UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
ENGENHARIA DE CONTROLE E AUTOMAÇÃO

PEDRO SCHNARNDORF - 00301979

ESTUDO DE VIABILIDADE DE ESTIMAÇÃO DE POSE DE RATOS EM AMBIENTE LABORATORIAL UTILIZANDO VISÃO COMPUTACIONAL

Porto Alegre
2026

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
ENGENHARIA DE CONTROLE E AUTOMAÇÃO

PEDRO SCHNARNDORF - 00301979

ESTUDO DE VIABILIDADE DE ESTIMAÇÃO DE POSE DE RATOS EM AMBIENTE LABORATORIAL UTILIZANDO VISÃO COMPUTACIONAL

Proposta de Trabalho de Conclusão de Curso submetida à COMGRAD/CCA da UFRGS como parte dos requisitos para a obtenção do título de *Bacharel em Engenharia de Controle e Automação.*

Orientadores:
Prof. Dr. Heraldo José de Amorim

Prof. Dr. Carlos Solon Soares Guimarães Jr.

Porto Alegre
2026

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
ENGENHARIA DE CONTROLE E AUTOMAÇÃO

PEDRO SCHNARNDORF - 00301979

ESTUDO DE VIABILIDADE DE ESTIMAÇÃO DE POSE DE RATOS EM AMBIENTE LABORATORIAL UTILIZANDO VISÃO COMPUTACIONAL

Esta Proposta de Trabalho de Conclusão de Curso foi julgada adequada para a obtenção dos créditos da Atividade de *Trabalho de Conclusão de Curso CCA - I* e aprovada em sua forma final pelos Orientadores e Banca Examinadora abaixo.

Orientadores:
Prof. Dr. Heraldo José de Amorim, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, Brasil

Prof. Dr. Carlos Solon Soares Guimarães Jr., UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, Brasil

Banca Examinadora:
Prof. Dr. Jakson Manfredini Vassoler, UFRGS
Doutor pela Politecnico di Torino – Torino, Itália

Prof. Dr. Marcelo Götz, UFRGS
Doutor pela Universität Paderborn - Paderborn, Alemanha

Fabiano Disconzi Wildner
Coordenador de Curso
Eng. de Controle e Automação

Porto Alegre
Junho - 2026

AGRADECIMENTOS

Agradeço à Universidade Federal do Rio Grande do Sul, UFRGS, pela oportunidade de realização de estudos e por dar suporte ao desenvolvimento da pesquisa nacional e ao Centro de Reprodução e Experimentação de Animais de Laboratório, CREAL, e à professora Fernanda Bastos de Mello por possibilitar a realização desta pesquisa.

À minha família e ao Schnaps, nosso cão, pelo amor incondicional e pelo suporte dado ao longo da minha vida, ainda mais ao longo da produção deste trabalho.

Ao meu amor, por ter me apresentado os ratos e dado suporte diretamente à esta pesquisa e por estar sempre ao meu lado.

Às amizades que venho cultivando desde a época da escola e que ainda caminham ao meu lado. Dividimos muitas emoções juntos, nossos momentos guardo para sempre na minha memória.

Aos colegas de curso, em especial ao Grupo de Apoio. Quando criamos o grupo não sabiamos que o nome ia ser tão representativo da nossa relação. Por muitas dificuldades passamos neste curso e com certeza afirmo que nossa amizade foi a ferramenta mais valiosa para superá-las. O curso agora chega ao fim, mas nossa amizade permanece.

À dois professores, em especial, que me educaram e incentivaram à pesquisa em diferentes etapas do ensino: Karliuza Bittencourt e Luís Gustavo Doninelli Mendes. O entusiasmo que trouxeram às aulas é reflexo de sua paixão pela educação.

Aos pesquisadores e professores que me orientaram e auxiliaram nesta pesquisa: Heraldo José de Amorim, Carlos Solon Soares Guimarães Jr. e Jonnas Espindula. Agradeço profundamente a orientação de vocês, do início ao fim do projeto.

O começo do conhecimento está na descoberta daquilo que ainda não compreendemos.

Frank Herbert, Duna

RESUMO

A análise de comportamento animal é um componente fundamental nas áreas de neurociência e pesquisa biológica, porém ainda depende de métodos intrusivos e de observação manual que estão sujeitos à variabilidade entre avaliadores. Neste trabalho foi investigada a aplicação de técnicas de estimação de pose sem marcadores para a análise de ratos em ambiente laboratorial utilizando visão computacional e aprendizado profundo. Para isso, foi empregada a ferramenta SLEAP para treinamento de modelos de detecção e estimação de pose baseados em redes neurais convolucionais. Os modelos foram inicialmente avaliados em um conjunto de dados de referência e posteriormente treinados e testados em um conjunto de dados próprio obtido em laboratório. A avaliação foi realizada por meio de métricas de detecção e estimação de pose para análise dos erros de localização dos *keypoints*. Os resultados obtidos no conjunto de referência demonstraram elevado desempenho tanto na detecção quanto na estimação de pose. Entretanto, nos conjuntos de dados laboratoriais observou-se uma redução significativa da capacidade de generalização dos modelos, principalmente em cenários contendo oclusões, múltiplos indivíduos e diferenças de domínio entre os dados de treinamento e teste. Apesar das limitações observadas, os resultados indicam que a estimação de pose apresenta potencial para aplicações em estudos comportamentais com roedores, desde que acompanhada de conjuntos de dados mais abrangentes e de uma seleção criteriosa dos *keypoints* utilizados.

Palavras-chave: Aprendizado profundo, Estimação de pose animal, Ratos, Visão computacional

ABSTRACT

Animal behavior analysis is a fundamental component of neuroscience and biological research. However, it still relies on intrusive methods and manual observation procedures that are subject to inter-observer variability. This work investigates the application of markerless pose estimation techniques for the analysis of laboratory rats using computer vision and deep learning. For this purpose, the SLEAP framework was employed to train detection and pose estimation models based on convolutional neural networks. The models were initially evaluated on a reference dataset and subsequently trained and tested on a custom dataset collected in a laboratory environment. The evaluation was performed using detection and pose estimation metrics to analyze keypoint localization errors. The results obtained on the reference dataset demonstrated high performance in both detection and pose estimation tasks. However, a significant reduction in the models' generalization capability was observed on the laboratory datasets, particularly in scenarios involving occlusions, multiple individuals, and domain differences between the training and testing data. Despite these limitations, the results indicate that pose estimation has potential for applications in behavioral studies involving rodents, provided that more comprehensive datasets are available and that the selected keypoints are carefully defined.

Palavras-chave: Animal Pose Estimation, Computer vision, Deep Learning, Rats

LISTA DE ILUSTRAÇÕES

1 Modelo de um neurônio artificial. 17
2 *Frameworks* para APE e HPE e metodologias utilizadas para reconhecimento de ações. 20
3 Resultados alcançados pelo *framework* SLEAP. 24
4 Frame do vídeo de referência do SLEAP. 25
5 Frame do vídeo capturado em vista superior no CREAL. 26
6 Frame do vídeo capturado para teste no CREAL. 26
7 Exemplo de esqueletos anotados. 27
8 *Pipeline* de treinamento. 29
9 Distribuição de tamanhos das anotações dos conjuntos de treinamento. 33
10 *Loss* do modelo centroide durante o treinamento (*dataset* SLEAP). . . . 37
11 *Loss* do modelo centralizado durante o treinamento (*dataset* SLEAP). . 37
12 Distribuição de tamanhos das anotações do conjunto de validação (*dataset* SLEAP). 38
13 Percentual de predições corretas com base na variação do limiar do PCK (*dataset* SLEAP). 39
14 Distribuição do erro para cada *keypoint* (*dataset* SLEAP). 39
15 *Loss* do modelo centroide durante o treinamento (*dataset* CREAL). . . . 40
16 *Loss* do modelo centralizado durante o treinamento (*dataset* CREAL). 41
17 Distribuição de tamanhos das anotações do conjunto de validação(*dataset* CREAL). 42
18 Percentual de predições corretas com base na variação do limiar do PCK (*dataset* CREAL). 42
19 Distribuição do erro para cada *keypoint* (*dataset* CREAL). 43
20 Exemplo de sobreposição dos animais (*dataset* CREAL). 44
21 *Precision* do modelo pela variação do limiar de aceite (*dataset* CREAL - teste). 45
22 Distribuição de tamanhos das anotações do conjunto de teste (*dataset* CREAL). 46
23 Percentual de predições corretas com base na variação extendida do limiar do PCK (conjunto de teste - CREAL). 47
24 Erros *Outliers* do torso (vermelho: predição, verde: anotação). 47
25 Distribuição do erro para cada *keypoint* (conjunto de teste - CREAL). 48
26 Exemplo de ratos escondidos (*dataset* CREAL). 49
27 Resultado de mOKS para os três conjuntos. 49
28 Frames com alto erro médio de inferência (vermelho: predição, verde: anotação). 50
29 Frames com baixo erro médio de inferência (vermelho: predição, verde: anotação). 51

LISTA DE TABELAS

1 *Precision* e *Recall* do modelo de detecção(*dataset* SLEAP). 37
2 *Precision* e *Recall* do modelo de detecção (*dataset* CREAL). 41
3 *Precision* e *Recall* do modelo de detecção (*dataset* CREAL - teste). . . . 44
4 Comparação entre as configurações dos modelos de detecção e estimação utilizados no conjunto SLEAP. 59
5 Comparação entre as configurações dos modelos de detecção e estimação utilizados no conjunto CREAL. 60

LISTA DE ABREVIATURAS

ANN *Artificial Neural Network*

ATRW *Amur Tiger Re-Identification in the Wild*

APE *Animal Pose Estimation*

CNN *Convolutional Neural Network*

CREAL Centro de Reprodução e Experimentação de Animais de Laboratório

COCO *Common Objects in Context*

FN *False Negative*

FP *False Positive*

fps *Frames per Second*

GUI *Graphical User Interface*

HOG *Histogram of Oriented Gradients*

HPE *Human Pose Estimation*

IQR *Interquartile Range*

LEAP *LEAP Estimates Animal Poses*

mAP *mean Average Precision*

MARS *Mouse Action Recognition System*

mOKS *mean Object Keypoint Similarity*

MPII *Max-Planck-Institute für Informatik*

NN *Neural Network*

OKS *Object Keypoint Similarity*

PCK *Percent of Correct Keypoints*

PCP *Percentage of Correct Parts*

PDJ *Percent of Detected Joints*

ReLU *Rectified Linear Unit*

SIFT *Scale-Invariant Feature Transform*

SLEAP *Social LEAP Estimates Animal Poses*

TP *True Positive*

LISTA DE SÍMBOLOS

x_n	Sinal de entrada de um neurônio artificial
w_n	Peso associado ao sinal de entrada de um neurônio artificial
b	Termo de viés (*bias*)
z	Termo de saída do neurônio
$R(z)$	Termo resultante da função de ativação
d	Distância entre pontos (ou elemento genérico do conjunto de distâncias)
x_p, x_r	Posição no eixo x do ponto predito e referência, respectivamente
y_p, y_r	Posição no eixo y do ponto predito e referência, respectivamente
d_i	Distância entre pontos predito e real da instância i
t	Limiar de distância
N	Total de instâncias
s	Escala do animal em pixels
k_i	Fator de incerteza do *keypoint* i
v_i	Visibilidade do *keypoint* i
$\delta(v_i)$	Função indicadora de visibilidade do *keypoint* i
Q_1	Primeiro quartil (percentil 25%) da distribuição de dados
Q_3	Terceiro quartil (percentil 75%) da distribuição de dados
$\mathcal{O}$	Conjunto de dados identificados como *outliers*
D	Conjunto contendo todas as distâncias calculadas
$\in$	Operador lógico de pertencimento ("pertence a")
$\mid$	Operador de restrição de conjunto ("tal que")
$\vee$	Operador lógico de disjunção ("ou")

SUMÁRIO

1 Introdução 14

2 Revisão da Literatura 16
2.1 Visão Computacional 16
2.2 Redes Neurais 17
2.3 Estimação de Pose. 19
2.3.1 *Datasets* para Estimação de Pose Animal 21
2.3.2 Métricas de Avaliação e Desempenho 22

3 Materiais e Métodos 24
3.1 *Framework* de Estimação de Pose 24
3.2 Construção do *Dataset* 25
3.2.1 Dados utilizados 25
3.2.2 Construção do Esqueleto 27
3.2.3 Anotação dos Dados 28
3.3 Treinamento dos modelos 28
3.3.1 Escolha da Arquitetura 29
3.3.2 Pré-processamento dos dados 29
3.3.3 Parâmetros de Treinamento 30
3.3.4 Modelo de Detecção (Centroide) 31
3.3.5 Modelo de Estimação (Centralizado) 32
3.4 Metodologia de Avaliação 32
3.4.1 Modelo de Detecção (Centroide) 32
3.4.2 Modelo de Estimação (Centralizado) 33
3.4.3 Análises Pontuais 34

4 Resultados e Discussão 36
4.1 Resultados com o *dataset* de referência 36
4.2 Resultados com o *dataset* obtido em laboratório 40
4.3 Resultados do conjunto de teste 44

5 Conclusões 52

Referências 54

Apêndices 58

Apêndice A - Parâmetros de treinamento: Modelo SLEAP 59

Apêndice B - Parâmetros de treinamento: Modelo CREAL 60

1 INTRODUÇÃO

A análise do comportamento animal constitui um pilar fundamental em diversas áreas da ciência, incluindo neurociência, farmacologia e ciência do bem-estar animal. Tradicionalmente, essa análise baseia-se na observação direta dos animais, atividades que são frequentemente onerosas em termos de tempo e condicionadas aos critérios e limitações do observador. Isso torna a análise suscetível a subjetividade, fadiga e vieses individuais, além da impossibilidade de observação ao longo das 24 horas do dia (CHEN; ZHU; NORTON, 2021; PÉREZ; TOLER-FRANKLIN, 2023). Outras abordagens podem envolver o uso de sensores ou marcadores visuais, as quais são muitas vezes intrusivas e podem induzir estresse e alterar o próprio comportamento que se pretende medir (MCILWAIN *et al.*, 2001; HANKENSON; BRADEN-WEISS; BLENDY, 2011). Dessa forma, a necessidade de se capturar a expressão comportamental natural e espontânea do animal, longe de interferências experimentais, levou a uma tendência crescente do monitoramento contínuo (GRIECO *et al.*, 2021).

A necessidade de superar essas limitações tem impulsionado o desenvolvimento de métodos automatizados baseados em visão computacional. Neste contexto, as técnicas de *Deep Learning* surgem como tecnologias transformadoras, oferecendo o potencial para automatizar a caracterização do comportamento de forma objetiva e não intrusiva (FAZZARI *et al.*, 2025). Especificamente, a estimação de pose animal, ou seja, a detecção precisa de pontos chave (*keypoints*) corporais a partir de imagens ou vídeos, representa um aprimoramento importante nas metodologias de análise comportamental, servindo como base para análises quantitativas do comportamento. A pose representa a configuração espacial das diferentes partes do corpo do animal ao longo do tempo. A partir da posição relativa dos *keypoints* e de sua evolução temporal, é possível extrair características relacionadas à orientação corporal, velocidade, postura e padrões de movimento, informações que podem ser utilizadas por sistemas automatizados para inferir diferentes comportamentos.

Embora a estimação de pose não realize diretamente a classificação comportamental, diversos trabalhos demonstram que ela fornece as informações necessárias para sistemas capazes de reconhecer comportamentos etologicamente relevantes, como locomoção, alimentação e interações sociais (CHEN; ZHU; NORTON, 2021; PÉREZ; TOLER-FRANKLIN, 2023). A capacidade de extrair automaticamente a pose do animal fornece uma representação quantitativa do movimento corporal que pode ser utilizada por métodos posteriores de classificação comportamental para identificar comportamentos ou níveis de atividade animal. Por exemplo, alterações nos ciclos de sono e vigília, que são biomarcadores críticos em modelos de doenças neurológicas como a doença de Huntington (MORTON, 2023), podem ser monitoradas de forma contínua através da análise automatizada de vídeo. Da mesma forma, a quantificação precisa de atividades como o autocuidado (*grooming*) e a exploração do ambiente na gaiola habitual (*home-cage*) fornece indicadores sensíveis do estado físico, emocional e de bem-estar do animal (GRIECO *et al.*, 2021).

Neste contexto, este trabalho propõe o desenvolvimento de um sistema automatizado de visão computacional para estimação de pose em pequenos roedores, utilizando redes neurais convolucionais adaptadas para animais. A abordagem inclui a utilização de um *framework* de estimação de pose baseado em arquiteturas consolidadas da literatura, dentre as quais se destaca o *Social LEAP Estimates Animal Poses* (SLEAP) (PEREIRA; TABRIS *et al.*, 2022), *framework* utilizado neste trabalho, a montagem de um ambiente controlado para captura de vídeo e a validação do sistema sob diferentes condições de treinamento. Embora a classificação automática de comportamentos represente uma aplicação natural da estimação de pose, este trabalho concentra-se exclusivamente no desenvolvimento e validação do sistema de estimação. A investigação sobre quais *keypoints* são mais adequados para a caracterização de diferentes comportamentos, bem como a implementação de classificadores comportamentais, é deixada como perspectiva para trabalhos futuros.

A automatização do processo de estimação de pose possibilita a análise de grandes volumes de dados com consistência, minimiza o estresse animal por meio da observação não intrusiva e aumenta a confiabilidade dos dados comportamentais, atendendo a uma demanda do campo das ciências biológicas e biomédicas (FAZZARI *et al.*, 2025; GRIECO *et al.*, 2021). Através deste sistema, espera-se fornecer uma ferramenta prática que auxilie pesquisadores na análise objetiva e de alto rendimento do comportamento animal.

Este trabalho visa contribuir para o avanço das práticas experimentais, implementando um sistema para estimação de pose de ratos em laboratório. Através deste, espera-se contribuir para futuras pesquisas, seja na área de estudos em animais ou no desenvolvimento de outras ferramentas de visão computacional, como classificadores de atividades baseados em estimação de pose (PÉREZ; TOLER-FRANKLIN, 2023). O trabalho também busca fornecer uma base para a construção de conjuntos de dados anotados, além dos próprios dados coletados neste estudo. Do ponto de vista da Engenharia de Controle e Automação, os dados de pose obtidos podem servir como entrada para sistemas inteligentes de monitoramento, geração de alertas e gêmeos digitais de animais de laboratório, possibilitando o acompanhamento automatizado do comportamento e do estado dos indivíduos em ambientes experimentais.

