UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANÁLISE DE DESEMPENHO DE MECANISMOS DE ATENÇÃO PARA ESTIMATIVA DE POSE 2D BASEADA EM RESNET-50

MARLON NANAEL LEITÃO MALHEIROS

DM 26/2025

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2025

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MARLON NANAEL LEITÃO MALHEIROS

ANÁLISE DE DESEMPENHO DE MECANISMOS DE ATENÇÃO PARA ESTIMATIVA DE POSE 2D BASEADA EM RESNET-50

Dissertação/Tese submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para obtenção do Grau de Mestre em Engenharia Elétrica na Área de Computação Aplicada.

Orientador: Prof.ª Dr.ª Adriana Rosa Garcez Castro

DM 26/2025

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2025

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

M249a Malheiros, Marlon Nanael Leitão.
Análise de desempenho de mecanismos de atenção para estimativa de pose 2D baseada em ResNet-50 / Marlon Nanael Leitão Malheiros. — 2025.
74 f. : il. color.

Orientador(a): Profª. Dra. Adriana Rosa Garcez Castro
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2025.

1. Estimação de Pose Humana. 2. Mecanismos de Atenção. 3. Redes Neurais Convolucionais. 4. Convolutional Block Attention Module. 5. Coordinate Attention. I. Título.

CDD 006.3

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANÁLISE DE DESEMPENHO DE MECANISMOS DE ATENÇÃO PARA ESTIMATIVA DE POSE 2D BASEADA EM RESNET-50

AUTOR: **MARLON NANAEL LEITÃO MALHEIROS**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA. .

APROVADA EM: 11/09/2025

BANCA EXAMINADORA:

Prof.ª Dr.ª Adriana Rosa Garcez Castro
(Orientadora – PPGEE/ITEC/UFPA)

Prof.ª Dr.ª Jasmine Priscyla Leite de Araújo
(Avaliadora Interna - PPGEE/ITEC/UFPA)

Prof.ª Dr.ª Maria da Conceição Pereira Fonseca
(Avaliadora Externa ao Programa – FEEB/ITEC/UFPA)

Prof. Dr. Orlando Shigueo Ohashi Júnior
(Avaliador Externo – UFRA)

VISTO:

Prof. Dr. Diego Lisboa Cardoso
(Coordenador do PPGEE/ITEC/UFPA)

Resumo

A estimação de pose humana 2D é um problema fundamental em visão computacional que visa identificar a localização de pontos anatômicos humanos. A evolução do aprendizado profundo, em particular das Redes Neurais Convolucionais (CNNs), tem proporcionado avanços significativos no campo. Recentemente, a introdução de mecanismos de atenção se destacou como uma abordagem eficaz para aprimorar o foco das CNNs em regiões importantes das imagens. Esta dissertação apresenta um estudo comparativo do impacto de seis mecanismos de atenção na tarefa de estimação de pose humana 2D, integrando-os a uma arquitetura CNN baseada em *ResNet-50*: *Convolutional Block Attention Module* (CBAM), *Coordinate Attention*, *Global Context Attention*, *Self-Attention*, *Multi-Head Attention* e *SimAM* (*Simple, Parameter-Free Attention Module*). O treinamento e a avaliação dos modelos utilizaram o conjunto de imagens *MS COCO* (*Common Objects in Context*) sob um protocolo experimental unificado. Os resultados quantitativos demonstraram que todos os mecanismos de atenção testados melhoraram o desempenho da arquitetura base. Os mecanismos CBAM e *Coordinate Attention* mostraram-se os mais eficazes, com os maiores ganhos na métrica principal *Average Precision (AP)*. O modelo com *Coordinate Attention* alcançou uma *AP* de 67,7% (+1,5 p.p.), enquanto o modelo com CBAM atingiu 67,6% (+1,4 p.p.), obtendo também a melhor pontuação na métrica secundária *AP75*. A análise de custo-benefício revelou que CBAM e *Coordinate Attention* alcançaram esses ganhos com acréscimo mínimo de parâmetros e *FLOPS*. Em contraste, *Self-Attention*, de maior custo computacional, apresentou um dos menores ganhos, enquanto *SimAM*, livre de parâmetros, obteve o menor ganho sem custo adicional. Em síntese, os resultados demonstram que a integração de mecanismos de atenção é uma estratégia eficaz para aprimorar modelos de estimação de pose, destacando-se abordagens com ênfase em informação espacial explícita, como CBAM e *Coordinate Attention*, por oferecerem um excelente equilíbrio entre desempenho e eficiência computacional.

Palavras-chave: Estimação de Pose Humana, Mecanismos de Atenção, Redes Neurais Convolucionais, ResNet, CBAM, *Coordinate Attention*.

Abstract

2D human pose estimation is a fundamental computer vision problem focused on locating human anatomical keypoints. While deep learning, particularly Convolutional Neural Networks (CNNs), has driven significant progress, attention mechanisms have emerged as an effective method to enhance a model's focus on salient image regions. This dissertation presents a comparative study analyzing the impact of six different attention mechanisms on 2D human pose estimation when integrated into a ResNet-50-based CNN baseline. The evaluated mechanisms are: Convolutional Block Attention Module (CBAM), Coordinate Attention, Global Context Attention, Self-Attention, Multi-Head Attention, and SimAM (Simple, Parameter-Free Attention Module). All models were trained and evaluated on the MS COCO dataset under a unified experimental protocol. Quantitative results show that all attention mechanisms improved performance over the baseline. Coordinate Attention and CBAM were the most effective, achieving an Average Precision (AP) of 67.7% (+1.5 p.p.) and 67.6% (+1.4 p.p.), respectively, with CBAM also leading in the AP75 metric. A cost-benefit analysis confirmed these two models offered the best performance gains with a minimal increase in parameters and FLOPS. Conversely, the computationally expensive Self-Attention yielded one of the smallest gains, while the parameter-free SimAM offered the lowest improvement at no extra cost. In conclusion, this work demonstrates that integrating attention mechanisms is an effective strategy for human pose estimation. Specifically, approaches emphasizing explicit spatial information, like CBAM and Coordinate Attention, provide an excellent balance between performance improvement and computational efficiency.

Keywords: Human Pose Estimation, Attention Mechanisms, Convolutional Neural Networks, ResNet, CBAM, Coordinate Attention.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

CNN	Convolutional Neural Networks
CBAM	Convolutional Block Attention Module
ViT	Vision Transformers
GPU	Graphics Processing Unit
ReLU	Rectified Linear Unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ResNet	Residual Network
VGGNet	Visual Geometry Group Network
HRNet	High-Resolution Network
MS COCO	Microsoft Common Objects in Context
MPII	Max Planck Institute for Informatics
OKS	Object Keypoint Similarity
AP	Average Precision
TP	True Positive
FP	False Positive
FN	False Negative
RNN	Recurrent Neural Network
MSE	Mean Squared Error
FLOPS	Floating Point Operations Per Second

Sumário

1 Introdução

A tarefa de estimativa de pose humana 2D é um dos problemas fundamentais do campo da visão computacional, sendo seu objetivo principal a localização precisa de pontos anatômicos chave do corpo humano em imagens ou sequências de vídeo. O progresso nesse problema abriu caminho para uma vasta gama de aplicações, como reconhecimento de ações, que identifica o tipo de ação baseado nos dados de estimação de pose, interfaces homem-máquina, prevenção de acidentes, reabilitação, jogos, animação, análise esportiva, tradução de linguagens de sinais, entre outras.

A tarefa de estimação de pose humana encontra desafios como a oclusão, onde partes do corpo estão ocultas por pessoas, objetos ou até mesmo outras partes do próprio corpo; a complexidade das posturas que seres humanos podem assumir; e fatores como variação de contraste, tom de pele e iluminação, que modificam drasticamente as características visuais.

As Redes Neurais Convolucionais (CNNs – *Convolutional Neural Networks*) permitiram significativos avanços no campo da Visão Computacional. Inspiradas pelo funcionamento do córtex visual de mamíferos, as CNNs conseguem aprender a reconhecer características visuais relevantes de maneira autônoma a partir dos dados. O primeiro modelo amplamente reconhecido como uma CNN, a *LeNet-5*, foi proposto para o reconhecimento de dígitos manuscritos, estabelecendo a base para o uso de convoluções e *pooling* em tarefas visuais (LECUN et al., 1998). Desde então, as CNNs evoluíram consideravelmente, tornando-se pilares fundamentais em diversas aplicações, incluindo classificação de imagens, detecção de objetos e, mais recentemente, estimação de pose humana 2D.

Os avanços na área de estimação de pose foram impulsionados por diversas abordagens baseadas em CNNs. Em um dos trabalhos pioneiros, a *DeepPose*, de Toshev e Szegedy (2014) formularam a estimação de pose como um problema de regressão, utilizando uma CNN para predizer diretamente as coordenadas dos pontos anatômicos. Posteriormente, surgiram arquiteturas multiestágio mais sofisticadas, capazes de refinar as estimativas e capturar relações espaciais de longo alcance, como as *Convolutional Pose Machines* de Wei et al. (2016) e as *Stacked Hourglass Networks* de Newell, Yang e Deng (2016). Essa linha de pesquisa evoluiu para modelos ainda mais robustos, como a *Cascaded Pyramid Network* de Chen et al. (2017). Paralelamente, foram desenvolvidos sistemas voltados para aplicações práticas, como o *OpenPose*, um popular sistema de estimação 2D multipessoa em tempo real que realiza a detecção e associação conjuntamente proposto por Cao et al. (2017); e arquiteturas focadas em preservar a qualidade das características, como a *HRNet* (*High-Resolution Network*) de Sun et al. (2019), que mantém representações de alta resolução para aumentar a precisão espacial.

Apesar dos bons resultados, arquiteturas tradicionais que empregam CNNs têm como limitação o fato de a operação de convolução atuar localmente, o que limita sua capacidade de capturar relações entre regiões mais distantes dos mapas de características. Nos últimos anos,

os mecanismos de atenção têm se destacado como uma abordagem eficaz para aprimorar a capacidade das redes neurais de destacar seletivamente as regiões mais relevantes dos dados de entrada, o que pode ser aplicado em CNNs como uma solução para esse problema.

Os mecanismos de atenção, inspirados pela capacidade humana de focar seletivamente em partes específicas de um estímulo visual, permitem um maior foco em regiões mais relevantes, ao mesmo tempo em que suprimem outras menos úteis. Essa propriedade é particularmente útil em tarefas que envolvem estrutura e dependências espaciais, como é o caso da estimação de pose humana.

A integração de mecanismos de atenção em arquiteturas baseadas em CNNs tem se provado uma estratégia valiosa para superar as limitações das operações de convolução local. Isso pode ser alcançado através de diversas abordagens, aplicadas com sucesso em várias tarefas na Área da Visão Computacional. O trabalho pioneiro de Hu, Shen e Sun (2017) aplicou o conceito de atenção na dimensão dos canais para recalibrar dinamicamente os mapas de características, conferindo aos modelos a capacidade de salientar os canais mais relevantes, melhorando a performance em tarefas de classificação e detecção. A partir disso, outros mecanismos foram propostos, como o *Convolutional Block Attention Module (CBAM)* de Woo et al. (2018), que combina sequencialmente atenção de canal e espacial para fornecer aos modelos a capacidade de ponderar sobre qual informação é mais útil e qual é a sua localização. A versatilidade dos mecanismos de atenção também permitiu o desenvolvimento de arquiteturas como a *Vision Transformers (ViT)* de Dosovitskiy et al. (2021), que dispensam o uso de convoluções e superam as CNNs em inúmeros *benchmarks* clássicos.

Dado o sucesso dos mecanismos de atenção em diversas áreas e considerando a natureza da tarefa de estimativa de pose, altamente dependente da captura de relações espaciais, a aplicação de tais mecanismos é um caminho de estudo promissor. No entanto, dada a grande variedade de estratégias de atenção, desde módulos focados em canais, relações posicionais ou interações globais mais genéricas, a eficácia relativa e o custo de cada abordagem em uma arquitetura de referência para estimativa de pose ainda são áreas que necessitam de uma análise comparativa detalhada.

1.1 Objetivo Geral e Específicos

Este trabalho tem como objetivo geral investigar, através de um estudo comparativo, o impacto da integração de seis diferentes mecanismos de atenção em um modelo de referência baseado em uma CNN *ResNet-50* para a tarefa de estimativa de pose 2D. O estudo visa fornecer uma análise sistemática, avaliando o impacto nas métricas quantitativas e investigando o comportamento de cada mecanismo de atenção através de uma análise qualitativa pontual.

Para isso, foram definidos os seguintes objetivos específicos:

- Estabelecer um modelo de referência sem atenção para comparação e treiná-lo usando a base de dados pública MS COCO (*Microsoft Common Objects in Context*).

- Integrar cada um dos seis mecanismos de atenção à arquitetura de referência e treinar todas as variantes sob as mesmas condições experimentais.
- Avaliar os desempenhos quantitativo e qualitativo utilizando as métricas padrão para o problema e as predições geradas pelos modelos.
- Analisar o custo-benefício de cada abordagem, para verificar a relação de ganho de desempenho, assim como suas limitações.

1.2 Estrutura do Trabalho

Este trabalho está estruturado em cinco capítulos. O presente capítulo apresenta a contextualização do problema da estimação de pose humana 2D e define os objetivos da pesquisa. Em seguida, o Capítulo 2 estabelece a fundamentação teórica necessária, abordando desde as arquiteturas de CNN até os mecanismos de atenção que são o foco deste estudo. O Capítulo 3 detalha a metodologia, apresentando o protocolo experimental, a arquitetura de referência e as condições controladas sob as quais os diferentes módulos foram avaliados. No Capítulo 4, os resultados são apresentados e analisados, combinando uma comparação quantitativa do desempenho com uma discussão qualitativa para interpretar o comportamento de cada modelo. Por fim, o Capítulo 5 conclui o estudo com a síntese dos principais achados, a discussão de suas implicações práticas e a proposição de caminhos para investigações futuras.

2 Fundamentação Teórica

2.1 Estimativa de Pose 2D

A estimativa de pose humana 2D pode ser formalmente definida como o processo de mapear uma imagem de entrada I para um conjunto de coordenadas 2D que representam a localização dos pontos anatômicos de interesse para uma ou mais pessoas presentes na imagem.

Seja $P = \{p_1, p_2, \ldots, p_K\}$ o conjunto de K pontos pré-definidos (como, por exemplo, cotovelo esquerdo, joelho direito, nariz), onde cada p_k corresponde a uma coordenada (x_k, y_k), no plano da imagem, o objetivo da estimativa de pose 2D é, portanto, encontrar uma função $f : I \rightarrow \{(x_1, y_1), \ldots, (x_K, y_K)\}$ que prediz a localização de cada ponto para cada indivíduo em uma imagem (TOSHEV; SZEGEDY, 2014; MUNEA et al., 2020; ZHENG et al., 2023). A Figura 1 apresenta um exemplo do processo de estimativa de pose humana, onde na esquerda da figura tem-se a imagem de entrada, no meio a imagem com os pontos anatômicos estimados, e na direita a reconstrução do esqueleto a partir dos pontos estimados, denominada pose.

Figura 1 – Exemplo do processo de estimativa de pose humana 2D.

Fonte: Elaborado pelo autor (2025).

A tarefa de estimação tem um conjunto de desafios intrínsecos que podem dificultar ou até mesmo impossibilitar sua realização, como:

- **Oclusões:** Partes do corpo podem estar ocultas na imagem, devido a objetos, outras pessoas, ou outras partes do próprio corpo (auto oclusão). A Figura 2(a) ilustra esse cenário, onde a pessoa no plano de fundo está parcialmente oculta.
- **Variações de iluminação:** Condições de iluminação extremas, como baixa luz ou superexposição, podem degradar a qualidade da imagem, dificultando a distinção com o plano de fundo, como na Figura 2(c).
- **Variações de aparência:** Diferenças na vestimenta (roupas largas, justas ou coloridas), características físicas (idade, gênero, altura, biotipo) e acessórios introduzem grande variabilidade na aparência das pessoas.

- **Poses complexas:** A flexibilidade do corpo humano o permite assumir uma ampla variedade de poses, incluindo aquelas que são incomuns, acrobáticas e que envolvem movimentos muito rápidos, conforme exemplificado na Figura 2(b).
- **Perspectiva:** A pose de uma pessoa pode parecer muito diferente dependendo do ângulo e distância da câmera.
- **Plano de fundo complexo:** Cenários muito ruidosos ou com objetos semelhantes ao corpo humano acabam gerando falsos positivos ou dificultando a segmentação.

Figura 2 – Desafios para estimativa de pose 2D.

Fonte: Elaborado pelo autor (2025).

Para lidar com esses desafios é necessário o desenvolvimento de abordagens mais robustas, capazes de gerar representações mais gerais, menos afetadas pela complexidade e variabilidade inerentes à tarefa.

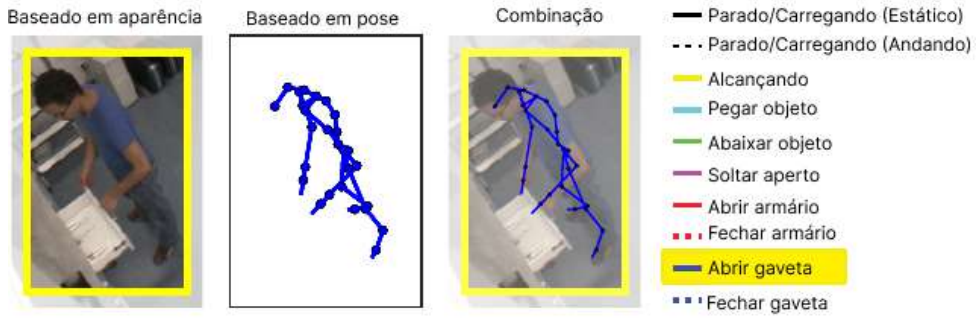
2.2 Aplicações da Estimativa de Pose

A versatilidade da estimativa de pose a torna uma ferramenta útil para vários domínios, sendo que as aplicações mais proeminentes são:

2.2.1 Reconhecimento de Ações

Analisando a sequência de poses estimadas ao longo do tempo, é possível identificar e classificar atividades humanas, como andar, correr, pular, ou ações mais complexas como interagir com algum objeto. A estimativa de pose pode ser uma etapa crucial para diversas aplicações como sistemas de vigilância inteligente, monitoramento de idosos e interfaces humano-computador. A Figura 3 apresenta um exemplo clássico em que dados de estimativa de pose são utilizados para identificar ações humanas comuns no cotidiano.

Figura 3 – Classificação de atividade humana baseado na sequência temporal de poses estimadas.

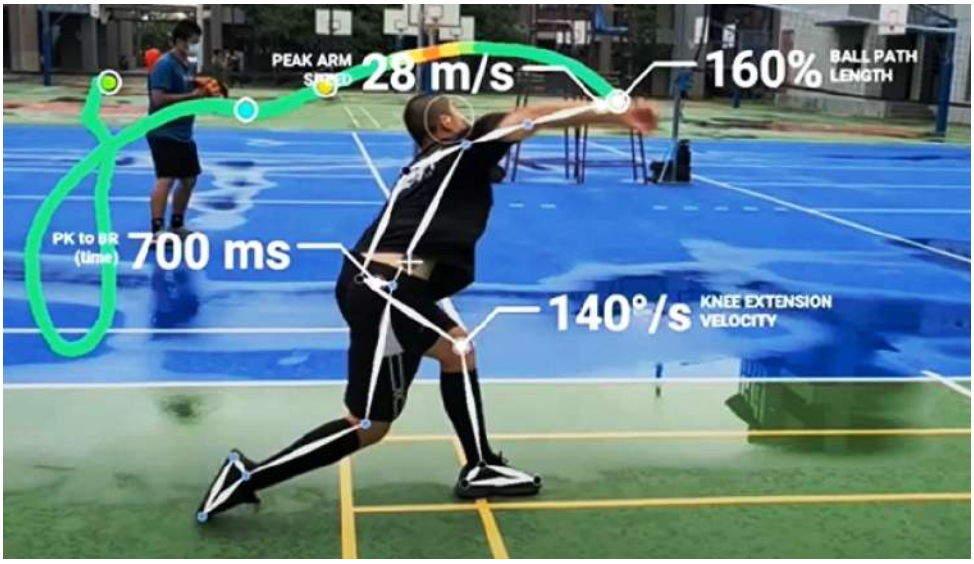


Fonte: Adaptado de (YAO et al., 2011).

2.2.2 Análise Biomecânica e Esportiva

A estimativa de pose permite uma análise detalhada do movimento humano em contextos esportivos e clínicos. Como ilustrado na Figura 4, atletas podem ter sua técnica aprimorada pela identificação de padrões de movimento ineficientes ou potencialmente lesivos. Na medicina, essa tecnologia também é aplicada no monitoramento e correção da progressão motora de pacientes.

