Pseudo-Box Target Sensitivity for Point-Supervised Cervical Cell Detection and Counting

Jean Carlos Rodrigues Sousa[1], **Karielly de Carvalho**[1],
João Antônio Leal de Miranda[1], **Romuere Rodrigues Veloso e Silva**[1]

[1]Campus Senador Helvídio Nunes de Barros (CSHNB)
Universidade Federal do Piauí (UFPI) – Picos, PI – Brasil

{jean.rodrigues,karielly.carvalho,joaoantonio,romuere}@ufpi.edu.br

Abstract. *Point annotations identify cervical cell centers but do not define object extent. This work studies the experimental consequences of converting such points into detection targets on the CRIC Cervix dataset, across ten pseudo-box sizes (48–192 pixels) and three YOLO11 scales with image-disjoint partitions. Validation F1 forms a broad plateau between 96 and 192 pixels, with distinct precision, recall and counting trade-offs; the 144-pixel target was kept as an intermediate operating point and YOLO11s gave the best balance. On the held-out test set it reached AP50=0.8723, F1=0.8190 and a mean absolute counting error of 4.68 cells per image. A point-distance criterion, independent of the synthetic box, yields an equivalent F1, confirming genuine nucleus localization. The study contributes an auditable target-construction analysis for point-supervised cervical cell detection; it claims neither anatomical bounding boxes nor diagnostic use.*

Resumo. *Anotações pontuais identificam centros celulares, mas não definem a extensão do objeto. Este trabalho estuda as consequências experimentais de converter esses pontos em alvos de detecção na base CRIC Cervix, sobre dez tamanhos de pseudo-caixa (48–192 pixels) e três escalas YOLO11, com partições separadas por imagem. O F1 de validação forma um platô entre 96 e 192 pixels, com compromissos distintos de precisão, revocação e contagem; o alvo de 144 pixels foi mantido como ponto operacional intermediário e o YOLO11s deu o melhor equilíbrio. No teste retido, alcançou AP50=0,8723, F1=0,8190 e erro absoluto médio de 4,68 células por imagem. Um critério de distância ao ponto, independente da caixa sintética, produz F1 equivalente, confirmando localização real dos núcleos. A contribuição é uma análise auditável da construção do alvo para detecção cervical supervisionada por pontos, sem reivindicar caixas anatômicas ou uso diagnóstico.*

1. Introdução

O rastreamento do câncer do colo do útero depende de programas capazes de identificar alterações precursoras antes da progressão da doença. Em 2020, foram estimados aproximadamente 604 mil novos casos e 342 mil óbitos no mundo, com maior carga onde vacinação, rastreamento e tratamento são menos acessíveis [Singh et al. 2023]. Na citologia convencional, a automação tem sido estudada em diferentes escalas: classificação de recortes, segmentação de estruturas, detecção de instâncias e análise de campos ou lâminas [Jiang et al. 2023, Fang et al. 2024].

Essas tarefas pressupõem formas distintas de anotação. Um rótulo de classe informa o que há no recorte; um ponto informa onde uma instância está; uma caixa acrescenta extensão aproximada; uma máscara descreve fronteira. Logo, transformar um ponto em caixa não é apenas conversão de formato: é uma decisão sobre qual região o detector deve aprender. Em campos com muitas células, o tamanho adotado também controla a sobreposição entre alvos e altera a relação entre localização e contagem.

A CRIC Cervix disponibiliza 400 imagens convencionais de Papanicolau e 11.534 células classificadas, cada uma associada a uma coordenada nuclear [Rezende et al. 2021]. A base permite estudar detecção sem recortar previamente as células, mas não fornece bounding boxes anatômicas. Trabalhos anteriores usaram a CRIC em detecção e classificação [Kalbhor et al. 2023]; ainda assim, a sensibilidade do resultado à geometria criada a partir dos pontos permanece uma questão experimental não investigada. A contagem automática de células por imagem é uma tarefa complementar ao diagnóstico — na contagem celular por redes de regressão, o erro absoluto médio depende tanto do método quanto da densidade dos campos [Xie et al. 2018] — e, como se verá, o tamanho da pseudo-caixa afeta localização e contagem de formas distintas.

Este artigo formula o problema como *engenharia do alvo supervisionado*: não se propõe uma nova arquitetura nem se usa a categoria Bethesda como saída. Todas as coordenadas são tratadas como instâncias de uma classe única e o objetivo é determinar em que medida um detector treinado com pseudo-caixas recupera os pontos anotados e preserva a contagem por campo.

