Universidade da Beira Interior
Departamento de Informática

P04 - 2025: *Human Recognition in Surveillance Settings – Database*

Elaborado por:

Carolina Castanheira Fernandes
nº50252

Orientador:

Professor Doutor Hugo Proença

24 de junho de 2025

Agradecimentos

Gostaria de expressar a minha sincera gratidão a todas as pessoas que aceitaram participar voluntariamente nas filmagens realizadas ao longo deste projeto, contribuindo de forma essencial para a construção do conjunto de dados. Agradeço igualmente aos três estudantes do 12.º ano do curso profissional de Informática da Escola Secundária Campos Melo, que colaboraram nas anotações manuais dos vídeos com dedicação e atenção ao detalhe. Um especial agradecimento ao meu orientador, Professor Hugo Proença, pelo acompanhamento contínuo, pela orientação técnica e pelo incentivo constante. Por fim, agradeço ao Kailash Hambarde, investigador em Visão Computacional e Inteligência Artificial no Instituto de Telecomunicações (IT), cujo apoio ao longo das várias etapas do projeto — desde a recolha de dados até à anotação — foi determinante para o sucesso deste trabalho.

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

UBI	Universidade da Beira Interior
IA	Inteligência Artificial
UAV	Unmanned Aerial Vehicle
IoU	Intersection over Union
WSL	Windows Subsystem for Linux
NMS	Non-Maximum Suppression

Capítulo

1

Introdução

1.1 Enquadramento

O progresso tecnológico na área da visão computacional tem impulsionado a aplicação de modelos de Inteligência Artificial (IA) em contextos cada vez mais exigentes. Uma das áreas com maior relevância prática é a *human recognition* (reconhecimento humano), que desempenha um papel essencial em sistemas de segurança, análise comportamental, vigilância aérea, entre outros.

Este relatório descreve o desenvolvimento de um projeto realizado no âmbito da unidade curricular de Projeto da Universidade da Beira Interior (UBI).

O trabalho divide-se em várias fases, desde a recolha de dados controlada até à avaliação objetiva e subjetiva de modelos de deteção. A recolha de dados envolveu a captura de imagens e vídeos de indivíduos em ambiente controlado e exterior, anotação manual com recurso ao *CVAT* e aplicação de metadados sob a forma de *soft bios*. Posteriormente, modelos como *YOLO11*, *YOLO-NAS* e *RT-DETR* foram testados em múltiplas condições com o intuito de compreender os seus limites e pontos fortes.

1.2 Motivação

A deteção humana em imagens captadas por drones apresenta-se como um desafio significativo devido à presença de múltiplas variáveis contextuais, como oclusões, variações de iluminação, ângulo de visão e distância ao sujeito. A motivação para este projeto surgiu da curiosidade em explorar até que ponto os modelos atuais conseguem lidar com estas variáveis.

Adicionalmente, o tema apresenta relevância prática para cenários reais como controlo de multidões e monitorização remota. A realização de uma recolha de dados própria e a análise crítica do comportamento dos modelos permite uma compreensão mais aprofundada do seu funcionamento e aplicabilidade no mundo real.

1.3 Objetivos

O objetivo principal deste projeto é recolher, anotar e analisar um conjunto de dados sobre pessoas a partir de imagens aéreas captadas por Unmanned Aerial Vehicle (UAV), com foco em longas distâncias e sob diferentes ângulos, poses, ações e altitudes do drone. Pretende-se integrar os dados recolhidos no conjunto de dados *PedX*, contribuindo com novas instâncias captadas em condições realistas e desafiantes.

Complementarmente, o projeto visa realizar uma análise comparativa entre diferentes modelos de deteção humana, avaliando a sua robustez e desempenho em contextos variados. A investigação centra-se especialmente nas situações em que os modelos conseguem ou não detetar indivíduos, analisando fatores como a oclusão, distância, orientação, iluminação e resolução da imagem.

De forma mais detalhada, os objetivos do projeto são:

- Recolher e anotar dados de vídeo e imagem de indivíduos em movimento, tanto em ambiente controlado como em voo de drone;
- Atribuir metadados aos vídeos e imagens através de atributos físicos e visuais (*soft bios*);
- Integrar os dados recolhidos no dataset *PedX*;
- Aplicar modelos de deteção humana (YOLO11, YOLO-NAS, RT-DETR) ao conjunto de dados recolhido;
- Avaliar os modelos com métricas objetivas como Intersection over Union (IoU), *precision*, *recall* e F1-score;
- Realizar uma análise visual e qualitativa dos resultados, identificando padrões de falha e sucesso;
- Discutir o potencial destes modelos em cenários reais de monitorização aérea.

1.4 Organização do Documento

De modo a refletir o trabalho desenvolvido, este documento encontra-se estruturado da seguinte forma:

1. O primeiro capítulo — **Introdução** — apresenta o enquadramento do projeto, a motivação que levou à sua escolha, os seus objetivos principais e a organização geral do relatório.

2. O segundo capítulo — **Estado da Arte** — inclui uma revisão crítica sobre a deteção de pessoas em imagens aéreas e a evolução dos modelos de *deep learning*, com foco especial nos métodos aplicados a imagens captadas por drones. São ainda abordadas comparações entre modelos e os principais desafios técnicos desta área.

3. O terceiro capítulo — **Tecnologias Utilizadas** — descreve as ferramentas e tecnologias aplicadas no projeto, incluindo o CVAT, Python e os três modelos de deteção avaliados: YOLOv11, YOLO-NAS e RT-DETR. São explicadas as principais características, funcionamento interno e a forma como cada modelo foi integrado no projeto.

4. O quarto capítulo — **Experiências e Resultados** — detalha a metodologia adotada para a recolha e anotação dos dados (fases *indoor* e *outdoor*), a preparação da base de dados (PedX), a implementação dos modelos e os testes realizados. Inclui ainda a **avaliação experimental**, tanto objetiva (com base em métricas como *precision*, *recall*, *F1-score* e IoU), como subjetiva (análise visual de frames com melhor e pior desempenho, por grupo/sessão, em situações extremas e síntese comparativa dos modelos).

5. O quinto capítulo — **Conclusões e Trabalho Futuro** — sintetiza as conclusões mais relevantes, destacando as forças e limitações de cada modelo, e propõe possíveis direções para trabalhos futuros, incluindo a progressão natural para tarefas de reconhecimento e estimativa de pose.

Capítulo

2

Estado da Arte

2.1 Introdução

Este capítulo apresenta uma visão geral sobre a evolução das técnicas de deteção de pessoas em imagens aéreas, particularmente captadas por veículos aéreos não tripulados (UAVs). O objetivo é contextualizar a problemática abordada neste trabalho, analisando os desafios existentes e as soluções mais relevantes desenvolvidas até à atualidade. Para além disso, realiza-se uma comparação entre diferentes abordagens, com destaque para os modelos YOLO11, YOLO-NAS e RT-DETR, cujas características detalhadas serão descritas no capítulo das Tecnologias Utilizadas.

2.2 Deteção Humana em Imagens Aéreas

A deteção humana em imagens captadas por UAVs (Unmanned Aerial Vehicles) representa um desafio significativo para a visão computacional, devido a diversos fatores técnicos e contextuais, entre os quais se destacam:

- **Variação acentuada de escala**, com pessoas a ocuparem áreas muito pequenas na imagem quando filmadas a grandes altitudes;
- **Oclusões e sobreposições**, especialmente em ambientes urbanos ou aglomerados;
- **Condições de iluminação variáveis**, causadas por sombra, reflexos ou meteorologia;
- **Perspetiva aérea incomum**, com ângulos de visão que dificultam a identificação de silhuetas humanas completas.

Apesar destas dificuldades, a deteção de pedestres em imagens aéreas tem vindo a ganhar importância em múltiplas aplicações práticas, incluindo:

- **Segurança e vigilância urbana**, com UAVs a sobrevoarem áreas públicas para monitorizar multidões, prevenir comportamentos suspeitos ou apoiar operações policiais;
- **Busca e salvamento**, permitindo a localização de pessoas em zonas de difícil acesso, como florestas, áreas montanhosas ou zonas afetadas por desastres naturais;
- **Análise de tráfego pedonal**, em ambientes como centros históricos ou eventos de grande escala, para estimar densidades de pessoas e padrões de movimento;
- **Apoio à resposta a emergências**, ajudando a identificar vítimas em cenários de catástrofe ou evacuação;
- **Planeamento urbano e estudos sociais**, através da recolha de dados sobre o uso do espaço público e interação de pedestres com o ambiente urbano.

Estes contextos reforçam a necessidade de modelos de deteção cada vez mais robustos, capazes de operar em tempo real e em condições adversas, tirando partido da flexibilidade e mobilidade dos drones. A combinação de UAVs com redes neuronais profundas tem assim potencial para transformar significativamente a forma como recolhemos e analisamos dados sobre o comportamento humano em espaços exteriores.

2.3 Evolução dos Modelos Baseados em Deep Learning

Os modelos de deteção de objetos evoluíram de pipelines baseados em múltiplas etapas (ex.: R-CNN) para arquiteturas unificadas e em tempo real (ex.: YOLO). Atualmente, as abordagens baseadas em transformadores (como RT-DETR) têm vindo a ganhar destaque.

- **YOLO**: séries sucessivas de versões com foco em desempenho em tempo real.
- **YOLO-NAS**: integração de NAS para otimização automática da arquitetura.

- **RT-DETR**: elimina a necessidade de pós-processamento com NMS, utilizando mecanismos de atenção globais.

2.4 Comparação de Modelos

A tabela seguinte apresenta uma comparação entre variantes representativas dos modelos **YOLOv11**, **YOLO-NAS** e **RT-DETR**, com base em métricas de desempenho como o **mAP@50–95**, avaliado no conjunto de dados COCO, e observações técnicas relevantes fornecidas pelos autores. Os dados foram retirados diretamente da documentação oficial da Ultralytics[1].

Modelo	mAP@50–95 (COCO)	Observações
YOLO11n	53.4	Excelente equilíbrio entre precisão e velocidade
YOLO-NAS-l	52.2	Otimização automática via NAS
RT-DETR-l	53.0	Detetor sem NMS, arquitetura baseada em Transformer

Tabela 2.1: Comparação de desempenho entre modelos recentes de deteção.

Todos os valores apresentados referem-se aos resultados reportados para o conjunto de validação **COCO 2017**, sendo os testes conduzidos com **GPU NVIDIA T4** para garantir comparabilidade em termos de desempenho computacional. Estas medições são úteis para avaliar a adequação dos modelos a aplicações em tempo real, como vigilância aérea ou navegação assistida por UAVs.

2.5 Conclusões

A análise do estado da arte permite concluir que, embora existam múltiplas abordagens para a deteção humana em imagens aéreas, os modelos mais recentes baseados em deep learning — em especial os das famílias YOLO e RT-DETR — representam o estado atual mais avançado. O seu uso em ambientes com limitações computacionais e em aplicações em tempo real torna-os especialmente relevantes para o contexto deste trabalho, justificando a sua seleção e análise comparativa nos capítulos seguintes.

