PRÉ-ANOTAÇÃO ASSISTIDA POR MODELOS FUNDACIONAIS PARA MONITORAMENTO ANIMAL EM BORDA

Fabiano Frank Novais Magalhães1 - fabianofrank@protonmail.com
[1]Programa de Pós-Graduação em Modelagem Computacional e Sistemas, Universidade Estadual de Montes Claros, UNIMONTES - Montes Claros, MG, Brazil

Abstract. *O monitoramento animal em borda exige modelos leves, mas a construção de bases rotuladas em vídeos reais ainda depende de anotação manual custosa, especialmente quando há oclusão e sobreposição entre indivíduos. Este trabalho apresenta uma avaliação exploratória de pré-anotação assistida por modelos fundacionais, tratando modelos pesados como professores offline para curadoria de máscaras, identidades e candidatos de rótulo, e não como modelos finais de implantacao em edge. Foram comparadas oito configurações de detecção, segmentação, pose e rastreamento em uma amostra curta de vídeo com ovelhas em vista superior, entre os frames f120 e f360. O melhor resultado preliminar foi obtido com SAM2.1 para rastreamento de máscaras em vídeo, inicializado por caixas automáticas geradas a partir de localização com GroundingDINO, mantendo dois objetos presentes em 100% dos 241 frames avaliados, com IoU temporal médio de 0,9594. Apesar disso, as máscaras ainda apresentam vazamentos em regiões de contato e não classificam postura. Os resultados indicam um caminho promissor para gerar candidatos auditáveis de dataset antes do treinamento de um aluno leve para borda.*

Keywords: *Visão computacional, Segmentação de vídeo, Rastreamento temporal, Edge AI, Monitoramento animal*

1. INTRODUÇÃO

Aplicações de visão computacional para monitoramento animal podem apoiar tarefas como acompanhamento de comportamento, detecção de eventos e gestão de bem-estar. Em cenários de borda, entretanto, há uma tensão prática: modelos leves são necessários para executar em hardware restrito, mas esses modelos dependem de dados rotulados com qualidade suficiente. Quando os vídeos contêm vista superior, fundo repetitivo, grades, mudança de postura e sobreposição entre animais, a anotação manual de caixas, máscaras ou keypoints se torna um gargalo relevante.

Este trabalho aborda esse gargalo como um problema de curadoria de dataset. A hipótese avaliada é que modelos fundacionais de visão podem ser usados como professores offline para

gerar pré-anotações auditáveis, reduzindo o esforço humano antes de treinar um aluno leve para execução em edge. Assim, modelos como GroundingDINO (Liu et al., 2023), YOLO-World (Cheng et al., 2024), MobileSAM (Zhang et al., 2023), DeepLabCut/SuperAnimal (Mathis et al., 2018; Ye et al., 2024) e SAM2 (Ravi et al., 2024) não são tratados como solução embarcada final, mas como componentes de uma etapa de preparação de dados.

A contribuição deste trabalho é um protocolo preliminar e auditável para comparar ferramentas de detecção, segmentação, pose e rastreamento temporal em um trecho curto de vídeo realista com oclusão. O foco não é demonstrar um classificador final de postura nem benchmark embarcado, mas identificar qual família de ferramenta produz melhores candidatos de máscara e identidade para revisão e posterior treinamento supervisionado.

2. MÉTODO EXPLORATORIO

O experimento utilizou o vídeo *ovelha1.mp4*, com recorte entre os frames globais f120 e f360. A amostra foi escolhida por conter dois animais em contato, vista superior, estruturas de grade e mudança de posição ao longo do tempo. Esses elementos representam um caso difícil para anotação automática porque caixas e máscaras isoladas tendem a misturar indivíduos no ponto de sobreposição.

Foram avaliadas oito configurações, agrupadas em quatro frentes: segmentação local, detecção/segmentação open-vocabulary no Colab, pose/keypoints e rastreamento temporal. Cada teste foi registrado com identificador de execução, ferramenta, amostra, métricas, falhas observadas e decisão técnica, seguindo o plano metodológico de testes sequenciais (fases box/mask, pose/keypoint e tracking) definido para o experimento. As decisoes foram classificadas como *keep*, *fallback* ou *reject*. A métrica *quality score* foi usada como nota operacional de 0 a 100, derivada de observação técnica e métricas disponíveis da fase. Para rastreamento, foram priorizadas consistência temporal, permanência dos dois objetos e estabilidade visual de identidade.

No teste de melhor desempenho, a localização inicial dos dois animais foi obtida por detecção com GroundingDINO (frame f120), e as caixas resultantes foram usadas para inicializar o rastreamento de máscaras em vídeo com SAM2.1, sem cliques manuais. O protocolo seguiu uma regra conservadora: pré-anotações automáticas permanecem como candidatos dentro do experimento e não são promovidas diretamente a dataset final sem revisão. Esse ponto é essencial porque uma máscara visualmente plausível pode ainda vazar para o animal vizinho, contaminar o rótulo e induzir erro no treinamento do aluno leve.