Figura 4 – Análise de movimento para otimização de técnica esportiva.

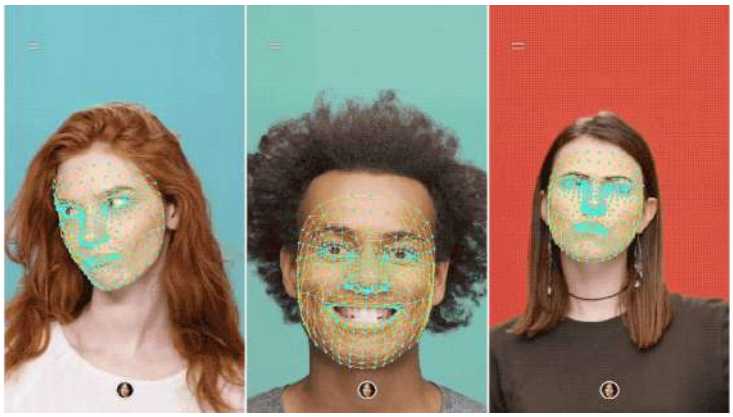


Fonte: (CHANGING. . . , 2023)

2.2.3 Realidade Aumentada e Realidade Virtual

Nestas aplicações, a estimativa de pose é usada para sobrepor elementos virtuais no corpo dos usuários de maneira mais realista, o que permite experiências mais imersivas. Uma aplicação prática amplamente difundida é o uso de filtros em redes sociais, como demonstrado na Figura 5, onde elementos gráficos são sobrepostos ao rosto do usuário com base na estimativa de pose facial.

Figura 5 – Filtro de realidade aumentada que utiliza estimativa de pose facial para alinhar elementos gráficos ao rosto do usuário em tempo real.

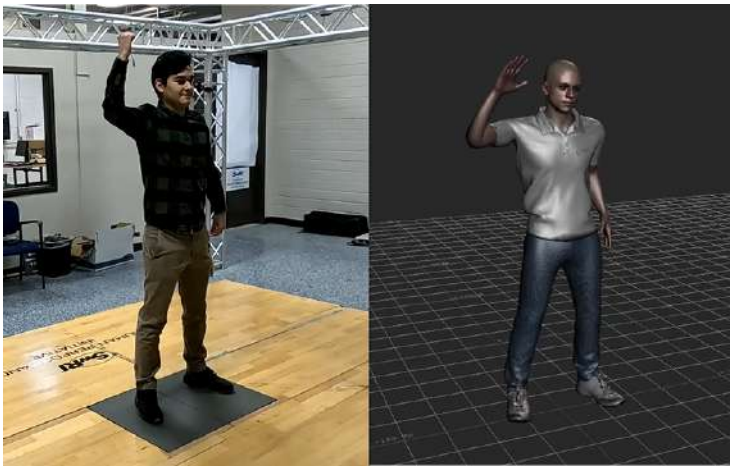


Fonte: (GOOGLE, 2025)

2.2.4 Animações e Efeitos Visuais

Na indústria do entretenimento, pode ser usada para animar personagens digitais usando o movimento de atores reais (*motion-capture*), e na criação de efeitos visuais realistas. Esse processo é ilustrado na Figura 6, que apresenta a animação de um personagem digital gerada a partir da captura de movimento via estimativa de pose.

Figura 6 – Animação de personagens digitais gerada a partir da captura de movimentos humanos via estimativa de pose.



Fonte: (SWRI, 2025)

2.3 Métodos de Estimativa de Pose 2D

Os métodos de estimativa de pose podem ser categorizados de acordo com diferentes critérios, como o número de pessoas na imagem de entrada (*single-person* ou *multi-person*), a ordem usada na execução das etapas (*top-down* ou *bottom-up*) e como a saída do modelo é representada

(regressão direta ou mapas de calor). Cada abordagem apresenta *trade-offs* de precisão, eficiência e robustez a cenários complexos.

No que diz respeito à capacidade do método em lidar com o número de indivíduos na imagem, tem-se:

- **Métodos de pessoa única (*Single-Person*):** Têm como objetivo realizar a estimativa de pose em um único indivíduo em uma imagem (MUNEA et al., 2020). Esse método pressupõe que existe apenas um indivíduo na imagem de entrada, o que geralmente demanda uma etapa anterior de detecção de pessoa para isolar o sujeito na imagem usando uma caixa delimitadora (*bounding-box*). Esse tipo de método é o mais simples e eficiente, porém limitado a cenários controlados, com a condição de que haverá somente uma pessoa na imagem utilizada.

- **Métodos multipessoa (*Multi-person*):** São abordagens mais complexas, pois precisam lidar com um número variável e potencialmente alto de indivíduos. O desafio não é apenas detectar os pontos, mas também associar corretamente cada ponto ao respectivo corpo. Esses métodos enfrentam dificuldades adicionais como oclusões entre pessoas, escalas variadas, gestos e interações imprevisíveis entre diferentes pessoas (CHEN et al., 2017).

As abordagens multipessoas se subdividem em duas estratégias principais de execução, sendo elas:

- ***Top-Down* (De cima para baixo):** Nesta estratégia, um detector de pessoas é empregado inicialmente para identificar a localização de cada indivíduo na imagem. Em seguida, um estimador de pose para pessoa única é aplicado sobre cada uma das regiões detectadas para estimar os pontos chave individualmente. Essa abordagem tende a obter um melhor desempenho por pessoa, porém seu custo computacional cresce linearmente com o número de pessoas na imagem (PAPANDREOU et al., 2017).

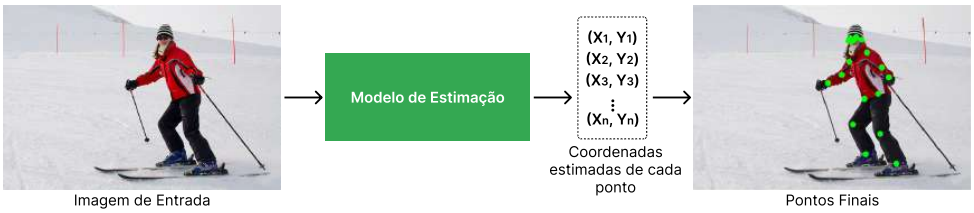
- ***Bottom-Up* (De baixo para cima):** Nesta estratégia, o modelo inicialmente detecta todos os pontos possíveis na imagem, sem considerar a qual pessoa pertencem, para posteriormente um algoritmo de associação (*matching*) agrupar os pontos correspondentes a cada indivíduo. Essa abordagem é mais eficiente em termos de tempo de execução, mas a etapa de associação pode introduzir erros, especialmente em casos com muitas sobreposições ou planos de fundo complexos (CAO et al., 2017; NEWELL; YANG; DENG, 2016).

Também é possível dividir os métodos de estimativa de pose de acordo com sua representação dos pontos localizados, sendo eles:

- **Modelos baseados em regressão:** Esse tipo de modelo trata o problema como uma tarefa de regressão coordenada, nesse caso ele é treinado para predizer diretamente as coordenadas (x, y) de cada ponto anatômico para uma dada imagem de entrada (conforme

exemplificado pela Figura 7). Apesar da simplicidade, essa abordagem tende a ser menos precisa, pois o uso de um único par de coordenadas limita a capacidade de lidar com pequenos erros.

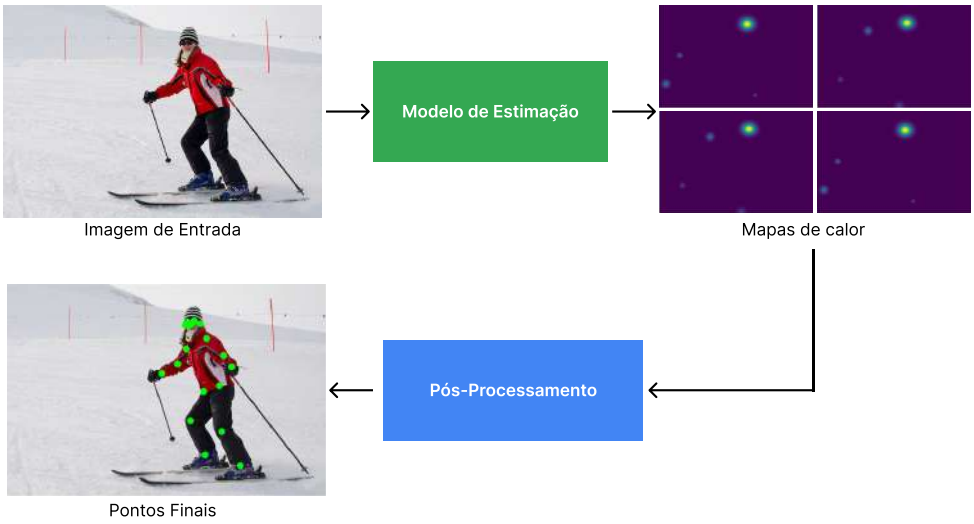
Figura 7 – Ilustração do processo de regressão direta de coordenadas para estimativa de pose.



Fonte: Elaborado pelo autor (2025).

- **Modelos baseados em mapas de calor (*heatmaps*):** Essa é a abordagem encontrada na maioria dos métodos considerados estado da arte e possui uma performance superior em comparação aos métodos de regressão direta, pois são capazes de preservar a estrutura espacial da imagem de entrada (QU et al., 2022). Nesse tipo de modelo, o objetivo é estimar um mapa de calor bidimensional para cada ponto, que funciona como uma distribuição de probabilidade indicando a localização mais provável de cada ponto (ZHENG et al., 2023). Cada mapa é processado para gerar a coordenada 2D (x, y) de cada um dos pontos anatômicos. O fluxograma apresentado na Figura 8 ilustra esse processo.

Figura 8 – Ilustração do processo de regressão baseado em mapas de calor para estimativa de pose.



Fonte: Elaborado pelo autor (2025).

2.4 Arquiteturas de CNN para Estimativa de Pose 2D

As Redes Neurais Convolucionais (CNNs) são inspiradas em propriedades observadas no córtex visual de mamíferos, como os campos receptivos locais, a organização hierárquica do processamento e o uso de filtros semelhantes aplicados em diferentes regiões da imagem. Diferentemente das redes neurais totalmente conectadas, as CNNs aproveitam a estrutura espacial dos dados visuais, extraindo padrões locais com o uso de filtros com pesos compartilhados. Essa estratégia reduz significativamente o número de parâmetros da rede, diminui o risco de *overfitting* e melhora a capacidade de generalização. A primeira CNN amplamente conhecida foi a LeNet-5, proposta por Lecun et al. (1998), voltada para o reconhecimento de dígitos manuscritos. Embora limitada em profundidade e aplicada a tarefas simples, a LeNet estabeleceu os fundamentos centrais da convolução, *pooling* e treinamento supervisionado com o algoritmo *backpropagation*.

Apesar do potencial das CNNs, elas enfrentaram um período de estagnação devido principalmente à limitação dos recursos computacionais disponíveis e à ausência de bancos de dados aproveitáveis na época. Esse cenário mudou drasticamente com a introdução da *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), que marcou o renascimento das CNNs em larga escala. Vencedora do desafio *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* de 2012, a *AlexNet* reduziu drasticamente o erro de classificação em relação ao segundo colocado, utilizando *GPUs (Graphics Processing Units)* para acelerar o treinamento, popularizando a função de ativação *ReLU* e incorporando técnicas como *dropout* e *data augmentation*. A arquitetura da rede *AlexNet* demonstrou de forma inequívoca a superioridade das CNNs sobre métodos tradicionais de visão computacional em grandes bases de dados.

A estrutura básica de uma CNN usada para tarefas de classificação de imagens, conforme ilustrado na Figura 9, é composta por quatro componentes principais:

1. **Camadas convolucionais:** aplicam filtros treináveis sobre regiões locais da imagem para extrair padrões como bordas, texturas e formas. Os resultados dessas operações são chamados de mapas de características e representam a informação extraída em cada estágio da rede.

2. **Funções de ativação não lineares:** aplicadas sobre os mapas de características para introduzir não linearidade ao modelo, permitindo que a rede aprenda representações mais complexas.

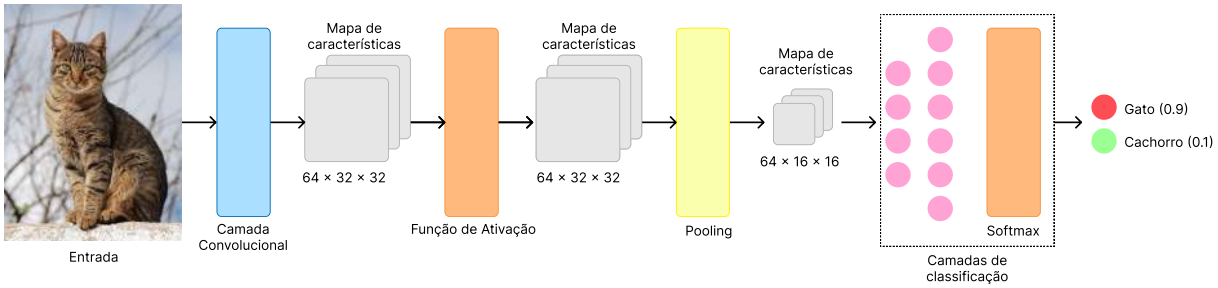
3. **Camadas de *pooling*:** realizam uma operação de redução de dimensionalidade espacial dos mapas de características. Seu objetivo principal é condensar as informações mais relevantes em uma representação compacta, reduzindo o custo computacional e acelerando o treinamento. Essa operação é realizada por meio de uma janela deslizante denominada *kernel*, que percorre o mapa de características e aplica uma função de agregação sobre os valores contidos na região da janela. As variações mais encontradas de *pooling* são:

    - *Max Pooling*: Seleciona o valor mais alto dentro do *kernel* em cada deslocamento.

- *Average Pooling*: Calcula a média dos valores dentro do *kernel*.
- *Global Pooling*: Utiliza um *kernel* com o tamanho da dimensão espacial do mapa de características e o transforma em um único valor.

4. **Camada de classificação:** composta por uma ou mais camadas totalmente conectadas. Responsável por gerar um vetor de valores correspondentes às possíveis classes, que são normalizados por outra função de ativação (como *sigmoide* ou *softmax*) produzindo uma distribuição de probabilidade sobre as classes, permitindo selecionar a mais provável saída para a rede.

Figura 9 – Arquitetura simplificada de rede para classificação de imagens.



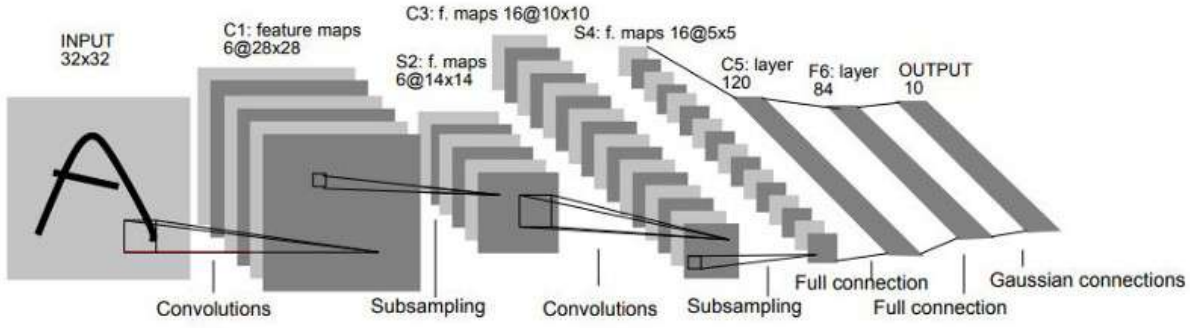
Fonte: Elaborado pelo autor (2025).

Arquiteturas modernas também incorporam camadas de normalização em lote (*Batch Normalization*), que padronizam os valores dentro de cada *mini-batch* para acelerar e estabilizar o treinamento. Além disso, muitas redes utilizam conexões residuais, introduzidas na arquitetura *ResNet* (HE et al., 2015), que permitem o fluxo direto do gradiente por meio da soma entre a entrada e a saída de blocos convolucionais, facilitando o treinamento de redes profundas. Por fim, técnicas de regularização, como o *dropout*, são amplamente empregadas para reduzir o risco de *overfitting*.

A Figura 10 apresenta a arquitetura da LeNet-5, uma das primeiras CNNs aplicadas com sucesso no campo de visão computacional. O modelo recebe como entrada uma imagem 32×32 de um dígito manuscrito, processando-a sequencialmente através de camadas convolucionais (C1, C3), *pooling* (S2, S4) e, por fim, camadas totalmente conectadas (C5, F6) responsáveis pela classificação. Esse arranjo já reflete a organização *encoder-decoder* presente em redes modernas, com extração progressiva de características e uma etapa final de interpretação densa.

Do ponto de vista computacional, as CNNs são extremamente mais eficientes do que redes totalmente conectadas para o processamento de imagens. Por exemplo, considerando uma imagem de entrada com resolução 32×32 e 1 canal. Uma camada densa conectando cada pixel da imagem a 120 neurônios exigiria 1.228.800 parâmetros. Em contraste, uma camada convolucional com 6 filtros de tamanho 5×5 pixels teria apenas 150. Essa economia é fundamental para a escalabilidade das redes e viabiliza o treinamento de arquiteturas profundas, mesmo com limitações computacionais.

Figura 10 – Arquitetura da CNN LeNet-5.



Fonte: (LECUN et al., 1998)

As arquiteturas modernas de CNNs para estimativa de pose humana 2D geralmente seguem o paradigma *encoder-decoder*, sendo compostas por dois módulos principais que atuam de forma sequencial: o *backbone* (espinha dorsal), responsável pela extração de características visuais relevantes da imagem, e a *estimation head*, que interpreta essas características para estimar a localização dos pontos anatômicos. Embora nem todos os trabalhos utilizem explicitamente essa nomenclatura, essa divisão é útil para compreender o funcionamento modular das redes e discutir estratégias específicas de aprimoramento, como a inserção de mecanismos de atenção em partes distintas do modelo.

2.4.1 *Backbone*

O *backbone* é o módulo responsável por processar a imagem de entrada e convertê-la em um mapa de características rico em informações relevantes para a tarefa de estimativa de pose. Geralmente, são arquiteturas de rede CNN bem estabelecidas, pré-treinadas em grandes bancos de dados como o *ImageNet*, o que lhes confere uma poderosa capacidade de generalização. Para uso como *backbone*, essas arquiteturas, que originalmente foram desenvolvidas para problemas de classificação, são adaptadas através da remoção de suas camadas classificadoras, preservando apenas a parte responsável pela extração de características. A seguir, serão apresentados exemplos de arquiteturas frequentemente utilizadas na literatura como *backbones* em tarefas de estimativa de pose, sendo elas: *ResNet*, *VGGNet* e *HRNet*.

2.4.1.1 *ResNet (Residual Network)*

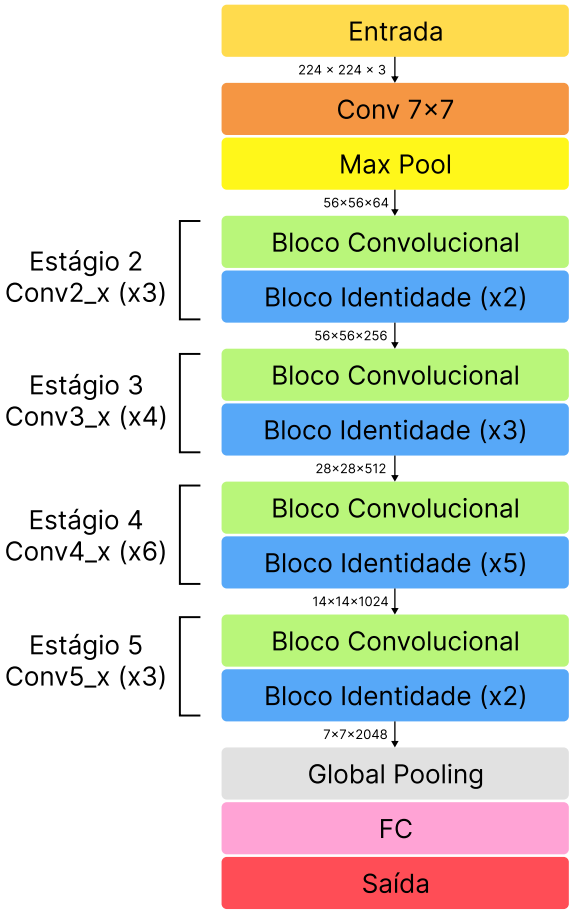
A *ResNet*, proposta por He et al. (2015), representa um marco na arquitetura de redes convolucionais profundas. Sua principal inovação foi solucionar o problema de degradação, um fenômeno onde o aumento da profundidade da rede leva a um erro de treinamento maior, indicando que o otimizador não consegue encontrar uma solução eficaz para as camadas adicionadas. Esse desafio de otimização foi atenuado pela introdução dos blocos residuais.