[1] `https://docs.ultralytics.com/pt/models/yolo11/`, `https://docs.ultralytics.com/pt/models/yolo-nas/`, `https://docs.ultralytics.com/pt/models/rtdetr/`

Capítulo

3

Tecnologias Utilizadas

3.1 Introdução

Neste capítulo são descritas as principais tecnologias, frameworks e ferramentas utilizadas ao longo do desenvolvimento do projeto. Estas tecnologias foram selecionadas com o objetivo de permitir a recolha, anotação, análise e avaliação do desempenho de modelos de deteção humana em imagens aéreas.

3.2 CVAT

Para a anotação dos vídeos e imagens foi utilizada a ferramenta *CVAT* (Computer Vision Annotation Tool), uma plataforma open-source desenvolvida pela Intel. Permite a criação de *bounding boxes* frame a frame, com recurso a tracking, em vídeos, atribuindo uma label a cada bounding box e permitindo assim diferencia-las de pessoa para pessoa, facilitando a geração de dados anotados para treino e validação de modelos.

3.3 Python

Os scripts desenvolvidos para processar os dados, aplicar os modelos, gerar estatísticas e métricas foram implementados em *Python*, devido à vasta disponibilidade de bibliotecas como `OpenCV`, `Pandas`, `NumPy` e `Matplotlib`.

3.4 Modelos de Deteção Humana

3.4.1 YOLO11

O *YOLO11* (You Only Look Once, versão 11) é um modelo de deteção de objetos em tempo real, projetado para identificar e localizar instâncias de objetos numa imagem com elevada precisão e eficiência. No contexto deste projeto, o modelo foi utilizado para detectar pessoas em imagens aéreas, captadas por UAVs, devolvendo caixas delimitadoras (*bounding boxes*) com as respetivas classes (classe = 0 para pessoas) e níveis de confiança [1].

3.4.1.1 Principais Características

O YOLO11 apresenta várias melhorias significativas que o distinguem das versões anteriores [2]:

- **Extração de caraterísticas melhorada**: graças a uma arquitetura otimizada do *backbone* e do *neck*, o modelo é capaz de capturar representações visuais mais ricas e robustas, aumentando a precisão na deteção de objetos, mesmo em condições desafiantes.
- **Eficiência e velocidade**: o YOLO11 foi desenvolvido com um foco particular na velocidade de processamento e na eficiência computacional. O modelo YOLOv11m, por exemplo, alcança uma precisão média (mAP) superior à do YOLOv8m, utilizando menos 22% de parâmetros, o que reduz significativamente o custo computacional.
- **Suporte a múltiplas tarefas**: o modelo suporta não só a deteção de objetos, mas também segmentação de instâncias, estimativa de pose, classificação de imagens e deteção orientada de objetos (OBB — Oriented Bounding Box), tornando-o extremamente versátil.
- **Adaptabilidade**: o YOLO11 pode ser facilmente implementado em diferentes ambientes, desde dispositivos periféricos até plataformas em nuvem com GPUs NVIDIA, o que o torna apropriado para aplicações móveis ou remotas, como as realizadas neste projeto com drones.

3.4.1.2 Desempenho

O desempenho do YOLO11 foi avaliado com base no dataset COCO, considerando diferentes variantes do modelo com distintos tamanhos e complexidades. A tabela seguinte resume os valores de mAP (precisão média), latência e complexidade computacional de cada versão:

Modelo	mAPval (50–95)	CPU ONNX (ms)	T4 TensorRT10 (ms)	Parâmetros (M)	FLOPs (B)
YOLO11n	39.5	56.1 ± 0.8	1.5 ± 0.0	2.6	6.5
YOLO11s	47.0	90.0 ± 1.2	2.5 ± 0.0	9.4	21.5
YOLO11m	51.5	183.2 ± 2.0	4.7 ± 0.1	20.1	68.0
YOLO11l	53.4	238.6 ± 1.4	6.2 ± 0.1	25.3	86.9
YOLO11x	54.7	462.8 ± 6.7	11.3 ± 0.2	56.9	194.9

Tabela 3.1: Desempenho das variantes do YOLO11 no conjunto de dados COCO.

3.4.1.3 Funcionamento do YOLO11

O funcionamento do YOLO11 pode ser descrito pelas seguintes etapas principais:

1. **Pré-processamento da imagem:** A imagem de entrada é redimensionada (tipicamente para 640 × 640 píxeis), normalizada e convertida para um tensor, adequado ao processamento por redes neuronais. Esta etapa garante consistência nos dados de entrada e facilita a inferência eficiente.

2. **Extração de características (Backbone):** A imagem é passada por uma rede convolucional profunda, o *backbone*, que extrai mapas de características relevantes. Inicialmente, são detetados padrões simples como arestas e texturas; posteriormente, a rede aprende a identificar formas complexas como silhuetas humanas. O YOLO11 introduz melhorias nesta etapa, aumentando a precisão da extração sem aumentar significativamente o número de parâmetros.

3. **Fusão de escalas (Neck):** Os mapas de características extraídos a diferentes profundidades são combinados para preservar informação espacial e semântica de múltiplas escalas. Esta etapa melhora a deteção de objetos de diferentes tamanhos, desde pessoas próximas até indivíduos distantes nas imagens aéreas. O neck do YOLO11 foi otimizado para maior eficiência.

4. **Cabeça de deteção (Detection Head):** A imagem é concetualmente dividida numa grelha, e para cada célula dessa grelha o modelo prevê:

    - As coordenadas da *bounding box* (x, y, largura, altura);
    - A classe prevista (por exemplo, *pessoa*);
    - Um score de confiança que reflete a certeza do modelo quanto à presença e classe do objeto.

Assim, para cada região da imagem, o YOLO11 tenta prever se há um objeto e, em caso afirmativo, onde está e o que é.

5. **Supressão de redundâncias (NMS – Non-Maximum Suppression):** Como o modelo pode gerar múltiplas *bounding boxes* para o mesmo objeto, é aplicado o algoritmo NMS. Este elimina previsões redundantes, mantendo apenas as caixas com maior score de confiança e menor sobreposição com outras (IoU inferior a um determinado limiar).

6. **Resultado final:** Após o NMS, o modelo devolve a lista final de deteções: as caixas delimitadoras, as classes e os respetivos scores. Estas caixas são desenhadas na imagem para facilitar a visualização e posterior análise.

Este pipeline altamente otimizado permite que o YOLO11 funcione em tempo real, mesmo em dispositivos com recursos limitados, mantendo uma precisão elevada na deteção de objetos como pessoas, mesmo em condições adversas como as captadas por drones em movimento e a grandes distâncias [1].

3.4.1.4 Aplicação no Projeto

No contexto deste projeto, o YOLO11 foi utilizado para a deteção de pessoas em imagens aéreas obtidas por UAVs, com o objetivo de avaliar a sua capacidade de identificar indivíduos em condições de distância, ângulo e altura variáveis. A escolha do modelo baseou-se na sua combinação de rapidez e precisão, bem como na facilidade de integração com o ecossistema `ultralytics`, que permite a realização de treino, validação e inferência com apenas algumas linhas de código em *Python*.

A estrutura modular do YOLO11 permitiu ainda a comparação direta com outros modelos como o YOLO-NAS e o RT-DETR, sob as mesmas condições experimentais, facilitando a análise crítica do desempenho relativo de cada um.

3.4.2 YOLO-NAS

O *YOLO-NAS* (You Only Look Once - Neural Architecture Search) é um modelo de deteção de objetos de última geração desenvolvido pela Deci AI, com suporte fornecido atualmente pela Ultralytics. Esta arquitetura representa um avanço em relação aos modelos anteriores da série YOLO, sendo o resultado de técnicas de pesquisa automática de arquitetura neuronal (NAS) [3].

O YOLO-NAS foi projetado para atingir uma precisão elevada com baixa latência, sendo particularmente eficaz em ambientes com recursos computacionais limitados, como aplicações em tempo real.

3.4.2.1 Principais Características

- **Arquitetura quantizável**: o YOLO-NAS inclui blocos básicos otimizados para quantização, permitindo converter os modelos para formatos como INT8 com perdas mínimas de precisão. Esta característica melhora significativamente o desempenho em dispositivos com restrições de memória e processamento.

- **Treino e quantização avançados**: utiliza esquemas de formação sofisticados combinados com quantização pós-treino, o que contribui para uma redução de latência sem comprometer a exatidão da deteção.

- **Otimização AutoNAC**: recorre à técnica AutoNAC (Automatic Neural Architecture Construction) para gerar arquiteturas eficientes automaticamente, treinadas previamente em datasets extensos como COCO, Objects365 e Roboflow 100 [4].

- **Variedade de tamanhos**: o YOLO-NAS está disponível em três variantes — pequena (s), média (m) e grande (l) — oferecendo um equilíbrio entre precisão e velocidade consoante os recursos disponíveis.

3.4.2.2 Desempenho

O desempenho do YOLO-NAS foi avaliado com base no dataset COCO. A tabela seguinte resume os valores de mAP (precisão média) e latência para cada versão:

Modelo	mAP	Latência (ms)
YOLO-NAS-s	47.5	3.21
YOLO-NAS-m	51.55	5.85
YOLO-NAS-l	52.22	7.87
YOLO-NAS-s (INT8)	47.03	2.36
YOLO-NAS-m (INT8)	51.0	3.78
YOLO-NAS-l (INT8)	52.1	4.78

Tabela 3.2: Desempenho das variantes do YOLO-NAS no conjunto de dados COCO.

3.4.2.3 Funcionamento do YOLO-NAS

O funcionamento do YOLO-NAS pode ser descrito pelas seguintes etapas principais:

1. **Pré-processamento da imagem:** A imagem de entrada é redimensionada (por exemplo, para 640 × 640 píxeis), normalizada e convertida para um tensor. Este formato é compatível com as operações das redes neuronais convolucionais e garante consistência na entrada, além de permitir o processamento eficiente em tempo real.

2. **Extração de características (Backbone):** A imagem é processada por uma rede convolucional otimizada com blocos estruturados para serem compatíveis com quantização. Estes blocos extraem características relevantes em várias camadas, desde formas simples até estruturas complexas como contornos humanos. A arquitetura do YOLO-NAS foi desenhada para manter alta precisão mesmo quando convertida para versões quantizadas (como INT8), tornando-a ideal para dispositivos com capacidade limitada.

3. **Fusão de escalas (Neck):** Os mapas de características obtidos em diferentes resoluções são combinados para melhorar a deteção de objetos em várias escalas. Esta etapa permite detetar tanto objetos grandes como pequenos, e é essencial para imagens aéreas em que as pessoas aparecem com tamanhos reduzidos. O neck foi otimizado automaticamente através do processo AutoNAC para maximizar a eficiência e o desempenho.

4. **Cabeça de deteção (Detection Head):** Tal como em outras variantes da série YOLO, a imagem é dividida numa grelha virtual, e para cada célula da grelha, o modelo prevê:

   - As coordenadas da *bounding box* (x, y, largura, altura);
   - A classe do objeto detetado (por exemplo, *pessoa*);
   - Um score de confiança associado à previsão.

   Estas previsões são feitas de forma paralela e em tempo real, permitindo múltiplas detecções numa única passagem pela rede.

5. **Supressão de redundâncias (NMS – Non-Maximum Suppression):** Para evitar caixas sobrepostas ou duplicadas para o mesmo objeto, o YOLO-NAS aplica o algoritmo de supressão não-máxima. Este seleciona apenas as caixas com maior confiança e menor sobreposição (definida por um limiar de IoU), removendo as restantes.