Três perguntas orientam o desenho experimental. **RQ1 (alvo):** quanto o tamanho da pseudo-caixa modifica F1, AP50 e erro de contagem? **RQ2 (capacidade):** aumentar a escala de YOLO11n para YOLO11s ou YOLO11m melhora o compromisso entre localização e contagem? **RQ3 (transferência):** a configuração selecionada na validação mantém desempenho no conjunto de teste retido, e como o erro varia com a densidade celular?

A principal contribuição é tornar mensurável uma decisão normalmente escondida na preparação dos dados: a geometria do alvo criada a partir de pontos. Todos os resultados são acompanhados de partições por imagem, limiar escolhido sem acesso ao teste e intervalos de confiança calculados por bootstrap sobre imagens completas.

2. Supervisão Espacial em Imagens Celulares

2.1. Do contorno completo ao ponto

Máscaras de instância permitem avaliar fronteiras e separar objetos adjacentes, mas exigem anotação detalhada. HoVer-Net, por exemplo, aprende segmentação e classificação nuclear em histologia a partir de contornos de instância [Graham et al. 2019]; Cellpose também aborda segmentação celular generalista com supervisão de máscaras [Stringer et al. 2021]. Esses métodos resolvem um problema espacial mais rico que o deste artigo, mas dependem de um tipo de ground truth ausente na CRIC.

Pontos reduzem custo de anotação e identificam instâncias sem descrever suas fronteiras. Ribera et al. estudaram localização de objetos sem bounding boxes [Ribera et al. 2019], e Cheng et al. investigaram segmentação de instâncias supervisionada por pontos [Cheng et al. 2022]. Em imagens biomédicas, Chen et al. pro-

puseram segmentação histopatológica fracamente supervisionada com pontos esparsos [Chen et al. 2021]; Xia et al. exploraram pontos e colorização para segmentação nuclear [Xia et al. 2023]. Em comum, esses estudos tratam o ponto como supervisão incompleta, não como caixa implícita. Uma abordagem alternativa é aprender a extensão diretamente por regressão, como faz o P2BNet [Chen et al. 2023]. O presente trabalho, ao contrário, torna *explícita* a geometria fixa e mede seu efeito — uma escolha que métodos como o P2BNet ocultam na otimização.

Neste artigo, pseudo-caixas são adotadas por pragmatismo: permitem usar um detector convencional mantendo saída por instância. O IoU mede concordância com um alvo sintético, não com a fronteira citológica; por isso, o tamanho do alvo é submetido a ablação e a contagem avaliada em paralelo.

2.2. Detecção em citologia cervical

Revisões da área descrevem resultados elevados em classificação celular, mas também heterogeneidade de bases, preparação, classes e protocolos de avaliação [Jiang et al. 2023, Fang et al. 2024]. Essa heterogeneidade impede interpretar métricas de tarefas diferentes como ranking direto. Classificar um recorte centrado, localizar uma célula no campo e segmentar núcleo/citoplasma não são problemas intercambiáveis.

O CerviCell-detector avaliou detectores de objetos em imagens de Papanicolau, incluindo experimentos na CRIC, com formulações binária e Bethesda [Kalbhor et al. 2023]. O presente estudo remove a decisão diagnóstica do alvo e pergunta se todas as instâncias pontuais podem ser localizadas e contadas. Essa escolha produz uma contribuição complementar: um detector de classe única pode anteceder um classificador, mas seu desempenho deve primeiro ser entendido como problema de cobertura espacial.

AP em diferentes limiares de IoU é adotada para medir rigor de localização, seguindo a tradição de benchmarks como COCO [Lin et al. 2014]. Entretanto, como o alvo é sintético, AP não é analisada isoladamente. F1 em um ponto operacional e erros de contagem completam a avaliação, enquanto o conjunto de teste permanece oculto durante a seleção.

3. Protocolo Experimental

3.1. Dados, unidade amostral e partições

Foram usadas as 400 imagens e as 11.534 coordenadas celulares da CRIC Cervix [Rezende et al. 2021]. As categorias Bethesda foram preservadas nos metadados para auditoria, mas não participaram do treinamento: a classe única *cell* agrega as seis categorias da base, de células normais (58,8%) a lesões de alto grau e carcinoma (HSIL 14,8%, LSIL 11,8%, ASC-H 8,0%, ASC-US 5,3% e SCC 1,4%). Assim, a tarefa é de localização agnóstica à classe sobre um conjunto clinicamente heterogêneo. Cada coordenada gerou uma instância dessa classe.

A imagem foi a unidade de partição, na proporção 80/10/10. Para evitar que a carga celular se concentrasse em uma partição, as imagens foram ordenadas pelo número de células e atribuídas de forma gulosa à partição mais distante de sua meta proporcional de objetos (semente 42). O procedimento destinou 320 imagens ao treino, 40 à validação

e 40 ao teste, com 9.236, 1.149 e 1.149 células, respectivamente; o balanceamento explica a igualdade exata de células entre validação e teste. Nenhuma imagem foi compartilhada entre partições. A CRIC não disponibiliza identificador de paciente; logo, a separação por imagem é o maior controle de independência suportado pelos metadados, não uma garantia de independência clínica.