Os blocos residuais implementam conexões de atalho (*skip connections*) permitindo que a entrada de um bloco seja somada à sua saída, criando um caminho direto para o fluxo do gradiente durante o processo de *backpropagation*. Com isso, em vez de forçar as camadas a aprenderem uma transformação completa, a rede é otimizada para aprender uma função residual mais simples, onde, na pior das hipóteses, se uma camada adicional não for útil, a rede pode aprender a zerar seus pesos, fazendo com que a camada se comporte como uma conexão de identidade e não prejudique o desempenho (HE et al., 2015).

Dentre as variantes da família *ResNet*, a *ResNet-50* é uma das arquiteturas mais populares e equilibradas em termos de desempenho e custo computacional. Conforme ilustrado na Figura 11, a *ResNet-50* é formada por uma camada inicial de processamento, seguida por quatro estágios sequenciais que aumentam progressivamente a profundidade do mapa de características. Cada estágio é composto por uma pilha de blocos de dois tipos: o Bloco de Identidade, utilizado quando as dimensões do tensor são mantidas, e o Bloco Convolucional, empregado na transição entre os estágios para realizar *downsampling* e ajustar o número de canais. Os detalhes e parâmetros de cada estágio da *ResNet-50* são apresentados na Tabela 1.

No contexto da estimativa de pose humana, a *ResNet* se estabeleceu como um *backbone* eficaz e amplamente utilizado. Sua capacidade de extrair mapas de características ricos, sem sofrer com o problema de degradação de gradiente em arquiteturas mais complexas, é fundamental para tarefas de localização mais granulares, como a estimativa de pose. Por essa razão, serve como *backbone* para arquiteturas influentes, como a *SimpleBaseline* de Xiao, Wu e Wei (2018), *Cascaded Pyramid Network* de Chen et al. (2017), ou é usada como base na *Residual Steps Network* (CAI et al., 2020) e na *HRNet* (SUN et al., 2019).

Figura 11 – Arquitetura simplificada da ResNet-50.



Fonte: Elaborado pelo autor (2025).

Tabela 1 – Composição da arquitetura ResNet-50.

Estágio	**Tamanho da saída**	**Estrutura do bloco**	**Repetições**
Conv1	112×112	7×7, 64, *stride* 2	1
MaxPool	56×56	3×3 *max pooling*, *stride* 2	–
Conv2_x	56×56	$[1 \times 1, 64] \rightarrow [3 \times 3, 64] \rightarrow$ $\rightarrow [1 \times 1, 256]$	×3
Conv3_x	28×28	$[1 \times 1, 128] \rightarrow [3 \times 3, 128] \rightarrow$ $\rightarrow [1 \times 1, 512]$	×4
Conv4_x	14×14	$[1 \times 1, 256] \rightarrow [3 \times 3, 256] \rightarrow$ $\rightarrow [1 \times 1, 1024]$	×6
Conv5_x	7×7	$[1 \times 1, 512] \rightarrow [3 \times 3, 512] \rightarrow$ $\rightarrow [1 \times 1, 2048]$	×3
AvgPool	1×1	Global *average pooling*	1
FC	1×1	Camada totalmente conectada	1

Fonte: Elaborado pelo próprio autor.

2.4.1.2 *VGGNet*

A *VGGNet* (*Visual Geometry Group Network*), proposta por Simonyan e Zisserman (2015), representou um avanço significativo na compreensão da importância da profundidade em redes neurais convolucionais. O modelo se destacou por sua arquitetura simples e homogênea, que utiliza exclusivamente filtros convolucionais pequenos de 3×3 empilhados sequencialmente. Essa abordagem contrasta com arquiteturas anteriores, como a *AlexNet*, que utilizavam filtros maiores (ex: 11×11 ou 5×5) nas camadas iniciais.

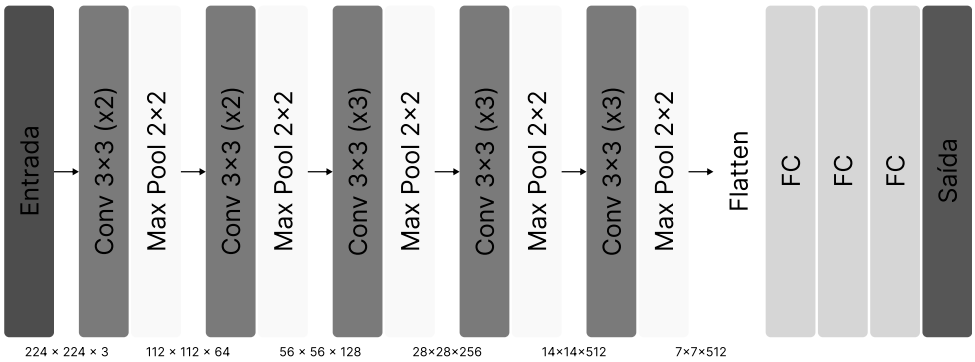
A principal contribuição da *VGGNet* foi demonstrar empiricamente que a profundidade e a uniformidade estrutural da rede são fatores cruciais para o desempenho. Ao empilhar múltiplas camadas convolucionais de 3×3, o modelo alcança um campo receptivo efetivo maior; por exemplo, três camadas consecutivas de 3×3 operam com campo equivalente a uma única de 7×7. Isso traz duas vantagens fundamentais: a introdução de mais não-linearidades, com a aplicação da função de ativação *ReLU* após cada convolução, e uma redução no número de parâmetros em comparação ao uso de uma única camada de filtro grande para o mesmo campo receptivo, tornando o treinamento mais eficiente.

Apesar de seu desempenho, que lhe garantiu o segundo lugar na competição *ILSVRC* de 2014, a *VGGNet* possui como desvantagem o alto custo computacional. Modelos como a VGG-16 contêm aproximadamente 138 milhões de parâmetros, limitando sua aplicação em cenários com restrição de recursos. Ainda assim, sua estrutura simples e a forte capacidade para extração de características a tornaram um *backbone* popular nos primórdios da estimativa de pose com técnicas de *deep learning*. Sua aplicação mais notável foi como *backbone* da arquitetura

OpenPose (CAO et al., 2017), uma das primeiras arquiteturas amplamente bem-sucedidas, que adaptou as primeiras camadas da *VGG-19* para extração de características.

A Figura 12 ilustra a arquitetura da VGG-16, uma das variantes mais populares juntamente com a VGG-19.

Figura 12 – Arquitetura simplificada da rede VGG-16.



Fonte: Elaborado pelo autor (2025).

2.4.1.3 *HRNet (High-Resolution Network)*

A *HRNet* (*High-Resolution Network*), proposta por Sun et al. (2019), é uma arquitetura desenvolvida especialmente para tarefas de localização espacial densa com precisão, como a estimativa de pose humana. Sua principal inovação é que, diferente de arquiteturas mais tradicionais como a *ResNet* e a *VGGNet*, que reduzem progressivamente a resolução espacial à medida que a profundidade da rede aumenta, a *HRNet* mantém uma representação de alta resolução ao longo de todo o processo de extração.

Sua estrutura é composta por múltiplos ramos paralelos que operam em diferentes resoluções, começando com um ramo principal de alta resolução; a cada novo estágio, é adicionado um novo ramo de resolução mais baixa (diminuindo pela metade). Ao final de cada estágio, é realizada uma operação de troca de informações denominada fusão multi-escala bidirecional, onde cada estágio se comunica com os adjacentes para melhorar a representação de alta resolução com informação proveniente dos ramos de baixa resolução (SUN et al., 2019).

A estratégia de fusão permite à rede combinar precisão espacial com contexto global, o que é particularmente útil na tarefa de estimativa de pose usando mapas de calor. Para a predição final, é utilizada apenas a saída do ramo de maior resolução, que já foi beneficiada com as informações dos outros ramos, e elimina a necessidade da operação de *upsampling* (SUN et al., 2019).

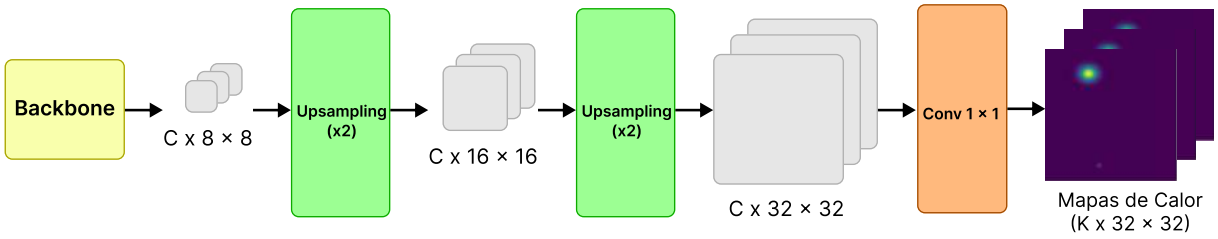
2.4.2 *Estimation Head*

A *estimation head* é o componente terminal das arquiteturas de estimativa de pose humana 2D. Sua principal função é transformar os mapas de características extraídos pelo *backbone* em predições de pontos anatômicos. Essa etapa final cumpre um papel análogo ao da camada de saída em modelos de classificação, onde o mapa de características final é convertido em

probabilidades. Dependendo da estratégia utilizada, essas predições podem assumir duas formas distintas: coordenadas (x, y) para cada ponto, ou mapas de calor que representam a distribuição espacial da probabilidade de localização dos pontos.

Em arquiteturas baseadas em mapas de calor, como apresentado na Figura 13, a *estimation head* é normalmente composta por camadas de *upsampling*, implementadas através de convoluções transpostas.

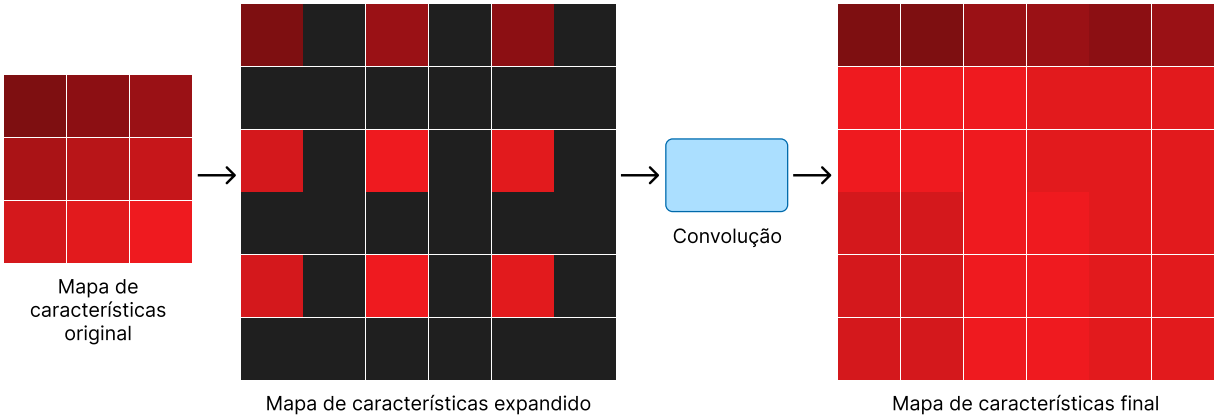
Figura 13 – Processo realizado por uma *estimation head* usando mapas de calor.



Fonte: Elaborado pelo autor (2025).

Diferentemente das operações de convolução tradicionais que costumam fazer uma redução da dimensão espacial (altura e largura), a convolução transposta realiza um aumento destas. No processo de convolução transposta, ilustrado na Figura 14, cada valor do mapa de características de entrada é expandido espacialmente através da inserção de zeros entre os elementos existentes, criando um mapa intermediário de maior resolução, porém com valores vazios. Este mapa vai então ser submetido a uma operação de convolução tradicional, onde filtros aprendíveis vão preencher os valores vazios, gerando o mapa de características final com a resolução aumentada.

Figura 14 – Ilustração do processo de convolução transposta.



Fonte: Elaborado pelo autor (2025).

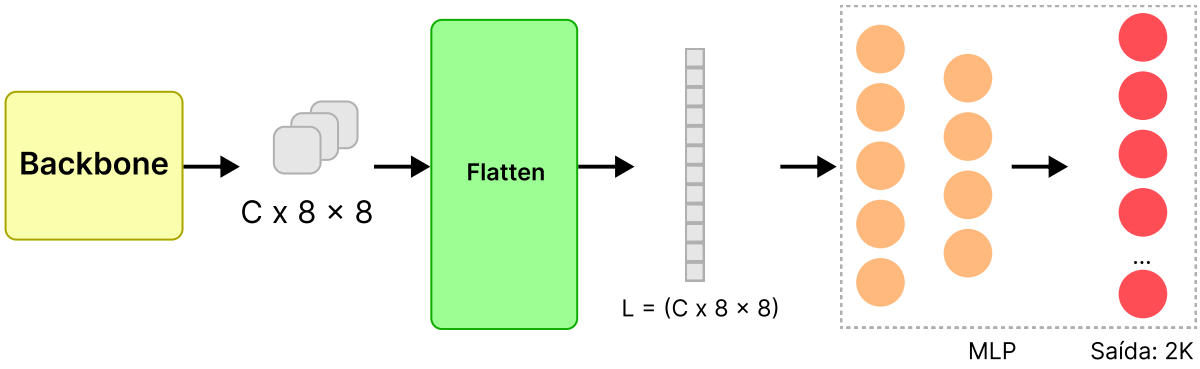
As camadas de *upsampling* restauram progressivamente a resolução espacial reduzida pelo *backbone* e são seguidas por alguma operação para redução do número de canais do mapa de características, de modo a gerar um mapa de calor por ponto anatômico. O resultado é um *tensor* $K \times H \times W$, onde K é o número de pontos, e cada canal encapsula a probabilidade da presença de um ponto em diferentes regiões da imagem.

Essa abordagem oferece maior precisão espacial e permite representar incertezas de maneira mais robusta, aparecendo em diferentes variações ao longo da literatura, como no modelo *SimpleBaseline*, proposto por Xiao, Wu e Wei (2018), onde a *estimation head* é composta por três camadas de deconvolução com *ReLU* e normalização de lote (*BatchNorm*), seguidas por uma convolução final para geração dos mapas de calor, ou na *HRNet*, proposta por Sun et al. (2019), onde os ramos de alta resolução mantêm a resolução espacial do mapa de características, e a estimativa final é gerada por uma única convolução para redução do número de canais, dispensando o uso de *upsampling* explícito.

Por outro lado, modelos baseados em regressão direta tratam o problema como uma tarefa de regressão supervisionada, prevendo diretamente os pares de coordenadas (x, y) correspondentes a cada ponto. Nessa configuração, o *backbone* é geralmente seguido por camadas totalmente conectadas, que transformam a representação extraída para um vetor plano de saída com $2K$ elementos, ilustrado na Figura 15.

Embora a estrutura da *estimation head* seja relativamente simples quando comparada ao restante da arquitetura, ela exerce um papel crucial na definição do tipo de saída, no modo de supervisão durante o treinamento e na forma como a precisão espacial é alcançada.

Figura 15 – Processo realizado por uma *estimation head* baseada em regressão direta.



Fonte: Elaborado pelo autor (2025).

2.5 Bancos de Dados e Métricas de Avaliação de Modelos de Estimativa de Pose 2D

O desenvolvimento e avaliação de modelos de estimativa de pose humana 2D requer conjuntos de dados anotados com precisão e métricas de avaliação objetivas e padronizadas. Estes componentes são essenciais para treinar os modelos e comparar seus desempenhos de forma justa. Nesta seção são apresentados alguns dos principais bancos de dados utilizados na literatura para desenvolvimento de modelos de estimativa de pose 2D; em seguida, são descritas as métricas de avaliação mais comuns para essa tarefa.

2.5.1 Bancos de Dados

Modelos de estimativa de pose humana 2D são tradicionalmente treinados e avaliados com bancos de dados de imagens com anotações detalhadas de pontos (pontos anatômicos) corporais. Esses conjuntos variam quanto ao número de pessoas por imagem, diversidade de posturas corporais representadas (como andar, sentar, correr, etc.), diversidade de cenário e formato das anotações.

2.5.1.1 *Microsoft Common Objects in Context* (MS COCO)

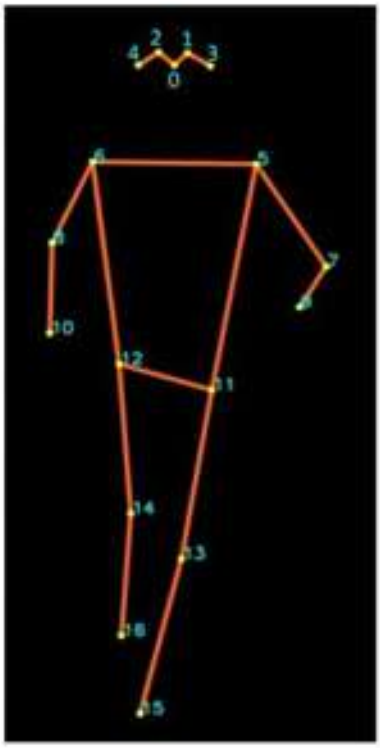
O banco de dados MS COCO (*Microsoft Common Objects in Context*) (LIN et al., 2015) é amplamente reconhecido como um dos *benchmarks* mais influentes e desafiadores para a tarefa de estimativa de pose humana 2D, principalmente devido à sua natureza *in-the-wild* (em ambientes não controlados, com grande variação de fundo, iluminação, poses e interações), necessária para o uso em aplicações reais (ANDRILUKA et al., 2014). Suas imagens retratam cenas cotidianas complexas, com grande variedade de fundos, condições de iluminação, poses e interações, tornando-o ideal para avaliar a robustez de modelos em cenários realistas.

A versão da base de dados MS COCO de 2017 é a mais utilizada atualmente e contém cerca de 118.000 imagens, com anotações para mais de 250.000 instâncias de pessoas. Cada instância pode ser anotada com até 17 pontos, que incluem articulações principais (como ombros, cotovelos e joelhos) e pontos faciais (nariz, olhos, orelhas), exibidos na Figura 16. Uma característica fundamental das anotações é a distinção explícita entre pontos visíveis e oclusos, o que é crucial para treinar e avaliar a capacidade do modelo em lidar com o desafio comum da oclusão.

A versão de 2017 conta com o conjunto primário de treino *train2017* e também disponibiliza outras partições, como a *val2017*, usada para testes dos modelos, e a *test2017*, que não possui anotações públicas e é destinada às submissões na plataforma oficial do desafio (COCO…, 2021). Essa organização também é empregada na versão original do banco de dados publicada em 2014, sendo esta pouco usada atualmente. É importante também destacar que esse banco de dados pode ser usado para tarefas de detecção, segmentação semântica, legendagem e outras; essa flexibilidade faz com que o MS COCO seja um dos conjuntos de dados mais utilizados e influentes no campo da visão computacional.

Figura 16 – Pontos anatômicos disponíveis no conjunto de dados MS COCO.

Índice	Ponto
0	Nariz
1	Olho esquerdo
2	Olho direito
3	Orelha esquerda
4	Orelha direita
5	Ombro esquerdo
6	Ombro direito
7	Cotovelo esquerdo
8	Cotovelo direito
9	Pulso esquerdo
10	Pulso direito
11	Quadril esquerdo
12	Quadril direito
13	Joelho esquerdo
14	Joelho direito
15	Tornozelo esquerdo
16	Tornozelo direito



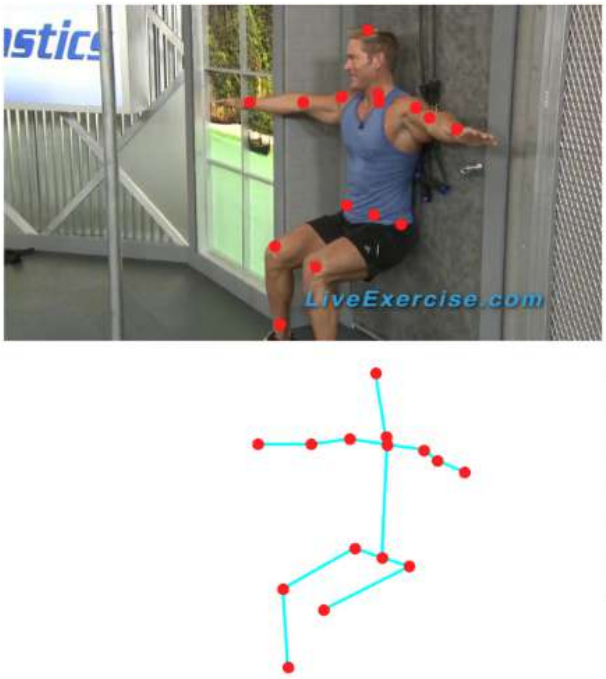
Fonte: Elaborado pelo autor (2025).