6. **Resultado final:** Após o NMS, o modelo devolve a lista final de detecções, composta por caixas delimitadoras, classes e scores de confiança. Estes resultados podem ser utilizados diretamente ou sobrepostos na imagem original para visualização e análise posterior.

Graças à sua arquitetura otimizada e blocos quantizáveis, o YOLO-NAS é capaz de realizar estas operações de forma rápida e eficiente, mantendo uma elevada precisão, mesmo em ambientes computacionalmente restritos [3, 4].

3.4.2.4 Aplicação no Projeto

Neste projeto, o YOLO-NAS foi utilizado como um dos modelos de comparação para deteção humana em imagens captadas por UAVs. A sua arquitetura otimizável, capacidade de quantização e desempenho elevado em latência permitiram avaliar como modelos eficientes se comportam em cenários desafiantes com oclusões, variações de distância e ângulo. A versão YOLO-NAS-l foi utilizada nas experiências, equilibrando bem velocidade e precisão.

3.4.3 RT-DETR

O *RT-DETR* (*Real-Time Detection Transformer*) é um modelo de deteção de objetos em tempo real desenvolvido pela Baidu, baseado na arquitetura de *Transformers de Visão*. Este modelo foi concebido para alcançar elevada precisão sem comprometer a velocidade de inferência, sendo particularmente adequado para aplicações em tempo real como condução autónoma, robótica ou vigilância aérea [5].

3.4.3.1 Principais Características

O RT-DETR distingue-se por diversas inovações arquitetónicas:

- **Codificador híbrido eficiente**: Processa caraterísticas multiescala através da dissociação entre a interação intra-escala e a fusão entre escalas, aumentando a eficiência computacional e a precisão.
- **Seleção de consultas com conhecimento da IoU**: Utiliza um mecanismo inteligente de seleção de consultas baseado em IoU, o que permite ao modelo focar-se nos objetos mais relevantes da cena.
- **Velocidade de inferência adaptável**: O RT-DETR permite ajustar a rapidez da inferência ao utilizar diferentes camadas do descodificador, sem necessidade de re-treino.

- **Estrutura sem NMS**: Elimina a tradicional etapa de supressão não máxima (Non-Maximum Suppression (NMS)), simplificando o pipeline e reduzindo o tempo de pós-processamento.

- **Deteção sem âncoras**: Não depende de caixas âncora pré-definidas, o que melhora a generalização e reduz a complexidade do modelo.

3.4.3.2 Desempenho

A Tabela 3.3 resume o desempenho das variantes principais do RT-DETR no conjunto de dados COCO:

Modelo	mAP	FPS (T4 GPU)
RT-DETR-L	53.0	114
RT-DETR-X	54.8	74

Tabela 3.3: Desempenho das variantes do RT-DETR no conjunto de dados COCO val2017.

3.4.3.3 Funcionamento do RT-DETR

O funcionamento do RT-DETR pode ser descrito pelas seguintes etapas:

1. **Pré-processamento da imagem:** A imagem de entrada é redimensionada (por exemplo, para 640 × 640 píxeis), normalizada e convertida para um tensor. Este tensor é usado como entrada para o modelo Transformer.

2. **Extração de características (Backbone):** A imagem é processada por uma rede convolucional (*ConvNet*) que extrai mapas de características em múltiplas escalas (por exemplo, S3, S4, S5). Estes mapas são usados como entrada para a parte transformer do modelo.

3. **Codificador híbrido (Hybrid Encoder):** Aplica mecanismos específicos para tratar separadamente:

    - **Interação intra-escala (AIFI)**: Para refinar a informação dentro de cada nível de escala;
    - **Fusão entre escalas (CCFM)**: Para combinar características entre diferentes níveis de detalhe.

4. **Seleção de consultas (Query Selection):** Um mecanismo consciente do IoU seleciona as regiões mais relevantes da imagem como *queries* para deteção. Isso torna a deteção mais eficiente e dirigida.

5. **Descodificador e Cabeça de Previsão:** O descodificador refina iterativamente as *queries* para gerar previsões finais — coordenadas da *bounding box*, classe do objeto e score de confiança. Este processo é feito sem necessidade de NMS.

6. **Resultado final:** O modelo devolve diretamente as caixas preditas com os respetivos scores e classes, sem necessidade de supressão posterior, o que reduz a latência total.

3.4.3.4 Aplicação no Projeto

No âmbito deste projeto, o RT-DETR foi utilizado para testar a sua eficácia na deteção humana em imagens captadas por UAVs. A sua arquitetura baseada em *transformers*, aliada à eliminação da etapa de NMS, tornou-o particularmente interessante para avaliação em cenários onde a velocidade de inferência e a precisão são igualmente importantes.

A comparação direta com modelos como o YOLO11 e o YOLO-NAS permitiu avaliar as vantagens e desvantagens da abordagem baseada em *transformers* em relação a arquiteturas CNN tradicionais, especialmente em contextos de deteção aérea a longa distância.

3.5 Métricas de Avaliação

Os resultados dos modelos foram avaliados com base em métricas padrão de visão computacional:

- **IoU** (*Intersection over Union*): mede a sobreposição entre a *bounding box* prevista pelo modelo e a *bounding box* real (anotada manualmente), também chamada de *ground truth*.

- **Precision**: representa a proporção de verdadeiros positivos (*true positives*) em relação ao total de deteções feitas pelo modelo. Mede a exatidão das deteções. Um valor alto de *precision* indica que a maioria das deteções do modelo são corretas.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

  Onde:

  - **TP** – Verdadeiros positivos (deteções corretas);
  - **FP** – Falsos positivos (deteções erradas, onde o modelo "viu"uma pessoa que não existia).

- **Recall**: representa a proporção de verdadeiros positivos em relação ao total de objetos reais presentes nas imagens. Mede a capacidade do modelo em encontrar todas as ocorrências relevantes. Um valor alto de *recall* indica que o modelo consegue detetar a maioria das pessoas presentes.

$$\text{Recall} = \frac{TP}{TP + FN}.$$

  Onde:

  - **FN** – Falsos negativos (pessoas reais que o modelo não detetou).

- **F1-score**: métrica combinada que pondera *precision* e *recall*, calculando a sua média harmónica. É particularmente útil quando se pretende encontrar um equilíbrio entre a exatidão (*precision*) e a capacidade de cobertura (*recall*) do modelo, especialmente em contextos onde existe um desbalanceamento entre as classes (por exemplo, quando há muitas mais regiões negativas do que pessoas nas imagens).

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

  O valor do F1-score varia entre 0 e 1, sendo 1 o valor ideal (deteção perfeita). Um F1-score elevado indica que o modelo tem simultaneamente uma boa capacidade de detetar corretamente as pessoas (alto *recall*) e baixa taxa de falsos positivos (alta *precision*).

3.6 Ambiente de Desenvolvimento

O projeto foi desenvolvido num ambiente local baseado em Windows Subsystem for Linux (WSL), utilizando CPU. Esta configuração proporcionou maior controlo sobre os testes, bem como compatibilidade com ferramentas e bibliotecas comuns em sistemas Linux, mantendo a integração com o sistema operativo Windows.

Devido às diferentes dependências e versões exigidas por cada modelo testado — nomeadamente *YOLO11*, *YOLO-NAS* e *RT-DETR* — foram criados ambientes virtuais distintos com o auxílio do módulo `venv` do *Python*. Esta

abordagem permitiu isolar completamente os requisitos de cada modelo, evitando conflitos de versões entre bibliotecas como `torch`, `transformers`, `opencv-python`, entre outras.

Cada ambiente foi configurado com os pacotes estritamente necessários para o modelo correspondente, garantindo maior estabilidade, facilidade de manutenção e reprodutibilidade dos testes. A separação dos ambientes também facilitou a deteção de problemas específicos durante a instalação ou execução de cada arquitetura, contribuindo para uma análise mais rigorosa dos resultados.

3.7 Conclusão

Neste capítulo foram apresentadas as principais tecnologias, modelos e ferramentas que sustentaram o desenvolvimento do projeto. Desde a anotação manual com a ferramenta *CVAT*, passando pela implementação em *Python* com bibliotecas de suporte à visão computacional, até à utilização de modelos avançados como o *YOLO11*, *YOLO-NAS* e *RT-DETR*. A escolha tecnológica foi sempre orientada por critérios de desempenho, versatilidade e compatibilidade com os objetivos da deteção aérea de pessoas.

Adicionalmente, foram descritas as métricas de avaliação utilizadas e o ambiente de desenvolvimento adotado, incluindo a separação de ambientes virtuais, que garantiu a reprodutibilidade dos testes e evitou conflitos de dependências.

Este enquadramento técnico e metodológico fornece a base sólida para os testes e análises realizados, que serão aprofundados no capítulo seguinte.

Capítulo

4

Experiências e Resultados

4.1 Introdução

Este capítulo descreve em detalhe o processo de implementação do sistema de deteção humana em imagens captadas por drones, bem como os testes realizados para avaliar e comparar o desempenho dos modelos utilizados.

Inicialmente, apresenta-se a preparação do conjunto de dados, incluindo a recolha e anotação manual dos vídeos, seguida da configuração de ambientes virtuais independentes para cada modelo, de forma a garantir a compatibilidade com as respetivas dependências.

Posteriormente, é descrita a execução dos testes com base em métricas objetivas como o IoU, *Precision*, *Recall* e *F1-score*. Estas métricas permitiram avaliar a eficácia dos modelos na deteção de pessoas em diferentes condições de distância, altura e oclusão.

Para além da análise individual dos resultados, foi realizada uma comparação entre os diferentes modelos através da **correlação de Pearson**, com o objetivo de medir o grau de semelhança nas deteções realizadas por cada modelo. Esta análise estatística permitiu identificar o alinhamento (ou divergência) entre os modelos, revelando padrões consistentes e eventuais comportamentos discrepantes na deteção humana em contextos aéreos.

4.2 Preparação do Conjunto de Dados

A construção do conjunto de dados foi uma etapa central deste projeto, cujo objetivo principal passa por explorar os desafios associados ao reconhecimento humano em contextos de vigilância — ambientes caracterizados por baixa qualidade de imagem, oclusões, poses variáveis e grandes distâncias.

Tais condições refletem cenários realistas em que os sujeitos não estão conscientes do processo de captação de dados.

O trabalho partiu da necessidade de colmatar a escassez de conjuntos de dados anotados que evidenciem, de forma controlada, as variações nos principais covariantes (distância, pose, iluminação, resolução). Assim, foram organizadas duas fases distintas de aquisição de dados — uma fase **indoor**, com recolha individual de alta qualidade, e uma fase **outdoor**, baseada em captação aérea com drones, onde as imagens têm uma baixa qualidade.

4.2.1 Fase Indoor: Captura Individual de Referência

Com o objetivo de contribuir para a construção da base de dados PedX, foi realizada uma fase inicial de recolha de dados em ambiente controlado (indoor), centrada na captura individual de cada participante. Esta recolha teve como principal finalidade fornecer dados de referência de alta qualidade, essenciais para futuras tarefas de reconhecimento humano, tais como identificação, reidentificação ou análise de marcha (*gait recognition*).

A captação decorreu em condições ideais — elevada resolução, iluminação estável e visibilidade total do corpo — de forma a compensar as limitações visuais das gravações aéreas realizadas com o drone. Esta fase fornece, assim, uma base sólida para comparações futuras, análise subjetiva e desenvolvimento de modelos que ultrapassem a mera deteção de presença.