3.2. Construção do alvo

Para um ponto nuclear (x_i, y_i) e lado s, definiu-se a pseudo-caixa

$$B_i(s) = \left[x_i - \frac{s}{2},\ y_i - \frac{s}{2},\ x_i + \frac{s}{2},\ y_i + \frac{s}{2}\right], \tag{1}$$

recortada aos limites da imagem. Centro, largura e altura foram normalizados no formato YOLO. A transformação fixa o centro fornecido pelo especialista, mas introduz uma hipótese sobre extensão. Essa hipótese foi testada para $s \in \{48, 64, 72, 96, 112, 128, 144, 160, 176, 192\}$ pixels.

A Figura 1 sobrepõe as pseudo-caixas geradas a uma imagem real da CRIC, permitindo inspecionar visualmente o alinhamento entre o alvo sintético e a citologia subjacente.

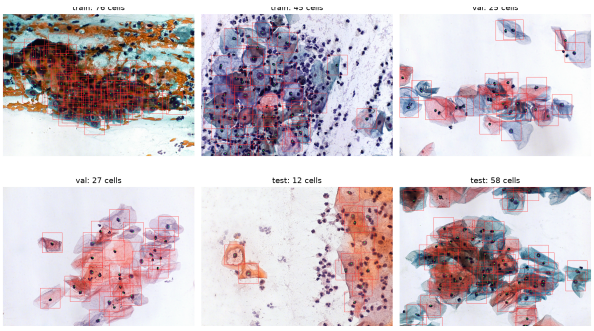


Figura 1. Auditoria visual de pseudo-caixas sobre imagem real da CRIC. Caixas são alvos computacionais, não contornos anatômicos.

3.3. Fases do experimento

A Figura 2 separa construção, seleção e teste. Na fase RQ1, YOLO11n [Khanam and Hussain 2024] foi treinado para cada valor de s, por até 35 épocas, para isolar a sensibilidade ao alvo. Na RQ2, fixou-se $s = 144$ e foram comparados YOLO11n, YOLO11s e YOLO11m. A RQ3 usou apenas o YOLO11s selecionado e o limiar definido na validação para abrir o teste uma única vez.

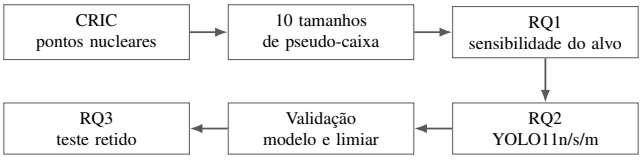


Figura 2. Fluxo experimental. O teste não participa da escolha de caixa, capacidade ou limiar.

Os modelos foram inicializados com pesos pré-treinados. Usou-se AdamW, taxa inicial 0,002, agendamento cossenoidal, *mosaic* moderado, variações geométricas/cromáticas e precisão mista. YOLO11n e YOLO11s receberam entrada nominal 832; YOLO11m, 864. Os lotes foram 12, 6 e 4, respectivamente. As comparações de capacidade tiveram 80 épocas máximas e paciência 20. O treinamento foi executado em uma RTX 3060 Laptop com 6 GB de VRAM.

3.4. Critérios de avaliação

As predições foram exportadas pela rotina de validação do Ultralytics, sem aumento no teste, e todas as métricas foram recomputadas a partir delas. A associação predição–alvo priorizou maior IoU e impôs unicidade de ambos os lados. A AP em IoU 0,30, 0,50 e 0,75 foi obtida por interpolação de 101 pontos, no estilo COCO [Lin et al. 2014]; precisão, revocação e F1 completam o conjunto. Como a pseudo-caixa é sintética, adotou-se ainda um critério independente da sua geometria: uma detecção é correta se o centro da predição estiver a até R pixels do ponto nuclear anotado, com associação gulosa única por menor distância. Esse critério avalia o detector contra a anotação original, não contra a caixa derivada dela. O limiar de confiança foi escolhido, em cada configuração, pelo maior F1 na validação em IoU 0,50. Na fase RQ1, o valor ótimo do YOLO11n variou entre 0,20 e 0,50 conforme o tamanho da pseudo-caixa, refletindo a calibração individual de cada treino. Na fase RQ2/RQ3, os três modelos comparados (YOLO11n, YOLO11s e YOLO11m) atingiram o pico em 0,35; o valor do YOLO11s foi usado na abertura do teste.