2.5.1.2 MPII Human Pose Dataset

A base de dados *MPII Human Pose* (ANDRILUKA et al., 2014), desenvolvida pelo *Max Planck Institute for Informatics*, é outro *benchmark* fundamental para a estimativa de pose humana, notório por sua rica diversidade de atividades. Diferente de bases de dados com poses mais controladas, o MPII é composto por aproximadamente 25.000 imagens extraídas de vídeos do *YouTube*, que retratam pessoas em uma vasta gama de contextos do mundo real, como esportes, trabalho e lazer. Essa origem garante uma grande variabilidade e complexidade nas posturas, tornando-o um teste robusto para a generalização dos modelos.

A base de dados contém anotações para cerca de 40.000 instâncias de pessoas, sendo que o subconjunto de treinamento possui aproximadamente 28.000 instâncias (ANDRILUKA et al., 2014). Cada instância é anotada com até 16 pontos, ilustrados na Figura 17, cobrindo as principais articulações do corpo. Além das coordenadas dos pontos, o MPII contém informações adicionais como indicadores de escala corporal e de oclusão para cada ponto.

Figura 17 – Pontos anatômicos disponíveis no conjunto de dados MPII.



Índice	Ponto
0	Tornozelo Direito
1	Joelho Direito
2	Quadril Direito
3	Quadril Esquerdo
4	Joelho Esquerdo
6	Pélvis
7	Tórax
8	Pescoço
9	Topo da Cabeça
10	Pulso Direito
11	Cotovelo Direito
12	Ombro Direito
13	Ombro Esquerdo
14	Cotovelo Esquerdo
15	Pulso Esquerdo

Fonte: Elaborado pelo autor (2025).

2.5.1.3 PoseTrack

Enquanto bases de dados como COCO e MPII focam primariamente na estimativa de pose em imagens estáticas, o PoseTrack (ANDRILUKA et al., 2017) é voltado para o processamento de vídeo, além de focar na tarefa de rastreamento de pose de múltiplas pessoas (*Multi-Person Pose Tracking*). Construído com base no formato de anotação do MPII, o PoseTrack é composto por sequências de vídeo densamente anotadas, extraídas de cenários complexos do mundo real.

O objetivo fundamental do PoseTrack é duplo: estimar com precisão a pose de cada indivíduo em um quadro (usando 15 pontos) e ser capaz de associar essas poses a uma identidade persistente ao longo de toda a sequência de vídeo. Isso o torna uma ferramenta indispensável para o desenvolvimento e a avaliação de algoritmos que possuem consistência temporal, uma propriedade crucial para aplicações práticas como análise de comportamento, interação humano-robô e vigilância inteligente. A base de dados captura explicitamente os desafios inerentes ao vídeo, como oclusões severas entre pessoas, desfoque de movimento, variações de escala e interações sociais complexas. A Figura 18 apresenta exemplos de quadros anotados do PoseTrack, onde múltiplas pessoas são identificadas e suas poses rastreadas ao longo de diferentes quadros.

Figura 18 – Exemplos de quadros anotados no conjunto de dados PoseTrack.

Fonte: (ANDRILUKA et al., 2017)

2.5.2 Métricas de Avaliação

A avaliação quantitativa do desempenho dos modelos de estimativa de pose humana requer métricas padronizadas que possam medir a precisão das coordenadas preditas em relação às anotações de referência. Neste trabalho, foi adotado o protocolo de avaliação padrão do *benchmark* MS COCO, que se baseia na métrica *Average Precision* (AP), calculada a partir da *Object Keypoint Similarity* (OKS).

2.5.2.1 *Object Keypoint Similarity* (OKS)

A *Object Keypoint Similarity* (OKS) é uma métrica fundamental para avaliar a proximidade entre um ponto predito e seu correspondente nos dados de referência. Diferentemente de uma simples distância Euclidiana, a OKS é uma medida normalizada que leva em consideração a escala do objeto (a pessoa), a dificuldade intrínseca de localizar diferentes tipos de pontos e a visibilidade dos pontos na anotação de referência. Por exemplo, erros na localização do nariz são penalizados mais severamente do que erros de mesma magnitude na localização do quadril, que é uma articulação com maior variabilidade natural.

A OKS para uma instância completa é definida como:

$$\text{OKS} = \frac{\sum_i \left[\delta(v_i > 0) \cdot \exp\left(\frac{-d_i^2}{2s^2k_i^2}\right)\right]}{\sum_i \delta(v_i > 0)} \tag{2.1}$$

Onde:

- d_i é a distância Euclidiana entre o ponto predito e o ponto de referência.
- s é a escala do objeto, calculada como a raiz quadrada da área do segmento do objeto na referência.
- k_i é uma constante para cada ponto que controla a penalidade de desvio. Essas constantes foram determinadas empiricamente a partir do desvio padrão das anotações no conjunto de dados COCO.
- v_i é o indicador de visibilidade do ponto i na anotação de referência.
- $\delta(v_i > 0)$ é a função delta de Kronecker que vale 1 quando o ponto está visível ou ocluso ($v_i > 0$), e 0 quando o ponto não está anotado ($v_i = 0$).

O valor da OKS para uma instância de pessoa completa varia de 0 (previsão completamente incorreta) a 1 (previsão perfeita). A inclusão do termo de visibilidade $\delta(v_i > 0)$ garante que apenas os pontos que estão presentes na anotação de referência sejam considerados no cálculo, tornando a métrica robusta para lidar com casos de oclusão onde alguns pontos anatômicos podem não estar visíveis ou anotados. O valor da OKS para uma instância de pessoa completa é a média dos valores de OKS de todos os seus pontos visíveis. O resultado varia de 0 (previsão completamente incorreta) a 1 (previsão perfeita).

2.5.2.2 *Average Precision* (AP)

A AP é a métrica primária utilizada para classificar e comparar os modelos. Ela sumariza a performance do modelo ao longo de diferentes níveis de dificuldade. O cálculo da AP se baseia na curva de *precision-recall*. Uma predição de pose para uma pessoa é considerada um Verdadeiro Positivo (TP) se seu valor de OKS for superior a um determinado limiar; caso contrário, é um Falso Positivo (FP).

Neste estudo, foram utilizadas as seguintes variantes da métrica AP, seguindo a convenção do desafio COCO:

- AP**:** É a métrica principal e mais robusta. Representa a média dos valores de AP calculados sobre dez limiares de OKS distintos, variando de 0.50 a 0.95 com incrementos de 0.05. Ela fornece uma avaliação abrangente do desempenho do modelo em um amplo espectro de exigências de precisão.
- $AP_{OKS=0.5}$ **(**AP_{50}**):** Calcula a AP utilizando um único e mais brando limiar de OKS, de 0.5. Esta métrica avalia a capacidade do modelo de realizar uma localização mais geral e aproximada dos pontos.

- $AP_{OKS=0.75}$ **(**AP_{75}**):** Calcula a AP utilizando um limiar de OKS mais rigoroso, de 0.75. Esta métrica é particularmente útil para avaliar a capacidade do modelo de localizar os pontos com alta precisão, sendo sensível a pequenos erros de localização.

Uma pontuação de AP mais alta indica um melhor desempenho geral do modelo. A análise dessas três métricas em conjunto permite uma compreensão mais completa das forças e fraquezas de cada abordagem de atenção avaliada.

2.6 Mecanismos de Atenção

A noção de atenção em redes neurais artificiais é inspirada em mecanismos da percepção visual humana. No cérebro humano, a atenção visual permite que o sistema nervoso central selecione de forma eficiente as regiões mais relevantes de uma cena, dedicando processamento mais profundo a essas áreas e ignorando estímulos menos significativos, processo esse essencial para lidar com a enorme quantidade de informações sensoriais recebidas continuamente.

Esse princípio é fundamentado por estudos clássicos da neurociência cognitiva. A *Feature Integration Theory*, proposta por Treisman e Gelade (1980), sugere que atributos visuais como cor, forma e orientação são processados separadamente e integrados em uma percepção coesa apenas quando há atenção focada. De maneira complementar, o modelo do foco atencional de Posner (1980) compara a atenção a um holofote que pode ser deslocado dinamicamente sobre o campo visual e a *Biased Competition Theory*, de Desimone e Duncan (1995), propõe que múltiplos estímulos competem por representação neural, e que a atenção atua para favorecer os mais relevantes. Todas essas ideias serviram de base para o desenvolvimento dos primeiros mecanismos de atenção em redes neurais artificiais, inicialmente aplicados a tarefas de natureza sequencial, como tradução automática.

Um marco importante se deu com o trabalho de Mnih et al. (2014) com o modelo *Recurrent Models of Visual Attention*, que introduziu o conceito de *hard attention*, uma forma não diferenciável de atenção que seleciona regiões específicas de uma imagem para análise, inspirada no comportamento de movimentos oculares humanos. Nesse modelo, uma rede recorrente explora a cena em etapas sucessivas, realizando *glimpses* (visões parciais) guiadas por uma política treinada via métodos de aprendizado por reforço. O principal avanço conceitual foi tratar a atenção como um mecanismo aprendível, capaz de decidir onde focar para realizar uma tarefa. Pouco tempo depois, Bahdanau, Cho e Bengio (2014) propuseram o modelo *Neural Machine Translation by Jointly Learning to Align and Translate*, que introduziu o mecanismo denominado *soft attention*. Em vez de selecionar uma única posição da entrada, como na *hard attention*, o modelo aprende a calcular uma distribuição contínua de pesos sobre todos os estados do *encoder*, permitindo que o *decoder* atenda a partes diferentes da entrada de forma fluida durante a geração da saída. Esse trabalho teve impacto seminal na área de tradução automática, resolvendo limitações dos modelos de codificação fixa e servindo de base para quase todos os avanços subsequentes em aplicações de processamento de linguagem natural com atenção.

Essa generalização dos mecanismos de atenção se ampliou com Xu et al. (2015), no trabalho *Show, Attend and Tell*, que adaptou o uso de atenção para a tarefa de descrição automática de imagens (*image captioning*). O modelo combina uma CNN extratora de características com um decodificador RNN (*Recurrent Neural Network*) que gera descrições em linguagem natural, aplicando atenção visual a regiões específicas da imagem a cada palavra gerada. Essa foi uma das primeiras aplicações bem-sucedidas de atenção em visão computacional, mostrando que a atenção poderia ser não só eficaz, mas também interpretável. Finalmente, o trabalho de Vaswani et al. (2017), *Attention Is All You Need*, consolidou o conceito com a introdução do *Transformer*, um modelo composto exclusivamente por blocos de atenção, sem usar operações de convolução ou de redes recorrentes. A arquitetura foi projetada inicialmente para tradução, mas sua eficiência, paralelismo e capacidade de modelar dependências globais o tornaram o novo paradigma para uma ampla gama de tarefas em processamento de linguagem, e posteriormente em visão computacional.

A partir dessas contribuições, mecanismos de atenção passaram a ser incorporados em diversas tarefas de visão computacional, incluindo a estimativa de pose humana. Na próxima seção será apresentada uma taxonomia funcional dos principais tipos de atenção, seguido então da seção com descrição detalhada de cada mecanismo avaliado experimentalmente neste trabalho.

2.6.1 Taxonomia dos Mecanismos de Atenção

Diversos mecanismos de atenção vêm sendo propostos para melhorar o desempenho de redes neurais profundas em tarefas de Visão Computacional, como classificação de imagens, detecção de objetos e estimativa de pose humana. Embora cada mecanismo tenha sua formulação específica, é possível agrupá-los com base na natureza da informação que modulam e na forma como essa modulação é aplicada. Essa organização não apenas ajuda a compreender suas diferenças estruturais, mas também facilita a escolha do tipo de atenção mais adequado para diferentes tarefas.

2.6.1.1 Atenção de Canal (*Channel Attention*)

Atua sobre os canais de um mapa de características, atribuindo pesos escalares a cada canal. Parte da premissa de que diferentes canais representam diferentes tipos de características visuais, como bordas, texturas ou partes do corpo, e que nem todas são igualmente relevantes em todos os contextos. Um exemplo clássico é o módulo *Squeeze-and-Excitation* (SE) de Hu, Shen e Sun (2017), que inspirou diversas variantes mais sofisticadas.

2.6.1.2 Atenção Espacial (*Spatial Attention*)

Modula diretamente as regiões espaciais da imagem (altura e largura), permitindo que a rede aprenda onde focar. É útil para destacar regiões importantes da imagem, como articulações ou

rostos, e se mostra especialmente eficaz em tarefas de localização densa, como a estimativa de pose.

2.6.1.3 Atenção Espaço-Temporal (*Spatio-Temporal Attention*)

Presente em tarefas que envolvem sequências de imagens, como vídeos. Esse tipo de atenção modela dependências espaciais e temporais simultaneamente, integrando informações de múltiplos quadros para tornar a predição mais robusta e consistente.

2.6.1.4 *Self-Attention*

Cada elemento da entrada interage com todos os outros, permitindo capturar relações de longo alcance globais. Essa atenção global é baseada na similaridade entre pares, e foi inicialmente proposta no contexto de tradução automática com o modelo *Transformer* (VASWANI et al., 2017). Em Visão Computacional, foi adaptada por meio de módulos como o *Non-Local Block* (WANG et al., 2018).

2.6.1.5 Atenção Não-Paramétrica ou Simples (*Parameter-Free Attention*)

Foca em simplicidade e eficiência computacional, dispensando convoluções adicionais e pesos treináveis. Um exemplo é o módulo SimAM de Yang et al. (2021), que utiliza uma função baseada em energia para atribuir pesos neurônio a neurônio, mantendo desempenho competitivo com baixo custo.

2.7 Mecanismos de Atenção Avaliados

2.7.1 CBAM – *Convolutional Block Attention Module*

O CBAM (*Convolutional Block Attention Module*) é um mecanismo de atenção leve e eficiente, proposto por Woo et al. (2018), que pode ser incorporado como um bloco independente em diversas arquiteturas convolucionais, com mínima modificação estrutural. O CBAM aplica atenção sequencialmente em duas dimensões distintas: canal e espaço. A ideia principal é permitir ao modelo que aprenda o que focar e onde focar, realçando tanto os canais relevantes quanto as regiões espaciais mais informativas da imagem. Esse módulo tem sua arquitetura dividida em dois submódulos:

2.7.1.1 Módulo de Atenção de Canal (*Channel Attention Module*)

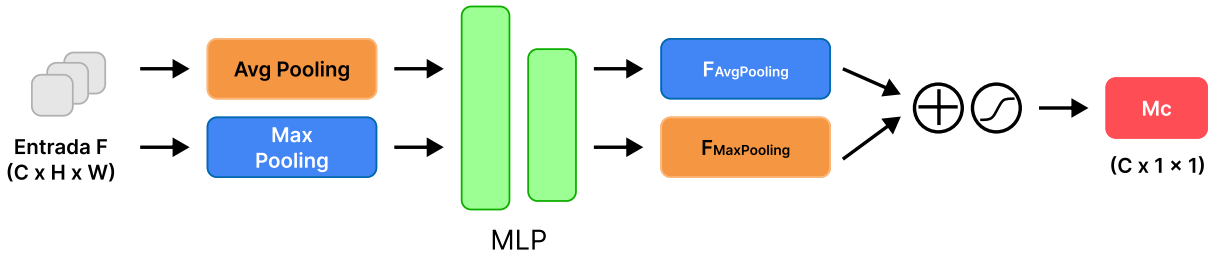
Responsável por gerar um vetor de pesos atencionais com tamanho igual ao número de canais de entrada, atribuindo uma importância relativa para cada canal. A intuição por trás disso é que cada canal codifica alguma característica, sendo assim possível calibrar quais características são mais relevantes. Para isso, conforme apresentado na Figura 19, são aplicadas duas operações de

pooling: *average pooling* e *max pooling*, ambas no eixo espacial, produzindo dois vetores de tamanho C $\times$ 1 $\times$ 1. Esses dois vetores são passados por uma rede MLP *(Multilayer Perceptron)* compartilhada com uma camada oculta, composta por duas camadas lineares com uma ativação ReLU intermediária. Em geral, a primeira camada reduz a dimensionalidade (com fator de redução r para os canais C $\rightarrow$ C/r), e a segunda a restaura. O resultado dos dois caminhos (operações de *average* e *max pooling*) é somado e submetido a uma função de ativação sigmoide, produzindo o vetor final de atenção de canal:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{2.2}$$

onde M_c é o mapa de atenção de canais gerado, que é multiplicado ponto a ponto com a entrada original F.

Figura 19 – Módulo de Atenção de Canal do CBAM.



Fonte: Elaborado pelo autor (2025).

2.7.1.2 Módulo de Atenção Espacial (*Spatial Attention Module*)

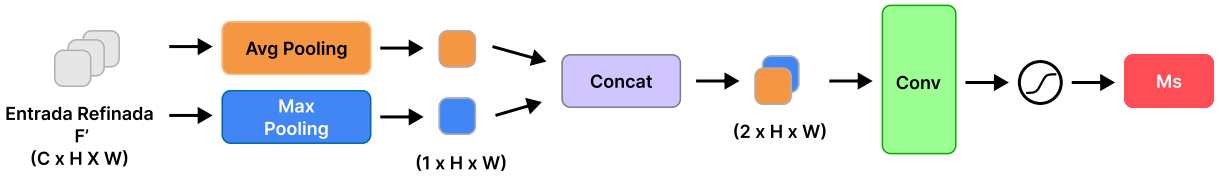
Após a aplicação da atenção na dimensão dos canais, o resultado é usado como entrada para o módulo de atenção espacial. Nesta etapa, conforme apresentado na Figura 20, é realizado *pooling* na dimensão dos canais, usando novamente *average pooling* e *max pooling*, resultando em dois mapas 2D de tamanho H $\times$ W. Esses mapas são concatenados e passados por uma convolução 2D 7$\times$7 com uma função de ativação sigmoide, resultando no mapa de atenção espacial:

$$M_s(F) = \sigma(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \tag{2.3}$$

que é então aplicado ao mapa de características F' (refinado pelo módulo de atenção de canal), permitindo ao modelo realçar regiões espacialmente relevantes da imagem.

O CBAM integra atenção de canal e espacial de forma sequencial, permitindo à CNN adaptar dinamicamente sua ênfase tanto em termos do tipo quanto da localização das informações extraídas. Essa estrutura foi projetada para ser leve e de simples uso, com custo computacional reduzido, o que a torna adequada para incorporação em arquiteturas convolucionais já estabelecidas. Os autores demonstraram empiricamente que a aplicação sequencial de atenção de canal seguida de atenção espacial produz melhores resultados do que abordagens paralelas ou inversas, o que reforça a importância da ordem na composição dos módulos (WOO et al., 2018). Em

Figura 20 – Módulo de Atenção espacial do CBAM.



Fonte: Elaborado pelo autor (2025).

virtude desse equilíbrio, entre eficácia e simplicidade, o CBAM tem sido amplamente adotado como módulo de atenção genérico em diversas tarefas de visão computacional, incluindo a estimativa de pose.

2.7.2 *Coordinate Attention*

O módulo de *Coordinate Attention* (Atenção Coordenada) foi proposto por Hou, Zhou e Feng (2021) como uma extensão leve e mais eficiente de mecanismos de atenção de canal, em especial ao implementado no módulo *Squeeze-and-Excitation* (SE) Hu, Shen e Sun (2017). A principal motivação dessa abordagem é tentar capturar informação nos eixos direcionais de forma eficiente, diminuindo a perda de informação espacial decorrente de operações de *pooling* convencionais, como a implementada no SE.

Enquanto mecanismos como o CBAM geram atenção espacial a partir de uma visão global da imagem, o *Coordinate Attention* preserva a direção da informação (altura ou largura) ao aplicar *pooling* separadamente, em apenas um eixo por vez. Isso permite que o módulo codifique informações espaciais em ambos os eixos.

O módulo de *Coordinate Attention*, conforme a Figura 21, gera os pesos de atenção em duas etapas principais:

- Incorporação de informação coordenada: Uma entrada $X \in \mathbb{R}^{C\times H\times W}$ é submetida a duas operações de *pooling* adaptativo em cada dimensão espacial: uma no eixo horizontal W, produzindo $F_h \in \mathbb{R}^{C\times H\times 1}$, e outra no eixo vertical H, produzindo $F_w \in \mathbb{R}^{C\times 1\times W}$. Esses dois tensores representam as respostas agregadas ao longo de cada eixo direcional, mantendo o alinhamento na outra dimensão espacial. Eles são então concatenados e passados por uma camada convolucional 1×1 com redução de dimensionalidade dada por um fator r ($C \to C/r$), seguida por uma uma ativação não-linear ReLU:

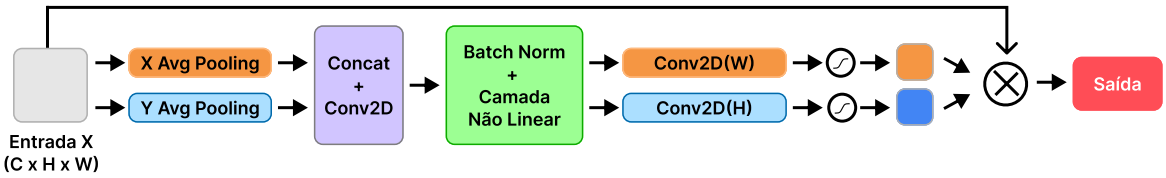
$$f = \delta(\text{Conv}_{1\times 1}([F_h; F_w])) \tag{2.4}$$

  O tensor intermediário $\mathbf{f} \in \mathbb{R}^{C/r\times 1\times (H+W)}$ é então separado novamente em dois componentes, $\mathbf{f}^h$ e $\mathbf{f}^w$, um para cada eixo.