Após a assinatura de um formulário de consentimento, cada participante foi individualmente registado através de:

- Três fotografias estáticas — frontal, perfil esquerdo e perfil direito — garantindo a visibilidade integral da aparência;

- Um vídeo, com cerca de 20 segundos de duração, onde o participante caminha naturalmente (para a frente e para os lados), permitindo análises posteriores de padrões de marcha (*gait recognition*).

Estas imagens e vídeos funcionam como referência visual clara, permitindo avaliar a correspondência entre os dados de alta qualidade (indoor) e as deteções realizadas sobre vídeos de baixa resolução (outdoor). Para além disso, foram recolhidos atributos biométricos suaves (*soft biometrics*) — como género, idade estimada, altura, peso e etnia — com o intuito de apoiar a nomeação organizada dos dados e futuras análises de caracterização e reconhecimento de identidade.

Importa reforçar que os dados indoor não foram utilizados nos testes aos modelos de deteção realizados neste trabalho. A sua principal finalidade foi a

integração na base de dados PedX, como suporte a tarefas mais complexas de reconhecimento humano.

4.2.1.1 Nomeação dos Ficheiros Indoor

Para garantir a rastreabilidade dos dados e facilitar a sua posterior utilização em tarefas de reconhecimento humano, foi adotada uma convenção de nomeação consistente para todas as imagens e vídeos captados na fase *indoor*.

Formato da nomenclatura:

```
PID_Gender_Age_Height_Weight_Ethnicity_Angle_View
```

Onde:

- `PID` — Identificador único da pessoa;
- `Gender` — Género (0: masculino, 1: feminino, 2: outro);
- `Age` — Grupo etário (0–11, 12–17, 18–24, ..., >65);
- `Height` — Altura (0: abaixo de 150cm, 1: 150cm - 160cm, 2: 160cm - 170cm, 3: acima de 170cm);
- `Weight` — Volume corporal (0: abaixo de 60kg, 1: 60kg - 80 kg, 2: acima de 80kg);
- `Ethnicity` — Etnia (0: indiano, 1: negro, 2: asiático, 3: branco.);
- `Angle` — Posição da câmara relativamente ao participante (F: frontal, L: perfil esquerdo, R: perfil direito);
- `View` — Tipo de ficheiro (para imagens estáticas ou vídeo de marcha).

Exemplos:

- `0001_0_2_1_1_3_F.jpg` — Imagem frontal do participante 0001, do sexo masculino, grupo etário 18–24, altura 150cm - 160cm, peso 60kg - 80kg, etnia branca.
- `0001_0_2_1_1_3.mp4` — Vídeo de marcha do mesmo participante.

Esta nomenclatura foi aplicada de forma sistemática às imagens captadas (F, L, R) e aos vídeos de marcha (.mp4), garantindo consistência e permitindo, no futuro, a filtragem dos dados com base nos atributos biométricos suaves (*soft biometrics*).

4.2.2 Fase Outdoor: Recolha Aérea com UAV

A fase mais desafiante do processo decorreu em ambientes exteriores, com recurso a veículos aéreos não tripulados (UAVs). Os dados foram captados em condições realistas e não controladas, simulando cenários de vigilância onde os sujeitos não estão conscientes da captação. Este contexto introduz vários fatores de degradação da qualidade visual:

- Resolução extremamente reduzida, com regiões de interesse (ROIs) frequentemente limitadas a apenas alguns píxeis;
- Iluminação natural variável e presença de sombras;
- Diversidade significativa de ângulos (30°, 60° e 90°), distâncias horizontais (10–120 m) e altitudes (5.8–120 m);
- Oclusões parciais e sobreposição entre indivíduos;

Cada participante foi filmado em duas sessões distintas, utilizando roupas diferentes em cada uma, de modo a simular variações realistas de aparência ao longo do tempo. Cada sessão contemplou 18 pontos de recolha por indivíduo, resultantes da combinação de diferentes pares de distância e altura (por exemplo, 10m/5.8m, 30m/17.3m, 120m/90m, entre outros). A duração média de cada vídeo foi de 20 segundos, e cada vídeo captava múltiplos indivíduos por grupo.

Importa sublinhar que a totalidade dos testes de deteção realizados neste trabalho foi efetuada exclusivamente com os vídeos captados em ambiente outdoor. Esta opção teve como objetivo avaliar o desempenho dos modelos em condições adversas e realistas, sem qualquer apoio de dados captados em condições ideais. A separação entre os dados de referência (indoor) e os dados de teste (outdoor) assegura uma avaliação mais rigorosa, refletindo os desafios práticos da deteção de pessoas em contexto de vigilância aérea com UAVs.

4.2.2.1 Nomeação dos vídeos Outdoor

As imagens extraídas dos vídeos captados com UAVs seguem um formato mais detalhado, codificando características do cenário, momento da recolha e atributos biométricos.

Formato da nomenclatura:

```
PID_Data_Hora_AlturaDrone_Distância_Ângulo_Local_HairColor_HairStyle
```

```
_Beard_Moustache_Glasses_HeadAccessories_UpperBodyCloths_LowerBodyCloths
          _Feet_Accessories_Action_DroneScene_Sessão
```

Exemplo:

```
0004_23_10_2024_09_18_5.8_10_30_3_0_1_1_0_-1_-1_6_0_0_6_152_178_S1
```

Este tipo de nomeação facilita não só a rastreabilidade das imagens em termos de condições de aquisição, como também permite selecionar subconjuntos de dados para experiências específicas, como testes por distância, altura, ação ou atributos visuais.

4.2.3 Anotação dos Dados

As anotações foram realizadas com a ferramenta *CVAT* (Computer Vision Annotation Tool), permitindo desenhar *bounding boxes* sobre cada indivíduo visível nos vídeos captados, frame a frame, tanto na fase *indoor* (vídeos de marcha) como *outdoor* (vídeos aéreos). Cada pessoa foi identificada com uma *label* única, assegurando consistência entre frames e sessões distintas.

Importa salientar que todas as anotações foram produzidas com o propósito exclusivo de integrar os dados na base de dados *PedX*, com vista a futuras tarefas de reconhecimento humano. No contexto da avaliação dos modelos de deteção, os vídeos originais (sem anotações visuais sobrepostas) foram utilizados como entrada para os modelos. Posteriormente, os resultados gerados por cada modelo foram comparados com as anotações feitas manualmente, possibilitando a avaliação objetiva da sua performance.

Este processo resultou num conjunto de dados anotado de forma precisa e robusta, suficiente para realizar testes quantitativos com diferentes modelos de deteção, através de métricas amplamente reconhecidas como *IoU*, *Precision*, *Recall* e *F1-score*.

4.2.4 Implementação dos Modelos

A implementação dos modelos foi realizada num ambiente local baseado em WSL (Windows Subsystem for Linux), tirando partido da flexibilidade dos sistemas Unix-like, mantendo compatibilidade com ferramentas e bibliotecas em Python. Todo o processamento foi efetuado com recurso a CPU, o que implicou a otimização do código e a escolha criteriosa dos parâmetros de execução, dada a limitação de recursos computacionais.

4.2.4.1 Ferramentas Utilizadas

O projeto recorreu às seguintes ferramentas e bibliotecas:

- **Python**: linguagem principal utilizada em todos os scripts de pré-processamento, inferência e análise de resultados;
- **Ultralytics**: biblioteca usada para carregar e aplicar os modelos *YOLO11* e *YOLO-NAS*;
- **PyTorch** e **Transformers**: utilizadas para a execução do modelo *RT-DETR*;
- **OpenCV**, **NumPy** e **Pandas**: para manipulação de vídeos, imagens e dados;
- **Matplotlib** e **Seaborn**: para visualização de métricas e construção de gráficos comparativos.

Ambientes virtuais distintos foram configurados com `venv`, garantindo o isolamento das dependências de cada modelo e evitando conflitos entre versões de bibliotecas como `torch` e `opencv-python`.

4.2.4.2 Carregamento e Avaliação dos Modelos

Os três modelos principais — *YOLO11*, *YOLO-NAS* e *RT-DETR* — foram utilizados exclusivamente para avaliação sobre os vídeos *outdoor*, previamente organizados por grupo, sessão, distância e altura. Cada modelo foi carregado a partir dos respetivos ficheiros de pesos, previamente treinados em COCO:

- **YOLO11**: carregado com a API da Ultralytics, usando o ficheiro `yolo11n.pt`;
- **YOLO-NAS**: carregado com a biblioteca `super-gradients`, utilizando o ficheiro `yolo-nas-l.pt`;
- **RT-DETR**: carregado através da interface da Ultralytics com o ficheiro `rtdetr-l.pt`.

Foram selecionados **1000 frames** dos vídeos anotados para teste, distribuídos proporcionalmente entre os diferentes grupos e sessões, de acordo com a quantidade de indivíduos. A organização seguiu a seguinte distribuição:

- **Grupo 1** (17 pessoas): 14 vídeos × 12 frames + 4 vídeos × 13 frames;

- **Grupo 2** (15 pessoas): 10 vídeos × 10 frames + 8 vídeos × 11 frames;
- **Grupo 3** (7 pessoas): 18 vídeos × 5 frames.

Cada frame selecionado foi submetido individualmente ao modelo correspondente. Os resultados das detecções foram armazenados em ficheiros `CSV`, possibilitando posterior análise **objetiva** (métricas como *IoU*, *Precision*, *Recall*, *F1-score*) e **subjetiva** (análise visual e comparação qualitativa).

4.2.4.3 Parâmetros Utilizados

Durante a execução, os seguintes parâmetros foram padronizados para garantir uma comparação justa entre os modelos:

- **Tamanho da imagem**: 640 × 640 píxeis (redimensionamento padronizado);
- **Confiança mínima**: 0.25 (limiar para considerar uma deteção válida);
- **IoU**: ≥ 0.5;
- **Classe de interesse**: apenas a classe `person` (ID = 0) foi considerada;
- **Batch size**: 1 (devido a limitações de CPU);
- **Dispositivo de execução**: `device='cpu'` em todos os testes.

4.2.4.4 Fluxo de Inferência

A sequência de execução foi comum aos três modelos e estruturada da seguinte forma:

1. Leitura e extração dos frames dos vídeos de teste;
2. Redimensionamento para 640×640 e normalização (quando aplicável);
3. Aplicação do modelo ao frame;
4. Armazenamento das caixas previstas (bounding boxes), classes e *scores* num ficheiro `CSV`;
5. (Opcional) Salvamento das imagens com *bounding boxes* desenhadas, utilizado apenas para suporte à análise subjetiva (ver Sec. 4.5). Estas imagens não foram usadas na avaliação quantitativa.

Este fluxo foi replicado de forma idêntica para os três modelos, assegurando condições equitativas de avaliação.

4.3 Métricas de Avaliação

As métricas utilizadas para avaliar o desempenho dos modelos foram previamente apresentadas na Sec. 3.5, no âmbito das tecnologias utilizadas. Para esta avaliação, foram consideradas as métricas padrão de visão computacional: *IoU*, *Precision*, *Recall* e *F1-score*, que permitiram quantificar a exatidão e robustez de cada modelo em condições de teste realistas.

A escolha destas métricas justifica-se pela sua capacidade de refletir não apenas a precisão das deteções (minimizando falsos positivos), mas também a sua abrangência (reduzindo falsos negativos), o que é essencial em cenários de vigilância com imagens aéreas, onde a resolução, o ângulo e a oclusão variam significativamente.