Para cada imagem j, com contagem anotada g_j e predita p_j, foram calculados

$$\mathrm{MAE} = \frac{1}{N}\sum_j |p_j - g_j|, \qquad \mathrm{bias} = \frac{1}{N}\sum_j (p_j - g_j). \tag{2}$$

RMSE e correlação complementaram a análise. Intervalos de 95% foram obtidos com 1.000 réplicas de bootstrap sobre imagens completas. Essa unidade de reamostragem preserva a dependência entre células do mesmo campo.

3.5. Rastreabilidade e controle de decisões

Cada etapa produziu artefatos persistentes: lista de imagens por partição, metadados das pseudo-caixas, configuração de treino, histórico por época, pesos selecionados, predições por imagem e tabelas de métricas. Figuras e valores reportados derivam diretamente desses arquivos, sem transcrição manual. A origem dos dados, o particionamento, o hardware, a seleção do limiar e as limitações de generalização foram explicitados em consonância com recomendações de relato para IA em imagens médicas [Mongan et al. 2020, Tejani et al. 2024, Collins et al. 2024].

As decisões foram encadeadas de modo unilateral: RQ1 definiu a região de pseudo-caixas; RQ2 comparou capacidades com o mesmo alvo; a validação selecionou confiança; e somente então o teste foi avaliado. Nenhuma métrica de teste foi usada para alterar pesos, lado da caixa ou limiar. Os resultados do sweep, por sua vez, são tratados como análise exploratória de sensibilidade, pois dez configurações e uma execução por configuração não permitem atribuir toda diferença observada exclusivamente ao tamanho do alvo.

4. Resultados

4.1. RQ1: o alvo altera o resultado?

A Figura 3 mostra que a resposta é não monotônica. F1 e AP50 aumentaram das menores caixas até uma região de estabilidade, mas o tamanho 160 apresentou queda acentuada e supercontagem. O maior F1 foi 0,8253 em 192 pixels. O menor MAE (3,65) e o maior AP50 (0,8776) ocorreram em 176 pixels. Em 144 pixels, F1 foi 0,8228, AP50 0,8748 e MAE 4,40.

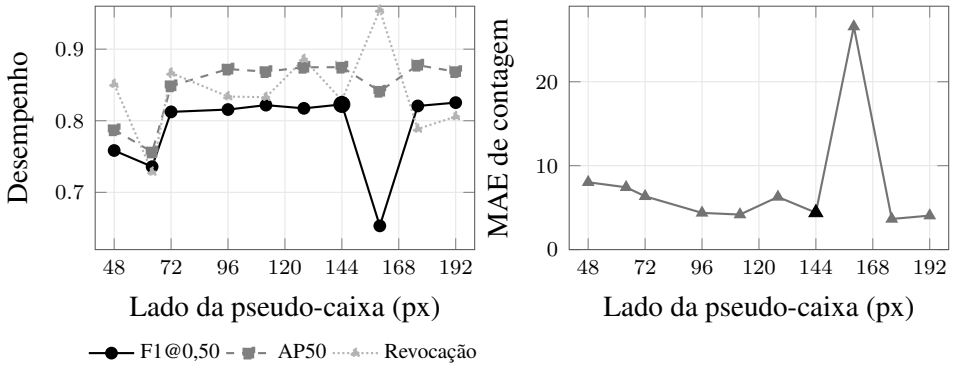


Figura 3. Sensibilidade ao tamanho do alvo na validação. O marcador maior indica a configuração de 144 pixels usada na comparação de capacidade.

Os valores entre 96 e 192 pixels, exceto 160, formam um platô estreito de F1 (0,8157–0,8253). Assim, os dados não sustentam um único tamanho universal. O ponto de 144 pixels foi mantido por combinar AP50 próxima do máximo (0,8748 vs. 0,8776 em 176), revocação superior a 176 e 192 (0,8285 vs. 0,7885 e 0,8059) e posição intermediária no platô. A maior revocação importa quando as células perdidas pelo detector representam potenciais falsos negativos diagnósticos; pagar um custo de MAE levemente maior (4,40 vs. 3,65 em 176) para manter cobertura foi o critério adotado. A instabilidade em 160 pixels impede interpretar a tendência como simples benefício de caixas maiores. Naquele ponto, o detector saturou o teto de detecções (300 por imagem em todas as 40 imagens de validação), com revocação 0,955 e precisão 0,496. Como caixas *maiores* (176 e 192 pixels) não reproduzem esse colapso, o efeito reflete a variância de um treino isolado, não uma consequência geométrica do tamanho do alvo. De fato, a mAP50 nativa desse modelo permaneceu saudável (0,883), indicando que o ranking das detecções foi preservado e que o colapso se restringiu à calibração do ponto operacional.