- Geração de atenção: Na segunda etapa, cada um dos componentes é submetido a uma convolução 1×1 seguida por uma função de ativação sigmoide, produzindo dois mapas de atenção: $M_h \in \mathbb{R}^{C \times H \times 1}$ e $M_w \in \mathbb{R}^{C \times 1 \times W}$. Esses mapas são aplicados ao tensor de entrada original X com multiplicação ponto a ponto, efetivamente aplicando os pesos de atenção.

Ao preservar a informação direcional durante a etapa de agregação e projetar separadamente os mapas de atenção vertical e horizontal, essa técnica consegue capturar padrões espaciais distribuídos sem colapsar a estrutura espacial da imagem. Isso a diferencia de mecanismos tradicionais que perdem essa estrutura ao aplicar *pooling* global completo.

Figura 21 – Arquitetura do bloco de *Coordinate Attention*.



Fonte: Elaborado pelo autor (2025).

2.7.3 *Global Context Attention*

O *Global Context Attention* (GC Attention) foi proposto por Cao et al. (2019) como uma alternativa leve e eficiente ao *Non-Local Block* (Bloco Não-Local), originalmente introduzido por Wang et al. (2018). O *Non-Local Block* busca capturar relações de longo alcance entre regiões da imagem por meio da computação de similaridade entre todos os pares de posições espaciais. Embora eficaz, esse mecanismo apresenta duas limitações principais: seu custo computacional quadrático e a redundância nos mapas de atenção gerados que, segundo demonstrado por Cao et al. (2019), tendem a ser muito similares para diferentes canais. Com base nessa observação, o módulo de *Global Context* substitui o cálculo explícito de pares por uma única agregação global do conteúdo espacial, seguida por uma transformação que redistribui adaptativamente essa informação ao longo dos canais. Dessa forma, ele preserva a essência da atenção global, mas com complexidade linear e menor redundância, combinando os princípios de atenção de canal e de contexto espacial em um único bloco eficiente.

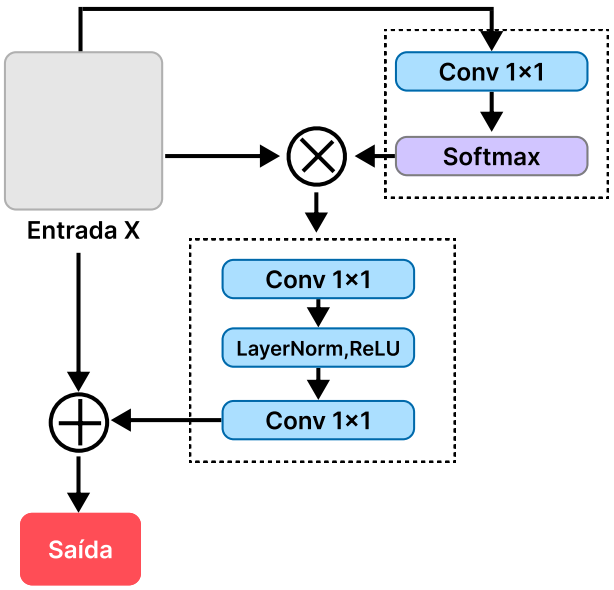
O processo de geração de atenção, ilustrado no diagrama da Figura 22, pode ser resumido em três etapas principais. A primeira é a de modelagem de contexto, onde o módulo de atenção busca identificar as regiões mais informativas da entrada, atribuindo pesos aprendidos:

- A entrada é projetada usando uma convolução 1x1, gerando um mapa escalar de *scores* de atenção;

- É gerado um mapa de atenção normalizado, aplicando a função *softmax* nesses *scores*;

- Esse mapa de atenção é então usado para realizar uma soma ponderada da entrada original, produzindo um vetor de contexto global que resume as informações relevantes da imagem.

Figura 22 – Arquitetura do bloco de Global Context Attention.



Fonte: Elaborado pelo autor (2025).

Após esta primeira etapa, ocorre a etapa de transformação, onde o vetor de contexto global obtido anteriormente é processado por um bloco do tipo *bottleneck*, composto por duas convoluções 1×1 intercaladas por uma função de ativação ReLU. A primeira convolução reduz o número de canais, criando uma representação intermediária mais compacta, enquanto a segunda expande novamente para a dimensão original. Essa operação visa refinar a informação contextual agregada, preparando-a para ser fundida com a entrada original na etapa final.

Na etapa final, o vetor de contexto transformado é fundido com o mapa de ativação original, geralmente por meio de uma operação de adição canal a canal. A utilização da conexão residual permite uma modulação suave da entrada com base na informação contextual aprendida, sem distorções agressivas, como as que poderiam ser introduzidas por multiplicações diretas.

Assim, o módulo captura relações de longo alcance entre diferentes regiões da imagem, sem depender de operações de similaridade entre pares de pixels (como no *Non-Local Block*). Em vez disso, ele aprende uma atenção global baseada em conteúdo, condensando a entrada em um vetor que redistribui sua informação para cada canal, configurando uma abordagem leve e escalável.

2.7.4 *Self-Attention*

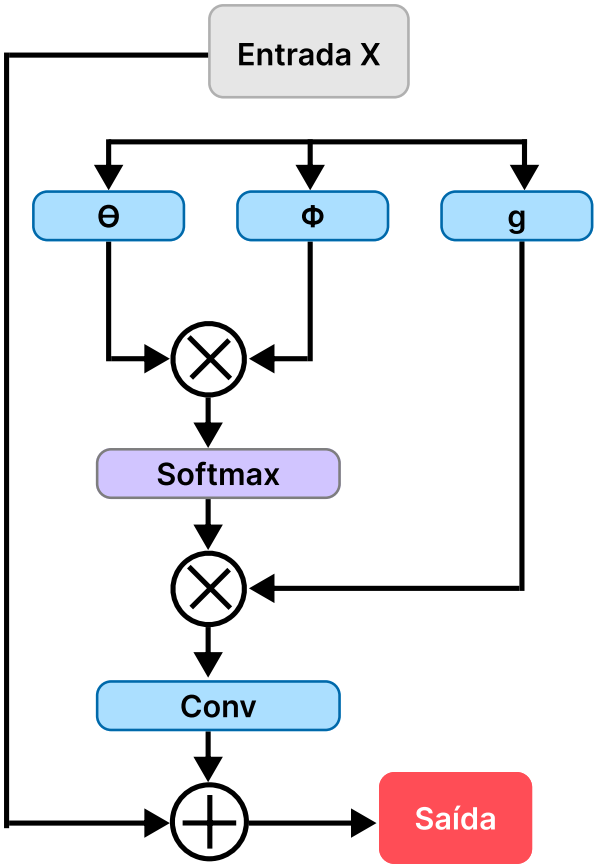
A ideia de *Self-attention* (auto-atenção) emergiu como um marco no aprendizado profundo com o trabalho seminal de Vaswani et al. (2017), que introduziu o *Transformer*, um modelo originalmente proposto para tarefas de processamento de linguagem natural. A proposta revolucionou o campo ao substituir estruturas recorrentes por operações de atenção global, permitindo que cada

elemento da entrada interaja diretamente com todos os demais por meio de ponderações baseadas em similaridade. Posteriormente, esse conceito foi adaptado ao domínio da visão computacional, originando mecanismos como o *Non-Local Block* (Bloco Não-Local), desenvolvido por Wang et al. (2018).

O *Non-Local Block* foi introduzido como uma forma de incorporar *self-attention* em redes neurais convolucionais. Enquanto as operações convolucionais tradicionais são restritas a regiões locais do campo receptivo, o *Non-Local Block* permite que cada ponto de uma imagem integre informações provenientes de qualquer outra posição espacial, o que é essencial para capturar dependências globais de longo alcance.

Conforme ilustrado no diagrama da Figura 23, dado um tensor de entrada X, três projeções lineares, $\theta(X)$, $\phi(X)$ e $g(X)$, são obtidas usando convoluções 1×1. Os mapas θ e ϕ são usados para calcular a similaridade entre todas as posições espaciais da imagem, gerando uma matriz de atenção $S \in \mathbb{R}^{N\times N}$, onde N é o número de posições espaciais ($H \times W$). Essa matriz é normalizada usando a função *softmax* e aplicada a $g(X)$, que representa as informações a serem agregadas. O resultado é um novo tensor com contexto global, transformado por uma convolução 1×1 para restaurar a dimensão original e somado à entrada com uma conexão residual.

Figura 23 – Arquitetura do bloco não-local.



Fonte: Elaborado pelo autor (2025).

Essa operação permite que a rede aprenda a reforçar conexões semânticas entre regiões

arbitrariamente distantes da imagem, o que se mostra especialmente vantajoso em tarefas como segmentação semântica, reconhecimento de ações e estimativa de pose humana, onde articulações ou regiões relevantes podem estar espacialmente separadas, mas semanticamente conectadas. Apesar de sua eficácia, o uso dos blocos não-locais é limitado por seu alto custo computacional e de memória, decorrente da necessidade de computar uma matriz de similaridade de tamanho quadrático em relação ao número de posições espaciais (WANG et al., 2018).

2.7.5 SimAM – A *Parameter-Free Attention Module*

O módulo SimAM (*Simple, Parameter-Free Attention Module*) foi proposto por Yang et al. (2021), como uma alternativa eficiente e biologicamente inspirada aos tradicionais mecanismos de atenção utilizados em redes convolucionais. Ao contrário de abordagens como CBAM e outras variantes que utilizam parâmetros aprendidos, o SimAM atribui pesos atencionais para cada unidade do mapa de características com base em uma formulação matemática livre de parâmetros, fundamentada em observações da neurociência cognitiva.

A motivação do SimAM surge a partir da constatação de que a atenção no cérebro humano não ocorre de maneira estritamente separada entre características (canais) e localizações (espaço), mas sim como um processo conjunto e distribuído. Carrasco (2011) revisa evidências comportamentais e neurofisiológicas que mostram que a atenção visual afeta simultaneamente tanto regiões específicas do espaço quanto atributos visuais distintos, como cor, orientação ou forma. Além disso, estudos clássicos como o de Hillyard, Vogel e Luck (1998), demonstram que a modulação atencional em cérebros de mamíferos frequentemente se manifesta como um ganho multiplicativo na resposta dos neurônios. Inspirado por essas observações, o SimAM aplica a atenção como um fator de escalamento local, e não como uma operação aditiva.

O ponto central da proposta é a formulação de uma função de energia associada a cada posição do mapa de características de entrada. Essa energia representa o custo de isolar uma posição específica, assumindo que posições mais importantes devem apresentar padrões de ativação distintos em relação à média da região onde estão inseridas. Em termos computacionais, os autores propõem que essa relevância pode ser modelada por meio da separabilidade linear entre uma posição e sua vizinhança, conceito que também está relacionado ao fenômeno de supressão espacial observado em redes neurais biológicas (WEBB et al., 2005).

Com base nessa ideia, os autores propõem a seguinte função de energia:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{2.5}$$

em que $\hat{\mu}$ e $\hat{\sigma}^2$ são, respectivamente, a média e a variância espacial do mapa de características no canal em que a posição t está inserida, e λ é um termo de regularização constante. Quanto maior o valor da energia, menor é a singularidade da posição em relação às demais.

Para aplicar a atenção, o valor inverso da energia é suavizado por uma função sigmoide, gerando um fator escalar contínuo entre 0 e 1 para cada posição. Esse fator é então aplicado

diretamente ao mapa de características original por multiplicação elemento a elemento:

$$\tilde{X} = \sigma\left(\frac{1}{E}\right) \odot X \tag{2.6}$$

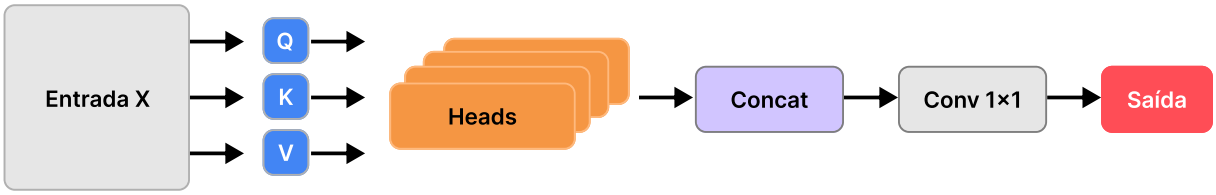
Esse processo simples, porém eficaz, permite que o módulo destaque automaticamente as posições mais informativas, com custo computacional mínimo e sem necessidade de parâmetros treináveis. A ausência de convoluções adicionais e de pesos aprendidos torna o SimAM especialmente interessante para redes compactas e aplicações com restrições de memória ou energia, como dispositivos móveis. Apesar de sua leveza e da fundamentação teórica sólida, o SimAM também apresenta limitações. Por não incorporar mecanismos aprendidos, ele não modela relações contextuais mais complexas nem interações entre canais, o que pode limitar seu impacto em redes mais profundas ou em tarefas que exigem raciocínio visual sofisticado. Ainda assim, sua proposta representa uma contribuição relevante para o design de módulos atencionais eficientes e interpretáveis.

2.7.6 *Multi-Head Attention*

O mecanismo de *Multi-Head Attention* (MHA) foi introduzido por Vaswani et al. (2017) como o componente central da arquitetura *Transformer*. Essa técnica revolucionou o campo de aprendizado profundo ao demonstrar que tarefas de processamento de sequências, como tradução automática, poderiam ser resolvidas com grande eficácia sem o uso de redes recorrentes ou convolucionais, apenas com operações de *self-attention* em múltiplas representações paralelas. Posteriormente o MHA também passou a ser adotado no campo da visão computacional, incorporado em arquiteturas como *Vision Transformers* (DOSOVITSKIY et al., 2021) e híbridos CNN-*Transformer*, por sua capacidade de capturar relações globais e estruturais com grande flexibilidade.

O princípio fundamental do MHA é permitir que o modelo avalie diferentes padrões de dependência entre elementos da entrada por meio de projeções lineares independentes, denominadas *heads*. Cada *head* executa uma instância separada de *self-attention*, e os resultados são concatenados e reprocessados, permitindo que a rede aprenda diferentes padrões de dependência entre elementos espaciais de forma paralela. Formalmente, conforme apresentado no diagrama de alto nível da Figura 24, dada uma entrada X, são geradas três projeções lineares: *queries* $Q = XW^Q$, *keys* $K = XW^K$ e *values* $V = XW^V$.

Figura 24 – Funcionamento geral do mecanismo de MHA.



Fonte: Elaborado pelo autor (2025).

Em cada cabeça de atenção (*Attention head*) é calculada a compatibilidade entre as *queries* e as *keys* através do produto escalar, seguido de normalização com *softmax*:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.7}$$

Onde d_k é a dimensão das *keys*. O número de *heads* h define quantas dessas operações são realizadas em paralelo, com projeções distintas. Ao final, os resultados de todas as cabeças são concatenados e transformados por uma última projeção linear, restaurando a dimensionalidade original da entrada:

$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.8}$$

Em tarefas de visão computacional, essa operação é aplicada sobre vetores extraídos de imagens. Em arquiteturas como os *Vision Transformers* (ViT), por exemplo, a imagem é dividida em sub-imagens (*patches*), que são linearizadas e tratadas como *tokens*, de maneira análoga às palavras em modelos de linguagem. Já em arquiteturas híbridas, como os blocos usados neste trabalho, o MHA pode ser aplicado sobre mapas de características extraídos por CNNs, convertendo regiões de interesse em sequências para capturar dependências de longo alcance entre elas.

A principal vantagem do MHA é sua capacidade de modelar relações contextuais globais sem depender da estrutura local imposta por convoluções. Como cada cabeça pode aprender a focar em padrões diferentes, o modelo obtém uma representação mais rica e diversificada da entrada. No entanto, essa expressividade vem com um custo: o MHA é computacionalmente intensivo, especialmente quando aplicado sobre entradas espaciais grandes, como imagens de alta resolução, e requer um número significativo de parâmetros, o que pode impactar o desempenho em ambientes restritos.

Apesar dessas limitações, o *Multi-Head Attention* permanece como um dos mecanismos mais poderosos e amplamente estudados na literatura, sendo utilizado tanto em arquiteturas puramente baseadas em atenção quanto em combinação com convoluções, como é o caso dos modelos explorados neste trabalho.

3 Metodologia

Este capítulo apresenta a metodologia adotada para investigar, de forma sistemática e comparativa, a contribuição de mecanismos de atenção para a tarefa de estimativa de pose humana 2D. Com o objetivo de garantir uma análise justa e controlada, foi utilizada uma arquitetura baseada na *SimpleBaseline* (XIAO; WU; WEI, 2018) como base para todos os modelos de estimativa desenvolvidos. Nessa arquitetura, baseada na *ResNet-50*, foram integrados e avaliados seis mecanismos de atenção distintos: *CBAM*, *Coordinate Attention*, *Global Context Attention*, *Self-Attention*, *SimAM* e *Multi-Head Attention*. Para isolar o impacto de cada técnica, todos os módulos de atenção foram inseridos no mesmo ponto da arquitetura base. Além disso, os experimentos de treinamento e avaliação foram conduzidos no conjunto de dados *MS COCO* sob um protocolo único, com os mesmos hiperparâmetros e ambiente computacional para todas as variantes do modelo, garantindo assim uma comparação imparcial e confiável entre os diferentes mecanismos de atenção.

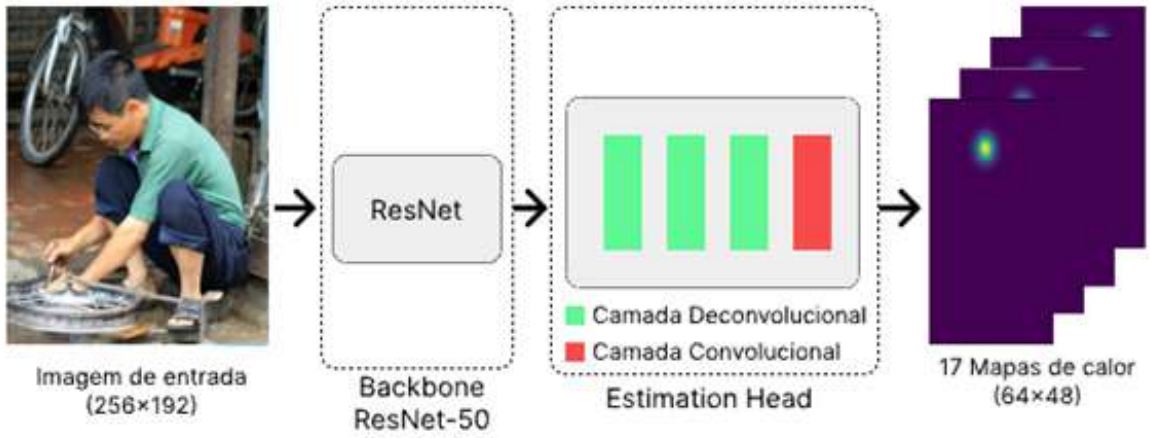
3.1 Arquitetura Base

A Figura 25 apresenta o modelo de rede neural base adotado neste trabalho para estimativa de pose 2D, baseado na arquitetura *SimpleBaseline*, proposta por Xiao, Wu e Wei (2018). Esta arquitetura foi selecionada por ser uma arquitetura padrão de referência amplamente utilizada em estimativa de pose humana 2D, combinando simplicidade estrutural com desempenho robusto. Sua natureza modular a torna um ponto de partida ideal para a avaliação isolada de componentes adicionais, como os mecanismos de atenção aqui estudados.