2 REVISÃO DA LITERATURA

2.1 Visão Computacional

A visão computacional é um campo interdisciplinar que envolve a ciência da computação, inteligência artificial e processamento de imagens, com o objetivo de permitir que os computadores analisem e interpretem imagens e vídeos. Ela busca desenvolver algoritmos e técnicas que permitam a análise automática de imagens e vídeos, possibilitando a extração de informações relevantes para diversas aplicações (SZELISKI, 2011).

Dentre as aplicações da área de visão computacional destacam-se a detecção e reconhecimento de objetos, segmentação de imagens, reconstrução tridimensional, rastreamento de movimento, inspeção industrial, entre outras (SZELISKI, 2011). Nos últimos anos, essas aplicações têm alcançado maior precisão e robustez devido ao avanço dos métodos de aprendizado profundo (LECUN *et al.*, 1998; GOODFELLOW; BENGIO; COURVILLE, 2016; HE *et al.*, 2016; REDMON *et al.*, 2015).

Além das técnicas de aprendizado profundo, a área da visão computacional também trabalha com técnicas de processamento de imagem, muitas das quais precedem o uso de aprendizado profundo na área. Operações como filtragem, detecção de bordas e extração de características locais constituem a base de muitos sistemas tradicionais, permitindo realçar estruturas relevantes, identificar pontos de interesse e descrever padrões presentes nas imagens. Técnicas de filtragem espacial e detecção de bordas são amplamente discutidas na literatura clássica de processamento digital de imagens (GONZALEZ; WOODS, 2008). No contexto da extração de características, detectores e descritores como o detector de cantos de Harris (HARRIS; STEPHENS, 1988), o método conhecido como *Scale-Invariant Feature Transform* (SIFT) (LOWE, 2004) e o *Histogram of Oriented Gradients* (HOG) (DALAL; TRIGGS, 2005) desempenharam papel fundamental em tarefas como correspondência de pontos, reconhecimento de objetos e análise de movimento.

Com o avanço do aprendizado profundo, técnicas baseadas em redes neurais convolucionais, do inglês *Convolutional Neural Networks* (CNNs), tornaram-se predominantes em diversas tarefas de visão computacional. As CNNs, inicialmente propostas por LeCun (LECUN *et al.*, 1998) permitem que a extração de características seja aprendida automaticamente a partir dos dados. Essa metodologia utilizada por modelos pioneiros como AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e ResNet (HE *et al.*, 2016), elevou significativamente a precisão e a robustez em tarefas de reconhecimento, detecção e segmentação. Outra característica importante das CNNs é a capacidade de reutilizar informações aprendidas em diferentes tarefas através de técnicas de Aprendizado por Transferência (*Transfer Learning*). Estas técnicas permitem que se adapte modelos já treinados em grandes bases de dados em situações específicas (GOODFELLOW; BENGIO; COURVILLE, 2016). Além disso, devido a necessidade do treinamento de redes neurais e anotação de dados que vem junto, novas abordagens têm proposto o uso de dados sintéticos

para treinamento em conjunto com dados experimentais, como é o caso de (ARENT *et al.*, 2021).

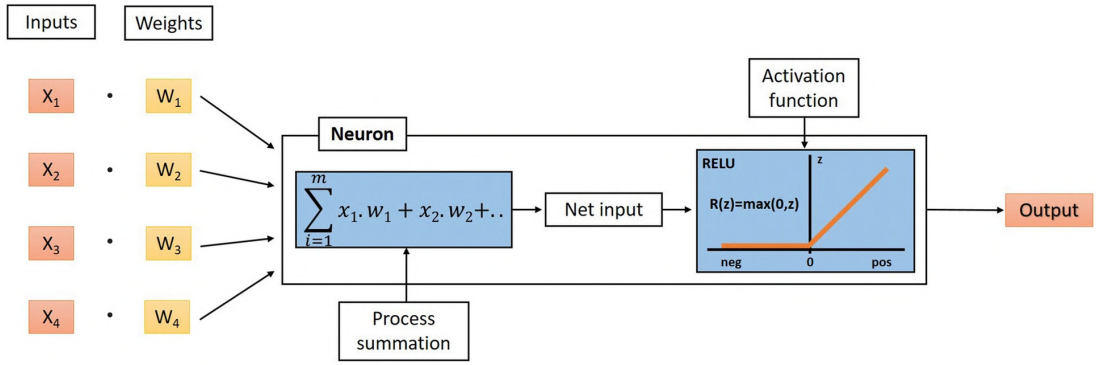
O desenvolvimento da visão computacional nas diversas áreas mencionadas permitiu sua aplicação em áreas que necessitavam dos vários sistemas integrados, como é o caso da estimação de pose. A estimação de pose é uma das áreas da visão que combina elementos de detecção e segmentação de estruturas corporais. Seu objetivo é determinar a configuração espacial, em duas ou três dimensões, de um corpo a partir de dados visuais. Diversas metodologias têm sido implementadas, principalmente com foco em estimação de pose humana, porém alternativas especializadas para estimação de pose animal de diversas espécies também são exploradas (ZAIDI; WAGAN, 2024).

2.2 Redes Neurais

As Redes Neurais Artificiais, do inglês *Artificial Neural Networks* (ANNs), são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano. Essa inspiração se reflete nas unidades básicas de processamento das ANNs, os neurônios artificiais, que, assim como os neurônios biológicos, são interconectados formando redes capazes de aprender padrões a partir de dados (LÓPEZ; LÓPEZ; CROSSA, 2022).

Um neurônio artificial, unidade fundamental das ANNs, consiste em um modelo matemático que simula, de forma simplificada, o comportamento de um neurônio biológico (LÓPEZ; LÓPEZ; CROSSA, 2022). Sua arquitetura básica, ilustrada na Figura 1, compreende as etapas do processo denominado de Propagação Direta (*Forward Propagation*):

Figura 1: *Modelo de um neurônio artificial.*



Fonte: Adaptado de Georgevici e Terblanche (2019)

Inicialmente, o neurônio recebe múltiplos sinais de entrada $(x_1, x_2, \ldots, x_n)$ provenientes das camadas anteriores ou, no caso da camada de entrada, os dados em si. Cada um desses sinais é ponderado por um coeficiente associado $(w_1, w_2, \ldots, w_n)$, denominado peso sináptico, que determina a influência relativa de cada conexão. Adicionalmente, é incorporado um termo de viés (b), que atua como um deslocamento no processo de ativação (HAYKIN, 2009; LÓPEZ; LÓPEZ; CROSSA, 2022).

Posteriormente, ocorre a aplicação da função de ativação. O valor obtido através do somatório ponderado dos sinais de entrada é somado ao termo de b de viés, conforme apresentado na Equação 1, e processado por uma função de ativação não linear.

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{1}$$

Dentre as funções utilizadas, destaca-se a *Rectified Linear Unit* (ReLU), definida pela Equação 2, amplamente adotada devido à sua simplicidade computacional (GOODFELLOW; BENGIO; COURVILLE, 2016). Esta função funciona atribuindo à saída o valor da entrada, quando este for positivo, e zerando-a quando este é igual ou menor que zero. O valor resultante dessa função é propagado às camadas subsequentes, dando continuidade ao processo de Propagação Direta.

$$R(z) = \max(0, z) \tag{2}$$

O valor resultante desta função é utilizado pelas camadas subsequentes como entrada, sendo propagado adiante ao longo da rede. O processo descrito é realizado para todas as camadas da rede. As camadas iniciais e finais são caracterizadas dessa forma por receberem os dados e computarem a resposta final, respectivamente, enquanto camadas intermediárias são denominadas camadas ocultas. No caso da última camada, a de saída, o valor resultante desta função constitui a resposta da rede ao dado de entrada (HAYKIN, 2009). Este processo, por si só, não garante que a rede neural se ajuste de acordo com os dados recebidos. Para isso, é necessário que ela aprenda e se adapte. Nas Redes Neurais, do inglês *Neural Networks* (NNs), este processo de aprendizado nada mais é do que o ajuste de forma iterativa de seus parâmetros (pesos e viés), com o objetivo de minizar o erro entre sua previsão e o resultado esperado. Para esse ajuste, a NN precisa realizar duas etapas: o cálculo da perda, com o uso de uma função de perda (*Loss Function*) e o ajuste de seus parâmetros, através do processo conhecido como Retropropagação (*Back Propagation*) (HAYKIN, 2009; LÓPEZ; LÓPEZ; CROSSA, 2022).

Após o processo de Propagação Direta, a saída prevista pela rede é comparada com o valor real esperado, por meio de uma função de perda. A função tem como objetivo quantificar quão errada estava a estimativa da rede em relação ao resultado verdadeiro (HAYKIN, 2009; MITCHELL, 1997).

Uma vez quantificado o erro de estimação, é realizado o processo de Retropropagação. Esse tem como objetivo propagar o erro através da rede no sentido contrário (da saída para a entrada), sendo possível calcular a direção em que os parâmetros devem ser ajustados, no que é conhecido como Descida do Gradiente (*Gradient Descent*). A velocidade de mudança dos parâmetros, também dita como a velocidade em que se realiza o *Gradient Descent* é definida através da velocidade de aprendizado (*Learning Rate*). O ciclo completo realizado pelas etapas descritas é o que permite a NN aprender com o tempo, melhorando seu desempenho na tarefa desejada (HAYKIN, 2009; MITCHELL, 1997).

Dentre os modelos existentes de redes neurais, destacam-se as Redes Neuras Convolucionais, *Convolutional Neural Networks* (CNNs), um tipo de NN muito utilizada para processamento de dados com estrutura de grade, como imagens e vídeos (LÓPEZ; LÓPEZ; CROSSA, 2022; LECUN *et al.*, 1998). O desenvolvimento das CNNs permitiu o avanço de tarefas como classificação de imagens, detecção de objetos e, como é o caso deste trabalho, estimação de pose (LÓPEZ; LÓPEZ; CROSSA, 2022; GRAVING *et al.*, 2019). As CNNs têm a capacidade de identificar aspectos importantes em imagens, como pontas e bordas e, a partir destas informações, identificar formas e contornos (GOODFELLOW; BENGIO; COURVILLE, 2016; LECUN *et al.*, 1998).

2.3 Estimação de Pose

A estimação de pose animal, do inglês *Animal Pose Estimation* (APE), constitui uma área emergente na visão computacional, com aplicações significativas em áreas que envolvam o estudo do comportamento animal (PEREIRA; TABRIS *et al.*, 2022). Diferentemente da estimação de pose humana, *Human Pose Estimation* (HPE), que conta com base de dados (*datasets*) extensos como *Common Objects in Context* (COCO) (LIN *et al.*, 2015) e *Max-Planck-Institute für Informatik* (MPII) (ANDRILUKA *et al.*, 2014), a APE enfrenta desafios únicos relacionados à diversidade morfológica entre espécies, escassez de dados anotados e variações consideráveis na aparência corporal (YU *et al.*, 2021; ZAIDI; WAGAN, 2024). Esse processo envolve a detecção e o rastreamento de pontos-chave específicos no corpo de um animal, como articulações e extremidades, os quais são usados para inferir sua postura e movimentos. Arquiteturas atuais de estimação de pose frequentemente utilizam redes neurais convolucionais, do inglês *Convolutional Neural Networks* (CNNs), e podem variar conforme o número de indivíduos presentes na imagem (ZAIDI; WAGAN, 2024).

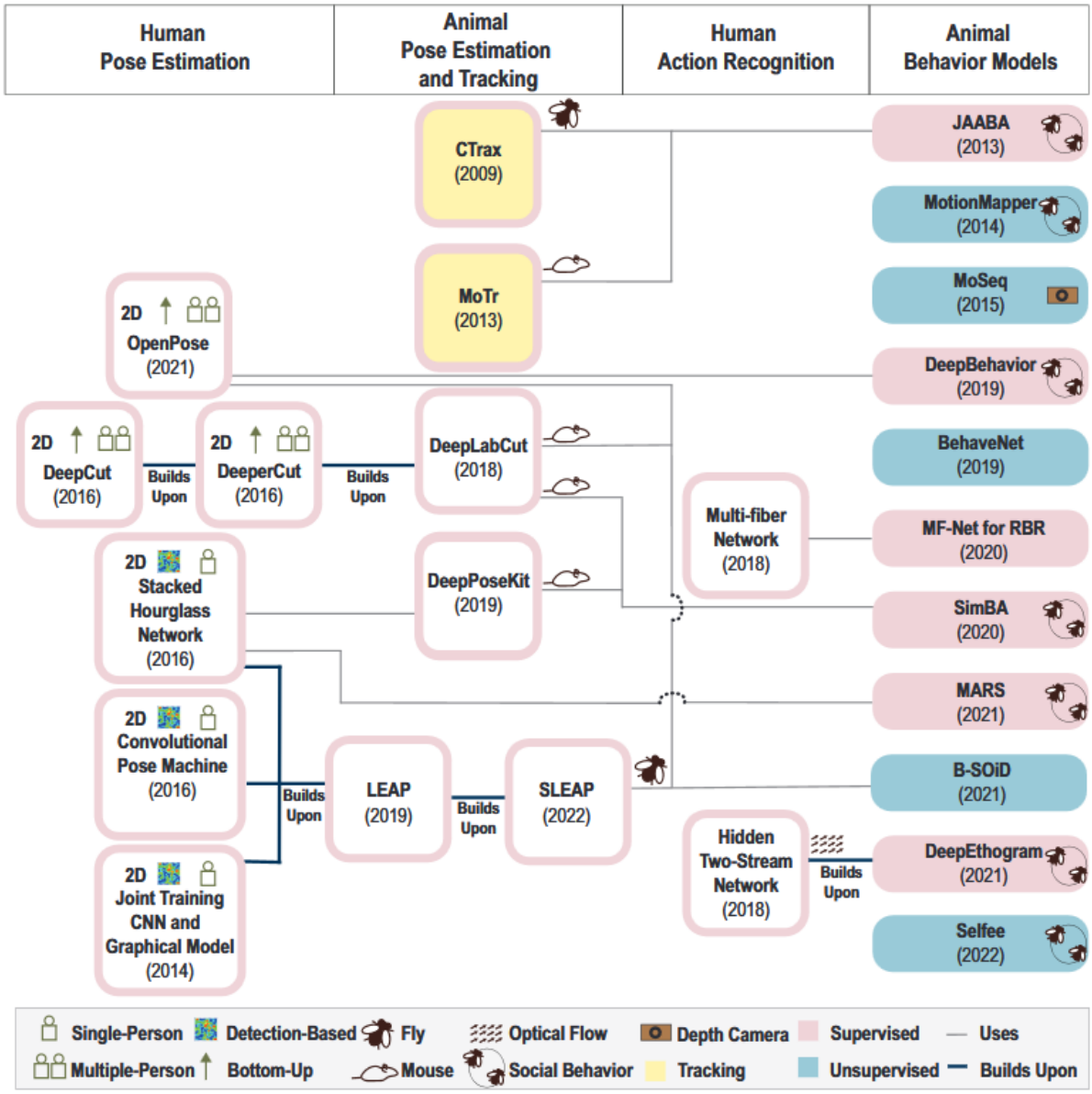
Além dos desafios singulares da estimação de pose animal, como quantidade de dados disponível e variações morfológicas entre animais, a APE também enfrenta dificuldades técnicas comuns a outras tarefas de visão computacional. Desafios incluem oclusões frequentes em ambientes naturais ou de laboratório, variações extremas de iluminação e fundo, e a necessidade de distinguir entre indivíduos em cenários multi-animal (MATHIS *et al.*, 2018; GRAVING *et al.*, 2019).

As abordagens desenvolvidas para APE frequentemente adaptam técnicas originalmente desenvolvidas para HPE (PÉREZ; TOLER-FRANKLIN, 2023), seguindo duas principais estratégias: *top-down* e *bottom-up*. Metodologias do tipo *top-down* começam com a detecção do objeto de interesse na imagem e a partir disto identificam seus pontos chaves (*keypoints*), geralmente as articulações corporais. Já as abordagens do tipo *bottom-up* primeiramente detectam todas as partes do corpo na imagem e depois agrupam estas partes para formar um corpo completo. Ambas abordagens têm apresentado resultados promissores, a depender da aplicação e do contexto (ZAIDI; WAGAN, 2024; PEREIRA; TABRIS *et al.*, 2022). A Figura 2 apresenta os principais modelos para HPE e APE e como estes se relacionam entre si. Além disso, é demonstrado o ferramentário necessário para alguns destes modelos, além do tipo de animal nos quais foram baseados.

Dentre os modelos para APE existem algumas abordagens de destaque (ZAIDI; WAGAN, 2024; PÉREZ; TOLER-FRANKLIN, 2023), que podem ser adotadas para estudos em ratos. Como observado na Figura 2, alguns desses modelos já foram validados para estudos com pequenos roedores (ratos e/ou camundongos) sob condições específicas (PÉREZ; TOLER-FRANKLIN, 2023; ZAIDI; WAGAN, 2024). A seguir, apresentam-se as principais metodologias utilizadas, com ênfase na sua aplicação prática e vantagens para a análise comportamental automatizada.

O DeepLabCut, primeiro modelo à adotar técnicas de HPE para animais (PÉREZ; TOLER-FRANKLIN, 2023), introduziu uma abordagem baseada em aprendizado por transferência que permitiu a adaptação de redes pré-treinadas em grandes bases de dados, como o ImageNet, para a estimativa de pose animal com conjuntos de anotação reduzidos. O modelo é ajustado com um número limitado de *frames* manualmente anotados, entre 100 e 200, provenientes de vídeos reais de um animal, permitindo a detecção de keypoints como focinho, orelhas, base da cauda e até articulações dos dedos de camundongos (MATHIS *et al.*, 2018). Além da possibilidade de utilizar o modelo com baixa quantidade de dados, o modelo mostrou-se robusto a desafios visuais como variações de iluminação, oclusões

Figura 2: *Frameworks para APE e HPE e metodologias utilizadas para reconhecimento de ações.*



Fonte: Pérez e Toler-Franklin (2023)

parciais e fundos dinâmicos, sendo validado em experimentais incluindo moscas-da-fruta e camundongos. DeepLabCut também demonstrou capacidade de estimar pose de multiplos indivíduos em um mesmo vídeo, embora com desempenho reduzido em comparação com a estimação de pose individual (MATHIS *et al.*, 2018; PÉREZ; TOLER-FRANKLIN, 2023). Apesar disso, os autores afirmam que o modelo pode ser adaptado para múltiplos animais, uma vez introduzido cenários com mais de um indivíduo durante o treinamento (MATHIS *et al.*, 2018).