4.2. RQ2: maior capacidade melhora a tarefa?

YOLO11s apresentou o maior F1 (0,8363), AP50 (0,8848) e o menor MAE (4,68) na validação. YOLO11n ficou próximo, com F1 0,8231 e MAE 4,83. YOLO11m atingiu a maior revocação (0,8921), mas sua precisão de 0,7158 elevou o MAE para 8,13 e reduziu F1 para 0,7943. A Figura 4 mostra que escala de parâmetros e desempenho operacional não cresceram juntos.

O ganho de YOLO11s sobre YOLO11n foi de 0,0132 em F1 e 0,0075 em AP50. Essa diferença é pequena e não possui intervalo de confiança entre treinos; portanto, suporta a escolha operacional do S neste experimento, mas não uma afirmação geral de superioridade arquitetural. O M foi descartado porque o aumento de revocação não compensou falsos positivos e erro de contagem.

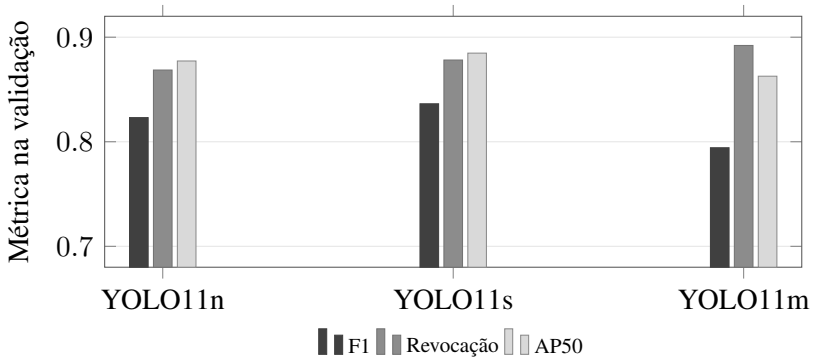


Figura 4. Comparação das escalas YOLO11 com pseudo-caixas de 144 pixels.

4.3. Seleção do ponto operacional

A confiança controla uma troca relevante para contagem. Na validação, limiares baixos preservaram quase todas as células, mas geraram muitas detecções excedentes: em 0,10, a revocação foi 0,9661 e a precisão 0,5187, com MAE 24,78. Em 0,50, a precisão subiu para 0,8853, mas a revocação caiu para 0,6919 e o viés tornou-se negativo ($-6,28$ células por imagem).

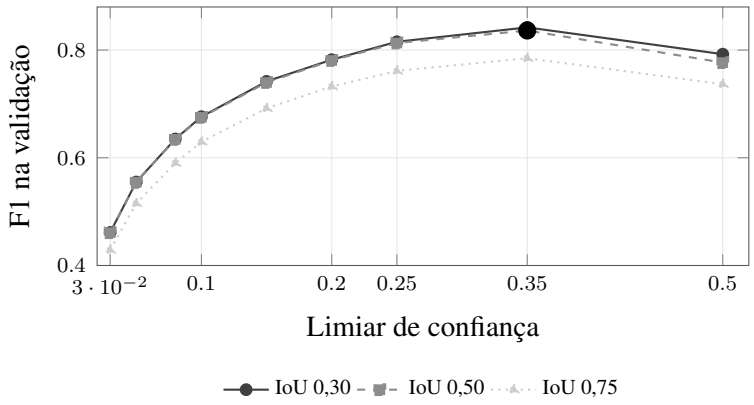


Figura 5. Curvas de F1 usadas para selecionar a confiança. O marcador maior identifica 0,35, escolhido exclusivamente na validação em IoU 0,50.

O limiar 0,35 maximizou F1@0,50 (0,8363), com precisão 0,7983, revocação 0,8782, MAE 4,68 e viés 2,88. A proximidade das curvas em IoU 0,30 e 0,50 indica que, nesse intervalo, a maioria das associações aceitas em 0,30 também satisfaz 0,50. A separação em IoU 0,75 evidencia a penalização adicional por desalinhamento fino com a pseudo-caixa. O teste recebeu 0,35 sem novo ajuste; portanto, seu resultado mede transferência do ponto operacional, não o melhor limiar retrospectivo para aquela partição.

4.4. RQ3: o resultado se mantém no teste?

Com YOLO11s, lado 144 e confiança 0,35, o teste produziu 975 verdadeiros positivos, 257 falsos positivos e 174 falsos negativos. A Tabela 1 reúne todas as medidas, cada uma com intervalo de 95% obtido por bootstrap de imagens.

O F1 de teste ficou 0,0173 abaixo da validação, enquanto AP50 caiu 0,0125. Essas diferenças são compatíveis com transferência moderada entre duas amostras de 40

Tabela 1. Resultado no teste retido. IC95% por bootstrap de imagens.