A arquitetura *SimpleBaseline* segue o paradigma *encoder-decoder*, dividido em duas partes principais:

- ***Backbone (encoder)*:** Este módulo é responsável pela extração de características visuais da imagem de entrada de dimensão 256×192. O módulo é composto por uma rede *ResNet-50* (HE et al., 2015) pré-treinada no conjunto de dados *ImageNet*, com a camada de classificação final removida. A rede processa a imagem de entrada e gera um mapa de características de alta dimensionalidade semântica e baixa resolução espacial, com dimensões de 2048 × 8 × 6 (canais × altura × largura).
- ***Estimation Head (decoder)*:** Este módulo tem a função de decodificar o mapa de características gerado pela *Backbone*, para então gerar as predições de pose. É composto por três camadas de deconvolução, que realizam o *upsampling* espacial do mapa de características, restaurando a resolução. Ao final, uma camada convolucional de 1×1 projeta a saída para gerar 17 mapas de calor, um para cada ponto anatômico definido pelo conjunto de dados *MS COCO*.

Figura 25 – Diagrama da arquitetura sem atenção utilizada como modelo base.



Fonte: Elaborado pelo autor (2025).

3.2 Integração dos Módulos de Atenção

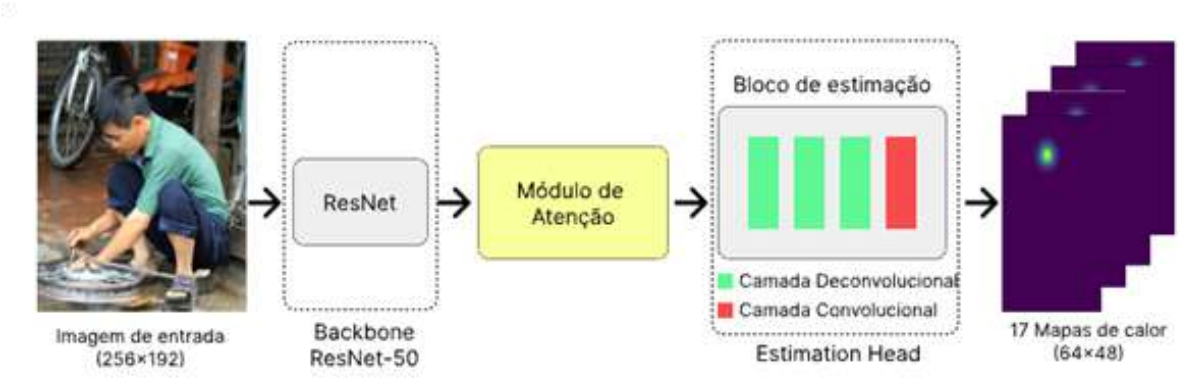
Com o objetivo de garantir uma avaliação justa e isolada dos mecanismos de atenção, conforme ilustrado na Figura 26, todos os módulos de atenção foram integrados ao modelo base no mesmo ponto da arquitetura: imediatamente após o *backbone ResNet-50* e antes da *estimation head*. Essa escolha garante que os módulos atuem sobre o mapa de características extraído pela *ResNet-50*, permitindo que a atenção module essas ativações antes que sejam decodificadas em mapas de calor. A escolha dos módulos buscou representar diferentes categorias funcionais de atenção (canal, espacial, *self-attention* e outros), com base em sua relevância na literatura e aplicabilidade em tarefas visuais. Os módulos de atenção avaliados durante os experimentos foram: *CBAM*, *Coordinate Attention*, *Global Context Attention*, *Self-Attention*, *SimAM* e *Multi-Head Attention*.

A estrutura geral do modelo permaneceu inalterada em todos os experimentos, com exceção do bloco de atenção inserido. Isso assegura que as comparações entre os modelos possam refletir exclusivamente o impacto dos mecanismos avaliados, sem interferência de variações na arquitetura, número de parâmetros ou profundidade da rede.

3.3 Protocolo Experimental

Para assegurar a reprodutibilidade dos experimentos e a validade das comparações entre os modelos, foi estabelecido um protocolo experimental consistente. Esta seção detalha os três principais componentes deste protocolo: definição e pré-processamento do conjunto de dados, configuração de treinamento e ambiente computacional.

Figura 26 – Diagrama do modelo base com a integração do módulo de atenção.



Fonte: Elaborado pelo autor (2025).

3.3.1 Definição e pré-processamento do conjunto de dados

Todos os modelos foram treinados e avaliados utilizando o banco de dados *MS COCO 2017*, especificamente a partição do banco de imagens com anotações de pontos anatômicos. A divisão dos dados seguiu uma metodologia comum para *benchmarks* da área, com os valores específicos apresentados na Tabela 2, sendo:

- **Treinamento e Validação:** O conjunto oficial *train2017* do *MS COCO 2017*, com 118.287 imagens, foi utilizado para o treinamento. Uma divisão de 80% deste conjunto foi usada para o ajuste dos pesos dos modelos, enquanto os 20% restantes serviram como conjunto de validação para monitoramento da convergência e para os mecanismos de ajuste da taxa de aprendizado e *early stopping*.
- **Teste:** O conjunto oficial *val2017* do *MS COCO 2017*, composto por 5.000 imagens, foi utilizado exclusivamente para teste e comparação final do desempenho dos modelos.

Tabela 2 – Partição de dados empregada no desenvolvimento dos modelos.

Subconjunto	**Número de imagens**
Treinamento	94.630
Validação	23.657
Teste	5.000

Fonte: Elaborado pelo próprio autor.

Com o objetivo de aumentar a robustez e a capacidade de generalização dos modelos, foi aplicado um protocolo de aumento de dados (*data augmentation*) durante a fase de treinamento. Cada imagem de entrada de treino foi submetida às seguintes transformações aleatórias:

- **Escalonamento:** Variação aleatória da escala da imagem na faixa de ±30%.

- **Rotação:** Rotação aleatória da imagem em um ângulo de ±40°.
- **Espelhamento Horizontal:** Aplicação de um espelhamento horizontal com 50% de probabilidade.

3.3.2 Configuração de Treinamento

A implementação de todos os modelos foi realizada na linguagem de programação *Python*, utilizando as bibliotecas *Numpy* (HARRIS et al., 2020), *Matplotlib* (HUNTER, 2007) e *PyTorch* (PASZKE et al., 2019). A configuração dos hiperparâmetros de treinamento, detalhada na Tabela 3, foi mantida igual para todos os experimentos.

Tabela 3 – Hiperparâmetros de treinamento utilizados para todos os modelos.

Parâmetro	**Valor**
Tamanho da imagem de entrada	256×192 pixels
Tamanho dos mapas de calor	64×48 pixels
Otimizador	*Adam*
Weight Decay	1×10^{-4}
Taxa de aprendizado inicial	1×10^{-3}
Agendador de Taxa de Aprendizado	Redução de 0.5 após 8 épocas sem melhora
Critério de Parada Antecipada	20 épocas
Batch size	128 imagens
Função de Perda	Erro Quadrático Médio (MSE)

Fonte: Elaborado pelo próprio autor.

3.3.3 Ambiente Computacional

Todos os experimentos foram executados em um ambiente computacional unificado para garantir a consistência, utilizando uma única GPU *NVIDIA GeForce RTX 4070* com 12 GB de VRAM. Para otimizar o uso de memória e acelerar o processo de treinamento, foi empregada a técnica de precisão mista de ponto flutuante (*mixed precision*). Além disso, a reprodutibilidade foi assegurada pela fixação da mesma semente aleatória (*random seed*) no início de cada execução, garantindo assim resultados consistentes e comparáveis.

3.4 Extração de Pontos Anatômicos

A saída gerada pelos modelos baseados em mapas de calor consiste em um conjunto de mapas de probabilidade, e não em coordenadas diretas dos pontos anatômicos. A conversão desses mapas em localizações (x, y) é uma etapa crítica que influencia significativamente o desempenho final do modelo de estimativa de pose. Uma abordagem simplificada, que consiste em apenas extrair a localização do valor máximo de cada mapa, tende a ser subótima e sensível a

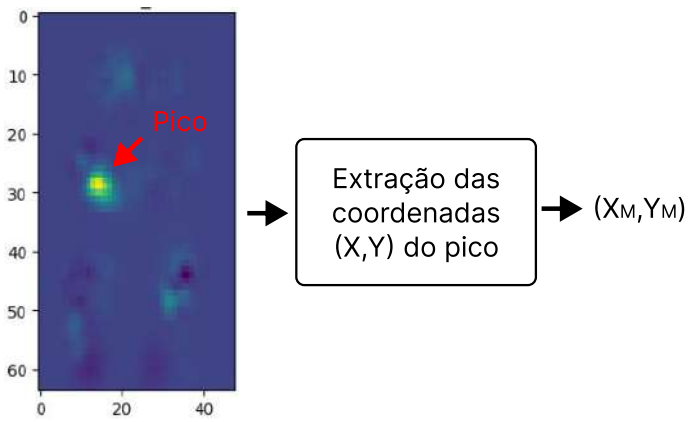
pequenas variações. Para obter predições mais robustas e precisas, foi implementado um *pipeline* de pós-processamento de múltiplos estágios, baseado em trabalhos com propostas semelhantes (CAI et al., 2020; CHEN et al., 2017; XIAO; WU; WEI, 2018). Este processo, ilustrado na Figura 29, é realizado de acordo com as seguintes etapas:

- ***Test-Time Augmentation* (TTA) e Agregação:** Para aumentar a robustez das predições, a inferência não é realizada apenas na imagem original. Uma versão espelhada horizontalmente da imagem também é fornecida como entrada para o modelo. Os mapas de calor resultantes desse espelhamento são então revertidos para corresponder à orientação original. Em seguida, os mapas de calor da imagem original e da imagem espelhada são agregados através de uma média elemento a elemento. Esta técnica ajuda a mitigar erros e a produzir uma estimativa de probabilidade mais estável.

- **Suavização com Desfoque Gaussiano:** Após a agregação, um filtro de desfoque gaussiano (com $\sigma = 2$ e um *kernel* de 3×3) é aplicado a cada mapa de calor resultante. O objetivo desta etapa é suavizar a distribuição de probabilidade, reduzindo o impacto de ruídos ou de picos de ativação errôneos e criando uma superfície mais contínua, o que facilita a localização precisa do pico de máxima confiança.

- **Detecção do Pico Máximo:** A localização primária de cada ponto anatômico p_i é determinada encontrando-se a coordenada com o valor de ativação máximo em seu respectivo mapa de calor suavizado. Este passo fornece uma estimativa inicial da localização do ponto com precisão de pixel. A Figura 27 ilustra um mapa de calor de saída utilizado para obtenção das coordenadas do pico.

Figura 27 – Processo de detecção de pico em mapa de calor.

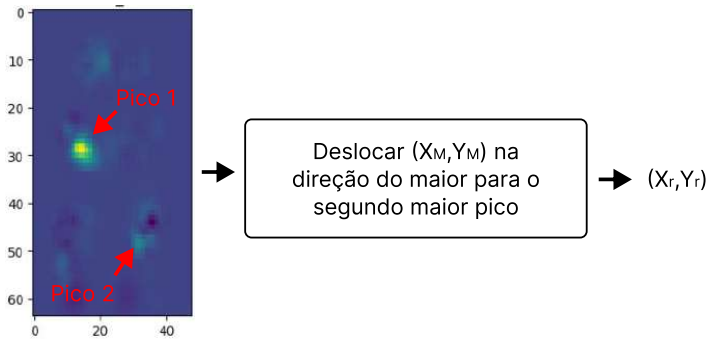


Fonte: Elaborado pelo autor (2025).

- **Refinamento de Localização Sub-pixel:** Dado que a localização real de um ponto raramente coincide perfeitamente com o centro de um pixel, um passo de refinamento é

aplicado para melhorar a precisão sub-pixel. As coordenadas p_i são ajustadas por um deslocamento de um quarto de pixel na direção do segundo pico de ativação mais alto no mapa de calor, como demonstrado na Figura 28. Este ajuste sutil, embora pequeno, é importante para melhorar a precisão em métricas rigorosas como a AP_{75} (NEWELL; YANG; DENG, 2016).

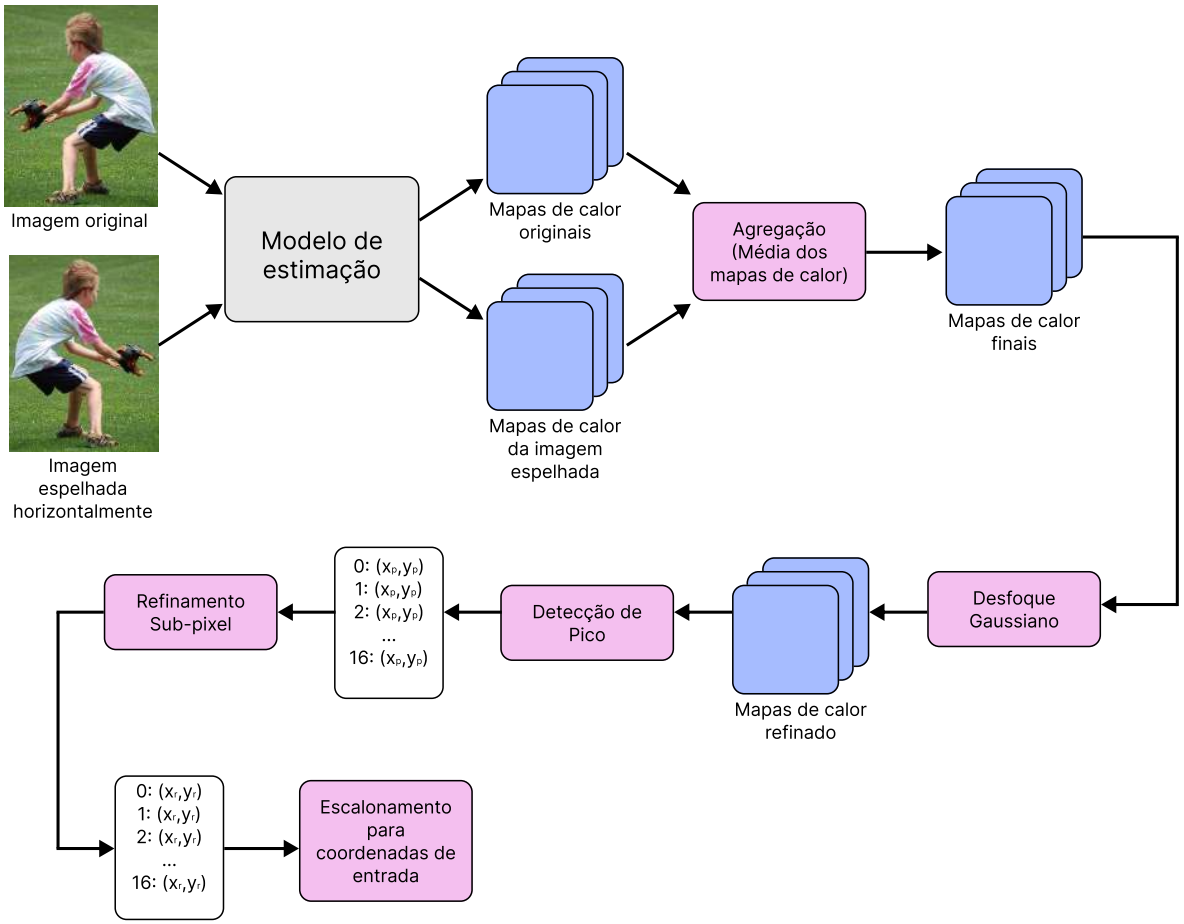
Figura 28 – Ilustração do processo de refinamento sub-pixel.



Fonte: Elaborado pelo autor (2025).

- **Escalonamento Final das Coordenadas:** Por fim, as coordenadas refinadas, que estão no sistema de coordenadas do mapa de calor (ex: 64×48), são linearmente reescalonadas para o sistema de coordenadas da imagem de entrada original (ex: 256×192). Este passo produz as coordenadas finais que serão utilizadas para a avaliação do desempenho do modelo.

Figura 29 – Pipeline de pós-processamento para extração de coordenadas a partir dos mapas de calor.



Fonte: Elaborado pelo autor (2025).

4 Resultados e Discussões

Este capítulo apresenta e analisa os resultados obtidos a partir dos experimentos descritos na metodologia. A avaliação quantitativa é apresentada inicialmente, com uma discussão detalhada sobre o desempenho de cada mecanismo de atenção seguida por uma análise do custo computacional.

4.1 Avaliação de Desempenho

O desempenho de cada modelo, avaliado no conjunto de teste de 5.000 imagens (*val2017*), é sumarizado na Tabela 4. A tabela apresenta os resultados para as métricas *Average Precision* (AP), AP_{50} e AP_{75} para o modelo base e para as seis variantes que integram diferentes mecanismos de atenção.

Tabela 4 – Desempenho dos modelos no conjunto de teste.

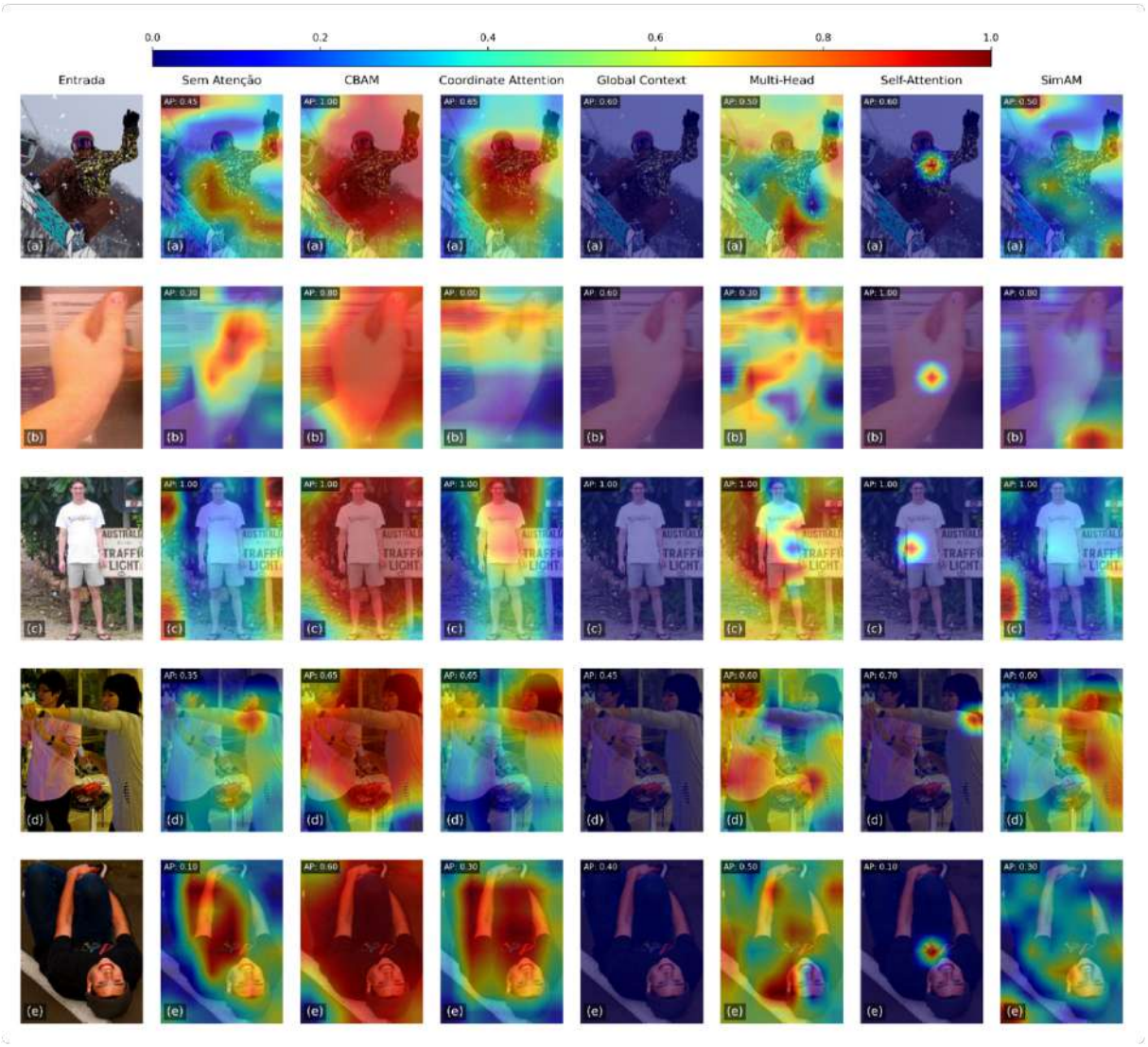
Modelo	**AP (%)**	**AP_{50} (%)**	**AP_{75} (%)**
Base (Sem atenção)	66.2	88.0	72.5
Base + CBAM	67.6	88.2	74.7
Base + *Coordinate Attention*	67.7	89.2	73.7
Base + *Global Context Attention*	66.8	88.2	73.5
Base + *Self-Attention*	66.6	88.1	72.6
Base + *Multi-Head Attention*	67.4	89.1	73.6
Base + SimAM	66.4	88.1	72.6

Fonte: Elaborado pelo próprio autor.