Outra abordagem, LEAP, foi desenvolvida para estimação de pose em vídeos de pequenos animais, como moscas-das-frutas e camundongos, adapatando CNNs desenvolvidas para HPE (PEREIRA; ALDARONDO *et al.*, 2019; PÉREZ; TOLER-FRANKLIN, 2023). Além de apresentar alta performance com baixa quantidade de dados de treino, apenas 100 *frames* anotados, o modelo também conseguiu taxas de estimação de pose de

185 Hz (PEREIRA; ALDARONDO *et al.*, 2019). Além do alto desempenho apresentado pela ferramenta, o LEAP também disponibiliza uma interface gráfica de usuário, *Graphical user interface* (GUI), que facilita o processo de anotação manual dos *keypoints*, tornando-o acessível para pesquisadores sem experiência prévia em aprendizado de máquina (PEREIRA; ALDARONDO *et al.*, 2019).

Baseado em LEAP, o SLEAP (*Social* LEAP), introduziu melhorias significativas em relação ao seu antecesseor, incluindo a capacidade de rastrear múltiplos animais simultaneamente e a combinação dinâmica de abordagens *top-down* e *bottom-up* para otimizar a precisão da estimação de pose em diferentes cenários (PEREIRA; TABRIS *et al.*, 2022; ZAIDI; WAGAN, 2024). Assim como LEAP, o SLEAP também oferece uma GUI para facilitar a anotação manual dos *keypoints* (PEREIRA; TABRIS *et al.*, 2022). SLEAP também se destaca por sua flexibilidade e velocidade de inferência, suportando diversos tipos de arquiteturas de rede neural e permitindo taxas de processamento em tempo real, quando em hardware adequado (PEREIRA; TABRIS *et al.*, 2022).

DeepPoseKit é outra abordagem desenvolvida para APE, validada em animais como moscas, zebras e peixes. Além de utilizar arquiteturas de CNNs adaptadas para a tarefa, o DeepPoseKit também incorpora técnicas de aumento de dados (*data augmentation* em inglês) para melhorar a robustez do modelo frente a variações visuais (GRAVING *et al.*, 2019). A ferramenta também foi testada em cenários com múltiplos indivíduos, demonstrando capacidade de separação e rastreamento de poses de indivíduos diferentes (GRAVING *et al.*, 2019).

2.3.1 Datasets para Estimação de Pose Animal

Como demonstrado por (ZAIDI; WAGAN, 2024), diferentemente da estimação de pose humana, a área de estimação de pose animal tem enfrentado desafios significativos relacionados à escassez de *datasets* anotados em larga escala. Nesta área são encontradas limitações, não apenas quanto à quantidade de dados disponíveis, mas também relacionadas a diversidade existente de espécies animais e também a suas variedades morfológicas.

Tendo este projeto o objetivo de estudar a estimação de pose em ratos, é importante destacar que poucos *datasets* públicos são focados em pequenos roedores, como ratos e camundongos. Apesar disto, existem alguns conjuntos de dados que incluem essas espécies, embora em quantidades limitadas (ZAIDI; WAGAN, 2024).

Dentre os *datasets* existentes, os seguintes conjuntos foram encontrados com potencial para uso em estimação de pose de ratos, são eles:

- **AP-10K**: O primeiro *dataset* em larga escala para estimação de pose de animais. Ele contém mais de 10.015 imagens, de 54 espécies diferentes, totalizando 23 famílias de animais. Além das imagens, apresenta também a anotação dos *keypoints*, validadas manualmente. Dentre os animais incluídos no conjunto, destaca-se a presença ratos e camundongos, os dois possíveis alvos desta pesquisa (YU *et al.*, 2021).

- **SLEAP**: O *dataset* SLEAP apresenta imagens e vídeos de moscas, abelhas, camundongos e ratos, assim como anotação de *keypoints* referentes. Diferentemente de outros conjuntos de dados, SLEAP apresenta vídeos com múltiplos indivíduos, o que apresenta um desafio adicional para a etapa de estimação de pose (PEREIRA; TABRIS *et al.*, 2022).

- **MARS**: o *Mouse Action Recognition System* (MARS)trata-se de uma base base de dados focada na classificação de comportamento de ratos em situações sociais. Ele apresenta vídeos com *keypoints* anotados por diversos anotadores diferentes com o intuito de garantir precisão do processo (SEGALIN *et al.*, 2021a).

Além destes, outros *datasets* foram desenvolvidos para diferentes espécies animais, como o Animal-Pose *Dataset*, contendo dados de cães, gatos, ovelhas, vacas e cavalos; Horse-10 e Horse-30, que apresentam um total de 8.000 *frames* de cavalos; MacaquePose desenvovlido para primatas não-humanos e *Amur Tiger Re-Identification in the Wild* (ATRW) para tigres (ZAIDI; WAGAN, 2024).

Tendo em vista as limitações referentes à quantidade de dados disponíveis para estimação de pose de animais, algumas abordagens exploraram a possibilidade de transferência de aprendizado entre espécies diferentes. A pesquisa realizada por (DOORNWEERD *et al.*, 2021) investigou a possibilidade de utilizar redes de estimação de pose treinadas e avaliadas em diferentes combinações de perus e frangos. Os modelos treinados em uma espécie e testados em outra apresentaram altos erros de predição, enquanto modelos especializados (treinados e testados sobre a mesma espécie), apresentaram ótimos resultados. Apesar disso, modelos treinados com dados mistos das duas espécies, apresentaram taxas de erro apenas levemente superiores em relação aos modelos especializados e exigiram menos da metade de *frames* anotados por espécie. O estudo concluiu que modelos dentro da mesma espécie fornecem um melhor desempenho geral para estimativas precisas, mas que modelos treinados com multiplas espécies podem reduzir a necessidade de anotações quando as espécies possuem morfologias comparáveis (DOORNWEERD *et al.*, 2021). Além disso, trabalhos como *Cross-Domain Adaptation for Animal Pose Estimation* (CAO *et al.*, 2019) abordaram o desafio da escassez de dados anotados para animais utilizando de métodos de adaptação cruzada entre domínios. Os autores utilizaram um *dataset* com 5.517 instâncias de cinco mamíferos (cão, gato, cavalo, ovelha e vaca) anotadas com até 20 *keypoints*. Esta abordagem mostrou-se particularmente eficaz para espécies com planos corporais similares aos humanos, indicando que a morfologia desempenha um papel crucial na transferência de aprendizado entre domínios (CAO *et al.*, 2019).

2.3.2 Métricas de Avaliação e Desempenho

As métricas de avaliação para estimativa de pose animal foram herdadas principalmente das metodologias de HPE, com adaptações para o contexto animal (ZAIDI; WAGAN, 2024; PÉREZ; TOLER-FRANKLIN, 2023). Entre as métricas mais comuns na literatura, destaca-se: *mean Average Precision*, *Percentage of Correct Parts*, *Percent of Detected Joints* e *Percent of Correct Keypoints*.

- ***mean Average Precision* (mAP)**: métrica utilizada não apenas por modelos de estimação de pose, mas também de detecção. Ela é calculada a partir da área sob a curva de *precision* e *recall* para diferentes limiares de aceite entre valores preditos e de referência.

- ***Percentage of Correct Parts* (PCP)**: avalia a detecção de membros. Um membro é considerado detectado corretamente se a distância entre a predição e a localização verdadeira do membro estiver dentro de um limiar definido como metade do comprimento do membro. Esta métrica costuma penalizar resultados com membros mais curtos (ZAIDI; WAGAN, 2024; PÉREZ; TOLER-FRANKLIN, 2023).

- ***Percent of Detected Joints* (PDJ)**: avalia a detecção de articulações. Uma articulação é considerada detectada corretamente se a distância entre a posição predita e a verdadeira da articulação estiver dentro de um limiar definido como uma fração específica do diâmetro do torso. Esta métrica foi desenvolvida para mitigar o as penalidades impostas por PCP em articulações localizadas em membros mais curtos (ZAIDI; WAGAN, 2024; PÉREZ; TOLER-FRANKLIN, 2023).
- ***Percent of Correct Keypoints* (PCK)**: avalia se a distância entre o *keypoint* predito e a localização verdadeira da articulação está dentro de um limiar variável, geralmente baseado no tamanho do torso ou da cabeça. Esta métrica também funciona para aliviar o impacto de penalidades em *keypoints* localizados em membros menores (ZAIDI; WAGAN, 2024).

3 MATERIAIS E MÉTODOS

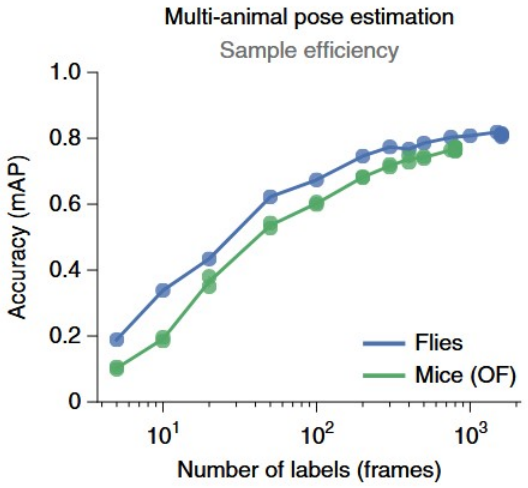
Para realização deste trabalho, o desenvolvimento e experimentação foram realizados em linguagem Python, devido ao suporte existente para desenvolvimento de modelos de visão computacional e também pela compatibilidade com o *framework* SLEAP, utilizado para estimação de pose.

A metodologia utilizada abordou as seguintes etapas: escolha do *framework* de estimação de pose, construção do *dataset*, definição dos parâmetros de treinamento e definição das métricas de avaliação. A seguir são apresentados os detalhes de cada uma dessas etapas.

3.1 *Framework* de Estimação de Pose

Com base na revisão da literatura, optou-se pela escolha do *framework* SLEAP para estimação de pose de ratos. Além de apresentar funcionalidades necessárias, como GUI para anotação de dados e possibilidade de construção do esqueleto para estimação de pose, os resultados apresentados pelo SLEAP se mostraram promissores para os objetivos deste trabalho. Conforme apresentado por (PEREIRA; TABRIS *et al.*, 2022), o *framework* consegue alcançar resultados de estimação elevados com uma baixa quantidade de dados anotados, como indica a Figura 3, que apresenta o desempenho dos modelos em função do número de *frames* anotados. Com aproximadamente 300 *frames* anotados o modelo alcança um mAP próximo de 0,7 para estimação de pose de camundongos.

Figura 3: *Resultados alcançados pelo framework SLEAP.*



Fonte: (PEREIRA; TABRIS *et al.*, 2022)

3.2 Construção do *Dataset*

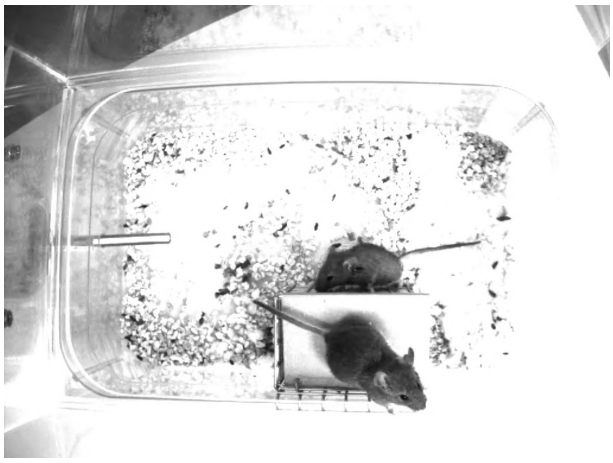
No contexto deste trabalho, o processo de construção do *dataset* compreende não apenas a definição dos dados utilizados (imagens ou vídeos), mas também a definição do esqueleto e a anotação dos dados.

3.2.1 Dados utilizados

Para construção dos *datasets* foram utilizados dados de duas fontes diferentes: um vídeo de dois camundongos disponibilizado pelo próprio *framework* SLEAP e dois vídeos gravados de ratos wistar no ambiente do Centro de Reprodução e Experimentação de Animais de Laboratório (CREAL), na UFRGS. O vídeo disponibilizado pelo SLEAP foi utilizado como base para validação dos resultados encontrados no artigo de referência (pereira; tabris *et al.*, 2022), e também para servir como referência inicial para estudo e desenvolvimento.

O vídeo monocromático fornecido pelo SLEAP, filmado em perspectiva superior e com iluminação controlada, apresenta dois camundongos pretos em um ambiente fechado, como observado na Figura 4. Além dos animais, o ambiente apresenta uma plataforma onde os camundongos podem subir e paredes levemente reflexivas. Destaca-se que os animais presentes são de coloração escura, o que os diferencia do fundo claro e facilita a detecção. O vídeo apresenta uma resolução de 1024x768 pixels, taxa de quadros de 47 fps e duração de 5 minutos, totalizando 14100 *frames*.

Figura 4: *Frame do vídeo de referência do SLEAP.*



Fonte: (pereira; tabris *et al.*, 2022)

Os vídeos coletados no CREAL foram obtidos com a câmera de um celular Samsung S23, sem nenhum tipo de zoom, com resolução de 1920x1080 pixels à 30 fps e mantendo a iluminação controlada. O celular utilizado foi fixado na superfície inferior de uma prateleira no laboratório onde se encontram os animais, de forma que fosse possível filmá-los com vista superior, assemelhando-se ao vídeo de referência. Na Figura 5 é apresentado um quadro do vídeo contendo dois indivíduos, com duração total de aproximadamente 12 minutos, sendo capturados aproximadamente 22.300 *frames*.

Destaca-se que, pelo ambiente estéril do laboratório, alguns cuidados adicionais foram necessários para realização das filmagens. O celular utilizado foi embalado em

plástico filme para impedir contaminação e quaisquer outros itens não eletrônicos foram higienizados com álcool antes de entrarem no ambiente.

Figura 5: *Frame do vídeo capturado em vista superior no CREAL.*



Fonte: Autor

Além do vídeo contendo dois indivíduos, foi coletado um vídeo adicional contendo apenas um animal. Este vídeo foi utilizado para avaliar o desempenho do modelo diante de um domínio similar, porém não observado, servindo como conjunto de teste. Na Figura 6 é exibido um *frame* de exemplo. Esse segundo vídeo apresenta diferenças que podem facilitar e também dificultar a tarefa de detecção. A presença de apenas um indivíduo remove a complexidade de interações entre diferentes animais, enquanto que a presença de ração na parte superior da gaiola proporciona um ambiente com possibilidade de obstrução, dificultando a detecção do animal. Este vídeo possui duração total de aproximadamente 9,5 minutos e 17.100 *frames*.

Figura 6: *Frame do vídeo capturado para teste no CREAL.*



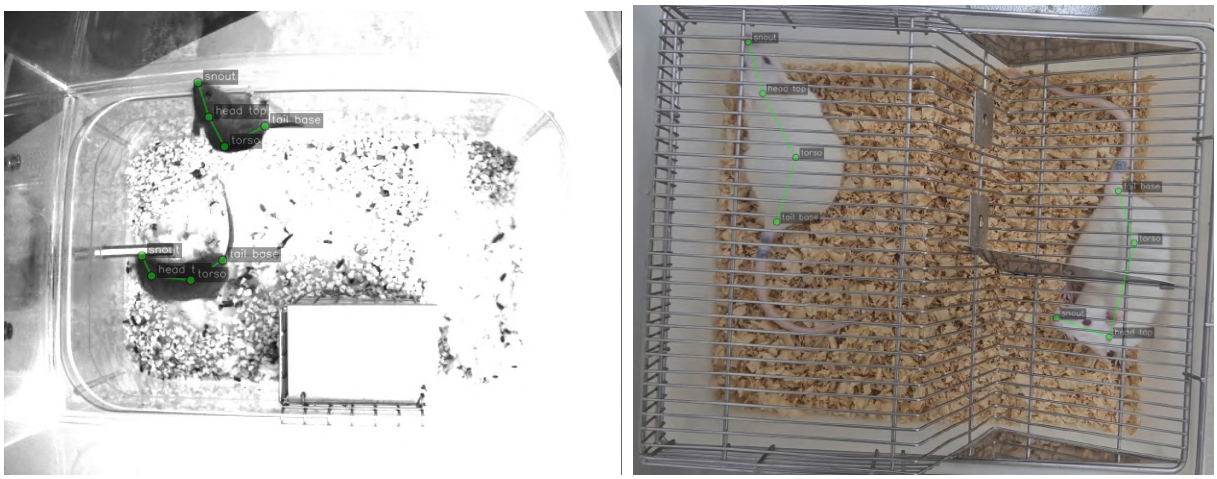
Fonte: Autor

3.2.2 Construção do Esqueleto

O processo de construção do esqueleto trata-se da definição dos *keypoints* que são anotados e avaliados durante estimação de pose e como estes se conectam. Esta escolha depende não apenas da aplicação mas também dos dados existentes e resultados esperados. Considerando as dificuldades de estimação de pose animal (em especial de pequenos roedores) apresentadas na Seção 2.3, como a variação morfológica e a fisionomia dos animais avaliados, optou-se por não adotar o uso de juntas de membros locomotores no esqueleto construído. Ambos os vídeos gravados e disponibilizados apresentaram baixa visibilidade destes membros, de forma que dificilmente seriam anotados com precisão e acabariam confundindo o modelo.

Considerando-se que o objetivo deste trabalho trata-se de um estudo inicial de estimação de pose, procurou-se pelo número mínimo de pontos que permitam a extração das informações de posição e movimento deste animal. Desse modo, o esqueleto proposto é composto de quatro *keypoints*, sendo eles: *snout* (focinho); *head top* (topo da cabeça); *torso* (torso); e *tail base* (base da cauda). Os pontos e suas respectivas conexões são apresentados na Figura 7.

Figura 7: *Exemplo de esqueletos anotados.*



(a) *dataset SLEAP.* **(b)** *dataset CREAL.*

Fonte: Autor.

Os quatro *keypoints* escolhidos permitem obter informações da curvatura do animal e também da direção para a qual está olhando, enquanto uma quantidade menor que esta não apresentaria estas mesmas informações ao mesmo tempo. Além disso, um número maior de *keypoints* aumentaria proporcionalmente o custo computacional do uso do modelo e o tempo necessário para anotação de dados, também sendo fatores considerados na decisão. Apesar disso, é possível aumentar o número de *keypoints* do esqueleto. Candidatos interessantes para isso seria pontos ao longo da cauda, para avaliar seu movimento, posições intermediárias ao longo da coluna, para uma precisão maior da sua curvatura, os ombros do animal, para avaliar sua locomoção ou até *keypoints* nos olhos ou orelhas.