4.4 Avaliação Objetiva

A avaliação objetiva foi conduzida com base em **1000 frames anotados**, distribuídos proporcionalmente entre os diferentes grupos e sessões (ver Sec. 4.2.4). Para cada modelo testado — *YOLO11*, *YOLO-NAS* e *RT-DETR* — foram calculadas as métricas ***Precision***, ***Recall***, **F1-score** e **IoU médio**, considerando um **threshold de IoU** mínimo de 0.5 para que uma deteção fosse considerada válida.

4.4.1 Resultados Globais

Os resultados globais foram obtidos através da contagem acumulada dos verdadeiros positivos (TP), falsos positivos (FP) e falsos negativos (FN) ao longo de todos os *frames* testados, sendo este processo realizado separadamente para cada modelo. A partir dessas somas, calcularam-se as métricas de avaliação globais, como *Precision*, *Recall*, *F1-score* e *IoU médio*. A Tabela 4.1 apresenta os resultados obtidos para os três modelos testados, enquanto a Figura 4.1 compara visualmente as principais métricas. Observa-se que o **YOLO11** alcança a maior ***precision*** e o melhor **IoU médio**, demonstrando maior exatidão nas deteções realizadas. O **YOLO-NAS** destaca-se nas métricas de ***recall*** e **F1-score**, sugerindo uma maior capacidade de deteção, mesmo com um maior número de falsos positivos. O **RT-DETR** apresenta resultados equilibrados, posicionando-se entre os dois modelos nas principais métricas.

Modelo	TP	FP	FN	*Precision*	*Recall*	*F1-score*	IoU Médio
YOLO11	1276	824	12210	0.608	0.095	0.164	0.666
YOLO-NAS	3953	3737	9533	0.514	0.293	0.373	0.66
RT-DETR	3743	5785	9743	0.393	0.278	0.325	0.647

Tabela 4.1: Resultados globais da avaliação objetiva para os modelos testados.

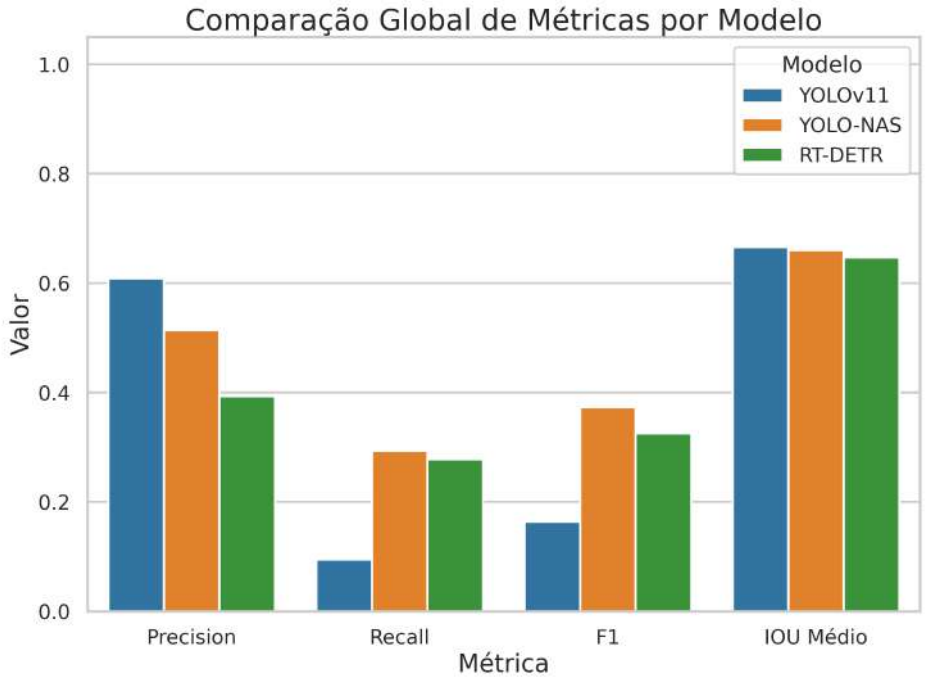

Figura 4.1: Comparação global das métricas entre os modelos YOLO11, YOLO-NAS e RT-DETR.

4.4.2 Resultados por Grupo e Sessão

As Figuras 4.2 e 4.3 ilustram os valores de *precision* e *recall* obtidos por grupo e sessão. Observa-se que o **Grupo 3** apresenta os melhores resultados em ambas as métricas, o que poderá estar relacionado com a menor densidade de pessoas e menor grau de oclusões nas suas gravações.

O **Grupo 2** revela resultados consistentes entre sessões, possivelmente por ter sido captado em condições ambientais semelhantes. Além disso, a **Sessão 2** apresenta, em média, um desempenho superior à Sessão 1 — fator que pode ser atribuído a uma iluminação mais favorável ou a um menor nível de movimento por parte dos participantes.

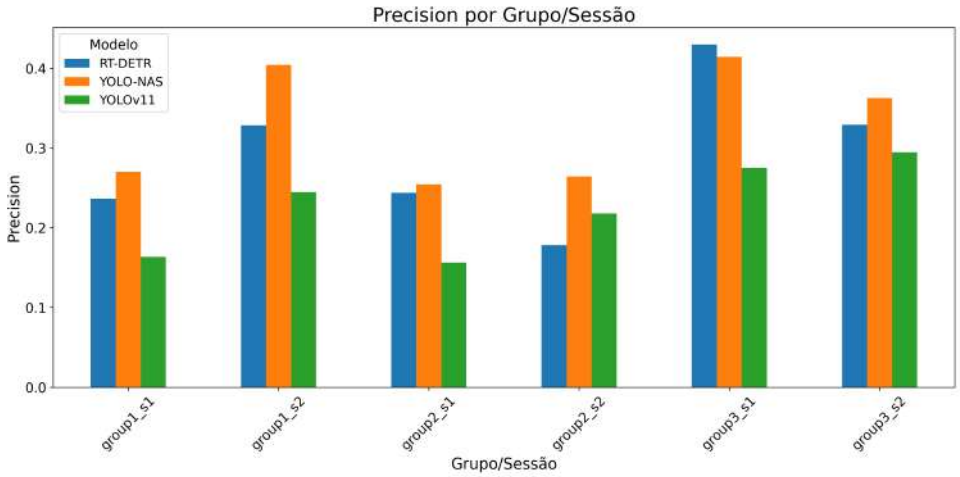

Figura 4.2: *Precision* por grupo e sessão.

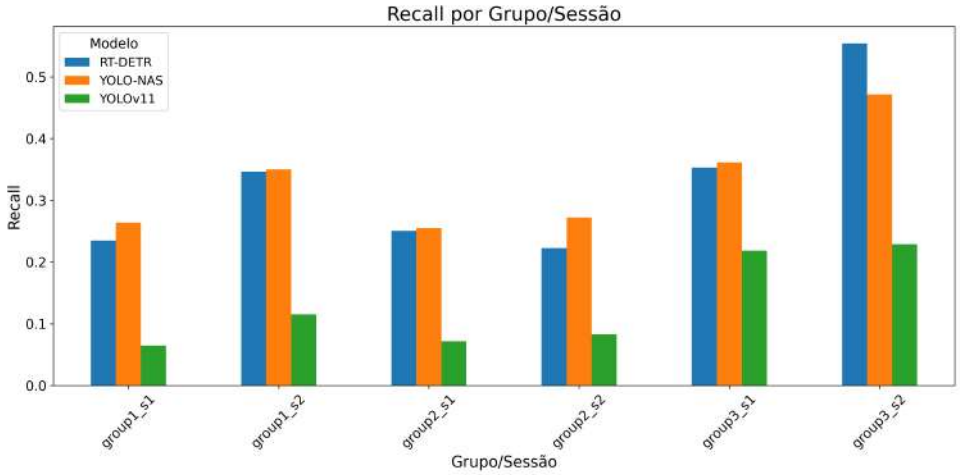

Figura 4.3: *Recall* por grupo e sessão.

4.4.3 Resultados por Ângulo

A orientação da câmara influencia significativamente o desempenho dos modelos de deteção, principalmente quando se trata de pessoas parcialmente ocluídas ou sobrepostas no campo de visão. Esta análise visa perceber se o ângulo de captação afeta métricas como *precision*, *recall* e *F1-score*.

A Tabela 4.2 apresenta os resultados médios por ângulo (30º, 60º e 90º) para cada modelo. Verifica-se que o desempenho é consistentemente superior para o ângulo de 30º. Este ângulo oferece uma perspetiva mais oblíqua, em que o corpo dos indivíduos é mais visível e menos sujeito a oclu-

sões. Por outro lado, o ângulo de 90º — correspondente a uma vista superior (top-down), onde apenas a cabeça ou os ombros são visíveis — conduz a uma queda acentuada no desempenho de todos os modelos, com destaque para o YOLO11, que apresenta valores residuais nesse cenário. Estes resultados reforçam o impacto do ponto de vista do drone na eficácia da deteção de pessoas.

Modelo	**Ângulo**	***Precision***	***Recall***	***F1-score***
YOLO11	30º	0.465	0.253	0.297
YOLO11	60º	0.158	0.106	0.118
YOLO11	90º	0.053	0.032	0.037
YOLO-NAS	30º	0.486	0.507	0.478
YOLO-NAS	60º	0.334	0.320	0.305
YOLO-NAS	90º	0.165	0.160	0.149
RT-DETR	30º	0.429	0.532	0.466
RT-DETR	60º	0.304	0.300	0.266
RT-DETR	90º	0.140	0.149	0.134

Tabela 4.2: Resultados médios por ângulo de captação para cada modelo.

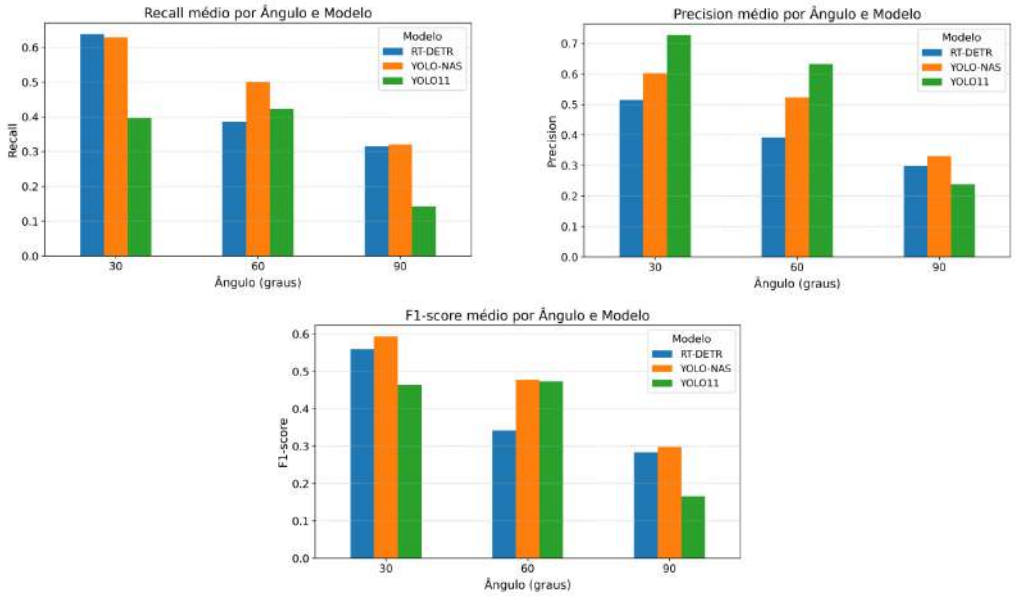

Figura 4.4: Comparação das métricas (*Precision*, *Recall* e *F1-score*) por ângulo (30º, 60º, 90º) entre os modelos testados.

