UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

NICOLLE CANUTO NUNES
Orientador: Prof. Dr. Guillermo Cámara Chávez
Coorientador: Me. Hugo Eduardo Ziviani

ROI NA DETECÇÃO DO USO DE EPI: UMA COMPARAÇÃO ENTRE AS REDES NEURAIS ARTIFICIAIS YOLO E DETR

Ouro Preto, MG
2024

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

NICOLLE CANUTO NUNES

ROI NA DETECÇÃO DO USO DE EPI: UMA COMPARAÇÃO ENTRE AS REDES NEURAIS ARTIFICIAIS YOLO E DETR

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Guillermo Cámara Chávez
Coorientador: Me. Hugo Eduardo Ziviani

Ouro Preto, MG
2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Nicolle Canuto Nunes

Region of Interest (ROI) na Detecção do uso de equipamentos de proteção individual (EPI): Uma comparação entre as redes neurais You Only Look Once (YOLO) e Detection Transformer (DETR)

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 8 de Outubro de 2024.

Membros da banca

Guillermo Cámara Chávez (Orientador) - Doutor - Universidade Federal de Ouro Preto
Hugo Eduardo Ziviani (Coorientador) - Mestre - PPGCC/UFOP
Pedro Henrique Lopes Silva (Examinador) - Doutor - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto

Guillermo Cámara Chávez, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 8/10/2024.

Documento assinado eletronicamente por **Guillermo Camara Chavez**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 10/10/2024, às 12:20, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0789758** e o código CRC **62F41A6A**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Resumo

Profissionais da Indústria da Construção Civil (ICC) enfrentam riscos constantes, o que os torna frequentemente suscetíveis a acidentes laborais. O uso inadequado de Equipamento de Proteção Individual (EPI) expõe esses trabalhadores a lesões graves, incluindo cortes, perfurações, quedas e, em casos extremos, fatalidades. Este estudo visa aprimorar a detecção do uso de EPIs na ICC, empregando abordagens baseadas em redes neurais. A pesquisa consiste no treinamento e avaliação das redes *You Only Look Once* (YOLO) e *DEtection TRansformer* (DETR) seguida da aplicação de diferentes *Region of Interest* (ROI), em busca da otimização do desempenho final. As avaliações incluem métricas relevantes, como precision, recall e f-score, que revelam uma superioridade da DETR em relação à YOLO em termos de arquitetura e adaptabilidade.

Palavras-chave: Equipamento de Proteção Individual. Redes Neurais Artificiais. Detecção de Objetos. Aprendizado profundo. YOLO. DETection TRansformer. Região de Interesse.

Abstract

Professionals in the Construction Industry face constant risks, making them frequently susceptible to occupational accidents. Improper use of Personal Protective Equipment (PPE) exposes these workers to serious injuries, including cuts, punctures, falls, and, in extreme cases, fatalities. This study aims to enhance the detection of PPE usage, employing neural network-based approaches. The study involves training and evaluating YOLO and DETR networks, and applying ROI differently generated for performance optimization. Evaluations will include relevant metrics such as precision, recall and F-score, which reveal the superiority of DETR to YOLO in terms of architecture and adaptability.

Keywords: Personal Protective Equipment. Artificial Neural Network. Object Detection. Deep Learning. YOLO. DETection TRansformer. Region of Interest.

Lista de Ilustrações

Lista de Tabelas

Nomenclatura

AP *Average Precision*. 3

CHVG *Colored Hats, Vest, Glass*. v, vi, 11, 16, 17, 19, 20

CIPA Comissão Interna de Prevenção de Acidentes. 1

CLT Consolidação das Leis de Trabalho. 1

CNN *Convolutional Neural Network*. v, viii, 3, 6, 7, 10, 12–14, 25

COCO *Common Objects in Context*. 3, 17

CPU *Central Processing Unit*. 16

DETR *DEtection TRansformer*. iii–vi, viii, 2–4, 10, 11, 15, 17, 22–25, 27–29

EPI Equipamento de Proteção Individual. iii, v, 1–5, 12–15, 18, 19, 23, 29, 30

FFN *Feed-Foward Network*. 10

FPS *Frames Per Second*. 13, 14

GPU *Graphics Processing Unit*. 16, 18

IA Inteligência Artificial. 2, 10, 13, 16

ICC Indústria da Construção Civil. iii, 1, 5

MLP *Multi-Layer Perceptron*. 9

NR Norma Reguladora. 1, 5

ResNet *Residual Neural Network*. 12

RFID *Radio-Frequency Identification*. 2

ROI *Region of Interest*. iii–v, viii, ix, 2, 4, 12, 14, 15, 17–20, 22–29

SHWD *Safety Helmet Wearing Dataset*. 11

VGG *Visual Geometry Group*. 12

ViT *Vision Transformer*. 14

YOLO *You Only Look Once*. iii–vi, viii, 2–4, 7, 8, 11–15, 17, 22–26, 29

Sumário

1 Introdução

Os profissionais que atuam na Indústria da Construção Civil (ICC) enfrentam frequentes exposições a diversos riscos, uma vez que esse setor é notoriamente propenso a acidentes laborais (Silveira, 2005). A utilização adequada de Equipamento de Proteção Individuals (EPIs) se mostra imprescindível para reduzir a incidência de acidentes e assegurar a segurança e proteção destes trabalhadores. Na Figura 1.1 são mostrados alguns exemplos de EPIs.