Os resultados quantitativos, expostos na Tabela 4, indicam que a integração de mecanismos de atenção pode melhorar a performance na tarefa de estimativa de pose. Contudo, a magnitude desse aprimoramento varia entre as diferentes abordagens, revelando uma sensibilidade da tarefa ao tipo de informação que cada mecanismo prioriza.

Para investigar como os mecanismos de atenção afetam o foco do modelo, foi feita uma análise dos mapas de ativação gerados por cada um para algumas amostras do subconjunto de testes. Para tal, foi empregada uma técnica de visualização baseada no método *Class Activation Mapping* (CAM) (ZHOU et al., 2015). Para o modelo sem atenção, o mapa de ativações foi gerado a partir da média das ativações da última camada do *backbone*, representando a atenção implícita do modelo. Para as variantes, foi visualizado o mapa de atenção explícito gerado pelo próprio módulo. A Figura 30 apresenta os mapas de ativação gerados para algumas imagens de teste. Este processo permite diagnosticar e comparar o foco espacial aprendido por cada arquitetura.

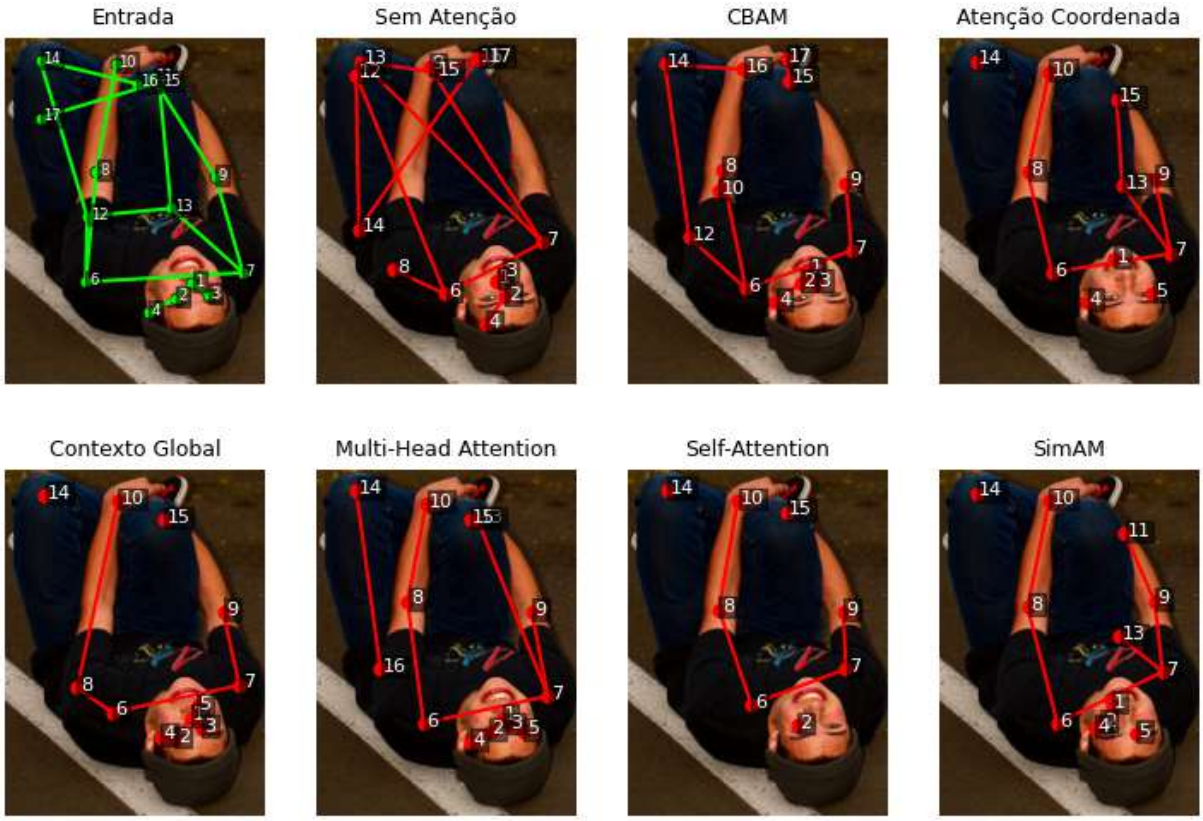
Figura 30 – Mapas de ativação para imagens de teste.



Fonte: Elaborado pelo autor (2025).

A Figura 31 apresenta as poses reconstruídas a partir das predições de cada modelo para a imagem de teste (e) da Figura 30, permitindo uma análise detalhada do resultado das estimativas. Adicionalmente, a Tabela 5 detalha o erro euclidiano para cada um dos pontos estimados nesta mesma imagem.

Figura 31 – Poses reconstruídas para a imagem de teste (e) da Figura 30.



Fonte: Elaborado pelo autor (2025).

Tabela 5 – Erro euclidiano (em pixels) das predições de cada modelo para a imagem de teste (e), por ponto.

Ponto	**Base**	**CBAM**	**Coord. Attention**	**Global Context**	**MHA**	**Self-Attention**	**SimAM**
1: Nariz (v=2)	5.32	11.10	20.05	2.28	2.28	N/D	13.53
2: Olho Esquerdo (v=2)	12.70	10.62	N/D	7.52	1.56	2.92	8.64
3: Olho Direito (v=2)	20.88	13.68	N/D	4.78	8.61	N/D	N/D
4: Orelha Esquerda (v=2)	20.95	7.88	7.88	10.46	7.16	N/D	14.45
5: Orelha Direita (v=0)	–	N/A	N/A	N/A	N/A	N/A	N/A
6: Ombro Esquerdo (v=2)	36.71	28.04	24.59	19.61	20.81	32.34	28.47
7: Ombro Direito (v=2)	23.85	19.43	14.96	14.45	7.80	30.06	26.47
8: Cotovelo Esquerdo (v=2)	73.19	2.38	2.23	64.07	1.53	4.48	3.30
9: Cotovelo Direito (v=2)	108.94	7.24	3.64	6.17	4.44	4.44	6.48
10: Pulso Esquerdo (v=2)	N/D	95.13	6.03	4.19	6.03	4.05	4.05
11: Pulso Direito (v=2)	N/D	N/D	N/D	N/D	N/D	N/D	33.56
12: Quadril Esquerdo (v=2)	109.10	21.85	N/D	N/D	N/D	N/D	N/D
13: Quadril Direito (v=2)	140.74	N/D	20.87	N/D	88.73	N/D	10.89
14: Joelho Esquerdo (v=2)	125.02	3.27	1.76	4.56	3.69	3.69	1.76
15: Joelho Direito (v=2)	27.77	5.52	21.28	5.50	4.47	3.65	N/D
16: Tornozelo Esquerdo (v=2)	28.61	16.16	N/D	N/D	124.07	N/D	N/D
17: Tornozelo Direito (v=2)	108.72	101.44	N/D	N/D	N/D	N/D	N/D

Fonte: Elaborado pelo próprio autor.

A seguir, cada modelo é analisado individualmente, discorrendo sobre os fatores que podem contribuir para o desempenho observado.

4.1.1 *Coordinate Attention*

O modelo com mecanismo de *Coordinate Attention* apresentou a melhor performance dentre as abordagens avaliadas neste estudo, alcançando o maior *score* na métrica principal AP, com um ganho de 1,5 p.p. sobre o modelo base (pontos percentuais), atingindo um valor de 67,7%, e na métrica secundária de detecção geral AP_{50} chegando a 89,2% (aumento de 1,2 p.p. sobre o modelo base). Contudo, este resultado pode ser entendido como um empate técnico com o mecanismo CBAM, separados por uma diferença de somente 0,1 p.p.. A eficácia dessa abordagem pode ser atribuída a seu viés espacial explícito, que é inerentemente benéfico a tarefas de localização espacial como a estimação de pose 2D. Conforme proposto por Hou, Zhou e Feng (2021), o mecanismo foi projetado para capturar dependências de longo alcance com informações posicionais precisas ao longo de ambos os eixos direcionais, uma característica fundamental para a compreensão da estrutura do corpo humano.

Na Figura 30, de maneira consistente, é possível observar, nas imagens selecionadas da base de teste, que o *Coordinate Attention* gera um mapa de ativação mais concentrado e preciso sobre a região da pessoa em comparação ao modelo sem atenção. Este refinamento espacial pode ser

interpretado como a causa direta da melhoria de desempenho observada, pois permite que o modelo dê mais atenção à região mais importante da imagem.

Além disso, analisando o impacto desse mecanismo no erro das predições para a imagem teste selecionada, a Tabela 5 revela que, com exceção do ponto 1 (nariz), este mecanismo proporcionou uma redução substancial de erro para todos os outros pontos detectados, sendo capaz de corrigir até mesmo erros estruturais grosseiros, como no ponto 14 (joelho esquerdo), onde o modelo base apresentou uma falha grave de localização, com um erro de 125 pixels, reduzido pelo mecanismo de atenção para somente 1,76 pixels. Isto reforça que, apesar das imperfeições, a implementação da atenção pode permitir melhorias de maneira significativa na pose reconstruída.

4.1.2 CBAM

O modelo com mecanismo de atenção CBAM também propiciou um desempenho competitivo, com um AP de 67,6% (ganho de 1,4 p.p. em relação ao modelo base), resultado muito próximo ao do mecanismo de *Coordinate Attention*, conforme já citado. Analisando a métrica secundária de alta precisão AP_{75}, o CBAM se destacou com a melhor performance dentre todos os modelos, alcançando 74,7%, representando um ganho de 2,2 p.p. em relação ao modelo base, conforme apresentado na Tabela 4. De maneira análoga ao *Coordinate Attention*, o CBAM possui um viés espacial explícito, combinando atenção espacial e de canal, logo, esperam-se resultados positivos em tarefas de localização espacial como a abordada neste trabalho.

Analisando os mapas de ativação da Figura 30 gerados nas figuras de teste selecionadas, é possível observar um foco maior na região da pessoa de interesse em comparação ao modelo sem atenção. Este mecanismo, juntamente com o de *Coordinate Attention*, gera os mapas de ativação mais concentrados na região esperada (onde a pessoa está), contudo também apresentam uma diferença perceptível. Todos os mapas de ativação gerados no modelo com CBAM são mais difusos e abrangentes em comparação com os gerados no *Coordinate Attention*; em outras palavras, o CBAM acaba focando em uma região maior.

Este padrão de atenção mais amplo, mas ainda assim focado no sujeito, parece ser particularmente benéfico para realizar uma localização mais precisa, refletido pelo valor alcançado na métrica AP_{75}. É possível observar na Figura 30(a) um caso em que essa diferença tem um maior impacto, pois o mapa de ativação gerado pelo CBAM abrange toda a região da pessoa, diferente da variante com *Coordinate Attention*, que deixa um membro fora da região mais ativada, sendo uma causa direta para o menor ganho de performance apresentado.

A Tabela 5, que apresenta erros para uma imagem teste selecionada, também revela que o CBAM conseguiu uma redução de erro em todos os pontos, com exceção do 1 (nariz), contudo com uma redução sutil em comparação ao modelo de *Coordinate Attention*. O CBAM aparenta ser um detector mais robusto, pois em múltiplos casos em que o mecanismo de *Coordinate Attention* falhou em produzir uma predição com confiança suficiente (pontos 2, 3, 12, 16 e 17), o CBAM foi capaz de produzir uma estimativa razoável. Isso reforça a hipótese de que os mapas

de ativação mais difusos do CBAM diminuem as chances de que alguma porção do corpo seja negligenciada.

Conclui-se que os dois mecanismos usam estratégias espaciais ligeiramente diferentes e complementares. Enquanto o *Coordinate Attention* se destaca em localizar o núcleo da pose, o que é excelente para a performance geral, o CBAM utiliza uma abordagem mais abrangente, considerando o corpo inteiro. Este contexto mais amplo parece ser mais impactante para a localização exata dos pontos, podendo explicar por que o CBAM obteve o melhor resultado na métrica de alta precisão (AP_{75}).

4.1.3 *Global Context Attention*

Nos resultados quantitativos apresentados na Tabela 4, o modelo com mecanismo de *Global Context Attention* obteve um ganho modesto de 0,6 p.p. em relação ao modelo base (atingindo 66,8% AP), um desempenho ligeiramente superior ao mecanismo de *Self-Attention* (66,6% AP), mas significativamente inferior aos mecanismos com viés espacial explícito.

A Figura 30, com imagens de teste selecionadas, fornece um indício relevante para o entendimento deste ganho de desempenho menor. Nela é possível observar que todos os mapas de ativação para o modelo de *Global Context Attention* são uniformes. Isto não é um artefato visual, mas uma consequência do design do mecanismo. É importante ressaltar que o bloco implementado, baseado no trabalho de Cao et al. (2019), captura atenção em três etapas principais, sendo que a primeira realiza uma operação de *Global Average Pooling*, que transforma o mapa de características de entrada ($H \times W$) em um vetor unidimensional (1×1), o que acaba por descartar informação posicional e provocar o padrão observado.

Em essência, este mecanismo trata todas as regiões espaciais da imagem de forma idêntica, diferente dos mecanismos com viés espacial, que ajudam o modelo a distinguir o que é mais importante. Apesar dessa limitação, o mecanismo de *Global Context Attention* ainda é capaz de oferecer um benefício perceptível, principalmente para imagens mais complexas.

Observando a Figura 31, que apresenta a pose reconstruída, é possível notar um aumento na precisão dos pontos gerados em relação ao modelo base. Esta conclusão também é corroborada pela análise do erro euclidiano na Tabela 5, onde todos os pontos estimados pelo modelo com *Global Context Attention* apresentam uma redução em relação ao modelo base. Isso sugere que, mesmo sem um foco espacial, o ajuste global realizado por esse mecanismo pode melhorar a qualidade geral da representação que alimenta a *estimation head*, levando a uma melhora na estimação dos pontos anatômicos.

4.1.4 *Self-Attention*

O modelo com *Self-Attention*, implementado através do *Non-Local Block* (WANG et al., 2018), tem como objetivo capturar um contexto global rico ao modelar as interações entre todos os pares de elementos do mapa de características. Teoricamente, essa capacidade deveria ser

vantajosa; no entanto, os resultados demonstram o segundo menor ganho de desempenho, de apenas 0,4 p.p. na métrica principal AP.

A análise qualitativa dos mapas de ativação na Figura 30 pode fornecer uma explicação para esse baixo ganho. Conforme observado, de forma consistente, o mecanismo de *Self-Attention* não aprende a distribuir o foco sobre a pessoa; em vez disso, ele exibe um padrão recorrente onde o mapa de calor se concentra intensamente em um único ponto no centro do corpo do sujeito, geralmente no torso ou pescoço (ou em torno do centro do membro identificado, como na Figura 30(b)). Este comportamento é compatível com a atenção redundante já apontada por Cao et al. (2019), onde o modelo gera mapas muito semelhantes independentemente da posição, aprendendo uma estratégia simples de usar um único ponto como contexto para estimar todos os outros.

Apesar dessa característica, o modelo com *Self-Attention* ainda oferece uma melhora em relação ao modelo base. Isso também pode ser verificado na Tabela 5, onde o modelo consegue reduzir o erro de estimação em diversos pontos. Contudo, uma observação pertinente é que esse modelo obteve os menores *scores* de confiança para suas predições, o que é evidenciado pelo maior número de pontos não detectados em comparação com os outros mecanismos.

4.1.5 *Multi-Head Attention*

O modelo com *Multi-Head Attention* apresentou um ganho de performance relevante de 1,2 p.p., alcançando o terceiro melhor resultado, com um AP de 67,4%, e o segundo maior em AP_{50}, atingindo 89,1%. Embora represente uma melhora significativa, este resultado foi ligeiramente inferior aos mecanismos com viés espacial explícito, como o CBAM e o *Coordinate Attention*.

A análise qualitativa dos mapas de ativação na Figura 30 fornece uma possível explicação. Ao contrário dos padrões focados observados nos mecanismos com viés espacial, os mapas gerados pelo MHA são consistentemente mais difusos. O foco do modelo acaba por se espalhar pela região da pessoa, mas frequentemente transborda para o fundo. Uma provável hipótese é que os mapas visualizados são uma superposição de estratégias distintas geradas por diferentes *heads*, produzindo um foco menos concentrado.

Apesar de seu potencial, a natureza mais genérica deste mecanismo pode oferecer menos benefícios para essa tarefa em comparação a abordagens com viés espacial explícito. Contudo, ele também oferece uma vantagem consistente sobre o modelo sem atenção, como verificado nos ganhos de AP e na redução do erro de estimação apresentada na Tabela 5.

4.1.6 SimAM

O modelo com SimAM, um mecanismo de atenção sem parâmetros, apresentou o ganho de desempenho mais baixo, com 66,4% de AP, uma melhora de apenas 0,2 p.p. sobre o modelo base.

A análise qualitativa dos mapas de ativação na Figura 30 revela que o foco do modelo não é tão diferente do modelo sem atenção. Para as imagens (b) e (e), é possível perceber uma piora, pois os mapas se encontram menos focados na pessoa. Apesar do ganho de performance indicado pela métrica AP e pela redução de erro na Tabela 5, os mapas de ativação sugerem que o mecanismo não aprendeu uma estratégia de atenção relevante.

Apesar do ganho mínimo, o fato de ter havido uma melhora sugere que a premissa teórica de destacar neurônios com base em sua separabilidade linear é válida. No entanto, o resultado indica que para uma tarefa complexa como a estimação de pose, a atenção derivada de uma função matemática fixa é significativamente menos eficaz do que a guiada por parâmetros aprendidos.

4.2 Análise do Custo Computacional

Além da precisão, a eficiência computacional é um fator crítico. Para avaliar o custo-benefício de cada mecanismo, foram analisados o número de parâmetros treináveis e o volume de operações de ponto flutuante (FLOPS). A Tabela 6 sintetiza os resultados, enquanto a Figura 32 oferece uma visualização comparativa.

Tabela 6 – Custo computacional e ganho de desempenho para cada modelo.

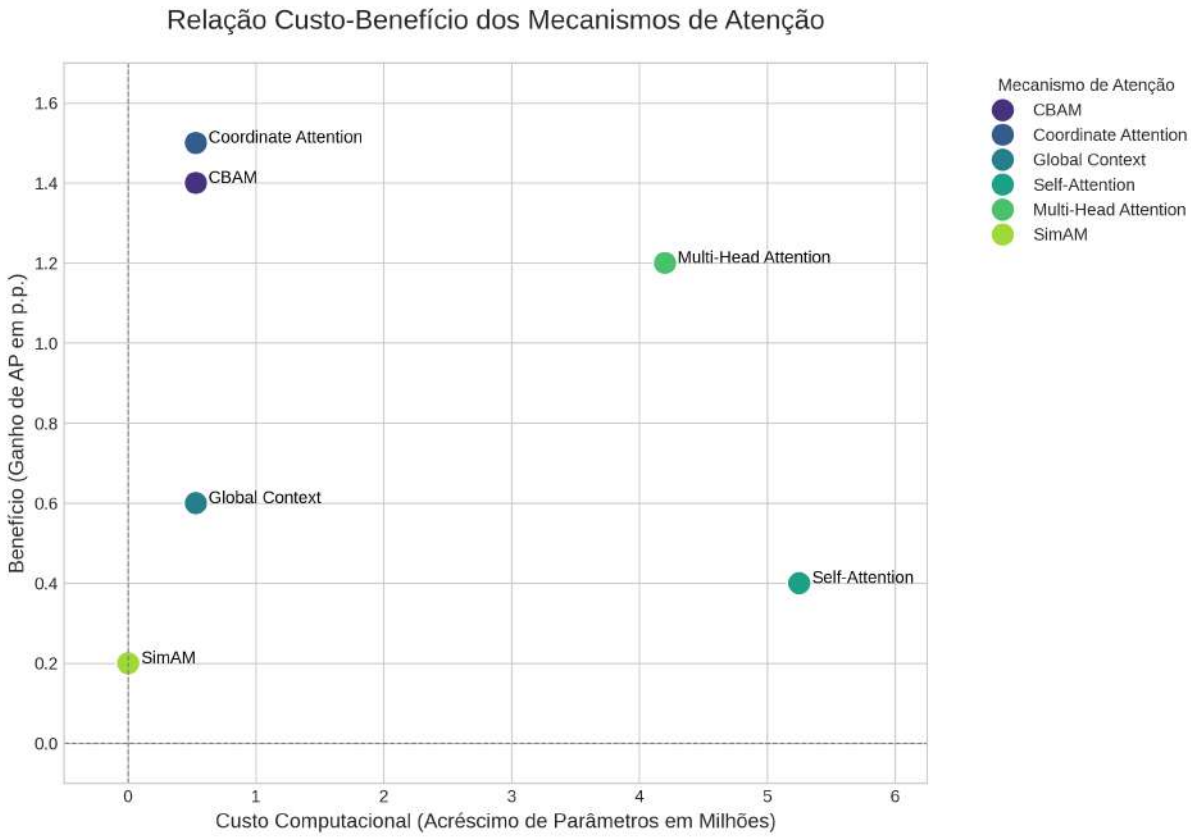
Modelo	**N. de Parâmetros (M)**	**FLOPS (G)**	**Ganho de AP (p.p.)**
Base (Sem atenção)	34.00	19.40	–
Base + CBAM	34.53	19.41	+1.4
Base + Coordinate Attention	34.53	19.42	+1.5
Base + Global Context	34.53	19.41	+0.6
Base + Self-Attention	39.25	19.91	+0.4
Base + Multi-Head Attention	38.20	19.81	+1.2
Base + SimAM	34.00	19.40	+0.2

Fonte: Elaborado pelo próprio autor.