Estes resultados ajudam a contextualizar os gráficos por distância e altura analisados nas próximas secções, uma vez que distâncias mais curtas nem

sempre correspondem a melhor desempenho, podendo isso dever-se ao ângulo em que os vídeos foram captados.

4.4.4 Resultados por Altura e Distância

A performance dos modelos deteriora-se de forma clara com o aumento da **altura** (Figura 4.5) e da **distância** (Figura 4.6). Enquanto todos os modelos mantêm boa precisão até aproximadamente 20 metros de altura, a partir dos 40 metros os resultados tornam-se altamente inconsistentes. O mesmo comportamento é observado para distâncias superiores a 40 metros. Este decréscimo de desempenho é esperado, uma vez que, com o aumento da altura e da distância, o número de píxeis que representam cada pessoa nas imagens diminui substancialmente.

Contudo, baixos desempenhos em distâncias ou alturas reduzidas nem sempre indicam melhores condições de deteção. Por exemplo, as alturas de **10 m** e **20 m** apresentam resultados inferiores ao esperado, o que pode ser explicado pelo facto desses registos terem sido captados sob um ângulo de **90º** — ou seja, com o drone na vertical, diretamente sobre o alvo. Esta perspetiva limita a visibilidade do corpo completo e aumenta a probabilidade de oclusões entre indivíduos. A mesma justificação aplica-se ao valor reduzido obtido para a distância **0 m**, igualmente associado ao ângulo de 90º e a uma vista de cima, desfavorável à deteção de pessoas.

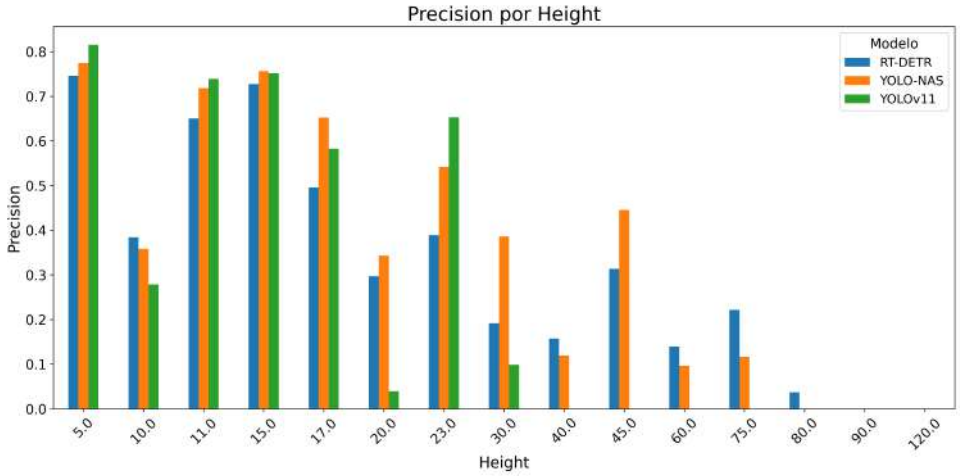

Figura 4.5: *Precision* média por altura do drone.

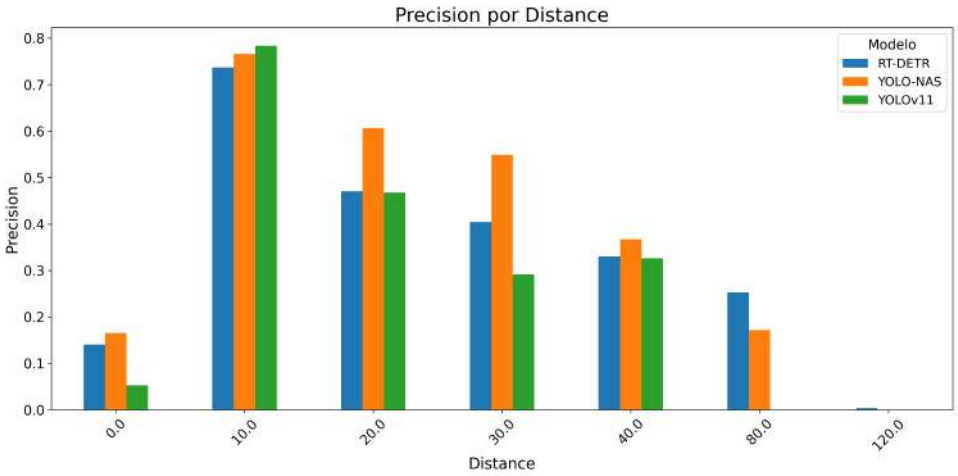

Figura 4.6: *Precision* média por distância horizontal do drone.

Foram obtidos resultados semelhantes para a métrica de ***recall***, conforme ilustrado nas Figuras 4.7 e 4.8.

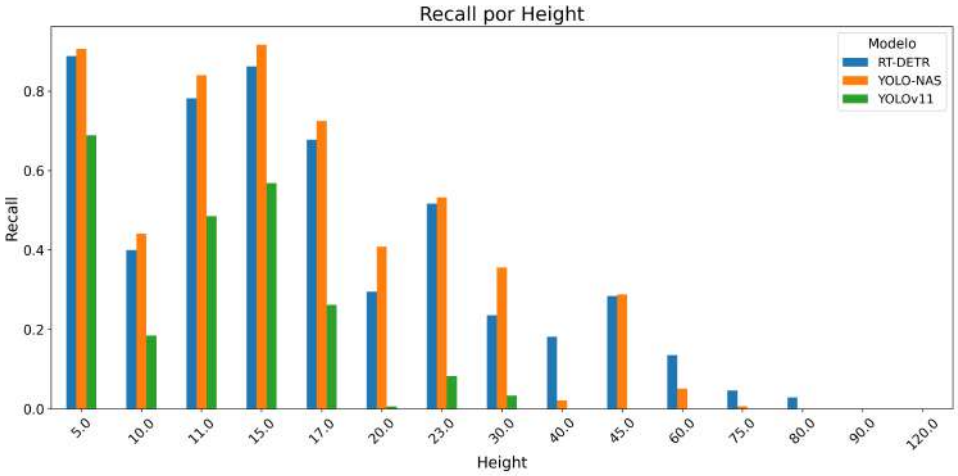

Figura 4.7: *Recall* médio por altura do drone.

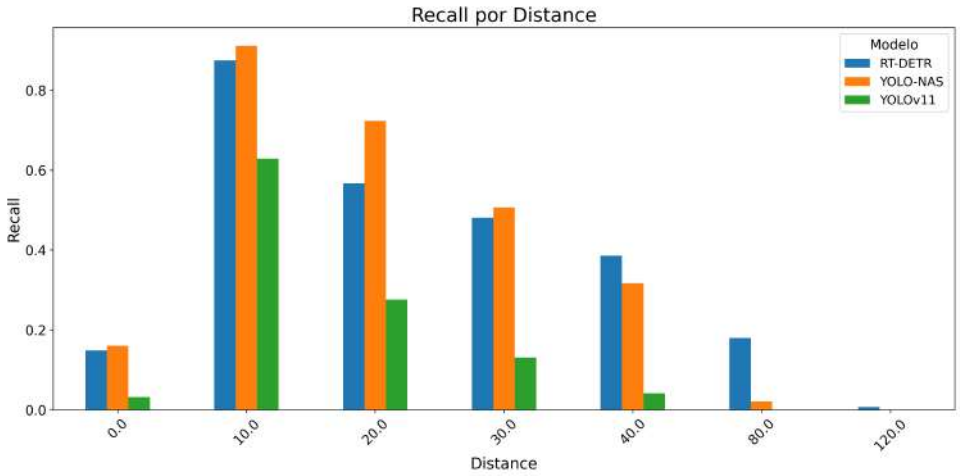

Figura 4.8: *Recall* médio por distância horizontal do drone.

O **YOLO-NAS** destaca-se como o modelo com maior sensibilidade, evidenciada pelos valores superiores de *recall*, sobretudo em alturas intermédias (entre 20 e 45 metros). No entanto, esta maior capacidade de deteção é acompanhada por um aumento no número de falsos positivos, comprometendo ligeiramente a precisão global.

Já o **RT-DETR**, embora não apresente os melhores resultados absolutos em termos de métricas globais, revela um **comportamento mais consistente em cenários de maior complexidade**, como em **alturas e distâncias mais elevadas**. Nestes contextos, onde a maioria dos modelos tende a falhar, o RT-DETR mantém um desempenho estável, demonstrando **robustez e adaptabilidade a condições visuais degradadas**.

Por sua vez, o **YOLO11** evidencia limitações claras à medida que os fatores adversos aumentam — com quebras notórias de precisão e *recall* a partir dos 30 metros. No entanto, em **condições mais favoráveis** (baixa altitude, curta distância e ângulos frontais), é o modelo com melhor **precisão**, o que o torna uma solução viável para aplicações controladas, onde a minimização de falsos positivos é prioritária.

4.4.5 Correlação de Pearson entre Modelos

A correlação de Pearson é uma medida estatística que quantifica a força e a direção da relação linear entre duas variáveis. O seu valor varia entre -1 e 1:

- $+1$ indica uma correlação linear positiva perfeita;
- 0 indica ausência de correlação linear;

- -1 indica uma correlação linear negativa perfeita.

A fórmula da correlação de Pearson entre duas variáveis X e Y, com n observações, é dada por:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Onde:

- X_i e Y_i são os valores individuais das variáveis X e Y;
- $\bar{X}$ e $\bar{Y}$ são as médias de X e Y;
- r representa o coeficiente de correlação de Pearson.

As Figuras 4.9, 4.10 e 4.11 apresentam a correlação de Pearson entre os três modelos para cada uma das métricas de avaliação.

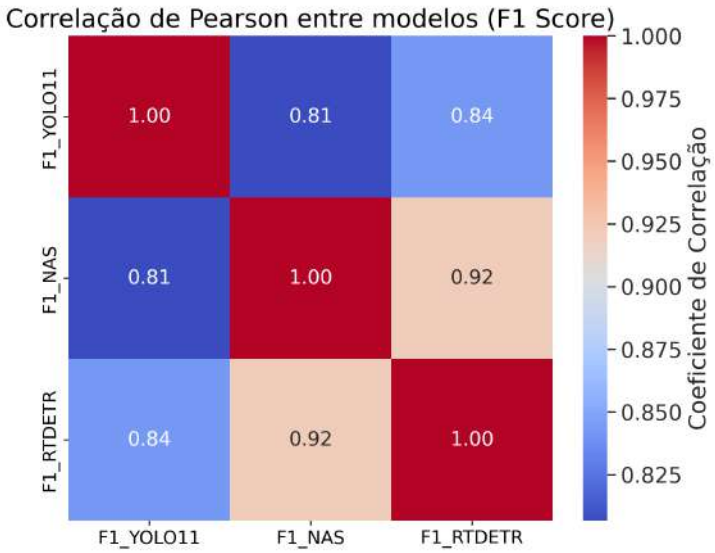

Figura 4.9: Correlação de Pearson entre modelos para a métrica F1-score.