Table 1: Resumo dos testes exploratorios em vídeo com sobreposição

Fase	Ferramenta	Score	Decisão	Evidência principal
Box/mask	YOLO-seg local	35	fallback	Roda localmente, mas mistura animais em overlap.
Box/mask	YOLO-World + MobileSAM	55	fallback	Melhora candidatos, mas duplica regiões e captura grade/laterais.
Box/mask	GroundingDINO + SAM2.1 + ROI/NMS	70	fallback	Foca no miolo e reduz falsos positivos, mas ainda vaza em contato.
Pose	RTMLib Animal Pose	20	reject	Keypoints cruzam indivíduos e não sustentam cabeça/corpo/quadril confiáveis.
Pose	DeepLabCut/SuperAnimal	25	reject	Pose genérica pre-treinada não entrega keypoints estáveis no caso top-view/overlap.
Tracking	LK optical flow + centroide	30	reject	Barato, mas as caixas não acompanham os animais até f360.
Tracking	GroundingDINO periódico + reassociação	55	fallback	Baseline útil, mas manteve ambos os tracks em apenas 55,56% dos keyframes.
Tracking	GroundingDINO (localização) + SAM2.1 vídeo tracking	82	keep	Manteve dois objetos em 100% dos 241 frames, com IoU temporal médio 0,9594.

3. RESULTADOS PRELIMINARES

A Tabela 1 mostra que os métodos baseados apenas em detecção ou pose não resolveram o gargalo central de identidade em sobreposição. O YOLO-seg local foi útil como baseline por executar localmente e gerar visualização rápida, mas misturou os animais centrais. YOLO-World com MobileSAM gerou candidatos mais ricos, porém com duplicação e falsos positivos. GroundingDINO combinado com SAM2.1 e restrições por região de interesse foi um bom gerador frame-a-frame de candidatos, mas ainda não produziu máscaras suficientemente limpas para uso direto como rótulo final.

Os testes de pose também foram insuficientes nesta amostra. RTMLib Animal Pose e DeepLabCut/SuperAnimal executaram, mas os keypoints não permaneceram confiáveis por indivíduo quando os corpos estavam em contato. Isso sugere que, para este tipo de vídeo, insistir inicialmente em keypoints genéricos pode ser menos produtivo do que estabilizar primeiro identidade, máscara e centroide.

O melhor resultado combinou GroundingDINO para localização inicial e SAM2.1 para segmentação e rastreamento em vídeo. O teste foi inicializado no frame f120 com caixas automáticas geradas por detecção com GroundingDINO, sem cliques manuais. O modelo propagou duas máscaras entre f120 e f360, totalizando 241 frames. Foram observados dois objetos presentes em 100% dos frames, IoU temporal médio de 0,9594 em relação ao frame anterior, tempo médio de 0,2283 s por frame e 4,38 FPS no ambiente Colab. A Figura 1 ilustra o frame original, a localização por GroundingDINO e o resultado do rastreamento por SAM2.1.