3.2.3 Anotação dos Dados

Para anotação dos dados utilizou-se a interface gráfica do SLEAP. As anotações foram realizadas manualmente por um único anotador, de forma a reduzir possíveis vieses de anotação advindos de observadores diferentes. Conforme apresentado por (PAULLADA *et al.*, 2021; SEGALIN *et al.*, 2021b), anotadores diferentes podem introduzir variação na forma como os dados são anotados, principalmente em regiões mais homogêneas e com menos referências visuais. Por outro lado, a presença de um único anotador pode influenciar o modelo com seu próprio viés.

Apesar dos vídeos utilizados apresentarem uma quantidade considerável de *frames*, optou-se por selecionar apenas parte destes dados para anotação. Esta restrição foi motivada por dois fatores principais: o tempo exigido para anotação dos dados e a similaridade entre os *frames* consecutivos.

Devido ao fato da anotação ter sido feita por um anotador apenas, o processo de anotação se mostrou bastante demorado, impossibilitando o uso de uma quantidade muito grande de dados. Estimou-se que o tempo gasto para anotação de um *frame* com dois indivíduos fosse de aproximadamente 30 segundos, podendo ser ainda maior para cenários mais complexos. Considerando os 300 *frames* anotados no vídeo de referência, o tempo gasto para anotação foi de aproximadamente 2 horas e 30 minutos por vídeo.

Destaca-se porém, que o fator mais determinante para a escolha de uma quantidade menor de dados anotados foi a similaridade entre *frames* consecutivos. Em momentos de menor atividade dos animais, a posição destes permaneceria relativamente estática, limitando a variação de poses, podendo gerar assim um *dataset* desbalanceado.

Portanto, a seleção dos *frames* utilizados foi feita através de intervalos regulares de tempo, de forma a garantir uma boa distribuição de pose dos animais. Além disso, para acelerar o processo de anotação e também para testar alguns parâmetros de treinamento, modelos intermediários foram treinados com quantidades crescentes de dados anotados, realizando-se a inferência sobre os demais *frames* e apenas corrigindo essas poses. Esse processo de anotação e treinamento iterativo é chamado de *human-in-the-loop* e é recomendado para acelerar o processo de anotação (PEREIRA; TABRIS *et al.*, 2022).

Devido à diferença de taxa de quadros entre os vídeos, o intervalo de tempo entre os *frames* anotados difere para cada um deles. Para o vídeo de referência, com taxa de quadros de 47 fps, este intervalo foi de aproximadamente 1 segundo, enquanto que para o vídeo coletado no laboratório, com taxa de quadros de 30 fps, este intervalo foi de aproximadamente 2.5 segundos. Obtendo-se assim, 300 *frames* para cada um dos vídeos e obtendo-se uma variedade mais alta de poses. Destaca-se que pela presença de dois indivíduos nos vídeos, apesar de apenas 300 *frames* anotados, o total de animais observados por vídeo foi de 600.

Para o vídeo de teste, optou-se por utilizar uma quantidade levemente maior de *frames*. Como o vídeo apresentava apenas um animal, para o mesmo número de *frames* anotados, haveria metade do número de instâncias. Assim, foram anotados 350 *frames*, com intervalo de 1.6 segundos entre eles.

3.3 Treinamento dos modelos

A seguir são apresentados detalhes sobre a arquitetura utilizada, configuração dos modelos de ambos *datasets* e processo de treinamento, além de técnicas de pré-processamento e *data augmentation*.

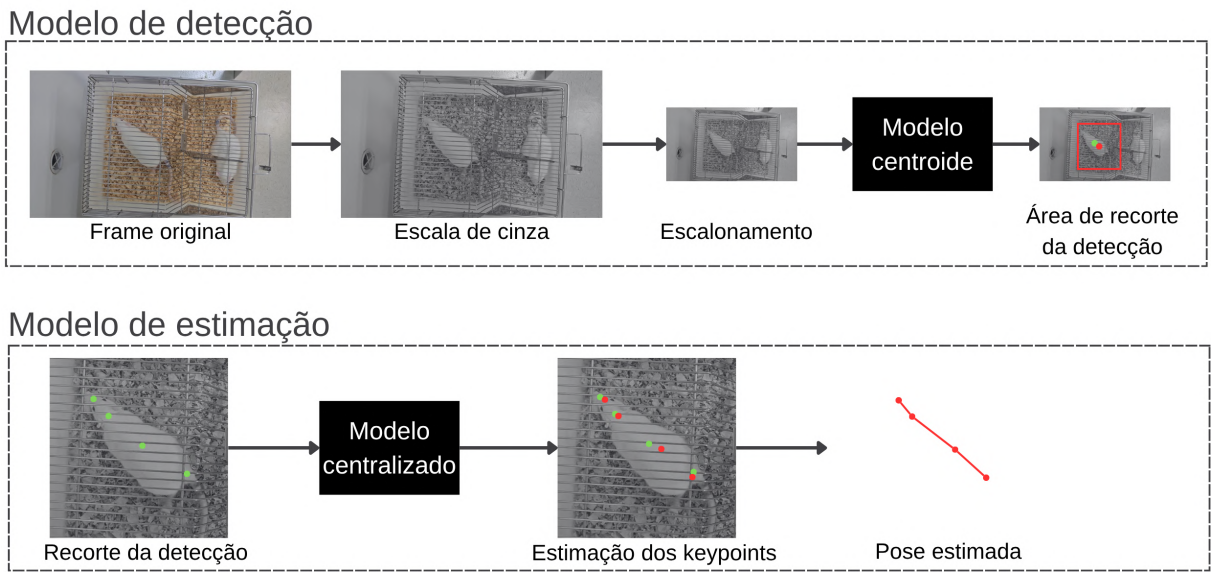
3.3.1 Escolha da Arquitetura

O *framework* SLEAP apresenta diferentes opções de arquitetura para estimação de pose, sendo elas a abordagem *top-down* e a abordagem *bottom-up*. A abordagem *top-down* é caracterizada por realizar a detecção dos animais presentes na imagem e, em seguida, realizar a estimação de pose para cada animal detectado. Já a abordagem *bottom-up* é caracterizada por realizar a estimação de pose para cada parte do corpo presente na imagem e, em seguida, associar as partes do corpo que pertencem ao mesmo animal (PEREIRA; TABRIS *et al.*, 2022).

Conforme apresentado por (PEREIRA; TABRIS *et al.*, 2022), a abordagem *top-down* foi a que teve, no geral, melhores resultados para estimação, enquanto a abordagem *bottom-up* se destacou em termos de escalabilidade do número de indivíduos presentes na imagem. Consequentemente, para este trabalho, optou-se pela abordagem *top-down*, uma vez que os vídeos utilizados apresentam uma quantidade relativamente baixa de indivíduos.

Devido ao seu funcionamento, a abordagem *top-down* é composta por dois modelos diferentes: um modelo de detecção, responsável pela detecção dos animais presentes na imagem, e um modelo de estimação, para estimar a pose de cada animal detectado. Dessa forma, o treinamento e avaliação de cada modelo é realizado de forma independete, utilizando métricas e parâmetros de treinamento específicos para cada um deles. A Figura 8 exemplifica o processo de treinamento descrito para a abordagem *top-down*.

Figura 8: *Pipeline de treinamento.*



Fonte: Autor.

3.3.2 Pré-processamento dos dados

Algumas técnicas de pré-processamento foram aplicadas aos vídeos, tanto para aumentar a robustez dos modelos a possíveis variações do ambiente e animais, quanto para reduzir o custo computacional do treinamento, considerando as limitações do hardware utilizado, uma NVIDIA GeForce RTX 4060. Os vídeos foram convertidos para escala de cinza, reduzindo a quantidade de informação processada e consequentemente o custo

computacional, sem que houvesse perda de informações relevantes para a tarefa de estimação de pose.

Para o modelo de detecção também foi aplicado um escalonamento da imagem de entrada com fator de 0,5, reduzindo a resolução das imagens pela metade, tanto para o *dataset* de referência quanto para o *dataset* coletado em laboratório. Essa escolha foi definida a partir de testes preliminares realizados durante o desenvolvimento, nos quais se observou que a redução da resolução permitia acelerar o treinamento e a inferência sem comprometer significativamente o desempenho do modelo de detecção. Além disso, a utilização da resolução original resultava em elevado consumo de memória de vídeo, ocasionando limitações durante o treinamento. Dessa forma, o fator de escala adotado representou um compromisso entre custo computacional e desempenho do modelo, não sendo necessariamente a configuração ideal.

Para o modelo de estimação, não foi aplicado nenhum tipo de escalonamento na entrada, mantendo-se a resolução original dos vídeos. Esta escolha foi motivada pela necessidade de garantir uma boa precisão na localização dos *keypoints*, o que poderia ser comprometido com a redução da resolução das imagens.

3.3.3 Parâmetros de Treinamento

Para realizar o treinamento e a avaliação do modelo de detecção, os dados anotados foram separados em conjuntos de treinamento e validação utilizando uma proporção de 80% e 20%, respectivamente, considerada um *split* padrão na literatura (GOODFELLOW; BENGIO; COURVILLE, 2016). Essa separação resultou em 240 *frames* e 480 instâncias de animais para treinamento, e 60 *frames* com 120 instâncias para validação, para cada vídeo utilizado no processo.

Além dos ajustes relacionados ao pré-processamento dos dados, o comportamento da rede é governado por duas categorias de variáveis: os parâmetros e os hiperparâmetros. Enquanto os parâmetros são elementos internos ajustados automaticamente pelo modelo durante o aprendizado (como os pesos e vieses das camadas), os hiperparâmetros são configurações definidas previamente pelo desenvolvedor para guiar e controlar o próprio processo de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

Neste trabalho, foram definidos parâmetros estruturais relacionados à arquitetura do modelo, como o *max stride*, e hiperparâmetros que moldam o aprendizado, como o tamanho do lote (*batch size*), o número de épocas e a taxa de aprendizado (*learning rate*). Assim como as etapas de pré-processamento, alguns hiperparâmetros foram determinados por meio de experimentação e validação dos resultados durante o processo iterativo. Outros foram condicionados a limitações técnicas do ambiente computacional, como o *batch size*. Os demais parâmetros e hiperparâmetros não mencionados explicitamente foram mantidos com os valores padrões fornecidos pelo *framework*.

Devido às especificidades da arquitetura *top-down* escolhida, as configurações não foram idênticas para ambos os modelos. Além disso, as características dos diferentes conjuntos de dados também impactaram a escolha de algumas variáveis, fazendo com que os valores variassem a depender do vídeo utilizado. Os valores exatos selecionados para cada caso podem ser consultados nos Apêndices A e B para os modelos dos *datasets* SLEAP e CREAL, respectivamente.

Em relação à dinâmica do treinamento, o *batch size* define a quantidade de amostras processadas simultaneamente antes da atualização dos pesos da rede. Este foi fixado em 8

para todos os modelos, garantindo que o treinamento ocorresse na GPU disponível, sem que houvesse estouro de memória de vídeo. Essa escolha justifica-se pela especificação de hardware da placa utilizada, que conta com uma capacidade moderada de memória dedicada de 8 GB VRAM, impondo restrições ao volume de dados que podem ser alocados simultaneamente. Já uma época corresponde a uma passagem completa por todas as amostras do conjunto de treinamento, determinando quantas vezes o modelo terá a oportunidade de ajustar seus pesos com base nos dados. O limite foi definido heuristicamente em 75 épocas, uma vez que, durante o processo iterativo, observaram-se tendências de sobreajuste (*overfitting*) em execuções mais longas. O *overfitting* caracteriza-se pelo ajuste excessivo do modelo aos dados de treinamento, o que compromete sua capacidade de generalização perante novos dados (GOODFELLOW; BENGIO; COURVILLE, 2016).

Além desses parâmetros, o processo de treinamento também é influenciado por hiperparâmetros relacionados ao processo de treinamento em si. A taxa de aprendizado (*learning rate*) controla a magnitude das atualizações realizadas nos pesos da rede a cada etapa de otimização, influenciando diretamente na velocidade e a estabilidade do treinamento. Valores elevados podem impedir a convergência do modelo, enquanto valores muito baixos podem resultar em um treinamento lento ou em mínimos locais indesejados. Para evitar o *overfitting* e interromper automaticamente o treinamento quando não houver melhorias significativas, empregou-se o mecanismo de parada antecipada (*early stopping*). Nesse contexto, o parâmetro *patience* define o número máximo de épocas consecutivas sem melhora da métrica monitorada antes que o treinamento seja encerrado. Caso houvesse uma parada antecipada do treinamento, foi definido que seria salva a versão com os melhores resultados.

3.3.4 Modelo de Detecção (Centroide)

O primeiro modelo utilizado é o modelo de detecção, responsável pela identificação dos animais presentes na imagem. Durante a execução do treinamento iterativo, e como evidenciado pelos resultados apresentados no Capítulo 4 com o conjunto de referência, o modelo de detecção apresentou facilidade para aprender a identificar os animais. Portanto, alguns dos parâmetros deste modelo foram escolhidos de forma a acelerar o processo de treinamento, como o *input scaling*, e aumentar a variabilidade dos dados através de métodos de *data augmentation*. Essa decisão impactou principalmente a escolha da escala de entrada da imagem (*input scaling*) e *max stride*. O *max stride* define o fator pela qual a imagem de entrada é reduzida conforme passa pelas camadas da rede (TALMO LAB, 2026b). Juntos, esses parâmetros são usados para calcular o valor do campo receptor (*receptive field*), que controla quanto contexto o modelo observa ao redor de cada pixel (TALMO LAB, 2026c).

Durante o treinamento, foi aplicada uma variação de escala, com o intuito de aumentar a robustez do modelo a possíveis variações de tamanho dos animais. Este escalonamento foi definido entre 90% e 110% do tamanho original do vídeo, de forma a não introduzir variações muito grandes, uma vez que, considerando a distância fixa e tamanho relativamente estático dos animais, não se esperam grandes variações de tamanho.

Técnicas de *data augmentation* foram também foram aplicadas durante o treinamento para aumentar sua capacidade de generalização, optando por adicionar variação de brilho e contraste. A variação de ambos contraste e brilho foram definidas entre 50% e 200% do valor original. A variação foi aplicada aleatoriamente durante o treinamento ao conjunto todo.

3.3.5 Modelo de Estimação (Centralizado)

Após a detecção dos animais na imagem, o modelo de estimação é aplicado num recorte da imagem ao redor do animal detectado. Este recorte é definido pelo parâmetro *crop size*, que determina o tamanho da janela utilizada e pelo parâmetro *anchor part* (parte âncora), que determina onde o centroide deste recorte será posicionado (PEREIRA; TABRIS *et al.*, 2022). Além do fator do tamanho, a recomendação da documentação do *framework*, conforme (TALMO LAB, 2026d), é que o tamanho do *crop size* seja divisível pelo parâmetro *max stride*, para evitar incompatibilidades dimensionais.

Para o modelo centralizado, também deve ser definido o *max stride*, juntamente com o *input scaling*, que são usados para cálculo do *receptive field* (TALMO LAB, 2026c). Neste caso, o contexto a ser observado é o animal em questão, portanto o tamanho deste campo deve ser aproximadamente o tamanho do animal médio. Como o *input scaling* já havia sido definido como 1.0, o *max stride* foi determinado como 32 e 64 para os modelos do SLEAP e CREAL, respectivamente, de forma que o *receptive field* não fosse muito menor que o animal médio.

Para escolha do tamanho do *crop size*, foi utilizado como referência o comprimento máximo dos esqueletos do conjunto de treinamento apenas, minimizando influências do conjunto de validação nas escolhas dos parâmetros. Uma margem foi adicionada para evitar que parte do animal ficasse fora do recorte. Já para o ponto âncora, optou-se por utilizar o *torso*, uma vez que este estava posicionado próximo do centro do animal, o que ajudaria a garantir que o animal inteiro estivesse presente no recorte, mesmo em situações de variação de pose. Além disso, o *torso* se destaca por ser um *keypoint* mais visível por ser maior que os demais.

Para calcular o tamanho dos esqueletos anotados, utilizou-se a a soma das distâncias euclidianas entre os nodos conectados, sendo eles: a distância do focinho até topo da cabeça; topo da cabeça até o torso; e do torso até a base da cauda. Como demonstrado pelo Figura 9, o tamanho dos esqueletos anotados dos conjuntos de treino varia a depender do *dataset*. Para o conjunto SLEAP, o tamanho máximo de esqueleto foi de 287 pixels, com boa parte das medidas estando concentradas em torno de 175 pixels. Já para o *dataset* CREAL, a medida máxima foi de 594 pixels, porém com a maioria dos dados variando entre 350 e 500 pixels.

Portanto, considerando as medidas do esqueleto, assim como uma margem de segurança, e a necessidade do *crop size* ser um múltiplo do *max stride*, este campo foi definido como 320x320 pixels para o modelo SLEAP e 640x640 pixels para o CREAL.

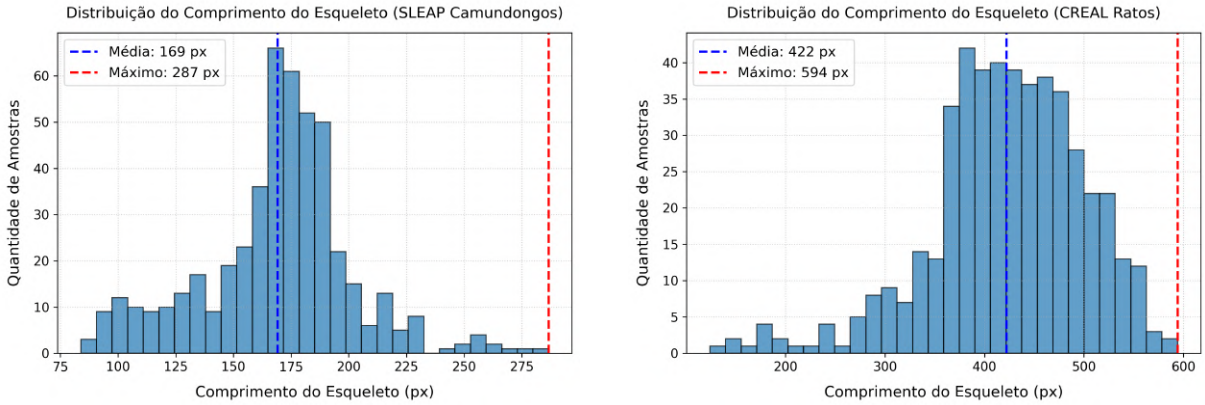
3.4 Metodologia de Avaliação

Uma vez treinados, os modelos de detecção e estimação de pose precisam ser avaliados para verificar a qualidade de suas predições. Para isso, são utilizadas métricas específicas para cada modelo, que levam em consideração as particularidades de cada um deles.