Métrica	Valor	IC95%	Métrica	Valor	IC95%
Precisão	0,7914	[0,746; 0,833]	AP30	0,8843	[0,858; 0,913]
Revocação	0,8486	[0,822; 0,877]	AP50	0,8723	[0,843; 0,904]
F1	0,8190	[0,796; 0,842]	AP75	0,7515	[0,695; 0,811]
MAE	4,675	[3,150; 6,425]	RMSE	7,216	[4,560; 9,850]
Viés	2,075	[0,050; 4,126]	Correlação	0,9434	[0,874; 0,979]

imagens, sem evidência de colapso. O viés positivo indica 2,08 detecções excedentes por imagem em média.

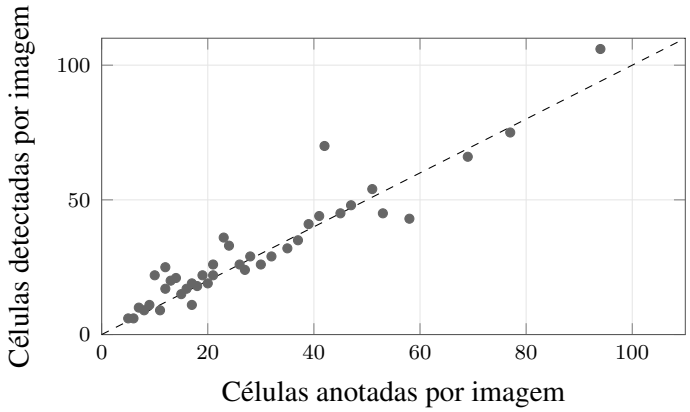


Figura 6. Concordância de contagem no teste; a linha tracejada representa $p_j = g_j$.

Na análise exploratória por tercis, o MAE foi 4,15 em imagens com 5–16 células, 3,69 entre 17–30 e 6,07 entre 32–94. A faixa mais densa concentrou maior erro absoluto, embora o F1 global permaneça 0,8190. A mediana global do erro foi 3 células. Esses estratos foram definidos depois da avaliação principal e não devem ser interpretados como hipóteses confirmatórias.

4.5. Avaliação independente da pseudo-caixa

Para verificar se o desempenho reflete localização real e não apenas aderência à caixa sintética, o teste foi reavaliado pelo critério de distância ao ponto. Como núcleos vizinhos distam em mediana 72 pixels, um raio R bem abaixo desse valor associa cada predição ao núcleo correto, não a um adjacente. A Figura 7 mostra o F1 crescendo e estabilizando: 0,812 em $R = 32$ pixels (metade do espaçamento típico) e 0,821 em 48.

Esse intervalo (0,812–0,821) coincide com o F1 por IoU 0,50 (0,819): duas medidas de natureza distinta concordam, indicando que a qualidade reportada é localização efetiva dos núcleos, não um artefato da convenção de alvo.

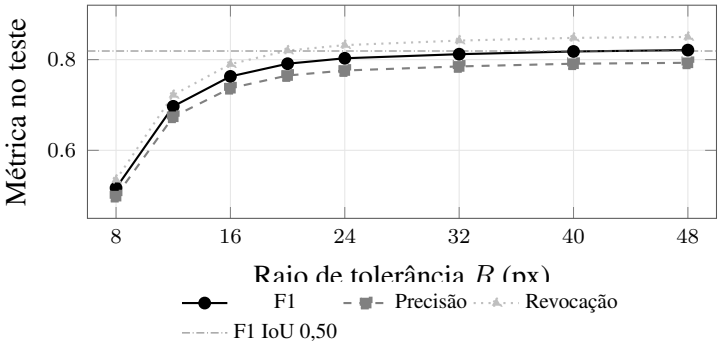


Figura 7. Avaliação por distância ao ponto anotado, independente da geometria da pseudo-caixa. A linha horizontal é o F1 obtido por IoU 0,50.

5. Discussão

5.1. O que a pseudo-caixa representa

O experimento não recupera fronteiras celulares; ele aprende uma vizinhança quadrada ao redor do núcleo. Essa distinção muda o significado das métricas. AP75 alta indica concordância com a regra da Equação 1, não segmentação precisa de núcleo ou citoplasma. Métodos com máscaras, como HoVer-Net e Cellpose, respondem a uma pergunta espacial mais detalhada [Graham et al. 2019, Stringer et al. 2021]. A objecção de que o IoU contra um alvo sintético seria igualmente artificial é respondida pela avaliação por distância: como o F1 por ponto converge ao F1 por IoU, a métrica não depende criticamente da geometria adotada, e o núcleo anotado — único ground truth da CRIC — é de fato recuperado.