Figura 1.1 – Exemplos de EPIs.

Fonte: da Autora, adaptado de Tarlengco (2023).

No Brasil, conforme a lei n.º 6.514/77 da Consolidação das Leis de Trabalho (CLT) e também com a Norma Reguladora (NR) 6 do Ministério do Trabalho, o uso dos equipamentos de proteção individual é mandatório sempre que o próprio ambiente de trabalho fornece algum risco à integridade física do profissional (Silveira, 2005). Ainda de acordo com a legislação brasileira, as empresas são obrigadas exercer uma supervisão contínua do uso de EPI (GEISEL, 1977), o que demanda tempo e recursos, como gastos com profissionais da segurança do trabalho e na estruturação de uma Comissão Interna de Prevenção de Acidentes (CIPA). Além disso, o não uso de EPIs pelos funcionários pode resultar em outros custos significativos, como multas, indenizações e licenças médicas (Scherer, 2017).

Diante dos possíveis riscos apresentados, a ausência ou utilização inadequada de EPIs não apenas expõe os trabalhadores a diversos riscos, como lesões físicas, cortes, perfurações, quedas com fraturas e escoriações, mas também pode ter consequências mais graves, chegando até mesmo a resultar em fatalidades em determinadas situações (Maia, 2014).

1.1 Justificativa

1.1.1 Cenário

Por diversas razões, os profissionais muitas vezes negligenciam o uso de EPIs, mesmo cientes de sua importância e obrigatoriedade, podendo inclusive enfrentar rescisão de contrato por justa causa em casos de descumprimento (Rojas; Mazur, 2015) (Guimarães; Raymundo, 2019).

Do ponto de vista das empresas, que tem o dever legal de fiscalizar o uso do equipamento, existem diversas dificuldades de fiscalizar e assegurar o uso desses equipamentos, devido ao grande número funcionários e localidades - o que aumenta os custos para manter pessoas fiscalizadoras em cada uma das obras e torna difícil a garantia da integridade do trabalhador.

Para enfrentar os desafios relacionados à fiscalização do uso de EPIs em obras, foram propstas diversas soluções ao longo dos últimos anos (Nath; Behzadan; Paal, 2020; Ferdous; Ahsan, 2022; Lo; Lin; Hung, 2022; Ahmed et al., 2023; Wang; Cai; Wu1, 2023). Dentre elas, usa-se a *Radio-Frequency Identification* (RFID) (Identificação de Rádio Frequência) para criar portais de identificação do uso de EPIs. Nesses casos, os portais são estrategicamente posicionados nas entradas dos canteiros de obras e utilizam as *tags* RFID, que são pequenos dispositivos eletrônicos de baixo custo e são normalmente fixados aos equipamentos, o que permite o controle automático quando os equipamentos ultrapassam o portal. No entanto, métodos como esse possuem algumas limitações, pois verificam apenas a presença do equipamento no momento da passagem pelo portal, não garantindo seu uso ao longo de todo o expediente e em tempo real (Kelm et al., 2013).

Soluções mais modernas, como aquelas apresentadas em estudos recentes (Ferdous; Ahsan, 2022; Lo; Lin; Hung, 2022; Ahmed et al., 2023; Wang; Cai; Wu1, 2023) propõem o uso de Inteligência Artificial (IA) para realizar a identificação do uso dos aparatos, com seus algoritmos implantados em câmeras distribuídas em diversos pontos das obras. Seguindo a tendência das abordagens mais avançadas, este trabalho propõe a aplicação de algoritmos de aprendizado profundo, utilizando redes neurais como *You Only Look Once* (YOLO) e *DEtection TRansformer* (DETR), visando aprimorar a eficácia e precisão na detecção e monitoramento do uso de EPIs em ambientes de construção.

Diante do cenário apresentado, considerando os riscos associados ao não uso de EPIs e dos altos custos que os empregadores enfrentam para manter a fiscalização adequada, surge a proposta de implementação de um sistema de reconhecimento do uso de EPIs rápido e eficiente, aplicando técnicas de pré-processamento com ROI na tentativa de aprimorar a aplicação.