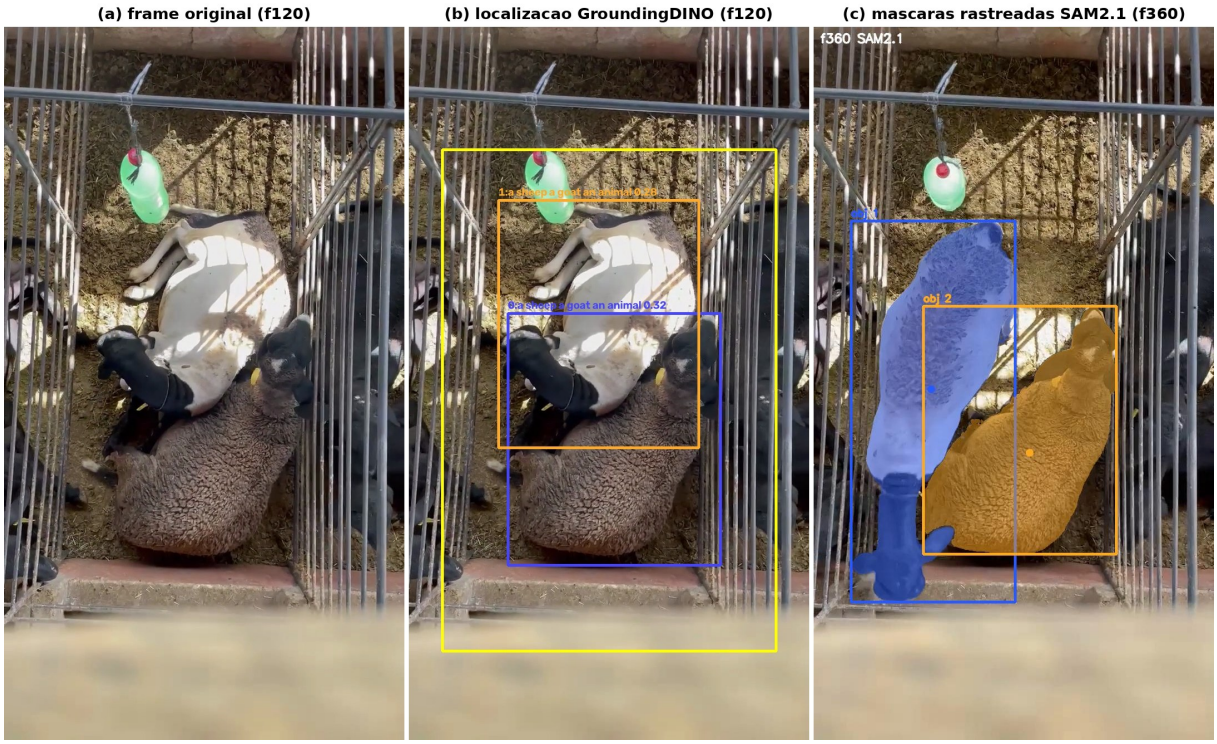

Figure 1: Painel comparativo do pipeline de pré-anotação: (a) frame original (f120); (b) localização inicial dos dois animais por GroundingDINO, com caixa de região de interesse (f120); (c) máscaras rastreadas por SAM2.1 mantendo identidade dos dois objetos (f360). As máscaras preservam identidade e centroides coerentes, mas ainda há vazamento em regiões de contato.

4. DISCUSSÃO E LIMITAÇÕES

O resultado principal não deve ser interpretado como solução final de monitoramento em borda. SAM2.1, GroundingDINO e ferramentas semelhantes possuem custo computacional superior ao esperado para execução direta em hardware restrito. O papel deles, neste protocolo, é operar offline como professores de curadoria: geram candidatos de máscara, identidade e regiões de interesse para revisão, auditoria e posterior treinamento de um modelo aluno menor.

A principal limitação é o tamanho da amostra. A avaliação foi feita em um trecho curto de um vídeo, sem conjunto amplo de rótulos manuais de referência. Portanto, o IoU temporal reportado mede estabilidade entre frames consecutivos, não acurácia contra uma verdade-terreno completa. Além disso, a classificação de postura ainda não foi implementada; o protocolo atual separa a etapa de localização/rastreamento da etapa posterior de rotulagem semântica. Tambem não houve treinamento do aluno leve, exportação para TensorFlow Lite, quantização INT8 ou benchmark em Coral/Edge TPU.

Ainda assim, a comparação é útil porque elimina caminhos pouco promissores para o gargalo atual. Pose pronta genérica e tracking simples por centroide falharam no caso de overlap. A redetecção periódica melhorou o baseline, mas não sustentou identidade suficiente. A combinação de GroundingDINO para localização com SAM2.1 para rastreamento de vídeo mostrou o melhor equilíbrio entre qualidade de candidato e auditabilidade, desde que acompanhada por filtros de falha, como salto de área, salto de centroide, aumento de sobreposição entre máscaras e queda de consistência temporal.

5. CONCLUSÕES

Este trabalho apresentou um protocolo exploratório para pré-anotação assistida por modelos fundacionais em vídeos de monitoramento animal com oclusão. A evidência preliminar indica que a combinação de GroundingDINO para localização e SAM2.1 para rastreamento de vídeo, inicializada por caixas automáticas, superou baselines de segmentação frame-a-frame, pose genérica e tracking simples na manutenção de identidade e máscaras entre f120 e f360. O melhor teste manteve dois objetos presentes em todos os 241 frames avaliados e obteve IoU temporal médio de 0,9594.

A conclusão operacional é que o caminho mais defensavel não é implantar modelos fundacionais pesados na borda, mas usá-los como professores offline para construir um dataset auditável. O próximo passo técnico é adicionar uma camada de detecção de falhas e revisão: marcar frames com vazamento, troca de identidade, salto de centroide ou queda de IoU temporal, e acionar GroundingDINO, VLM ou revisão humana apenas nesses casos. Somente depois dessa curadoria o trabalho deve avançar para treinamento de um aluno leve e benchmark em hardware edge.

REFERENCES

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X. and Shan, Y. (2024), "YOLO-World: Real-Time Open-Vocabulary Object Detection", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J. and Zhang, L. (2023), "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection", *arXiv preprint arXiv:2303.05499*.

Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W. and Bethge, M. (2018), "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning", *Nature Neuroscience*, 21, 1281–1289.

Ravi, N.; Gabeur, V.; Hu, Y.T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Radle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K.V.; Carion, N.; Wu, C.Y.; Girshick, R.; Dollar, P. and Feichtenhofer, C. (2024), "SAM 2: Segment Anything in Images and Videos", *arXiv preprint arXiv:2408.00714*.

Ye, S.; Filippova, A.; Schneider, S.; Gemuend, A.; Osorio, D.; Lee, M.; Mathis, M.W. and Mathis, A. (2024), "SuperAnimal pretrained pose estimation models for behavioral analysis", *Nature Communications*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S. and Hong, C.S. (2023), "Faster Segment Anything: Towards Lightweight SAM for Mobile Applications", *arXiv preprint arXiv:2306.14289*.