Ainda assim, pontos contêm informação de instância que rótulos globais não possuem. A literatura de supervisão pontual mostra que essa informação pode orientar localização e segmentação com menor custo de anotação [Ribera et al. 2019, Chen et al. 2021, Xia et al. 2023]. O resultado da RQ1 acrescenta uma observação prática: quando pontos são convertidos em caixas, a geometria escolhida deve ser reportada e testada. Caso contrário, uma parte relevante do desempenho permanece escondida na preparação do dado.

O platô de 96–192 pixels é mais informativo que a escolha isolada de 144. Ele mostra que diversas extensões produzem F1 parecido, mas não o mesmo compromisso de precisão, revocação e contagem. A queda em 160 pixels alerta contra uma leitura determinista da curva: cada ponto corresponde a um treino e inclui variabilidade de otimização. Repetições por semente seriam necessárias para separar efeito geométrico de variação estocástica.

5.2. Capacidade e custo computacional

YOLO11s foi o melhor compromisso observado, mas sua vantagem sobre YOLO11n foi pequena. Sem repetições independentes, não há base estatística para afirmar que o S é intrinsecamente superior. A conclusão suportada é mais restrita: sob os pesos, partição e orçamento executados, S gerou o melhor ponto operacional.

O M ilustra por que mAP de treinamento não deve ser o único critério. Sua revocação foi a maior, mas o aumento de falsos positivos deteriorou contagem e F1. Esse

comportamento importa em campos densos: detectar mais candidatos pode ser desejável em uma etapa de busca, mas transfere custo para supressão de duplicidades ou revisão posterior. O artigo, portanto, não trata escala do modelo como proxy automática de qualidade.

5.3. Escopo da comparação com a literatura

Os valores de AP deste artigo não constituem um ranking direto contra HoVer-Net ou Cellpose: esses métodos estimam máscaras, foram avaliados em outros conjuntos e respondem a uma definição de alvo distinta [Graham et al. 2019, Stringer et al. 2021]. Tampouco se deve comparar apenas o número final com trabalhos de classificação Bethesda. O CerviCell-detector, por exemplo, investiga detecção de células cancerosas, enquanto aqui toda coordenada celular pertence a uma classe binária [Kalbhor et al. 2023]. Métrica, unidade de análise, partição e semântica do rótulo precisam coincidir para sustentar superioridade numérica; as revisões da área mostram justamente grande heterogeneidade nesses elementos [Jiang et al. 2023, Fang et al. 2024].

A contribuição comparável á literatura de supervisão fraca é outra: quantificar o efeito de transformar um ponto em uma região de treinamento e mostrar como essa escolha repercute simultaneamente em localização e contagem. Ribera et al. aprendem localização sem caixas, enquanto Chen et al. e Xia et al. exploram pontos para segmentação [Ribera et al. 2019, Chen et al. 2021, Xia et al. 2023]. Este estudo não substitui essas formulações; fornece um protocolo simples e auditável para o cenário em que um detector de caixas é treinado a partir de coordenadas. Assim, a novidade reivindicada é metodológica e experimental, não uma alegada superioridade universal do YOLO11s.

5.4. Contagem e fontes de erro

A correlação 0,9434 confirma que o modelo acompanha a variação de carga celular entre imagens; o MAE de 4,68 e o viés de 2,08 revelam, porém, erro residual e tendência de supercontagem. Para referência de ordem de grandeza, Xie et al. aplicaram redes de regressão totalmente convolucionais à contagem celular em microscopia e obtiveram MAE competitivo em diferentes densidades [Xie et al. 2018]. O valor obtido aqui (MAE 4,68) situa-se na mesma casa de magnitude, embora a comparação seja apenas indicativa: os conjuntos de dados, a supervisão e a formulação diferem. Em termos absolutos, a faixa mais densa concentrou o maior erro (MAE 6,31): pequenas taxas de duplicidade ou omissão acumulam-se quando o campo contém dezenas de células.

Há pelo menos três explicações, não separadas por este experimento: duplicidade em grupos sobrepostos, quando o IoU entre pseudo-caixas adjacentes fica abaixo do limiar de supressão e dois hits sobrevivem para o mesmo núcleo; estruturas não celulares confundidas com o alvo; e células visíveis sem ponto correspondente na anotação. Distinguir essas causas exige revisão cega de falsos positivos por especialista. O colapso isolado em 160 pixels é de natureza distinta: ali o detector saturou o teto de detecções por imagem, um efeito de instabilidade de treino que não se repete nas demais escalas e não deve ser confundido com a supercontagem residual do modelo final.