3.4.1 Modelo de Detecção (Centroide)

Para avaliação do resultado de detecção foram utilizadas as métricas de *precision* e *recall*. Estas avaliam, respectivamente, a proporção de detecções classificadas corretamente

Figura 9: *Distribuição de tamanhos das anotações dos conjuntos de treinamento.*



(a) *dataset SLEAP.* **(b)** *dataset CREAL.*

Fonte: Autor.

e a capacidade do modelo de identificar todos os animais presentes nas imagens.

A métrica *precision* representa a razão entre o número de verdadeiros positivos, *true positives* (TP) e o total de predições positivas realizadas pelo modelo, incluindo os falsos positivos, *false positives* (FP), conforme demonstrado na Equação 3:

$$precision = \frac{TP}{TP + FP} \tag{3}$$

A métrica *recall* representa a razão entre o número de verdadeiros positivos e o total de exemplos positivos reais, incluindo os falsos negativos, *false negatives* (FN), conforme demonstrado na Equação 4:

$$recall = \frac{TP}{TP + FN} \tag{4}$$

Para o contexto de detecção, é necessário definir o que constitui um acerto e um erro. No fluxo de funcionamento do SLEAP, o acerto é calculado com base na distância euclidiana entre o centroide, ou ponto âncora, da referência e da predição, de forma que esta distância deve estar abaixo de um limiar parametrizado (definido como 50 pixels, conforme sugerido por (TALMO LAB, 2026a)). Os erros de detecção abrangem uma variedade maior de casos, englobando situações em que não houve detecção onde devia haver (FN) e detecções acima do limiar definido (FP).

3.4.2 Modelo de Estimação (Centralizado)

Diferentemente do modelo de detecção, o modelo de estimação de pose é avaliado com base na proximidade da predição do dado real e visibilidade dos *keypoints*, ao invés da detecção de indivíduos. Dessa forma, as métricas de *precision* e *recall* não são empregadas para avaliar a qualidade deste, apesar de ainda ser possível avaliar se houve ou não detecção de cada um dos *keypoints*.

Portanto, para compreensão da qualidade do modelo de estimação de pose, são empregadas métricas específicas deste que levam em consideração a distância euclidiana entre os *keypoints* preditos e os *keypoints* de referência, a escala do objeto e a visibilidade

dos pontos. A distância euclidiana pode ser utilizada para quantificar o erro de localização dos *keypoints* preditos em relação aos *keypoints* de referência (MATHIS *et al.*, 2018), sendo calculada conforme a Equação 5, onde (x_p, y_p) são as coordenadas do ponto predito e (x_r, y_r) são as coordenadas do ponto de referência.

$$d = \sqrt{(x_p - x_r)^2 + (y_p - y_r)^2} \tag{5}$$

A partir da medida de distância, é possível avaliar a precisão do modelo através da métrica *Percentage of Correct Keypoints* (PCK), que quantifica a proporção de *keypoints* preditos dentro de um limiar de distância em relação às anotações de referência (ZAIDI; WAGAN, 2024; YANG; RAMANAN, 2013). O limiar é ajustado levando em consideração o contexto da aplicação e o tamanho dos objetos avaliados, podendo ser aumentado ou reduzido para diferentes níveis de precisão. A métrica PCK é calculada conforme a Equação 6, onde d_i é a distância entre o ponto predito e o ponto de referência, t é o limiar de distância e N é o número total de instâncias.

$$\text{PCK} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(d_i < t) \tag{6}$$

Aliada a estas métricas, está a *Object Keypoint Similarity* (OKS), que mede a similaridade entre as poses predita e de referência considerando a distância entre *keypoints*, a escala do objeto e a incerteza de anotação dos pontos (PEREIRA; TABRIS *et al.*, 2022; RONCHI; PERONA, 2017). A métrica OKS é calculada conforme demonstrado na Equação 7, onde d_i é a distância entre o ponto predito e o ponto de referência, s é a escala do objeto, k_i é um parâmetro de incerteza específico para cada *keypoint*, v_i é a visibilidade do *keypoint* i e $\delta(v_i)$ é uma função indicadora da visibilidade desse. Ressalta-se que o valor utilizado de k_i foi de 0,025 para todos os *keypoints*, conforme padrão do *framework*. Esse valor foi estipulado com base na incerteza de anotação de olhos em *datasets* com humanos (PEREIRA; TABRIS *et al.*, 2022), podendo ser investigado com mais detalhes em trabalhos futuros, uma vez que, como demonstrado por (SEGALIN *et al.*, 2021b), alguns *keypoints* apresentam maior variabilidade na anotação.

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i)}{\sum_i \delta(v_i)} \tag{7}$$

O valor obtido pela métrica OKS varia entre 0 e 1, onde valores mais próximos de 1 indicam maior similaridade entre as poses predita e de referência e valores mais próximos de 0 indicam menor similaridade. Esta métrica pode ser expandida para o *mean* OKS (mOKS), que calcula a média do OKS para todas as poses estimadas, gerando um valor final de similaridade para o modelo.

3.4.3 Análises Pontuais

Além das métricas, é importante observar as situações em que o modelo mais apresentou dificuldade para estimação. Entender o contexto desses erros pode providenciar informações do porque o modelo teve dificuldade para predizê-las. Dessa forma, torna-se necessário obter os erros *outliers*, que podem ser calculados com base no intervalo entre quartis, *Interquartile Range* (IQR) (SMITI, 2020). As Equação 8 determina, a partir da distribuição normal dos dados, o intervalo entre os quartis Q_1 e Q_3, que representam os

dados mais próximos da mediana. A partir desses, é possível calcular os *outliers* ($\mathcal{O}$) a partir de um limite superior e inferior dos quartis, conforme Equação 9. No contexto deste trabalho o conjunto D representa as distâncias entre pontos anotados e preditos, sendo d uma distância pertencente a este conjunto.

$$\text{IQR} = Q_3 - Q_1 \tag{8}$$

$$\mathcal{O} = \{d \in D \mid d < Q_1 - 1.5 \cdot \text{IQR} \ \vee \ d > Q_3 + 1.5 \cdot \text{IQR}\} \tag{9}$$

4 RESULTADOS E DISCUSSÃO

Os experimentos realizados tiveram como objetivo avaliar o desempenho dos modelos treinados em diferentes conjuntos de dados e consequentemente a viabilidade de implementação da estimação de pose com ratos em ambiente laboratorial. Para isso, os modelos foram analisados utilizando-se as métricas relacionadas à detecção dos indivíduos e à precisão da localização dos *keypoints*. Os dados disponibilizados pelo *framework* SLEAP serviram como base para o treinamento e avaliação dos modelos, uma vez que constituem um cenário já avaliado de referência.

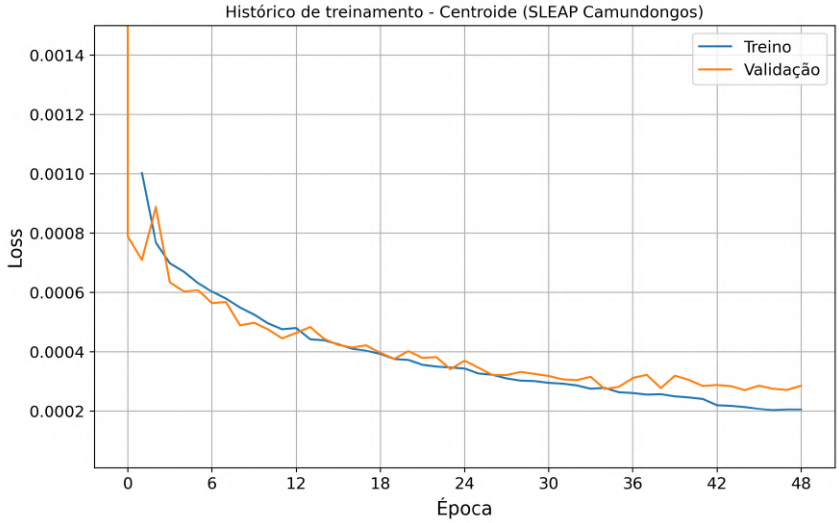
4.1 Resultados com o *dataset* de referência

O processo de treinamento utilizando o *dataset* de referência apresentou resultados satisfatórios tanto para a etapa de detecção quanto para a etapa de estimação de pose. Os modelos foram avaliados utilizando o conjunto de validação separado previamente. Para este caso, não foi utilizado um conjunto de teste, uma vez que esse *dataset* serviu como referência para o desenvolvimento e avaliação.

Durante o treinamento, o modelo de detecção apresentou uma curva de *loss* decrescente ao longo das épocas, porém, apresentando evidências de *overfitting* ao final, como evidenciado pela separação entre as curvas de treinamento e validação na Figura 10. Observa-se que nas épocas finais, a *loss* do treinamento continua reduzindo enquanto a da validação permanece relativamente estática, evidenciando um comportamento de *overfitting*. Isso sugere que o modelo podia estar se ajustando excessivamente aos dados de treino, podendo comprometer sua capacidade de generalização. Por isso, o treinamento foi interrompido antes que esse comportamento se agravasse.

Os resultados para *precision* e *recall* para este modelo de detecção, apresentados na Tabela 1, indicam alta capacidade de detecção para o modelo. A taxa de acertos elevada chama atenção devido à pequena quantidade de dados utilizados. Esse resultado pode ser justificado pelo fato do conjunto de dados utilizado estar contido no mesmo domínio, apresentando fundo relativamente estático, iluminação controlada e ambos os animais de mesma coloração e tamanho aparente, além do alto contraste entre animal e fundo. Outro fator importante é a tendência do modelo a *overfitting*, que considerando a similaridade entre os conjuntos de treino e validação pode implicar em métricas mais elevadas. Esses resultados, apesar disso, estão alinhados com os valores apresentados por (pereira; tabris *et al.*, 2022). Outro fator que influenciou os altos valores para esta métrica foi o limiar definido para uma detecção correta de 50 pixels. Esse limiar é considerado tolerante para a escala dos animais avaliados, pois aceita grandes distâncias de detecção. No conjunto analisado, o limite de 50 pixels representa cerca de 30% do tamanho médio do animal. A variação deste patamar não foi avaliada no cenário de referência, dado que o objetivo

Figura 10: *Loss do modelo centroide durante o treinamento (dataset SLEAP).*



Fonte: Autor.

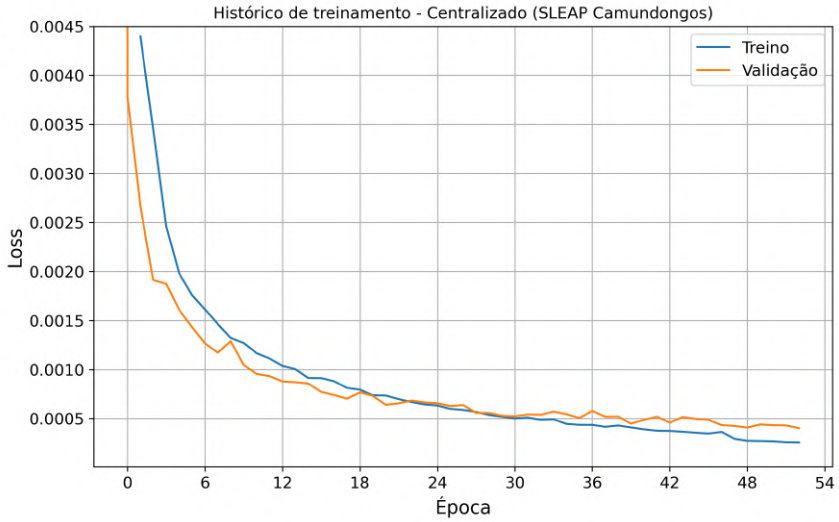
Tabela 1: *Precision e Recall do modelo de detecção(dataset SLEAP).*

Precision	Recall
0,99	0,96

principal da pesquisa restringiu-se a validar a viabilidade de aplicação no contexto do CREAL.

Na etapa de estimação de pose, o modelo precisou ser treinado por um número maior de épocas, sendo novamente interrompido antecipadamente devido à evidência de *overfitting*, como apresentado na Figura 11.

Figura 11: *Loss do modelo centralizado durante o treinamento (dataset SLEAP).*

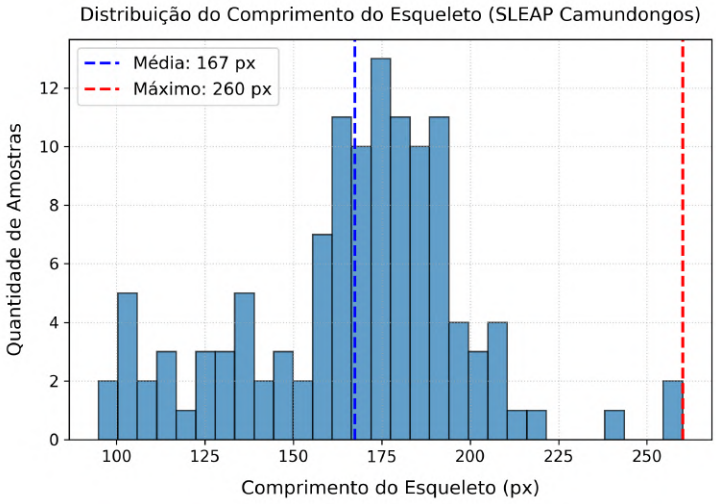


Fonte: Autor.

Para análise dos resultados de estimação, é necessário observar a distribuição do

tamanho dos esqueletos das anotações, apresentados na Figura 12, que foram determinados a partir da soma da distância entre *keypoints*. Neste caso, os esqueletos foram restritos apenas para o conjunto de validação, uma vez que é em cima dele que são analisados os resultados. A partir do histograma, é possível observar que uma parte considerável das anotações esta contida num tamanho entre 160 e 200 pixels. As instâncias com menor tamanho do esqueleto são geradas principalmente por poses em que o animal está na vertical.

Figura 12: *Distribuição de tamanhos das anotações do conjunto de validação (dataset SLEAP).*



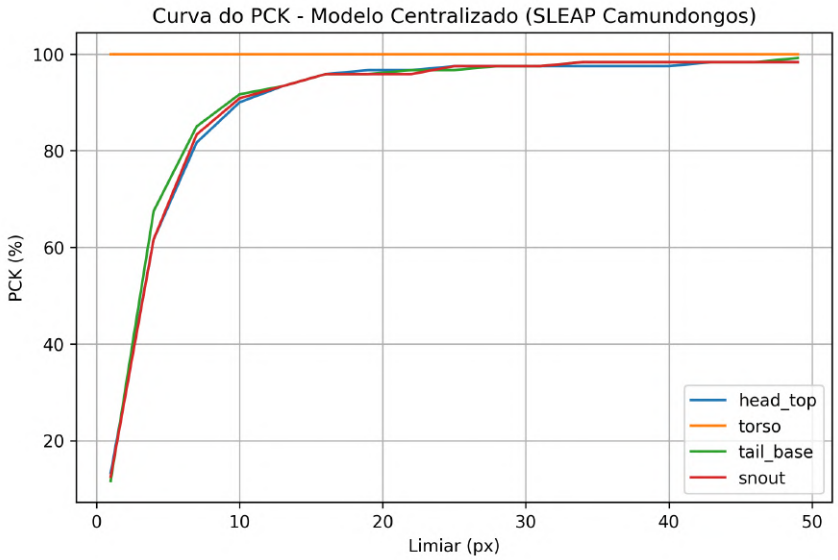
Fonte: Autor.

Com base na informação de tamanho dos animais, é possível contextualizar os resultados apresentados na Figura 13, onde o percentual de predições corretas é apresentado em função da variação do limiar do PCK. Observa-se que, para um limiar de 10 pixels, o modelo conseguiu localizar corretamente aproximadamente 90% dos *keypoints* presentes nas imagens de validação, com excessão do torso, por ser o *keypoint* âncora, para o qual ele teve 100% de acerto. Este resultado para o torso é esperado no conjunto de validação devido à forma como o SLEAP treina os modelos. O modelo de estimação treinado usa os recortes feitos em cima dos dados reais, que são centralizados no *torso*, fazendo com que o erro do *torso* seja nulo.

Os resultados apresentados na curva do PCK indicam que a maioria dos *keypoints* preditos estão localizados a uma distância inferior a 10 pixels dos *keypoints* de referência. Quando colocados em perspectiva do tamanho médio de 167 pixels dos animais, este resultado se apresenta satisfatório, considerando que equivale à 6% do tamanho do animal.

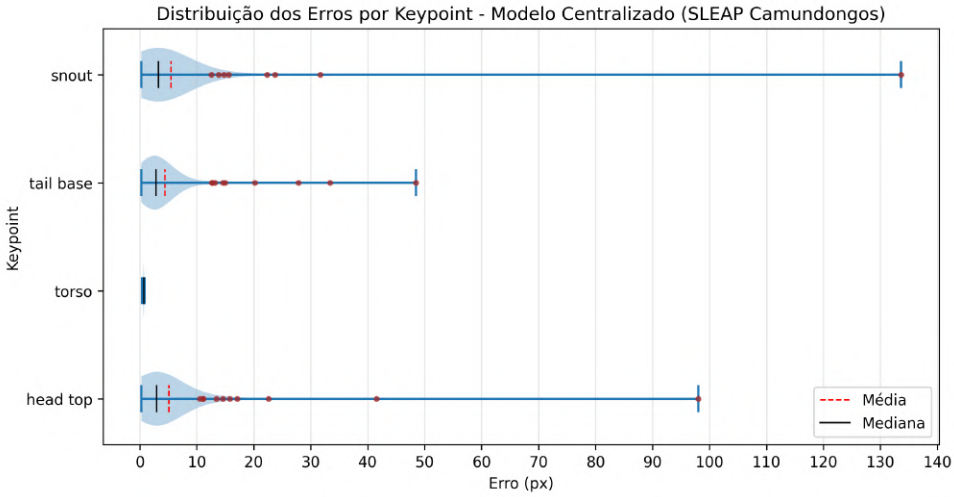
Uma demonstração mais detalhada dos erros para cada *keypoint* é apresentada na Figura 14 nos gráficos de violino. A distribuição da distância dos erros indicam que, não apenas a mediana, mas também a maioria dos erros de localização estão numa faixa abaixo de 10 pixels, o que é consistente com os resultados apresentados na curva do PCK.