Essa abordagem oferecerá uma forma prática e econômica para que as empresas possam monitorar efetivamente e garantir o uso adequado dos equipamentos durante o expediente. Dado que algumas empresas já possuem câmeras de monitoramento em suas obras, a implantação da solução proposta teria custos adicionais reduzidos, visto que parte da infraestrutura necessária

para aplicação de modelos de aprendizado de máquina estará disponível.

1.1.2 Histórico Técnico

Nos últimos dez anos, as *Convolutional Neural Network* (CNN) têm desempenhado um papel central e de muito destaque nas tarefas de detecção de objetos. Porém uma nova abordagem recente e também promissora é a aplicação de redes *transformers* (Arkin et al., 2022). Elas são capazes de capturar informações contextuais globais e construir dependências de longo alcance nos objetos para extrair características.

Na tarefa de detecção de objetos (Tarefa *Common Objects in Context* (COCO) 2017), redes *transformers* como a *DEtection TRansformer* (DETR) e suas variações são bem populares. Em Carion et al. (2020), os autores comparam a DETR com CNNs competitivas, como a *Faster-RCNN* e *EfficientDet*. Os resultados mostraram que a detecção de objetos com *transformers* foram mais precisos em relação às arquiteturas *Faster-RCNN* ao utilizar os mesmos *backbones* (*ResNet* 50, *ResNet* 101) para extração de características. Em particular, eles alcançam uma melhoria de até 4,7% em *Average Precision* (AP). No entanto, detectores de objetos baseados em CNNs state-of-the-art, como o EfficientDet-D7, ainda são superiores, ultrapassando os transformers em 3,5 pontos no COCO 2017 vision task.

Nesse estudo, será comparado o desempenho de uma rede CNN, mais especificamente a YOLO, com o de uma rede *transformer*, no caso da DETR. Em seguida, serão submetidas à algumas adaptações, na busca de otimizar sua performance.

Diante dos experimentos, é esperado que a DETR possa performar melhor que a YOLO, visto que estas redes usam uma atenção global, isso é, capturam informações de toda a imagem ao invés de somente uma pequena vizinhança como a CNN. Por exemplo, caso um trabalhador segure o seu capacete nas mãos, a depender do treinamento das redes, pode ser que a CNN classifique erroneamente essa imagem com o rótulo *Worker* + *Hat*, e que a *transformer*, por possuir o módulo de *Multi-Head Attention*, identifique que, apesar de ser um capacete, esse tipo de situação não é de nosso interesse, já que o EPI não está devidamente equipado, rotulando o trabalhador apenas como *Worker*. Essa capacidade de discernimento pode representar uma vantagem crucial em cenários de detecção de conformidade com o uso de EPIs.

1.2 Objetivos

O objetivo principal deste trabalho é comparar o desempenho das redes em sua forma mais básica, para, em seguida, realizar diferentes pré-processamentos na base de dados e testar algumas hipóteses, em busca da otimização da atuação das redes neurais. A proposta final dessa pesquisa visa inovar o processo e diminuir os gastos com a fiscalização do uso de EPIs por meio da mão de obra humana e garantir integridade física e riscos mínimos para os trabalhadores.

Os maiores desafios enfrentados por softwares de reconhecimento de objetos se devem às diferentes angulações e distâncias de captura das imagens, ambientes com baixa iluminação, oclusão por outros objetos e também fundos de imagens muito detalhados (Wu et al., 2023a). Nesse trabalho, serão desenvolvidas algumas técnicas que possam contribuir no combate a estes problemas, como, por exemplo, treinar o modelo com alguns EPIs oclusos e imagens distantes.

1.2.1 Objetivos Específicos

- Remover possíveis ruídos da base de dados, como rótulos mal posicionados;
- Treinar as redes YOLO e DETR para serem capazes de reconhecer EPIs em uma imagem;
- Testar a aplicação de ROI (região de interesse) definida pela *bounding box* da classe de pessoa, gerada a partir do uso de informações da própria base de dados, ou seja, dos *labels* (etiquetas) do *ground truth* (verdade fundamental).
- Testar a aplicação de ROI definida pela *bounding box* da classe de pessoa, gerada a partir da identificação feita previamente pela própria rede, para validar sua autossuficiência.
- Comparar as métricas obtidas em cada rede para cada adaptação, entender os impactos das diferentes abordagens de utilização do ROI e seu papel para cada arquitetura.

1.2.2 Estrutura da Monografia

Este trabalho monográfico é organizado em 4 capítulos, a saber:

- **Capítulo 2:** Revisão Bibliográfica - apresenta os conceitos essenciais utilizados no desenvolvimento deste trabalho, arquitetura das redes e trabalhos com temas relacionados.
- **Capítulo 3:** Experimentos e Resultados - descreve as fases de treinamento das redes, tal como suas métricas. Em seguida, são apresentadas novas experimentações e a interpretação dos seus respectivos resultados.
- **Capítulo 4:** Conclusão - apresenta considerações finais sobre a melhoria das redes e propostas para trabalhos futuros.

