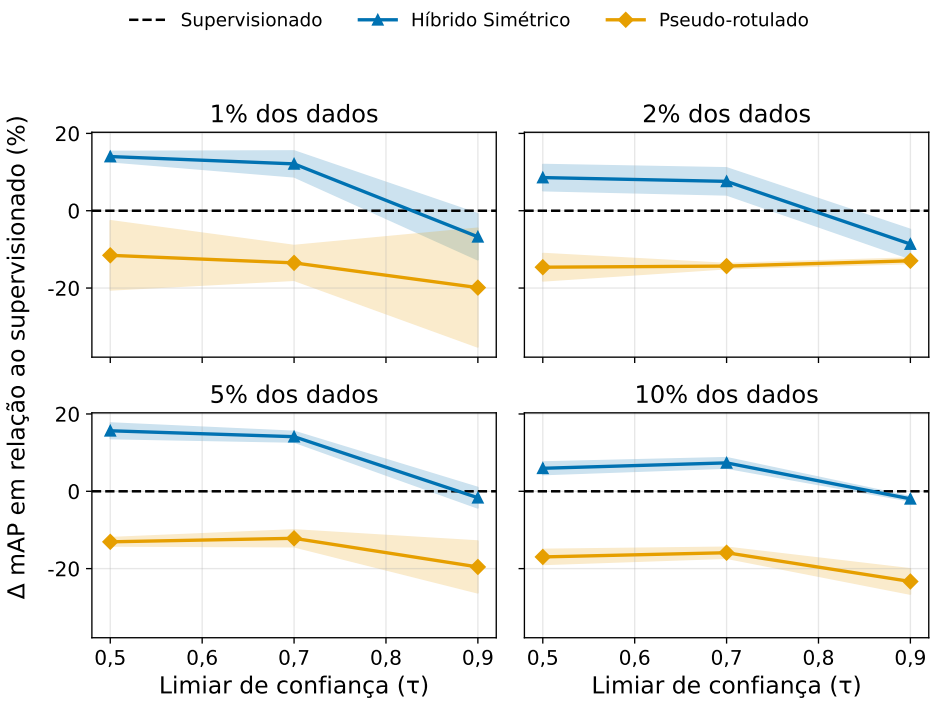


Figura 4. Variação relativa de desempenho em relação ao cenário supervisionado para diferentes valores do limiar de confiança τ considerando o uso de 1%, 2%, 5% e 10% de dados. Valores positivos indicam desempenho superior ao cenário supervisionado considerando a mesma proporção de dados rotulados manualmente, enquanto valores negativos indicam desempenho inferior.

e o procedimento de amostragem dos dados utilizados.

Nos cenários de disponibilidade de apenas 1% e 2% de dados com rótulos, a baixa disponibilidade de dados já introduz instabilidade e alta sensibilidade a efeitos provenientes de aleatoriedades do treinamento. Por outro lado, os casos com maior disponibilidade de dados ilustram cenários mais estáveis, em que a informação obtida pelos dados é suficiente para reduzir o impacto de efeitos externos. Mesmo assim, observa-se que, ao alcançar valores de τ mais elevados, os intervalos de confiança das configurações pseudo-rotulada e híbrida simétrica aumentam. Esses resultados indicam 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.

Tabela 4. 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

No cenário híbrido simétrico, observa-se uma sensibilidade ainda mais acentuada

à escolha do limiar de confiança. Para os valores de $\tau = 0{,}5$ e $\tau = 0{,}7$, a contribuição dos pseudo-rótulos ao processo de aprendizado é positiva, mantendo o desempenho médio da configuração superior ao do caso puramente supervisionado e com intervalos de confiança que não se sobrepõem aos obtidos nesse cenário. Entretanto, ao adotar o valor mais restritivo de $\tau = 0{,}9$, o modelo híbrido sofre uma degradação no seu desempenho em todos os volumes avaliados, obtendo resultados médios iguais ou inferiores ao do caso supervisionado e com intervalos de confiança que passam a se sobrepor aos obtidos nesse cenário.