A métrica **F1-score** apresenta uma correlação elevada entre todos os modelos, sugerindo que, apesar das diferenças em valores absolutos, os seus padrões de desempenho são consistentes. No entanto, a análise da ***precision*** (Figura 4.10) revela uma maior variação entre os modelos, indicando abordagens distintas na deteção e gestão de falsos positivos.

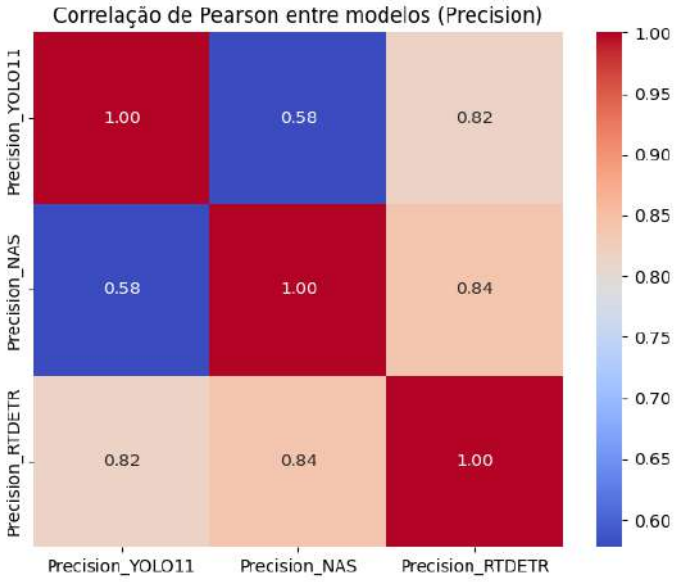

Figura 4.10: Correlação de Pearson entre modelos para a métrica *Precision*.

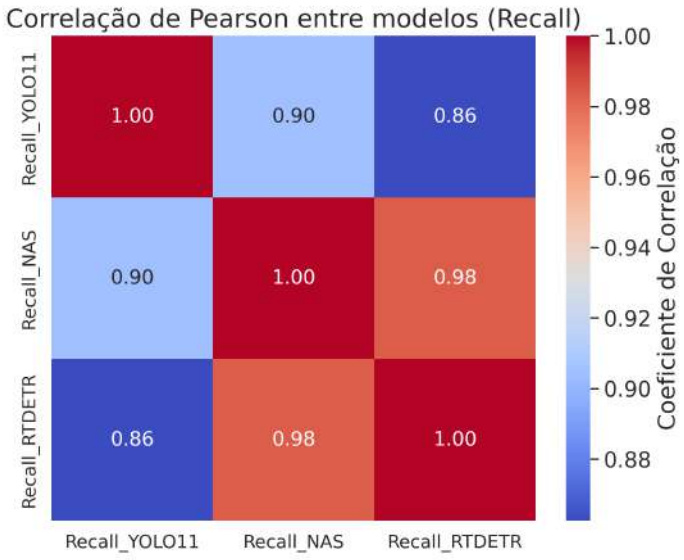

Figura 4.11: Correlação de Pearson entre modelos para a métrica *Recall*.

A maior correlação foi observada entre o **YOLO-NAS** e o **RT-DETR** (0.96), o que indica comportamentos de deteção muito semelhantes entre estes dois modelos, sobretudo em termos de *F1-score*. Por sua vez, o **YOLO11**, embora

também apresente correlações elevadas (0.92 com ambos), evidencia um padrão de desempenho ligeiramente distinto — possivelmente resultado de uma estratégia mais conservadora e centrada na maximização da precisão.

4.4.6 Conclusão da Avaliação Objetiva

A partir da análise objetiva dos resultados, é possível tirar as seguintes conclusões sobre os modelos avaliados:

- O **YOLO-NAS** destacou-se como o modelo com maior ***Recall*** e ***F1-score***, evidenciando-se pela sua elevada sensibilidade na deteção de pessoas, especialmente em alturas intermédias (20–45 metros). No entanto, essa maior capacidade de deteção resultou também num maior número de falsos positivos, o que compromete ligeiramente a sua precisão global;
- O **YOLO11** apresentou a melhor **precisão**, sendo eficaz em cenários com condições mais controladas (baixa altura, curta distância, ângulos frontais), onde a minimização de falsos positivos é essencial. No entanto, demonstrou forte degradação à medida que os fatores de complexidade aumentam (acima dos 30 metros), com quedas acentuadas de desempenho;
- O **RT-DETR**, embora não tenha atingido os melhores resultados globais, foi o modelo com comportamento mais **estável e consistente** em condições de maior desafio (alturas e distâncias elevadas). A sua robustez em cenários visuais degradados — como ângulos verticais (90º) ou pessoas parcialmente ocluídas — sugere um melhor potencial de generalização;
- Todos os modelos sofreram significativa degradação de desempenho com o aumento da **altura** e da **distância**, devido à diminuição do número de píxeis que representam as pessoas. Ainda assim, observou-se que desempenhos fracos em alturas/distâncias curtas podem ser explicados por ângulos desfavoráveis (ex: 90º), onde a vista vertical limita a visibilidade do corpo completo;
- A correlação de Pearson entre os modelos revelou valores elevados (acima de 0.90), indicando padrões de comportamento semelhantes entre os três modelos nos mesmos contextos. A maior correlação foi observada entre o **YOLO-NAS** e o **RT-DETR**, reforçando a proximidade dos seus desempenhos ao longo dos cenários avaliados.

De forma geral, os resultados demonstram o impacto significativo que fatores como altura, distância e ângulo de captação têm na deteção humana

em imagens aéreas, bem como as diferentes estratégias adotadas por cada modelo para lidar com essas variáveis.

4.5 Avaliação Subjetiva

Apesar das métricas objetivas fornecerem uma base sólida para a comparação entre modelos, a análise subjetiva permite compreender melhor o comportamento dos modelos em situações reais, complementando a avaliação quantitativa com observações qualitativas.

Nesta secção, são apresentados exemplos concretos de desempenho dos modelos em diferentes cenários, com especial foco nos seguintes aspetos:

- **Frames com melhor e pior desempenho (F1-score)** – para cada modelo, foram selecionados os **10 melhores** e **10 piores** *frames*, permitindo observar as características visuais associadas a acertos e falhas.

- **Avaliação por grupo e sessão** – foi realizada uma análise qualitativa dos **frames mais representativos** por grupo e sessão, selecionados com base numa combinação fixa de parâmetros (ângulo de 30º, distância de 10m e altura de 5.8m), permitindo avaliar o comportamento dos modelos em diferentes contextos visuais, sob condições geometricamente controladas.

- **Casos extremos com deteção bem-sucedida** – selecionaram-se, para cada modelo, os frames com **maior distância** e **maior altura** onde foi registado pelo menos um verdadeiro positivo (`TP > 0`), de forma a destacar os limites operacionais práticos dos modelos.

Através desta análise visual, pretende-se compreender melhor os fatores que influenciam o desempenho dos modelos — como o número de pessoas na cena, oclusões, iluminação, ângulo de captura, ou a resolução aparente dos indivíduos.

4.5.1 Frames com melhor e pior desempenho

Para além das métricas quantitativas, foi conduzida uma análise visual aos **10 melhores** e **10 piores** frames de cada modelo, com base no *F1-score* obtido em cada imagem testada. Esta avaliação subjetiva permitiu identificar padrões nos sucessos e falhas de deteção, bem como inferir limitações associadas a ângulos, sobreposições e contexto visual.

Nos melhores exemplos, é evidente que os modelos atingem a sua máxima eficácia em condições favoráveis — ângulos oblíquos (30º), curta distância, boa iluminação e separação clara entre os indivíduos, tal como já se havia verificado na avaliação objetiva. A maioria dos **top 10** frames pertence ao grupo 3, sessão 1, com ângulos de 30º e distâncias de 10 metros, onde o número de pessoas é reduzido e a visibilidade do corpo completo está assegurada. É igualmente relevante destacar que, entre todas as sessões, esta apresentava as condições meteorológicas mais favoráveis.

Importa salientar que, mesmo em alguns desses frames de melhor desempenho, existem casos de **sobreposição parcial entre indivíduos**. Ainda assim, os modelos conseguiram efetuar a deteção com sucesso, o que se justifica pelas condições de **baixa altitude e curta distância**, que favorecem uma melhor resolução e delineação das silhuetas.

Por outro lado, os piores exemplos ocorrem em cenários com múltiplos fatores de complexidade, como:

- **Ângulos de 90º** (vista vertical), onde apenas a cabeça ou os ombros estão visíveis;
- **Alturas elevadas**, com forte redução da resolução da pessoa na imagem;
- **Oclusões e sobreposição de indivíduos**, especialmente em agrupamentos densos (Grupo 1 e 2);
- **Forte variabilidade de vestuário ou sombra**, que dificulta a distinção entre pessoas e fundo.

As Figuras 4.12, 4.13 e 4.14 ilustram um exemplo representativo de um frame com deteção correta e outro com deteção falhada para cada um dos três modelos analisados.

Figura 4.12: Comparação entre um bom e um mau exemplo de deteção com o modelo RT-DETR.

Figura 4.13: Comparação entre um bom e um mau exemplo de deteção com o modelo YOLO-NAS.

Figura 4.14: Comparação entre um bom e um mau exemplo de deteção com o modelo YOLOv11.

4.5.2 Avaliação por Grupo e Sessão

De modo a complementar a análise subjetiva, foi realizada uma avaliação qualitativa dos **frames mais representativos** (com deteções reais, mas fora dos extremos de desempenho) em cada **grupo** e **sessão**, utilizando uma combinação fixa de parâmetros: ângulo de 30º, distância de 10 m e altura de 5.8 m. Esta combinação corresponde aos valores mínimos testados para essas variáveis, precisamente porque já foi previamente demonstrado que o aumento da distância e da altura tende a degradar significativamente o desempenho dos modelos. Assim, ao fixar estes parâmetros em condições mais favoráveis, pretende-se eliminar esse fator de influência e focar a análise em aspetos mais subtis e contextuais, como a iluminação, a presença de sombras, a densidade de pessoas, a oclusão parcial e a composição visual da cena.

Esta abordagem permite isolar o desempenho dos modelos de fatores externos, como iluminação, densidade populacional ou complexidade visual do cenário. Para cada grupo/sessão, foi gerado um conjunto de três imagens, cada uma correspondente à *prediction* do modelo RT-DETR, YOLO-NAS ou YOLO11 sobre o mesmo frame.

As Figuras 4.15 mostram exemplos comparativos lado a lado entre os três modelos, para diferentes grupos e sessões.

RT-DETR YOLO-NAS

YOLO11

Figura 4.15: *Predictions* dos três modelos no grupo 1, sessão 1, em condições de iluminação fraca.