Figura 13: *Percentual de predições corretas com base na variação do limiar do PCK (dataset SLEAP).*



Fonte: Autor.

Figura 14: *Distribuição do erro para cada keypoint (dataset SLEAP).*



Fonte: Autor.

Conforme evidenciado nos gráficos de violino, foram identificados *outliers* na distribuição das distâncias de erro. Embora os resultados gerais sejam positivos, a presença desses valores extremos indica a ocorrência de erros elevados em situações específicas. Tais casos sugerem cenários nos quais o modelo apresentou maior dificuldade para localizar corretamente determinados *keypoints*, possivelmente devido a fatores como oclusões, poses mais complexas ou semelhanças visuais entre diferentes regiões anatômicas dos animais.

Uma avaliação global do desempenho da estimação de pose pode ser obtida por meio do mOKS, cujo valor foi de 0,7467 para o conjunto de validação. Esse resultado indica um desempenho geral satisfatório, refletindo a capacidade do modelo de localizar os *keypoints* com precisão razoável na maioria das amostras. Entretanto, esse valor é influenciado pela elevada precisão observada para o *torso*, utilizado como ponto de referência durante o processo de centralização das imagens, o que pode mascarar erros mais expressivos em outros

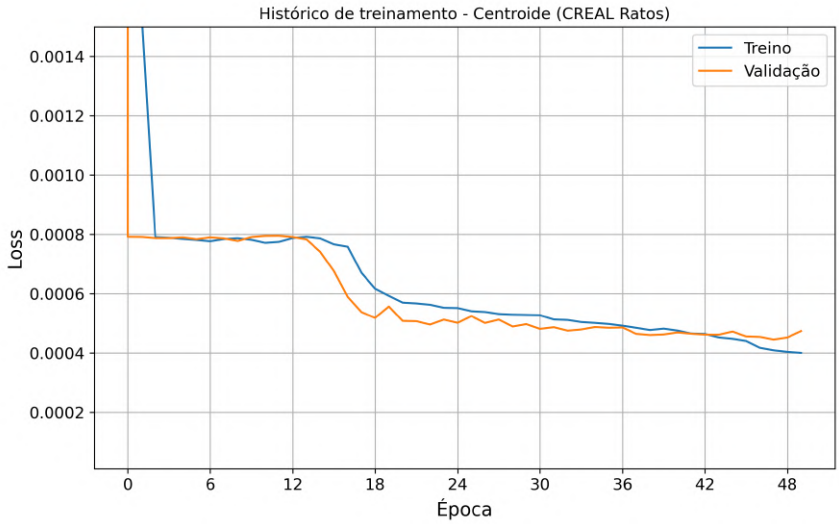
keypoints. Adicionalmente, a ocorrência de falhas catastróficas ocasionais, evidenciada pela presença dos *outliers*, contribui para explicar a discrepância entre a métrica global e a distribuição dos erros observados individualmente.

4.2 Resultados com o *dataset* obtido em laboratório

Os modelos treinados com o *dataset* obtido em ambiente laboratorial apresentaram comportamento semelhante entre si, embora com desempenho inferior ao observado no modelo treinado com o *dataset* de referência. A seguir, são apresentados os resultados referentes à detecção dos indivíduos e à precisão na localização dos *keypoints* para o *dataset* CREAL.

Assim como o modelo de detecção treinado com o *dataset* de referência, o modelo treinado com o *dataset* CREAL apresentou uma curva de *loss* decrescente ao longo do treinamento. Porém, evidências de *overfitting* foram observadas após a época 42, como indicado na Figura 15, causando o interrupção antecipada do treinamento para evitar ajustes excessivos aos dados de treinamento. Outro fenômeno observado durante o treinamento foi a trava inicial que durou até a época 12, indicando uma dificuldade do modelo para convergir em épocas iniciais.

Figura 15: *Loss do modelo centroide durante o treinamento (dataset CREAL).*



Fonte: Autor.

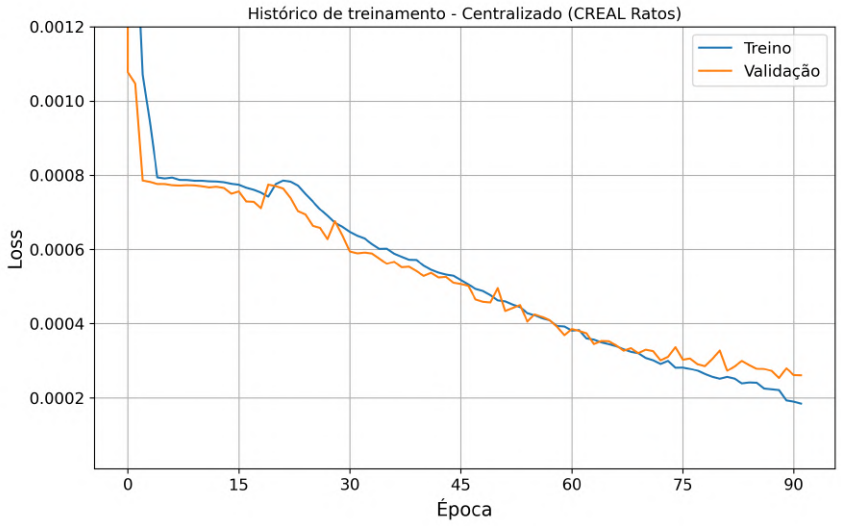
Destaca-se que, em comparação ao resultado de referência, o modelo de detecção para o *dataset* CREAL necessitou de um número maior de épocas para treinamento. Mesmo assim ele não obteve os mesmos resultados, como demonstrado na Tabela 2. Os valores de *precision* e *recall* obtidos indicam que o modelo apresentou dificuldade em detectar os animais. Esse resultado, aliado ao cenário mais complexo devido a presença das grades e da serragem e à redução da imagem de entrada, contribuem para uma queda de performance do modelo neste *dataset*. Além disso, deve-se levar em consideração a métrica utilizada para calcular o acerto na detecção. Para o conjunto de referência, os 50 pixels representavam uma proporção bem maior do animal do que para este conjunto, o que significa uma medida mais restritiva para o *dataset* do CREAL.

Tabela 2: *Precision e Recall do modelo de detecção (dataset CREAL).*

Precision	Recall
0,675	0,64

Assim como os demais modelos, o modelo centralizado também apresentou uma curva de *loss* decrescente durante o treinamento, porém, com evidências de *overfitting* mais ao final, como demonstrado na Figura 16. O modelo de estimação de pose para o *dataset* CREAL também necessitou de um número maior de épocas, e mesmo assim não obteve resultados tão bons quanto o treinado com o *dataset* de referência. Nota-se novamente a presença de um platô nas épocas iniciais do treinamento, reforçando a dificuldade de convergência do modelo.

Figura 16: *Loss do modelo centralizado durante o treinamento (dataset CREAL).*



Fonte: Autor.

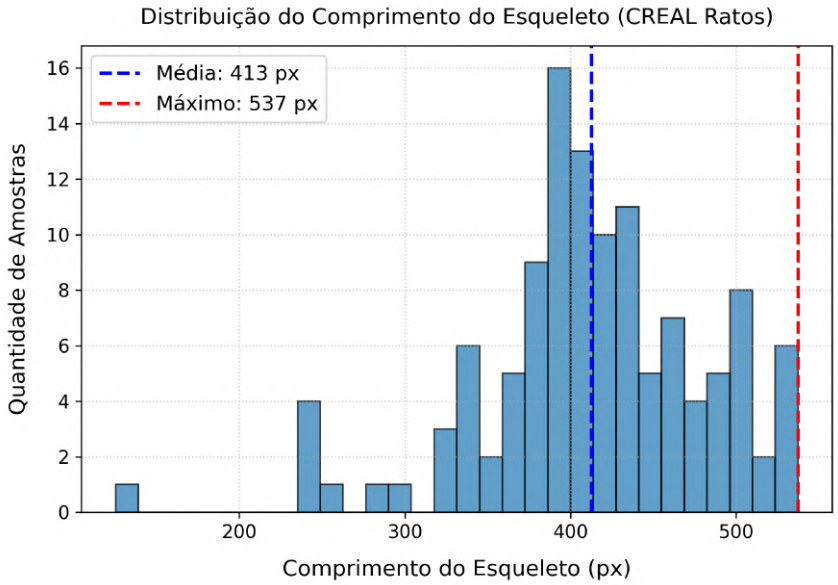
Devido às diferenças de domínio entre os conjuntos de dados, torna-se necessário reavaliar o contexto dos animais para uma interpretação adequada das métricas obtidas. A Figura 17 apresenta a distribuição do tamanho dos esqueletos no conjunto de validação. Observa-se que, diferentemente do *dataset* de referência, os animais do *dataset* CREAL apresentam esqueletos consideravelmente maiores em termos de número de pixels, concentrando-se predominantemente na faixa entre 350 e 500 pixels.

Esse comportamento pode ser explicado não apenas por variações no tamanho físico dos animais, mas também por fatores experimentais, como a distância entre a câmera e os indivíduos e a resolução das imagens adquiridas. Essas diferenças de escala podem influenciar diretamente o desempenho do modelo, especialmente em tarefas de estimação de pose sensíveis à variação de tamanho, além de impactar na interpretação das métricas.

A partir da distribuição de tamanhos das anotações, observa-se o percentual de predições corretas em função da variação do limiar do PCK, conforme apresentado na Figura 18. Para um limiar de 10 pixels, o modelo foi capaz de localizar corretamente entre 40% e 50% dos *keypoints* presentes nas imagens de validação, com exceção do *torso*, utilizado como âncora.

Para se atingir uma taxa de acerto equivalente à obtida pelo modelo treinado com

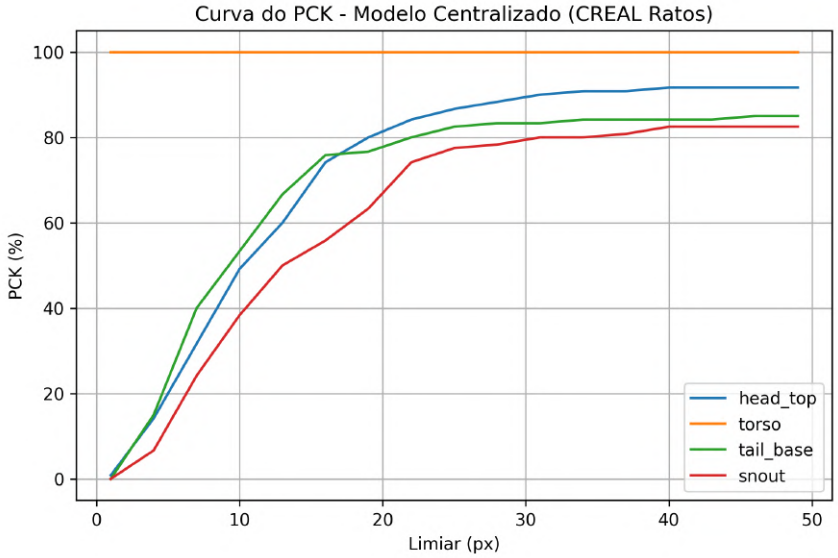
Figura 17: *Distribuição de tamanhos das anotações do conjunto de validação(dataset CREAL).*



Fonte: Autor.

o *dataset* de referência, seria necessário utilizar um limiar de aproximadamente 30 pixels, o que corresponde ao triplo do valor observado no modelo de referência.

Figura 18: *Percentual de predições corretas com base na variação do limiar do PCK (dataset CREAL).*



Fonte: Autor.

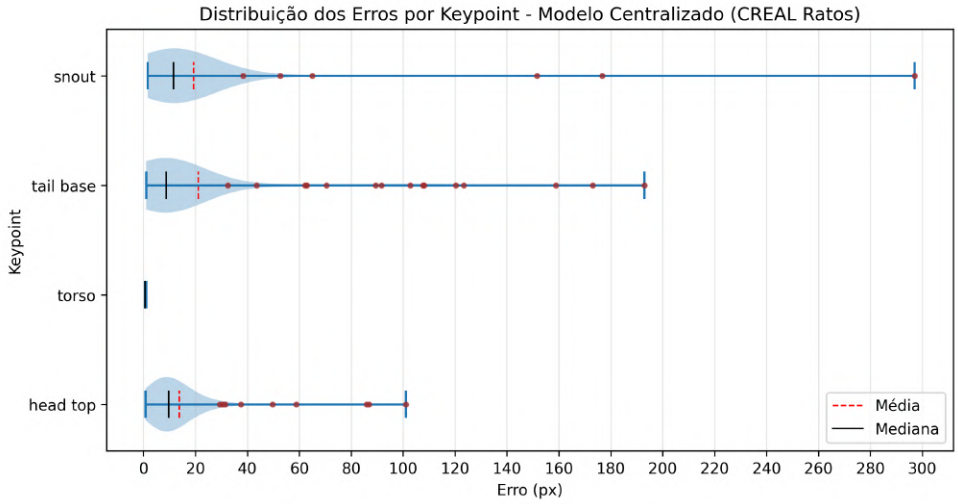
Esse resultado indica uma redução de desempenho em relação ao conjunto de referência, o qual apresentou aproximadamente o dobro de acertos para o limiar de 10 pixels. Observa-se, portanto, a importância de considerar as diferenças de domínio dos dados para uma interpretação adequada dos resultados.

Ao analisar a distribuição do tamanho dos esqueletos, verifica-se que os animais do *dataset* CREAL apresentam, em média, esqueletos aproximadamente 2,5 vezes maiores

do que aqueles do *dataset* de referência. Esse fator auxilia na interpretação das métricas, uma vez que um limiar de 10 pixels no conjunto de referência representa uma fração relativamente maior do tamanho do esqueleto do que o mesmo limiar aplicado ao *dataset* CREAL, configurando, portanto, uma condição mais restritiva neste último caso. Dessa forma, apesar da redução observada em relação ao modelo de referência, os resultados obtidos para o *dataset* CREAL ainda podem ser considerados satisfatórios, visto que a maioria dos *keypoints* preditos apresenta erro inferior a 10% do tamanho do esqueleto dos animais.

Adicionalmente, analisando-se a distribuição dos erros por *keypoint*, apresentada na Figura 19, observa-se que, embora um número menor de *keypoints* esteja abaixo do limiar de 10 pixels, a maior parte dos erros permanece inferior a 40 pixels. O *keypoint head top* se destaca em relação aos demais por apresentar quase todos seus erros concentrados numa faixa abaixo de 20 pixels. Os *outliers* destacados pelos pontos coloridos auxiliam na interpretação dos resultados, ao indicar que o conjunto em sua maioria apresenta um erro aceitável, com excessão de algumas predições catastróficas.

Figura 19: *Distribuição do erro para cada keypoint (dataset CREAL).*



Fonte: Autor.

Um fator que contribui para o erro mais elevado dos *keypoints* deste *dataset* é a presença de grades na gaiola, que podem gerar confusão visual para o modelo na hora de realizar a estimação da pose (DUTTA *et al.*, 2025). Embora neste trabalho não tenham sido empregadas técnicas para lidar com esse problema, existem abordagens voltadas à remoção dessas estruturas antes da estimação de pose. Entre elas destaca-se o UnCageNet (DUTTA *et al.*, 2025), que utiliza uma etapa de pré-processamento para reconstruir a imagem do animal sem a interferência visual das grades, contribuindo para melhorar a qualidade da estimação.

Entretanto, o objetivo desta pesquisa foi avaliar a viabilidade da estimação de pose utilizando imagens obtidas diretamente do ambiente de experimentação, sem modificações no cenário ou aplicação de técnicas adicionais de pré-processamento. Dessa forma, optou-se por manter as grades presentes nas imagens, permitindo que os resultados refletissem as dificuldades encontradas em uma situação real de aquisição.

Os resultados do conjunto de validação do *CREAL* podem ser sumarizados pelo resultado do mOKS, que alcançou 0,6468 para tal. De forma similar ao restante dos

resultados, isso representa uma queda significativa em relação ao modelo treinado em cima dos dados do SLEAP.

4.3 Resultados do conjunto de teste

A fim de validar a capacidade de generalização do modelo para ambientes similares ao do *dataset* de treino, foi realizada uma avaliação utilizando o conjunto de teste separado. Embora similar, este conjunto apresentava apenas um indivíduo da mesma espécie e um cenário com obstrução devido a presença de alimentos na gaiola.

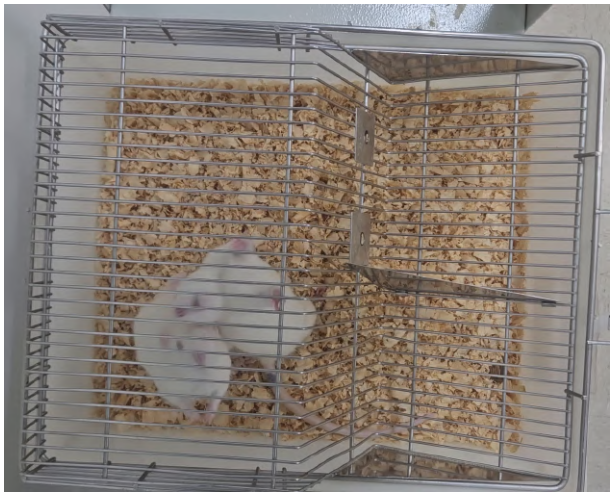
O modelo de detecção, avaliado sobre o conjunto de teste, apresentou valores de *precision* e *recall* superiores aos observados no conjunto de validação, conforme apresentado na Tabela 3. Esse resultado é, à primeira vista, inesperado, uma vez que o conjunto de teste pertence a um domínio relativamente distinto daquele utilizado durante o treinamento.