Apesar das diferenças observadas no comportamento das configurações avaliadas, os resultados indicam que a redução severa da quantidade de anotações disponíveis para o treinamento, provocada pelo aumento do limiar de confiança, compromete os benefícios promovidos pela introdução de maiores volumes e maior diversidade pelos pseudo-rótulos.

5. Conclusão

Este trabalho investigou o uso de modelos de fundação como fonte de pseudo-rótulos para o treinamento de detectores de objetos de um estágio em cenários de escassez de dados rotulados. Em particular, foi avaliado o uso do *Segment Anything Model 3* (SAM3) como gerador automático de anotações para o treinamento de um modelo de detecção YOLO11.

Os resultados obtidos indicam que o uso de pseudo-rótulos gerados pelo SAM3 é capaz de gerar ganho de desempenho para o treinamento do detector quando combinado a uma pequena quantidade de dados rotulados manualmente. Nos experimentos realizados, estratégias híbridas que combinam dados rotulados e pseudo-rotulados apresentaram desempenho superior ao treinamento puramente supervisionado quando considerada a mesma quantidade de dados rotulados manualmente.

Adicionalmente, observou-se que detectores treinados exclusivamente com pseudo-rótulos gerados pelo SAM3 são capazes de aprender mesmo na ausência completa de dados rotulados por humanos. Embora o desempenho obtido nesse cenário seja inferior ao observado quando se utilizam anotações manuais, os resultados indicam que modelos de fundação podem ser utilizados como ponto de partida para o treinamento de detectores em cenários de *cold start*.

Como trabalhos futuros, pretende-se investigar o impacto de diferentes estratégias de construção de *prompts* textuais na qualidade dos pseudo-rótulos gerados, assim como explorar o uso de modelos de fundação para geração de pseudo-rótulos em outros cenários de aprendizado com poucos rótulos, como nos casos de aprendizado federado e aprendizado ativo.

6. Agradecimentos

O presente trabalho foi realizado com apoio do CNPq (310234/2025-5, 407304/2025-8, 408255/2023-4 e 405940/2022-0); da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001 e 88887.954253/2024-00; da FAPERJ (E-26/200.438/2026 e E-26/210.778/2025); da FAPESP (2023/00673-7 e 2023/00811-0) e da Fundação de Desenvolvimento da Pesquisa - Fundep - Rota 2030 em conjunto dos nossos parceiros Stellantis e Mobway.

Referências

Bhaskar, U., Bhattacharya, R., Patel, A., Khoche, S., Kulkarni, P. A., and Manwani, N. (2025). Robust object detection with pseudo labels from vlms using per-object co-teaching.

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2022). On the opportunities and risks of foundation models.

Carion, N., Gustafson, L., Hu, Y.-T., et al. (2025). Sam 3: Segment anything with concepts.

Chapelle, O., Scholkopf, B., and Zien, Eds., A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Figueiredo, C. and Melo, T. (2025). Explorando o uso de vlms para classificação zero-shot de imagens. In *Anais do XVII Simpósio Brasileiro de Computação Ubíqua e Pervasiva*, pages 1–10, Porto Alegre, RS, Brasil. SBC.

Jin, Q., Yuan, M., Li, S., Wang, H., Wang, M., and Song, Z. (2022). Cold-start active learning for image classification. *Information Sciences*, 616:16–36.

Jocher, G. and Qiu, J. (2024). Ultralytics yolo11. `https://github.com/ultralytics/ultralytics`.

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Roboflow (2023). autodistill. `https://github.com/autodistill/autodistill`. Version 0.1.0, MIT license.

Roy, A., Cobb, A., Kaur, R., Jha, S., Bastian, N., Berenbeim, A., Thomson, R., Cruickshank, I., Velasquez, A., and Jha, S. (2025). Zero-shot detection of out-of-context objects using foundation models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 9168–9177.

Sohn, K., Zhang, Z., Li, C.-L., Zhang, H., Lee, C.-Y., and Pfister, T. (2020). A simple semi-supervised learning framework for object detection. In *arXiv:2005.04757*.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Y., Yao, X., Liu, C., Chen, F., Song, X., Xing, T., Hu, R., Chai, H., Xu, P., and Zhang, G. (2022). S4od: Semi-supervised learning for single-stage object detection.

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276.