Neste exemplo concreto, destaca-se a influência significativa da sombra sobre o desempenho dos modelos. A imagem contém uma área extensa de sombra projetada sobre o grupo de pessoas. Tanto o **YOLO-NAS** como o **RT-DETR** foram capazes de detetar a maior parte das pessoas de forma correta, isto é, conseguiram marcar a *bounding box* no sítio adequado. No entanto, podemos verificar que, no caso do YOLO-NAS, houve não só um falso positivo (um vaso de flores ao fundo), como também um falso negativo (um homem em frente ao carro). Já o RT-DETR não foi capaz de captar o indivíduo que se encontrava agachado. Por fim, o **YOLO11** revelou um desempenho inferior neste cenário, identificando apenas três pessoas bem iluminadas e falhando completamente nas restantes, o que reflete uma clara limitação em condições visuais menos favoráveis, neste caso, devido à sombra. Esta análise reforça a importância de considerar fatores ambientais, como a presença de sombra, na avaliação da robustez de modelos de deteção em contextos reais.

Para reforçar esta observação, foi analisado um frame semelhante do mesmo grupo, mas noutra sessão, captado com as mesmas condições geométricas

(ângulo de 30º, distância de 10 m, altura de 5.8 m), mas com iluminação uniforme e sem sombra significativa. Neste segundo exemplo, todos os modelos — incluindo o YOLO11, que anteriormente tinha falhado — conseguiram detetar corretamente os indivíduos presentes, o que confirma que a falha anterior esteve associada principalmente ao impacto visual da sombra.
Contudo, é importante ressaltar que, mesmo neste cenário iluminado, o modelo **YOLO11** apresentou falhas pontuais na deteção, atribuídas sobretudo à sobreposição de pessoas. Este comportamento revela a sua limitação em lidar com situações de oclusão entre indivíduos, sugerindo uma menor robustez em contextos de densidade populacional elevada.

RT-DETR

YOLO-NAS

YOLO11

Figura 4.16: *Predictions* dos três modelos no grupo 1, sessão 2, em condições de iluminação uniforme.

Outro fator que se revelou relevante na análise qualitativa foi a oclusão parcial. Mesmo em cenários com boa iluminação e distância reduzida, a presença de obstáculos físicos, como árvores ou postes, pode comprometer a deteção de indivíduos. No exemplo abaixo, um dos indivíduos encontra-se parcialmente ocultado por uma árvore na lateral da imagem. Este caso permite avaliar até que ponto cada modelo é capaz de lidar com o desafio da visibilidade reduzida, mantendo uma deteção eficaz mesmo com a silhueta incompleta ou parcialmente distorcida.

RT-DETR YOLO-NAS YOLO11

Figura 4.17: Comparação das *predictions* dos três modelos num cenário com oclusão parcial (indivíduo parcialmente escondido por uma árvore).

Como se pode observar, os modelos apresentam comportamentos distintos face à oclusão. O modelo que melhor lidou com a situação foi o **RT-DETR**, que, apesar da oclusão, foi capaz de captar totalmente o indivíduo. Já o modelo **YOLO-NAS** mostrou-se capaz de reconhecer que havia uma pessoa naquela zona, mas apenas detetou as pernas, ignorando o resto do corpo por este se encontrar parcialmente tapado pelo ramo da árvore. Por último, o **YOLO11** revelou-se incapaz de realizar qualquer deteção naquela área, tendo ignorado por completo a presença de alguém.

Este resultado sugere que a arquitetura do RT-DETR poderá ser mais robusta na extração de padrões parciais ou menos evidentes, mantendo uma deteção fiável mesmo em situações de visibilidade reduzida. Tal característica pode ser vantajosa em aplicações reais, como vigilância urbana ou análise de multidões, onde oclusões são frequentes.

4.5.3 Situações Extremas: Deteções Isoladas

Para além dos exemplos representativos e dos cenários com oclusão parcial, foi realizada uma análise direcionada a **situações extremas**, nas quais os modelos falham em detetar a maioria das pessoas presentes na imagem. Nestes casos, os resultados foram marcados por **baixa cobertura**, com apenas uma ou duas deteções realizadas, apesar da presença de vários indivíduos visíveis no frame.

O objetivo desta análise foi compreender **porque razão, mesmo num contexto de falha generalizada**, os modelos conseguiram identificar algumas pessoas de forma pontual. Essa avaliação focou-se em fatores como a pose, o isolamento espacial, o contraste entre a pessoa e o fundo, ou a ausência de oclusão.

As imagens apresentadas correspondem a capturas realizadas a partir de grandes distâncias, razão pela qual foram **ampliadas** no relatório, de modo a facilitar a visualização das deteções e permitir uma análise mais precisa das características associadas aos indivíduos detetados.

RT-DETR YOLO-NAS YOLOv11

Figura 4.18: Exemplo de situação extrema: deteções isoladas num cenário de altitude elevada e grande distância.

Nestes exemplos, verifica-se que as poucas deteções realizadas pelos modelos ocorrem tipicamente em indivíduos que apresentam características visuais favoráveis, nomeadamente:

- localização em zonas centrais ou próximas da câmara, com boa visibilidade;
- silhuetas bem definidas, frequentemente em pé e orientadas para a câmara;
- separação espacial em relação aos restantes indivíduos, evitando sobreposição;
- contraste elevado entre o vestuário e o fundo, facilitando a segmentação;
- ausência de elementos perturbadores como sombra, vegetação ou objetos urbanos.

Estes padrões sugerem que, mesmo sob condições adversas — como elevada distância, baixa resolução ou ângulos inclinados — os modelos conseguem, pontualmente, realizar deteções quando os alvos se encontram visualmente destacados e isolados. Contudo, o desempenho global permanece fortemente condicionado pela qualidade da informação visual e pela integridade da forma humana observável.

4.5.4 Síntese Comparativa dos Modelos

A análise qualitativa realizada permite identificar diferenças marcantes no comportamento dos três modelos avaliados.

O **RT-DETR** destacou-se pela sua **robustez em cenários complexos**, demonstrando capacidade para detetar indivíduos mesmo em condições de oclusão parcial, baixa visibilidade ou silhuetas incompletas. Esta competência parece refletir a sua arquitetura baseada em transformadores, com maior atenção a padrões espaciais distribuídos.

O **YOLO-NAS**, por sua vez, evidenciou uma **elevada sensibilidade**, sendo eficaz na deteção de pessoas em diversas poses e posições. Esta característica resultou numa cobertura geralmente elevada, ainda que acompanhada de **alguns falsos positivos**, sobretudo em contextos visuais ruidosos (ex.: objetos de forma humanoide ou sombras complexas).

Já o **YOLO11** apresentou um desempenho mais limitado fora de condições ideais. O modelo revelou-se eficaz em cenários bem iluminados e com indivíduos isolados, mas **falhou consistentemente** perante desafios como sombras intensas, oclusões ou aglomerações de pessoas, sugerindo uma menor capacidade de generalização a contextos adversos.

Estas observações mostram que, para além dos resultados quantitativos, a análise subjetiva é crucial para revelar nuances do comportamento dos modelos em situações do mundo real. Com base nos dados obtidos, concluise que o RT-DETR oferece maior consistência em ambientes desafiantes, enquanto o YOLO-NAS privilegia uma abordagem sensível mas menos precisa, e o YOLO11 funciona adequadamente apenas sob condições visuais controladas.

Capítulo

5

Conclusões e Trabalho Futuro

5.1 Conclusões Principais

Através deste trabalho foi possível realizar uma análise abrangente e detalhada do desempenho de três modelos de deteção humana em imagens aéreas: **YOLO11**, **YOLO-NAS** e **RT-DETR**. A avaliação englobou tanto métricas objetivas (*precision*, *recall*, *F1-score*, IoU) como subjetivas (análise visual de frames em diferentes cenários), o que permitiu obter uma compreensão completa dos comportamentos de cada arquitetura.

Para tal, foi necessário desenvolver todo um processo de recolha de dados adaptado ao cenário específico de observação aérea, com recurso a um drone. A captação foi cuidadosamente planeada para abranger diferentes grupos, sessões, distâncias, ângulos e alturas, permitindo uma caracterização variada dos contextos visuais. Posteriormente, as imagens recolhidas foram manualmente anotadas, seguindo um processo rigoroso, com o objetivo de reforçar e estender a base de dados **PedX**.

Os resultados da avaliação revelaram que:

- O **RT-DETR** foi o modelo com melhor desempenho geral, evidenciando robustez em condições difíceis, como oclusões, sombras e silhuetas parciais, bem como nas situações extremas onde apenas uma pessoa foi detetada.

- O **YOLO-NAS** destacou-se pela sua elevada sensibilidade, sendo capaz de detetar pessoas em diversas poses e condições, ainda que à custa de um maior número de falsos positivos em contextos visualmente complexos.

- O **YOLO11**, embora eficaz em ambientes ideais (boa iluminação, baixa densidade populacional), demonstrou fragilidades em cenários com sobreposição, sombras ou baixa resolução, tendo apresentado dificuldades significativas nas situações extremas de deteção isolada.

A análise subjetiva, em particular, permitiu perceber que o desempenho dos modelos está fortemente condicionado por fatores como iluminação, contraste com o fundo, grau de oclusão e densidade populacional. Esta abordagem qualitativa complementou eficazmente as métricas globais, realçando limitações que não seriam visíveis apenas através dos valores médios.

No conjunto, o trabalho desenvolvido permitiu não só testar e comparar modelos de última geração, mas também contribuir para a comunidade com um conjunto de dados anotados de elevada relevância para cenários de reconhecimento humano em contextos aéreos, abrindo caminho a investigações futuras mais especializadas.

5.2 Trabalho Futuro

Apesar da abrangência da análise realizada, vários caminhos interessantes foram identificados para trabalhos futuros, nomeadamente:

- **Reconhecimento Humano** após a deteção: sendo o reconhecimento dependente da qualidade das *bounding boxes* geradas na fase de deteção, este trabalho constitui uma base essencial para etapas mais avançadas. O reconhecimento é um passo natural a seguir, mas só deverá ser abordado após assegurar deteções fiáveis e bem localizadas.

- **Estudos intermédios** como **estimativa de pose**, **normalização** ou **frontalização** das pessoas detetadas poderão ser investigados antes do reconhecimento propriamente dito. Estes processos contribuem para melhorar a consistência da representação visual dos indivíduos detetados, especialmente em ângulos inclinados ou condições adversas.

- **Treino de modelos especializados** para contextos aéreos: todos os modelos utilizados foram generalistas. Um desenvolvimento futuro promissor seria o **re-treino ou afinação** (*fine-tuning*) dos modelos com conjuntos de dados captados por drones, mais adequados às condições específicas deste cenário.

- **Aplicações práticas**: o sistema de deteção pode ter utilidade em diversos contextos reais — vigilância urbana, monitorização de eventos pú-

blicos, operações de busca e salvamento — desde que adaptado às especificidades de cada cenário.

Assim, este trabalho constitui não só uma análise rigorosa ao estado da arte em deteção humana em imagens aéreas, mas também uma base sólida para evoluções futuras no sentido de reconhecimento individual, análise comportamental e aplicação em sistemas inteligentes mais complexos.

Bibliografia

[1] Ultralytics. Yolov11 documentation, 2024. Acedido em junho de 2025.

[2] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

[3] Shay Aharon, Louis-Dupont, Ofri Masad, Kate Yurkova, Lotem Fridman, Lkdci, Eugene Khvedchenya, Ran Rubin, Natan Bagrov, Borys Tymchenko, Tomer Keren, Alexander Zhilko, and Eran-Deci. Super-gradients, 2021.

[4] Ultralytics. Yolo-nas documentation, 2025. Acedido em junho de 2025.

[5] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. Detrs beat yolos on real-time object detection, 2023.