Tabela 3: *Precision e Recall do modelo de detecção (dataset CREAL - teste).*

Precision	Recall
0,775	0,79

Entretanto, uma diferença importante entre os dois conjuntos deve ser considerada. Enquanto o conjunto de validação continha cenas com dois indivíduos, o conjunto de teste apresentava apenas um animal por imagem. Dessa forma, é possível que a presença simultânea de múltiplos indivíduos tenha aumentado a complexidade da tarefa de detecção, especialmente em situações de proximidade ou sobreposição parcial entre os animais. Essa hipótese é consistente com os resultados obtidos. Os valores semelhantes de *precision* sugerem que quando uma detecção é realizada ela corresponde ao animal na maioria dos casos. Por outro lado, o aumento observado no *recall* pode indicar que a presença de apenas um indivíduo no conjunto de teste reduz ambiguidades durante a detecção, tornando menos provável que o modelo deixe de identificar o animal presente. Um exemplo desse tipo de situação é apresentado na Figura 20, na qual a sobreposição parcial entre os indivíduos dificulta a distinção entre as instâncias.

Figura 20: *Exemplo de sobreposição dos animais (dataset CREAL).*

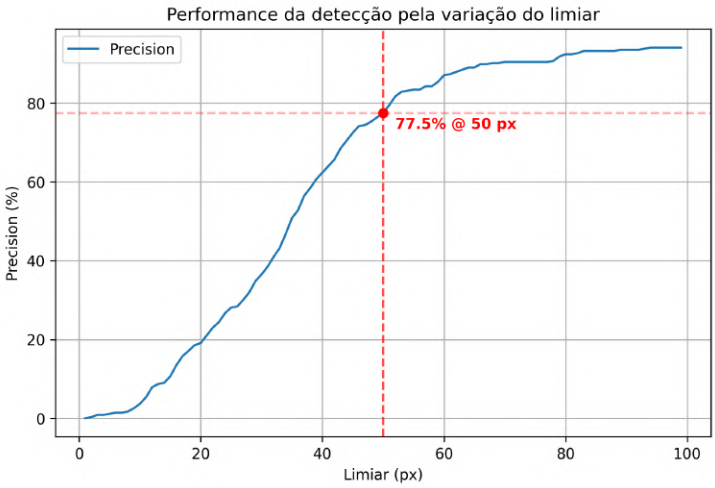


Fonte: Autor.

Além da maior dificuldade para a detecção, a presença simultânea de múltiplos indivíduos também impõe desafios relacionados ao rastreamento da identidade dos animais ao longo do tempo. Em cenários com frequentes interações e sobreposições, torna-se necessário empregar estratégias capazes de preservar a associação entre as diferentes instâncias ao longo do tempo. Embora essa questão não tenha sido um problema a ser resolvido neste trabalho, ela representa um desafio importante para aplicações futuras. O próprio SLEAP oferece suporte a rastreamento de identidades (PEREIRA; TABRIS *et al.*, 2022), mas existem outras abordagens que fazem uso de algoritmos como o *Hungarian algorithm* (GIRDHAR *et al.*, 2018) para associar instâncias ao longo do vídeo.

Apesar dos valores calculados apresentarem uma ideia geral do resultado, o impacto do limiar de detecção também deve ser observado. Na Figura 21 observa-se o comportamento da métrica de *precision* conforme é variado o limite máximo de distância entre o torso real e predito do animal. Esse resultado indica que o modelo não deixou de posicionar o *keypoint* do torso na imagem, mas sim que falhou em posicioná-lo de forma precisa.

Figura 21: *Precision do modelo pela variação do limiar de aceite (dataset CREAL - teste).*

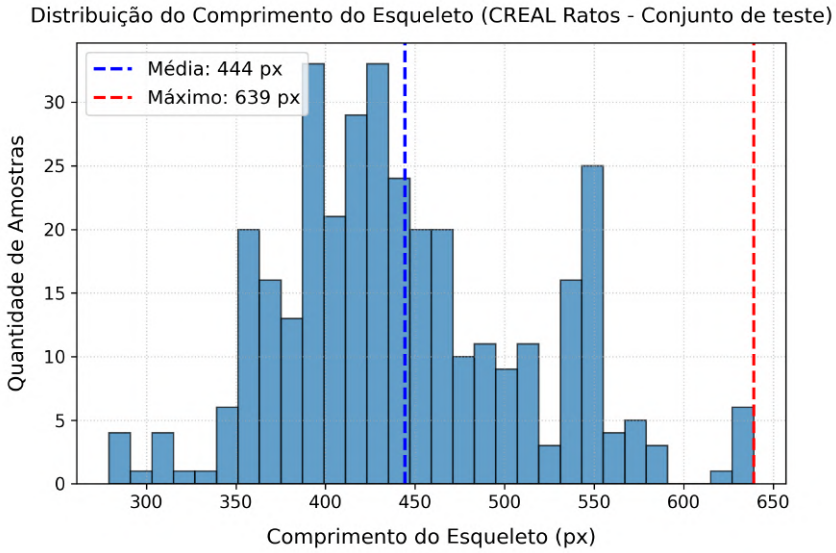


Fonte: Autor.

Para análise dos resultados de estimação, uma análise da distribuição do tamanho dos esqueletos anotados no conjunto de teste (Figura 22) torna-se necessária devido à mudança do indivíduo avaliado em relação aos conjuntos utilizados durante o treinamento e a validação. Observa-se que os esqueletos presentes no conjunto de teste apresentam dimensões superiores às observadas no conjunto de validação, com tamanho médio de 444 pixels, representando um aumento de aproximadamente 7,5%. Essa diferença de escala deve ser considerada na interpretação das métricas de desempenho, uma vez que influencia a relação entre os erros de localização e o tamanho do animal nas imagens. Apesar disso, observando-se as métricas de estimação de pose, o modelo apresentou uma queda em relação ao conjunto de validação, como apresentado na curva do PCK da Figura 23. O resultado anterior apresentava uma taxa de acertos próxima de 80% para um limiar de 20 pixels para todos *keypoints*, com excessão do torso. Neste conjunto, esses mesmos *keypoints*, só alcançaram um resultado semelhante a partir dos 30 pixels de erro.

Um aspecto que merece destaque é a degradação observada no desempenho associado ao *torso*. Diferentemente do comportamento observado no conjunto de validação, o PCK desse *keypoint* permaneceu relativamente baixo ao longo de grande parte da curva, atingindo

Figura 22: *Distribuição de tamanhos das anotações do conjunto de teste (dataset CREAL).*



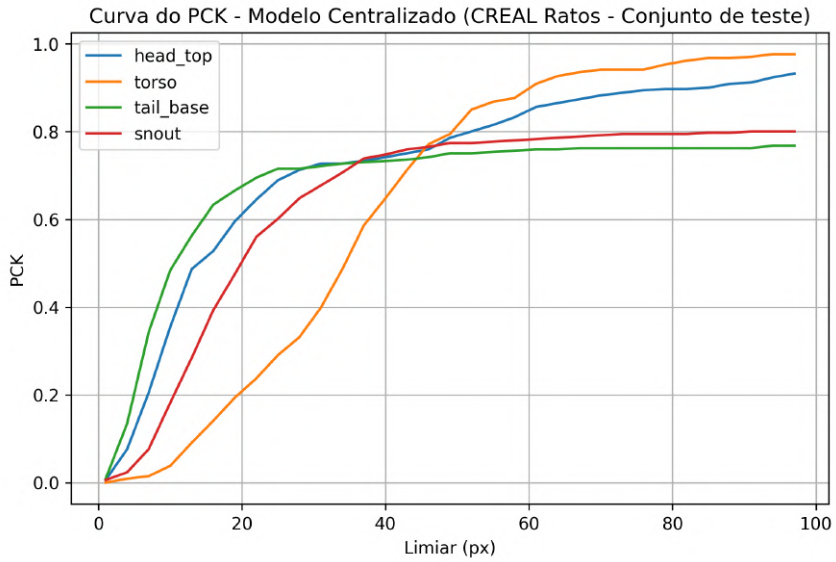
Fonte: Autor.

valores elevados apenas para limiares superiores a 40 pixels. Quando a análise é estendida para limiares de até 100 pixels, observa-se um aumento acentuado do PCK do *torso*. Esse comportamento, em conjunto com os resultados obtidos na etapa de detecção, sugere que o modelo frequentemente é capaz de detectar corretamente a presença do animal, porém posiciona o *keypoint* do *torso* a uma distância considerável de sua localização real. Tal interpretação é reforçada pelo valor de *precision* calculado para um limiar de 50 pixels. Nessa condição, o modelo localizou corretamente 77,5% dos *keypoints* do *torso*, valor idêntico à taxa de *precision* obtida na tarefa de detecção. Esse resultado sugere que uma parcela significativa dos erros observados na estimação de pose está associada à imprecisão na localização do centro do animal, e não necessariamente à falha na sua detecção.

Os resultados da Figura 23 evidenciam a importância da escolha do *keypoint* utilizado como âncora na etapa de detecção. Embora o *torso* corresponda a uma região mais extensa e, a princípio, mais fácil de ser detectada, a ausência de referências anatômicas pode dificultar o posicionamento preciso desse *keypoint*. Conforme discutido por (SEGALIN *et al.*, 2021b), determinadas regiões corporais apresentam maior ambiguidade de localização, resultando em maior variabilidade entre anotadores humanos. Em contraste, regiões associadas a características visuais mais distintas tendem a apresentar anotações mais consistentes, o que pode explicar o melhor desempenho observado.

Os dois cenários apresentados na Figura 24 exemplificam a ambiguidade mencionada anteriormente. Em ambos os casos, observam-se erros classificados como *outliers* para o *torso*, evidenciando como a ausência de referenciais visuais e determinadas posturas do animal podem dificultar sua localização precisa. Observa-se ainda que, no cenário ilustrado pela Figura 24a, apesar do erro mais elevado na detecção do *torso*, os demais *keypoints* foram estimados com boa precisão. Esse resultado sugere que uma falha na localização do ponto âncora não implica necessariamente em uma falha generalizada na estimação de pose. Além disso, os erros de localização observados foram de 81 e 78 pixels, respectivamente, valores superiores ao limiar de 50 pixels adotado. Dessa forma, ambos os casos são contabilizados como FP, embora as predições permaneçam visualmente coerentes e próximas da região esperada. Esse comportamento evidencia a influência da

Figura 23: *Percentual de predições corretas com base na variação extendida do limiar do PCK (conjunto de teste - CREAL).*



Fonte: Autor.

escolha do limiar de avaliação sobre os resultados obtidos e reforça a importância de sua definição adequada, uma vez que pequenas variações nesse parâmetro podem alterar significativamente a interpretação do desempenho do modelo.

Figura 24: *Erros Outliers do torso (vermelho: predição, verde: anotação).*



(a) *Frame 929 - erro: 81 px.* **(b)** *Frame 5661 - erro: 78 px.*

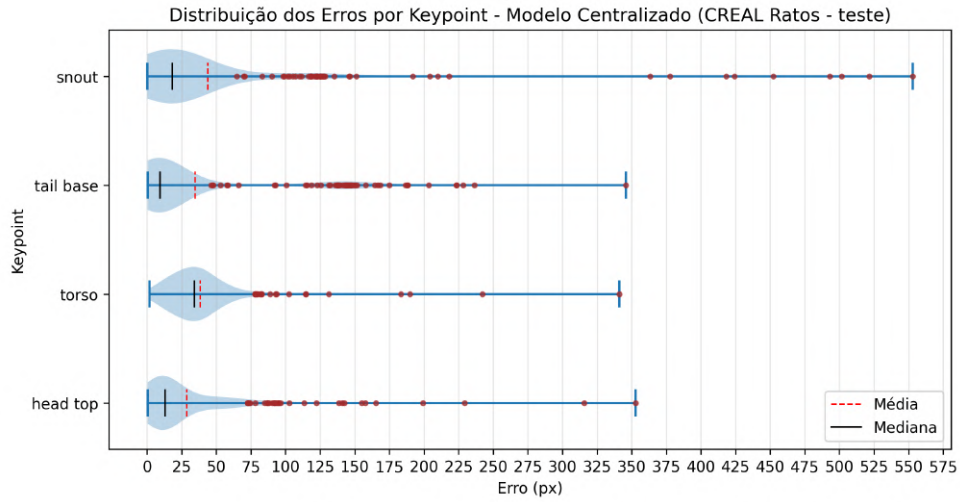
Fonte: Autor.

Ao analisar a distribuição dos erros apresentada na Figura 25, observa-se que o *keypoint head top* apresentou uma distribuição mais concentrada em erros reduzidos, indicando maior consistência em sua localização. Esse comportamento sugere que a presença de características visuais na região da cabeça, como os olhos e orelhas, fornecem referências mais estáveis para o processo de anotação e para o aprendizado do modelo, podendo ser uma alternativa para âncora do modelo de detecção. Em contraste, o *torso* corresponde a uma

região mais homogênea visualmente, o que pode contribuir para uma maior variabilidade em sua localização.

Apesar do melhor resultado de estimação do *head top*, a troca do *keypoint* âncora não pode ser decidida em cima deste dado apenas e deve ser avaliada com cautela, uma vez que essa escolha impacta diretamente o parâmetro de *crop size* utilizado pelo modelo de estimação de pose. O aumento deste parâmetro altera a região da imagem observada pelo modelo após a etapa de centralização, o que pode influenciar o contexto disponível ao redor do animal, a capacidade de localizar corretamente os demais *keypoints* e o custo computacional exigido para treinamento.

Figura 25: *Distribuição do erro para cada keypoint (conjunto de teste - CREAL).*



Fonte: Autor.

Outro fator importante a ser observado na Figura 25 é erro apresentado pelo *snout*, que apresenta *outliers* com maiores erros que os demais *keypoints*. Este pode ser explicado pela dificuldade de diferenciar visualmente esta parte do animal da serragem presente no chão, além de ser obstruída em situações em que os animais estão próximos um do outro, dormindo ou demais posições em que o focinho não esteja visível, como demonstrado na Figura 26. Esse problema evidencia a dificuldade na escolha das partes a serem utilizadas como *keypoints*. Referências visuais insuficientes podem comprometer não apenas a capacidade do modelo, mas também as anotações realizadas por humanos(o que, em consequência, afeta negativamente o modelo). Dessa forma, a escolha de partes do corpo que apresentem uma referência visual clara e consistente pode ser fundamental para o desempenho do modelo.

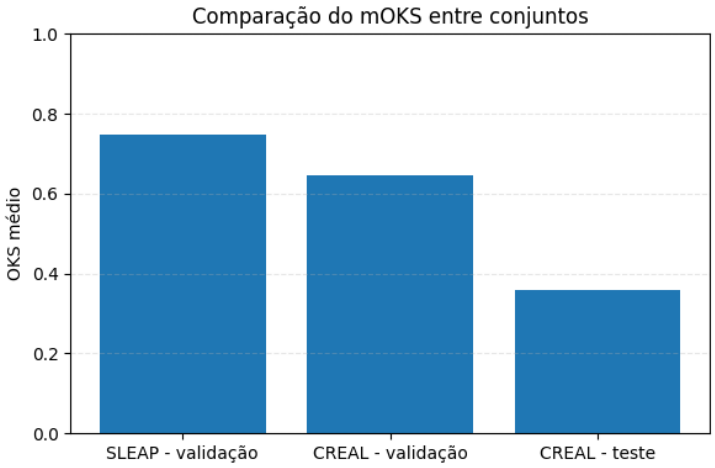
Figura 26: *Exemplo de ratos escondidos (dataset CREAL).*



Fonte: Autor.

Como nota geral do modelo, o mOKS apresentou um valor de 0,3598, o que se encontra aquém de um resultado adequado. Em comparação com os demais resultados apresentados na Figura 27 observa-se um queda considerável. Apesar disso, devem ser levados em consideração os resultados de PCK do torso, que influenciaram fortemente o mOKS dos conjuntos de validação. Como o esqueleto do animal apresenta apenas quatro *keypoints*, cada um desses tem um peso de 25% sobre o mOKS final, indicando que essa piora pode ser fortemente impactada pelos resultados do modelo de detecção, que para os conjuntos de validação não era observado.

Figura 27: *Resultado de mOKS para os três conjuntos.*



Fonte: Autor.

Alguns exemplos de predições com alto erro médio entre os *keypoints* estimados e anotados foram destacados na Figura 28, e auxiliam no entendimento dos erros. Uma das diferenças de domínio (*domaing gap*) entre os conjuntos de validação e teste é a presença de ração na gaiola, que introduz uma dificuldade para o modelo na hora de detectar animais obstruídos, mesmo que parcialmente, como é o caso da Figura 28a.

Outra situação que afetou o resultado foi a inversão de *keypoints*, como demonstrado na Figura 28b, onde o *snout* foi posicionado na região do *tail base*. Esse caso evidencia os resultados mencionados anteriormente relativos à grande quantidade de *outliers* presentes

na estimação do *snout*. Erros como estes podem ser tratados através de limites de distância e ângulo entre os *keypoints* baseados na anatomia do animal, como sugerido por (GOSZTOLAI *et al.*, 2021).

Figura 28: *Frames com alto erro médio de inferência (vermelho: predição, verde: anotação).*

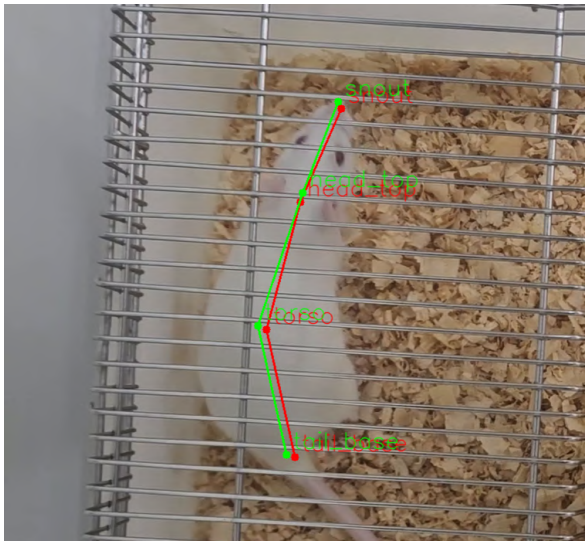


(a) *Frame 11248 - obstrução.* **(b)** *Frame 2196 - inversão.*

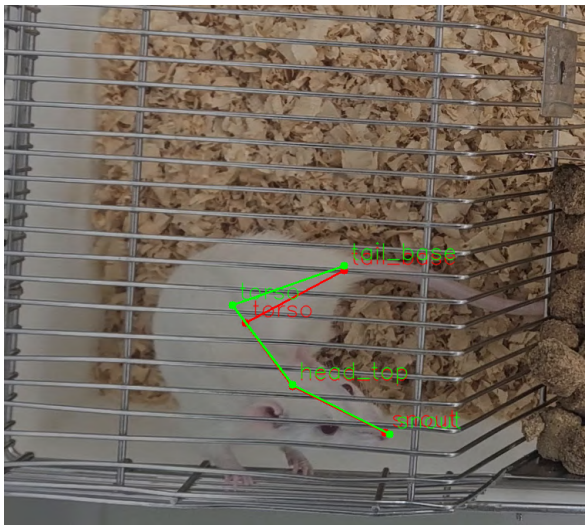
Fonte: Autor.