5.5. Limitações e perspectivas

Quatro limitações delimitam a interpretação. *Construção:* métricas de IoU descrevem consistência com um alvo sintético, não com a fronteira citológica. *Interna:* cada configuração foi treinada uma vez; repetições por semente seriam necessárias para separar

efeito de s de ruído de otimização. *Externa:* avaliação em base única, separação por imagem (não por paciente), sem teste em outro laboratório ou equipamento. *Conclusão:* os intervalos de bootstrap quantificam variação entre imagens, mas não incerteza de treinamento; diferenças no platô de F1 não devem ser generalizadas sem repetições.

A extensão imediata é comparar pseudo-caixas a formulacões que aprendem extensão a partir dos pontos [Chen et al. 2023], repetir o sweep com múltiplas sementes e anotar uma amostra com contornos manuais para calibrar a avaliação atual.

6. Conclusão

Este estudo tratou a conversão de pontos nucleares em caixas como variável científica. Para a RQ1, observou-se um platô de F1 entre 96 e 192 pixels, com compromissos distintos de AP50, revocação e contagem. Para a RQ2, YOLO11s foi o melhor ponto operacional, embora a vantagem sobre YOLO11n tenha sido pequena e YOLO11m tenha apresentado maior revocação. Para a RQ3, o modelo selecionado obteve AP50 0,8723, F1 0,8190 e MAE 4,68 no teste retido; o erro absoluto aumentou em imagens mais densas.

A conclusão principal não é que 144 pixels ou YOLO11s sejam universais. O resultado relevante é que a definição espacial criada a partir de pontos altera de forma mensurável detecção e contagem e, portanto, precisa fazer parte do protocolo reportado. O pipeline produzido oferece uma referência reprodutível para esse tipo de auditoria na CRIC. Sua extensão natural é substituir a geometria fixa por supervisão diretamente pontual e comparar com uma amostra de contornos anotados.

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq) – Processo nº 308950/2026-7.

Referências

Chen, P., Cai, X., Chen, Y., Chen, G., Zheng, F., Keutzer, K., Lu, T., and Li, X. (2023). P2BNet: Point-to-box network for object detection with point annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12244–12259.

Chen, Z., Chen, Z., Liu, J., Zheng, Q., Zhu, Y., Zuo, Y., Wang, Z., Guan, X., Wang, Y., and Li, Y. (2021). Weakly supervised histopathology image segmentation with sparse point annotations. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1673–1685.

Cheng, B., Parkhi, O., and Kirillov, A. (2022). Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2617–2626.

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378.

Fang, M., Liao, B., Lei, X., and Wu, F.-X. (2024). A systematic review on deep learning based methods for cervical cell image analysis. *Neurocomputing*, 610:128630.

Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., and Rajpoot, N. (2019). HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563.

Jiang, P., Li, X., Shen, H., Chen, Y., Wang, L., Chen, H., Feng, J., and Liu, J. (2023). A systematic review of deep learning-based cervical cytology screening: From cell identification to whole slide image analysis. *Artificial Intelligence Review*, 56(Suppl 2):2687–2758.

Kalbhor, M., Shinde, S., Wajire, P., and Jude, H. (2023). CerviCell-detector: An object detection approach for identifying the cancerous cells in pap smear images of cervical cancer. *Heliyon*, 9(11):e22324.

Khanam, R. and Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Mongan, J., Moy, L., and Kahn, C. E. (2020). Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2):e200029.

Rezende, M. T., Silva, R., Bernardo, F. d. O., Tobias, A. H. G., Oliveira, P. H. C., Machado, T. M., Costa, C. S., Medeiros, F. N. S., Ushizima, D. M., Carneiro, C. M., and Bianchi, A. G. C. (2021). CRIC searchable image database as a public platform for conventional pap smear cytology data. *Scientific Data*, 8(1):151.

Ribera, J., Guera, D., Chen, Y., and Delp, E. J. (2019). Locating objects without bounding boxes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6489.

Singh, D., Vignat, J., Lorenzoni, V., Eslahi, M., Ginsburg, O., Lauby-Secretan, B., Arbyn, M., Basu, P., Bray, F., and Vaccarella, S. (2023). Global estimates of incidence and mortality of cervical cancer in 2020: A baseline analysis of the WHO global cervical cancer elimination initiative. *The Lancet Global Health*, 11(2):e197–e206.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106.

Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S., Kahn, C. E., and Panel, C. . U. (2024). Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence*, 6(4):e240300.

Xia, L., Qu, Z., An, J., and Gao, Z. (2023). A weakly supervised method with colorization for nuclei segmentation using point annotations. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11.

Xie, W., Noble, J. A., and Zisserman, A. (2018). Microscopy cell counting and detection with fully convolutional regression networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):283–292.