Os modelos com CBAM e *Coordinate Attention* demonstram alta eficiência, conforme a Figura 32. Ambos introduzem um pequeno acréscimo de parâmetros e foram responsáveis pelos maiores ganhos de desempenho, posicionando-se no ponto ótimo do gráfico. O modelo com *Global Context*, apesar do mesmo custo extra, apresentou um ganho mais discreto (+0,6 p.p).

O modelo com *Self-Attention* representa a abordagem menos eficiente, o que é evidenciado por sua posição isolada na extrema direita do gráfico. Ele impôs o maior custo computacional (+15,4%), obtendo um dos menores ganhos de AP (+0,4 p.p). Este resultado reforça a hipótese de que a computação densa deste módulo é redundante para a granularidade exigida pela estimação de pose.

Figura 32 – Relação entre o Custo Computacional (Acréscimo de Parâmetros) e o Ganho de Desempenho (AP) para cada mecanismo de atenção em relação ao modelo base.



Fonte: Elaborado pelo autor (2025).

O modelo com *Multi-Head Attention* se posiciona em um meio-termo; embora seu custo seja considerável, mostrou-se mais eficiente que o *Self-Attention*, obtendo um ganho de desempenho significativamente maior (+1,2 p.p).

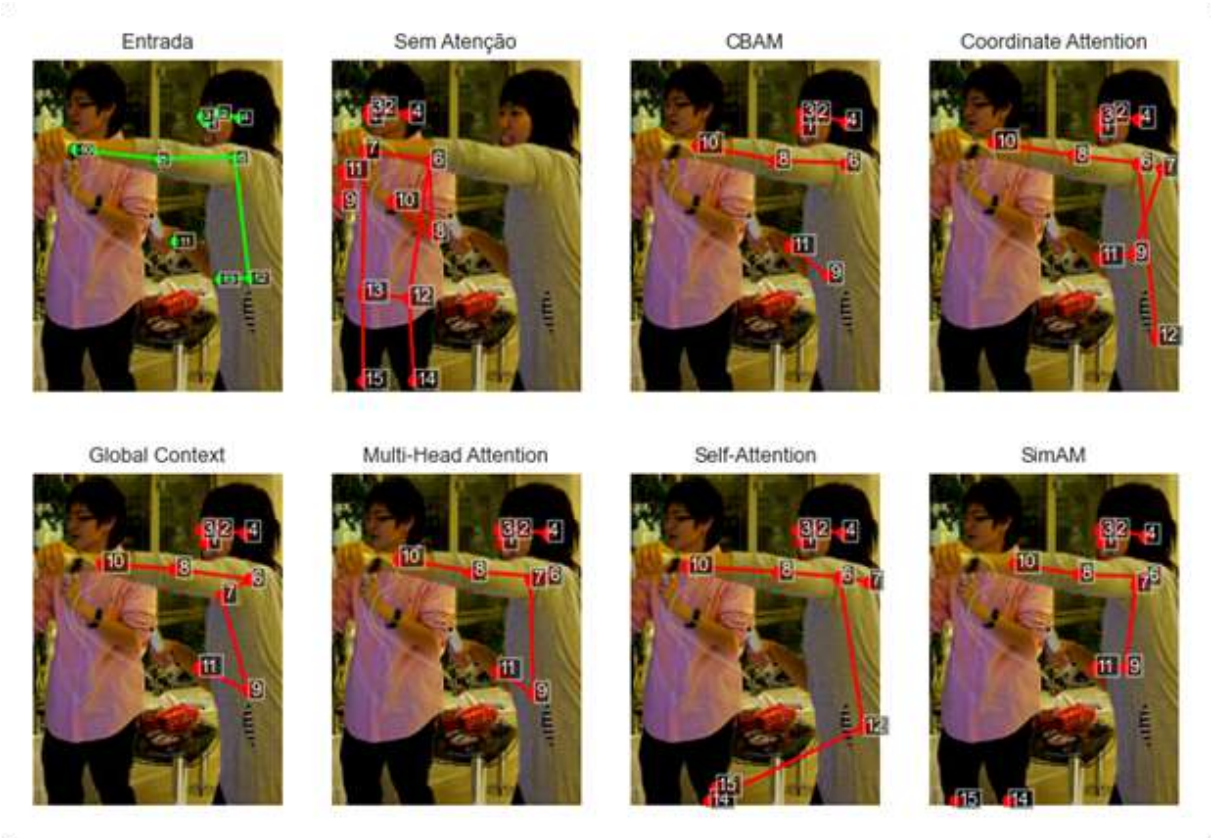
Finalmente, o modelo com SimAM é um caso único por ser livre de parâmetros. Na Figura 32, ele se localiza na origem do eixo de custo. O fato de ter produzido uma melhora sobre o modelo base, ainda que modesta (+0,2 p.p.), é importante, pois valida que é possível refinar a atenção de forma puramente analítica. Contudo, seu impacto limitado o torna uma opção de menor relevância prática em comparação com alternativas que, com um custo mínimo, apresentaram ganhos até sete vezes maiores.

4.3 Análise qualitativa geral

Analisando amostras mais complexas, com mais de uma pessoa na imagem, é possível perceber alguns padrões recorrentes. A Figura 33 apresenta a pose estimada para o exemplo da Figura 30(d). Este caso exige que o modelo faça a estimativa dos pontos e uma distinção sobre a quem eles pertencem. Nessas situações, os modelos com algum mecanismo de atenção

conseguem fornecer um resultado melhor, pois lidam com esse tipo de ambiguidade de maneira mais robusta.

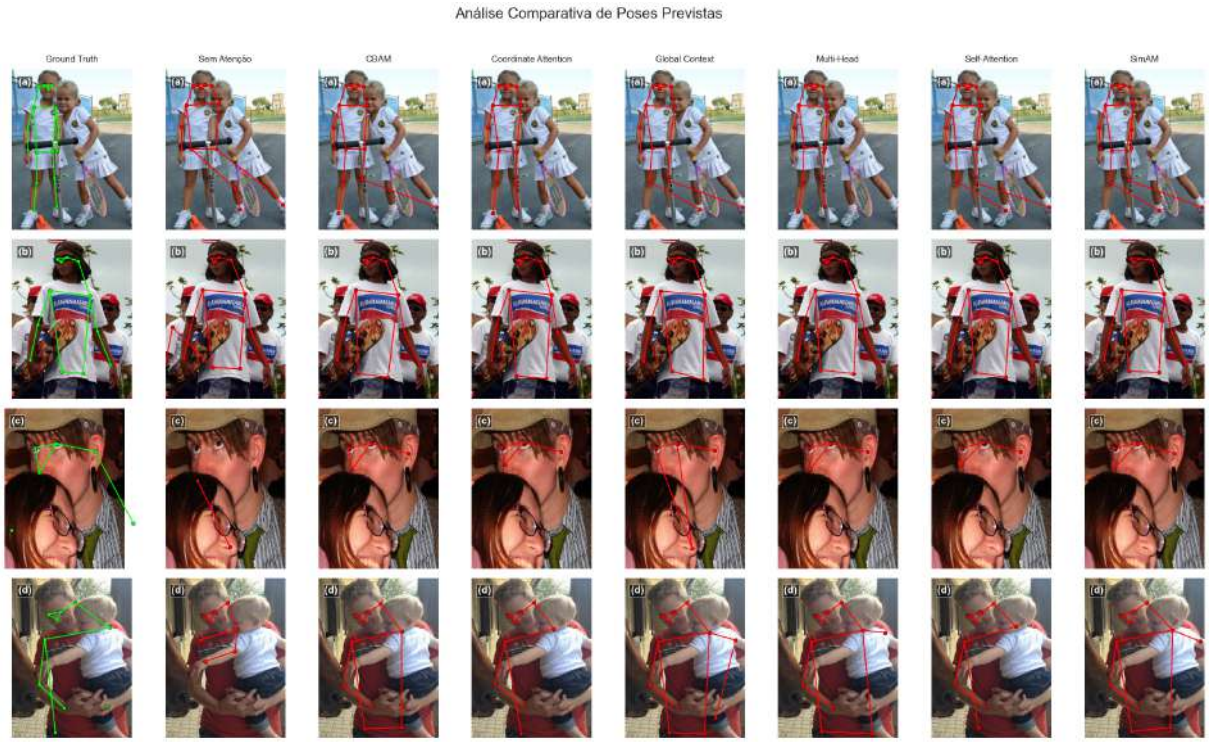
Figura 33 – Reconstrução de pose para a imagem da Figura 30(d).



Fonte: Elaborado pelo autor (2025).

Este padrão de melhora em cenas ambíguas é consistente em outras amostras. A Figura 34 apresenta outros casos com essa mesma característica. A partir da pose reconstruída, é possível observar um maior número de pontos estimados na pessoa correta pelos modelos com mecanismo de atenção, observável nas Figura 34(a), (b) e (c). Além disso, também é possível inferir que os mecanismos de atenção conseguem aumentar o grau de confiança das predições, constatado a partir dos pontos que o modelo base "tenta"estimar.

Figura 34 – Poses reconstruídas para imagens com ambiguidade.



Fonte: Elaborado pelo autor (2025).

5 Conclusão e Trabalhos Futuros

A estimativa de pose humana é uma das tarefas fundamentais do campo da visão computacional, e o desenvolvimento de técnicas capazes de aprimorar modelos para essa tarefa é crucial para uma vasta gama de aplicações. Este estudo teve como objetivo central investigar, por meio de análise comparativa, o impacto da integração de seis diferentes mecanismos de atenção em uma arquitetura baseada em *ResNet-50* para a tarefa de estimativa de pose humana 2D. Para tal, cada uma das arquiteturas avaliadas foi treinada usando o banco de dados público *MS COCO*, seguindo o mesmo protocolo experimental para garantir uma comparação justa. Cada um dos mecanismos: *Coordinate Attention*, *CBAM* (*Convolutional Block Attention Module*), *Global Context Attention*, *Self-Attention*, *Multi-Head Attention* e *SimAM* (*Simple Parameter-Free Attention Module*), foi integrado na mesma localização da arquitetura base, usando a mesma configuração e hiperparâmetros de treinamento, garantindo que as diferenças de desempenho pudessem ser atribuídas exclusivamente ao impacto de cada mecanismo de atenção.

A integração de mecanismos de atenção de modo geral, demonstrou ser uma estratégia eficaz para melhorar o desempenho dos modelos na tarefa de estimativa de pose humana, com todos os mecanismos usados apresentando uma melhora sobre o modelo base. Notavelmente os mecanismos que incorporam um viés espacial explícito, como o CBAM e *Coordinate Attention* foram os mais eficazes, oferecendo os maiores ganhos de performance na métrica primária. Isso, juntamente com seu aumento de custo computacional mínimo, posiciona essas abordagens como as melhores soluções.

Em contrapartida, mecanismos com uma abordagem de atenção mais genérica, como *Self-Attention* e *Multi-Head Attention*, embora tenham apresentado ganhos de desempenho, foram mais modestos e acompanhados de um custo computacional significativamente maior comparados aos outros. O *Self-Attention*, em particular, apresentou a pior relação custo-benefício, sugerindo que sua computação global pode ser redundante para a granularidade exigida pelo problema. Por fim, o mecanismo não paramétrico SimAM, embora não tenha adicionado complexidade ao modelo, gerou um ganho marginal, indicando que a capacidade de aprender atenção com parâmetros treináveis é mais adequado para a tarefa.

Em suma, este estudo demonstra que a escolha do mecanismo de atenção é um fator crítico, onde abordagens leves e com foco espacial são estratégias de baixo custo e alto impacto, oferecendo um caminho promissor para o aprimoramento de modelos de estimação de pose.

Apesar da consistência dos resultados, este estudo possui limitações que abrem caminho para futuras investigações. A análise foi feita sobre uma única arquitetura de *backbone* e um único conjunto de dados. O desempenho relativo dos mecanismos de atenção pode variar com *backbone* diferente, e com outras características de imagem (outros bancos de dados). Adicionalmente, o ponto de inserção dos módulos de atenção foi fixo, e a exploração de diferentes arranjos para a arquitetura da rede poderia render resultados distintos.

A partir dessas limitações, surgem direções para a realização de trabalhos futuros. Primeiramente, a replicação deste estudo em diferentes *backbones* e conjuntos de dados é essencial para validar e generalizar os achados aqui apresentados. Investigações futuras também podem explorar a combinação de diferentes mecanismos de atenção em estágios distintos da rede. E por fim, a aplicação e adaptação das conclusões deste trabalho para tarefas mais complexas, como a estimação de pose em vídeos (espaço-temporal) e a estimação de pose 3D, representam caminhos promissores para o desenvolvimento desta solução.

Referências

ANDRILUKA, M. et al. *PoseTrack: A Benchmark for Human Pose Estimation and Tracking*. 2017. Https://arxiv.org/abs/1710.10000v2. Citado 2 vezes nas páginas 21 e 22.

ANDRILUKA, M. et al. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 3686–3693. ISSN 1063-6919. Citado 2 vezes nas páginas 19 e 20.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, set. 2014. Citado na página 24.

CAI, Y. et al. *Learning Delicate Local Representations for Multi-Person Pose Estimation*. [S.l.]: arXiv, 2020. Citado 2 vezes nas páginas 13 e 39.

CAO, Y. et al. *GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond*. [S.l.]: arXiv, 2019. Citado 2 vezes nas páginas 29 e 47.

CAO, Z. et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 1302–1310. ISSN 1063-6919. Citado 3 vezes nas páginas 1, 8 e 16.

CARRASCO, M. Visual attention: The past 25 years. *Vision Research*, v. 51, n. 13, p. 1484–1525, jul. 2011. ISSN 1878-5646. Citado na página 32.

CHANGING the Game: Improving Athletic Performance with Vision-Based Motion Tracking and Pose Estimation. 2023. Https://www.theimagingsource.com/en-us/resource/news/2023/08/03/. Citado na página 6.

CHEN, Y. et al. *Cascaded Pyramid Network for Multi-Person Pose Estimation*. 2017. Https://arxiv.org/abs/1711.07319v2. Citado 4 vezes nas páginas 1, 8, 13 e 39.

COCO - Common Objects in Context. 2021. Https://cocodataset.org/#home. Citado na página 19.

DESIMONE, R.; DUNCAN, J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, v. 18, p. 193–222, 1995. ISSN 0147-006X. Citado na página 24.

DOSOVITSKIY, A. et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. [S.l.]: arXiv, 2021. Citado 2 vezes nas páginas 2 e 33.

GOOGLE. *Introdução a rostos aumentados | ARCore*. 2025. Https://developers.google.com/ar/develop/augmented-faces?hl=pt-br. Citado na página 7.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Nature Publishing Group, v. 585, n. 7825, p. 357–362, set. 2020. ISSN 1476-4687. Citado na página 38.

HE, K. et al. *Deep Residual Learning for Image Recognition*. [S.l.]: arXiv, 2015. Citado 4 vezes nas páginas 11, 12, 13 e 35.

HILLYARD, S. A.; VOGEL, E. K.; LUCK, S. J. Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 353, n. 1373, p. 1257–1270, ago. 1998. ISSN 0962-8436. Citado na página 32.

HOU, Q.; ZHOU, D.; FENG, J. *Coordinate Attention for Efficient Mobile Network Design*. [S.l.]: arXiv, 2021. Citado 2 vezes nas páginas 28 e 45.

HU, J.; SHEN, L.; SUN, G. *Squeeze-and-Excitation Networks*. [S.l.]: arXiv, 2017. Citado 3 vezes nas páginas 2, 25 e 28.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, v. 9, n. 3, p. 90–95, maio 2007. ISSN 1558-366X. Citado na página 38.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2012. v. 25. Citado na página 10.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, nov. 1998. ISSN 1558-2256. Citado 3 vezes nas páginas 1, 10 e 12.

LIN, T.-Y. et al. *Microsoft COCO: Common Objects in Context*. [S.l.]: arXiv, 2015. Citado na página 19.

MNIH, V. et al. *Recurrent Models of Visual Attention*. [S.l.]: arXiv, 2014. Citado na página 24.

MUNEA, T. L. et al. The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access*, v. 8, p. 133330–133348, 2020. ISSN 2169-3536. Citado 2 vezes nas páginas 4 e 8.

NEWELL, A.; YANG, K.; DENG, J. *Stacked Hourglass Networks for Human Pose Estimation*. 2016. Https://arxiv.org/abs/1603.06937v2. Citado 3 vezes nas páginas 1, 8 e 40.

PAPANDREOU, G. et al. *Towards Accurate Multi-person Pose Estimation in the Wild*. [S.l.]: arXiv, 2017. Citado na página 8.

PASZKE, A. et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. [S.l.]: arXiv, 2019. Citado na página 38.

POSNER, M. I. Orienting of Attention*. *Quarterly Journal of Experimental Psychology*, SAGE Publications, v. 32, n. 1, p. 3–25, fev. 1980. ISSN 0033-555X. Citado na página 24.

QU, H. et al. *Heatmap Distribution Matching for Human Pose Estimation*. [S.l.]: arXiv, 2022. Citado na página 9.

SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [S.l.]: arXiv, 2015. Citado na página 15.

SUN, K. et al. *Deep High-Resolution Representation Learning for Human Pose Estimation*. [S.l.]: arXiv, 2019. Citado 4 vezes nas páginas 1, 13, 16 e 18.

SWRI. *SwRI Launches BEAMoCap™ Markerless Motion Capture for 3D Animation in Gaming, Film | Southwest Research Institute*. 2025. Https://www.swri.org/newsroom/press-releases/swri-launches-beamocap-markerless-motion-capture-3d-animation-gaming-film. Citado na página 7.

TOSHEV, A.; SZEGEDY, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1653–1660. Citado 2 vezes nas páginas 1 e 4.

TREISMAN, A. M.; GELADE, G. A feature-integration theory of attention. *Cognitive Psychology*, v. 12, n. 1, p. 97–136, jan. 1980. ISSN 0010-0285. Citado na página 24.

VASWANI, A. et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2017. v. 30. Citado 4 vezes nas páginas 25, 26, 30 e 33.

WANG, X. et al. *Non-Local Neural Networks*. [S.l.]: arXiv, 2018. Citado 5 vezes nas páginas 26, 29, 31, 32 e 47.

WEBB, B. S. et al. Early and Late Mechanisms of Surround Suppression in Striate Cortex of Macaque. *The Journal of Neuroscience*, v. 25, n. 50, p. 11666–11675, dez. 2005. ISSN 0270-6474. Citado na página 32.

WEI, S.-E. et al. *Convolutional Pose Machines*. 2016. Https://arxiv.org/abs/1602.00134v4. Citado na página 1.

WOO, S. et al. *CBAM: Convolutional Block Attention Module*. [S.l.]: arXiv, 2018. Citado 3 vezes nas páginas 2, 26 e 27.

XIAO, B.; WU, H.; WEI, Y. *Simple Baselines for Human Pose Estimation and Tracking*. [S.l.]: arXiv, 2018. Citado 4 vezes nas páginas 13, 18, 35 e 39.

XU, K. et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: *Proceedings of the 32nd International Conference on Machine Learning*. [S.l.]: PMLR, 2015. p. 2048–2057. ISSN 1938-7228. Citado na página 25.

YANG, L. et al. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In: *Proceedings of the 38th International Conference on Machine Learning*. [S.l.]: PMLR, 2021. p. 11863–11874. ISSN 2640-3498. Citado 2 vezes nas páginas 26 e 32.

YAO, A. et al. Does Human Action Recognition Benefit from Pose Estimation? In: *Procedings of the British Machine Vision Conference 2011*. Dundee: British Machine Vision Association, 2011. p. 67.1–67.11. ISBN 978-1-901725-43-8. Citado na página 6.

ZHENG, C. et al. *Deep Learning-Based Human Pose Estimation: A Survey*. [S.l.]: arXiv, 2023. Citado 2 vezes nas páginas 4 e 9.

ZHOU, B. et al. *Learning Deep Features for Discriminative Localization*. [S.l.]: arXiv, 2015. Citado na página 42.