Apesar dos erros encontrados, alguns casos surpreenderam de forma positiva, por apresentarem resultados de sucesso na estimação, os quais foram apresentados na Figura 29. Nela destacam-se exemplos de situações de maior facilidade, como o caso das Figuras 29a e 29b, onde o modelo foi capaz de acertar com boa precisão até mesmo o *torso* e *snout*. E também são apresentadas situações mais complexas, onde o animal apresentou torção, na Figura 29c, e na Figura 29d, onde mesmo estando de pé e com a presença de ambos os membros próximos ao focinho, a pose foi estimada corretamente.

Figura 29: *Frames com baixo erro médio de inferência (vermelho: predição, verde: anotação).*



(a) *Frame 10562 - parado.*



(b) *Frame 3400 - curvado.*



(c) *Frame 6759 - roendo.*



(d) *Frame 12928 - de pé.*

Fonte: Autor.

5 CONCLUSÕES

Os resultados apresentados neste trabalho indicam que, com o atual conjunto de dados, a estimação de pose não atingiu um desempenho suficientemente robusto para aplicação direta no cenário proposto. A queda observada nas métricas de localização dos *keypoints*, especialmente no conjunto de teste, evidencia as limitações impostas pela quantidade de dados disponíveis e pelas diferenças de domínio entre os ambientes analisados. Apesar disso, observou-se que a interpretação das métricas de avaliação depende fortemente do contexto da aplicação, tornando fundamental a definição criteriosa dos limiares utilizados nos cálculos de desempenho, de forma que os resultados reflitam a precisão necessária para o problema estudado.

Mesmo não estando pronto para aplicação, os resultados do modelo demonstram o potencial da abordagem. As análises realizadas indicam que o aumento da quantidade e da diversidade dos dados de treinamento, aliado a um estudo mais aprofundado sobre a seleção dos *keypoints* utilizados e do ponto âncora empregado na detecção, pode contribuir significativamente para a melhoria do desempenho dos modelos. Os resultados obtidos indicaram que o modelo conseguiu estimar o *head top* com erro médio próximo de 25 pixels no conjunto de teste, tornando-o uma alternativa a ser considerada como ponto âncora para o modelo de detecção. No entanto, destaca-se a necessidade de avaliar o impacto na determinação dos parâmetros do modelo, em especial ao *crop size*, uma vez que este determina o tamanho da janela de recorte que é centralizada no ponto âncora, podendo deixar parte do animal fora deste recorte. Esse resultado também sugere que a escolha de regiões do corpo com referências visuais pode auxiliar tanto a anotação quanto o aprendizado do modelo. Para trabalhos futuros, sugere-se a escolha de *keypoints* que possuam referências anatômicas, como é o caso da região escolhida entre as orelhas e também entre os olhos.

Outro aspecto que merece investigação refere-se à escolha dos parâmetros utilizados ao longo da metodologia. Neste trabalho, os limiares utilizados na avaliação e outros hiperparâmetros do treinamento foram definidos com base em recomendações da literatura, testes preliminares e nas características do conjunto de dados utilizado. Entretanto, não foi realizada uma análise sistemática da influência desses parâmetros sobre o desempenho dos modelos, de modo que não é possível afirmar que a configuração adotada representa a solução ótima para o problema. Dessa forma, recomenda-se que trabalhos futuros realizem uma análise de sensibilidade desses parâmetros, investigando o impacto de suas variações nas métricas de detecção e estimação de pose, bem como sua interação com diferentes escolhas de *keypoints* e pontos âncora.

Os resultados observados no trabalho indicam que o modelo ainda não possui alta precisão na estimação das poses. Apesar disso, eles demonstram uma capacidade de posicionamento dos *keypoints* nas regiões próximas às desejadas. Isso permite a aplicação do modelo para outras abordagens que não necessitem de tanta precisão, como por exemplo, identificação de momentos de maior atividade do animal. Monitorando-se os *keypoints*

é possível avaliar a variação de posicionamento destes no tempo e, consequentemente, identificar períodos com diferentes níveis de movimentação do animal. Essa hipótese é sugerida para trabalhos futuros que tenham em vista o uso das informações de pose do animal.

Embora o objetivo deste trabalho tenha sido o desenvolvimento de uma metodologia de estimação de pose, a utilização de apenas quatro *keypoints* limita a aplicação direta dessas poses para tarefas futuras de classificação automática de comportamento. Com essa representação simplificada, comportamentos que dependem da movimentação das patas ou da postura detalhada do animal, como *grooming*, não podem ser caracterizados de forma confiável. A inclusão de *keypoints* adicionais, como ombros e patas, poderia ampliar significativamente o conjunto de comportamentos passíveis de identificação. No entanto, a anotação desses pontos adicionais demandaria câmeras posicionadas em ângulos diferentes e possivelmente técnicas de triangulação para a reconstrução da pose em 3D.

Conclui-se que a estimação de pose para ratos em ambiente laboratorial apresenta potencial para aplicações futuras, mas ainda demanda avanços na construção do conjunto de dados e na configuração dos modelos para que sua utilização ocorra de forma confiável. Entre as possibilidades de evolução da metodologia destacam-se a aplicação de técnicas de pré-processamento para remoção das grades da imagem, a incorporação de métodos de rastreamento de identidade entre múltiplos indivíduos e a utilização de técnicas de pós-processamento, para suavizar alguns dos erros considerados *outliers* e também preencher falsos negativos a partir da observação de frames anteriores e subsequentes.

Além disso, alterações na metodologia de aquisição das imagens, como a taxa de aquisição, a resolução e a distância da câmera em relação ao animal, também podem contribuir para a melhoria dos resultados obtidos. Em especial, para a etapa de treinamento, a redução da taxa de aquisição pode aumentar a diversidade das amostras anotadas, uma vez que *frames* consecutivos tendem a apresentar poses muito semelhantes. Dessa forma, espera-se que os resultados e as discussões apresentados neste trabalho contribuam para o desenvolvimento de futuras metodologias de estimação de pose aplicadas a animais, fornecendo uma base para pesquisas voltadas ao monitoramento automatizado e à análise comportamental.

REFERÊNCIAS

ANDRILUKA, M. *et al.* 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *In:* IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [*S. l.: s. n.*], jun. 2014.

ARENT, I. *et al.* Marker-Less Motion Capture of Insect Locomotion With Deep Neural Networks Pre-trained on Synthetic Videos. *Frontiers in Behavioral Neuroscience*, v. 15, abr. 2021. ISSN 1662-5153. DOI: 10.3389/fnbeh.2021.637806.

CAO, J. *et al.* Cross-Domain Adaptation for Animal Pose Estimation. *In:* 2019 IEEE/CVF International Conference on Computer Vision (ICCV). [*S. l.: s. n.*], 2019. p. 9497–9506. DOI: 10.1109/ICCV.2019.00959.

CHEN, C.; ZHU, W.; NORTON, T. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. *Computers and Electronics in Agriculture*, Elsevier, v. 187, p. 106255, ago. 2021. ISSN 0168-1699. DOI: 10.1016/J.COMPAG.2021.106255.

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). [*S. l.: s. n.*], 2005. v. 1, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

DOORNWEERD, J. E. *et al.* Across-Species Pose Estimation in Poultry Based on Images Using Deep Learning. *Frontiers in Animal Science*, v. 2, 2021. ISSN 26736225. DOI: 10.3389/fanim.2021.791290.

DUTTA, S. *et al. UnCageNet: Tracking and Pose Estimation of Caged Animal.* [*S. l.: s. n.*], 2025. arXiv: 2512.07712 [cs.LG]. Disponível em: https://arxiv.org/abs/2512.07712.

FAZZARI, E. *et al.* Animal behavior analysis methods using deep learning: A survey. *Expert Systems with Applications*, v. 289, p. 128330, set. 2025. ISSN 09574174. DOI: 10.1016/j.eswa.2025.128330.

GEORGEVICI, A. I.; TERBLANCHE, M. Neural networks and deep learning: a brief introduction. *Intensive Care Med*, v. 45, p. 712–714, 2019. DOI: 10.1007/s00134-019-05537-w. Disponível em: https://doi.org/10.1007/s00134-019-05537-w.

GIRDHAR, R. *et al.* Detect-and-Track: Efficient Pose Estimation in Videos. *In:* PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [*S. l.: s. n.*], jun. 2018.

GONZALEZ, R.; WOODS, R. *Digital Image Processing*. [*S. l.*]: Prentice Hall, 2008. ISBN 9780131687288. Disponível em: https://books.google.com.br/books?id=8uGOnjRGEzoC.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [*S. l.*]: MIT Press, 2016. http://www.deeplearningbook.org.

GOSZTOLAI, A. *et al.* LiftPose3D, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature Methods*, Springer Science e Business Media LLC, v. 18, p. 975–981, 2021. DOI: 10.1038/s41592-021-01226-z. Disponível em: https://doi.org/10.1038/s41592-021-01226-z.

GRAVING, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, v. 8, out. 2019. ISSN 2050-084X. DOI: 10.7554/eLife.47994.

GRIECO, F. *et al.* Measuring Behavior in the Home Cage: Study Design, Applications, Challenges, and Perspectives. *Frontiers in Behavioral Neuroscience*, Volume 15 - 2021, 2021. ISSN 1662-5153. DOI: 10.3389/fnbeh.2021.735387. Disponível em: https://www.frontiersin.org/journals/behavioral-neuroscience/articles/10.3389/fnbeh.2021.735387. Acesso em: 11 jun. 2026.

HANKENSON, F. C.; BRADEN-WEISS, G. C.; BLENDY, J. A. Behavioral and activity assessment of laboratory mice (Mus musculus) after tail biopsy under isoflurane anesthesia. *Journal of the American Association for Laboratory Animal Science : JAALAS*, v. 50, p. 686–94, 5 set. 2011. ISSN 2769-6677.

HARRIS, C. G.; STEPHENS, M. J. A Combined Corner and Edge Detector. *In:* ALVEY Vision Conference. [*S. l.: s. n.*], 1988. Disponível em: https://api.semanticscholar.org/CorpusID:1694378.

HAYKIN, S. S. *Neural networks and learning machines*. [*S. l.*]: Prentice Hall/Pearson, 2009. p. 906. ISBN 0131471392.

HE, K. *et al.* Deep Residual Learning for Image Recognition. *In:* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [*S. l.: s. n.*], 2016. p. 770–778. DOI: 10.1109/CVPR.2016.90.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, v. 25, jan. 2012. DOI: 10.1145/3065386.

LECUN, Y. *et al.* Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. DOI: 10.1109/5.726791.

LIN, T.-Y. *et al. Microsoft COCO: Common Objects in Context*. [*S. l.: s. n.*], 2015. arXiv: 1405.0312 [cs.CV].

LÓPEZ, O. A. M.; LÓPEZ, A. M.; CROSSA, J. Fundamentals of Artificial Neural Networks and Deep Learning. *In:* MULTIVARIATE Statistical Machine Learning Methods for Genomic Prediction. [*S. l.*]: Springer International Publishing, 2022. p. 379–425. DOI: 10.1007/978-3-030-89010-0_10.

LOWE, D. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, v. 60, p. 91–110, nov. 2004. DOI: 10.1023/B%3AVISI.0000029664.99615.94.

MATHIS, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, v. 21, 9 2018. ISSN 15461726. DOI: 10.1038/s41593-018-0209-y.

MCILWAIN, K. L. *et al.* The use of behavioral test batteries: Effects of training history. *Physiology & Behavior*, v. 73, p. 705–717, 5 ago. 2001. ISSN 00319384. DOI: 10.1016/S0031-9384(01)00528-5.

MITCHELL, T. M. *Machine Learning*. [*S. l.*]: McGraw-Hill, 1997. p. 414. ISBN 0070428077.

MORTON, A. J. Sleep and Circadian Rhythm Dysfunction in Animal Models of Huntington's Disease. *Journal of Huntington's Disease*, v. 12, p. 133–148, 2 jul. 2023. ISSN 1879-6397. DOI: 10.3233/JHD-230574.

PAULLADA, A. *et al.* Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, v. 2, n. 11, p. 100336, 2021. ISSN 2666-3899. DOI: https://doi.org/10.1016/j.patter.2021.100336. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666389921001847.

PEREIRA, T. D.; ALDARONDO, D. E. *et al.* Fast animal pose estimation using deep neural networks. *Nature Methods*, v. 16, p. 117–125, 1 jan. 2019. ISSN 1548-7091. DOI: 10.1038/s41592-018-0234-5.

PEREIRA, T. D.; TABRIS, N. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, v. 19, 4 2022. ISSN 15487105. DOI: 10.1038/s41592-022-01426-1.

PÉREZ, M.; TOLER-FRANKLIN, C. *CNN-Based Action Recognition and Pose Estimation for Classifying Animal Behavior from Videos: A Survey*. [*S. l.*], 2023.

REDMON, J. *et al.* You Only Look Once: Unified, Real-Time Object Detection, jun. 2015. DOI: 10.48550/arXiv.1506.02640.

RONCHI, M. R.; PERONA, P. Benchmarking and Error Diagnosis in Multi-instance Pose Estimation. *In:* DOI: 10.1109/ICCV.2017.48.

SEGALIN, C. *et al.* The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife*, v. 10, nov. 2021a. ISSN 2050-084X. DOI: 10.7554/eLife.63720.

SEGALIN, C. *et al.* The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. Edição: Gordon J Berman, Kate M Wassum e Asaf Gal. *eLife*, eLife Sciences Publications, Ltd, v. 10, e63720, nov. 2021b. ISSN 2050-084X. DOI: 10.7554/eLife.63720. Disponível em: https://doi.org/10.7554/eLife.63720.

SMITI, A. A critical overview of outlier detection methods. *Computer Science Review*, v. 38, p. 100306, 2020. ISSN 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2020.100306. Disponível em: https://www.sciencedirect.com/science/article/pii/S1574013720304068.

SZELISKI, R. *Computer Vision*. [*S. l.*]: Springer London, 2011. ISBN 978-1-84882-934-3. DOI: 10.1007/978-1-84882-935-0.

TALMO LAB. *Evaluation Metrics - SLEAP Documentation*. [*S. l.: s. n.*], 2026a. Disponível em: https://nn.sleap.ai/latest/reference/evaluation%5C_metrics/%5C#detection-metrics. Acesso em: 5 jun. 2026.

TALMO LAB. *Model config - SLEAP Documentation*. [*S. l.: s. n.*], 2026b. Disponível em: https://nn.sleap.ai/latest/api/config/model_config/%5C#sleap_nn.config.model_config.ConvNextBaseConfig. Acesso em: 5 jun. 2026.

TALMO LAB. *Model Types - SLEAP Documentation*. [*S. l.: s. n.*], 2026c. Disponível em: https://nn.sleap.ai/latest/reference/models/%5C#receptive-field. Acesso em: 5 jun. 2026.

TALMO LAB. *receptivefield - SLEAP Documentation*. [*S. l.: s. n.*], 2026d. Disponível em: https://docs.sleap.ai/latest/api/gui/learning/receptivefield/?h=crop_size%5C#sleap.gui.learning.receptivefield.find_instance_crop_size. Acesso em: 5 jun. 2026.

YANG, Y.; RAMANAN, D. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 12, p. 2878–2890, 2013. DOI: 10.1109/TPAMI.2012.261.

YU, H. *et al.* AP-10K: A Benchmark for Animal Pose Estimation in the Wild. *In:* THIRTY-FIFTH Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). [*S. l.: s. n.*], 2021. Disponível em: https://openreview.net/forum?id=rH8yliN6C83.

ZAIDI, S. A. A.; WAGAN, A. I. Deep Learning Based Approaches for Animal Pose Estimation (APE): A Survey. *In:* 2024 International Conference on Emerging Trends in Smart Technologies (ICETST). [*S. l.: s. n.*], 2024. p. 1–6. DOI: 10.1109/ICETST62952.2024.10737984.

Apêndices

Apêndice A - PARÂMETROS DE TREINAMENTO: MODELO SLEAP

Tabela 4: *Comparação entre as configurações dos modelos de detecção e estimação utilizados no conjunto SLEAP.*

Parâmetro	**Detecção**	**Estimação**
Divisão treino/validação	80% / 20%	80% / 20%
Formato das imagens	Escala de cinza	Escala de cinza
Tamanho original	1024×768 px	1024×768 px
Escalonamento de entrada	0,5	1,0
Tamanho escalonado	512×384 px	1024×768 px
Crop size	–	320×320 px
Ponto âncora	torso	torso
Max stride	32	64
Receptive field	156 px	316 px
Rotação aleatória	$[-180°, 180°]$	$[-180°, 180°]$
Escalonamento aleatório	$[0,9; 1,1]$	$[0,9; 1,1]$
Ajuste de brilho	$[0,5, 2,0]$	$[0,8, 1,2]$
Ajuste de contraste	$[0,5, 2,0]$	$[0,8, 1,2]$
Taxa de aprendizado inicial	10^{-4}	10^{-4}
Batch size	8	8
Épocas máximas	75	100
Early stopping	10 épocas	10 épocas

Apêndice B - PARÂMETROS DE TREINAMENTO: MODELO CREAL

Tabela 5: *Comparação entre as configurações dos modelos de detecção e estimação utilizados no conjunto CREAL.*

Parâmetro	**Detecção**	**Estimação**
Divisão treino/validação	80% / 20%	80% / 20%
Formato das imagens	Escala de cinza	Escala de cinza
Tamanho original	1920×1080 px	1920×1080 px
Escalonamento de entrada	0,5	1,0
Tamanho escalonado	960×540 px	1920×1080 px
Crop size	–	640×640 px
Ponto âncora	torso	torso
Max stride	64	64
Receptive field	316 px	316 px
Rotação aleatória	$[-180°, 180°]$	$[-180°, 180°]$
Escalonamento aleatório	$[0,9; 1,1]$	$[0,9; 1,1]$
Ajuste de brilho	$[0,5, 2,0]$	$[0,8, 1,2]$
Ajuste de contraste	$[0,5, 2,0]$	$[0,8, 1,2]$
Taxa de aprendizado inicial	10^{-4}	10^{-4}
Batch size	8	8
Épocas máximas	75	100
Early stopping	10 épocas	10 épocas