2 Revisão Bibliográfica

A contextualização através do estudo teórico é parte essencial de toda e qualquer pesquisa, portanto, a seguir, são apresentados conceitos que irão propiciar uma sólida base para o entendimento integral do estudo aqui desenvolvido.

2.1 Fundamentação Teórica

Inicialmente, serão apresentados os conceitos-chave usados durante a estruturação deste estudo. A compreensão dos seguintes tópicos é indispensável para um entendimento sólido do desenvolvimento.

2.1.1 Equipamentos de Proteção Individual

Ao nível mundial, as indústrias e os ambientes de trabalho podem apresentar condições hostis que podem comprometer a segurança dos trabalhadores. Dentre as principais causas de acidentes, destaca-se, principalmente, o não uso dos equipamentos de segurança, os EPIs, que, segundo a NR6 do Ministério do Trabalho, é todo dispositivo ou produto de uso individual utilizado pelo trabalhador e destinado à proteção de riscos suscetíveis de ameaçar sua segurança na execução de seu trabalho, sejam óculos, protetores auriculares, capacetes, dentre outros (Sehsah; El-Gilany; Ibrahim, 2020).

Os acidentes mais comuns em construções civis são as quedas de grandes alturas, choques elétricos e de queda de materiais nas áreas junto às fachadas, e é a negligência no uso dos EPIs que ocasiona acidentes com ferimentos e lesões de maior gravidade (Silveira, 2005).

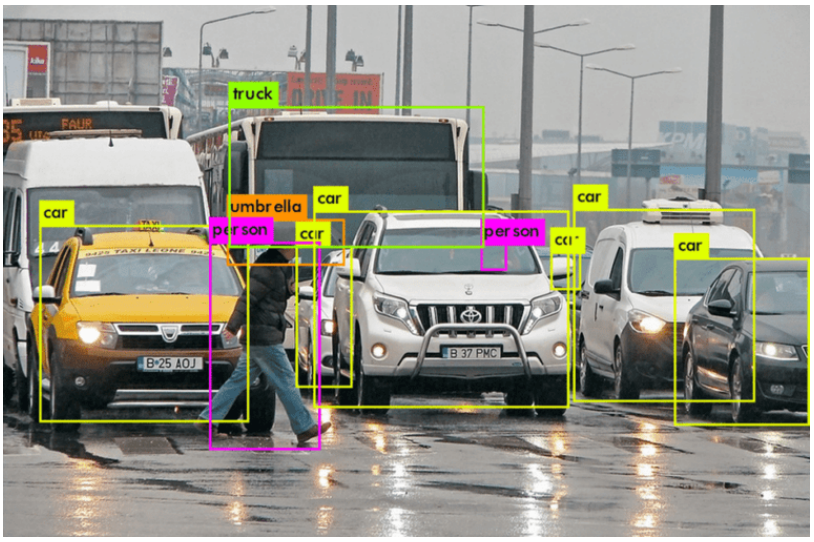
Por mais que as empresas tenham o dever de fornecer EPIs a seus empregados e assegurar que estes estejam sendo utilizados corretamente para resguardar a saúde e a integridade física dos trabalhadores, a participação da Indústria da Construção Civil (ICC) no total de acidentes fatais registrados no Brasil passou de 10,1%, em 2006, para 16,5%, em 2013 (Filgueiras et al., 2015). Já em 2022, aconteceram 612,9 mil acidentes, com cerca de 2000 óbitos registrados, apresentando a maior taxa dos últimos dez anos (TRT, 2022) - dados que evidenciam a necessidade de uma fiscalização mais rigorosa e efetiva do uso correto dos equipamentos de segurança.

2.1.2 Classificação digital de imagens

A classificação de imagens consiste na nomeação da classe de uma instância de um objeto - é o momento em que o computador define a categoria à qual aquela parte da imagem pertence, ver Figura 2.1.

Um marco revolucionário na classificação digital de imagens se deu com o surgimento da rede neural LeNET, desenvolvida por LeCun et al. (1998) na década de 1990, que foi um dos primeiros modelos de rede neural a ser aplicado com sucesso na classificação de dígitos escritos à mão. Em sua publicação, o autor sugere que, no futuro, sistemas de reconhecimento seriam mais eficazes se construídos através do aprendizado de máquina e utilizassem menos heurísticas manuais. Confirmando a hipótese de LeCun et al. (1998), nas últimas décadas, diversos métodos de classificação de objetos foram desenvolvidos utilizando o aprendizado de máquina, tais como redes neurais artificiais e métodos de aprendizado supervisionado não paramétrico (Deepan; Sudha, 2020).

Figura 2.1 – Exemplo de detecção e classificação de objetos.



Fonte: Joshi (2022)

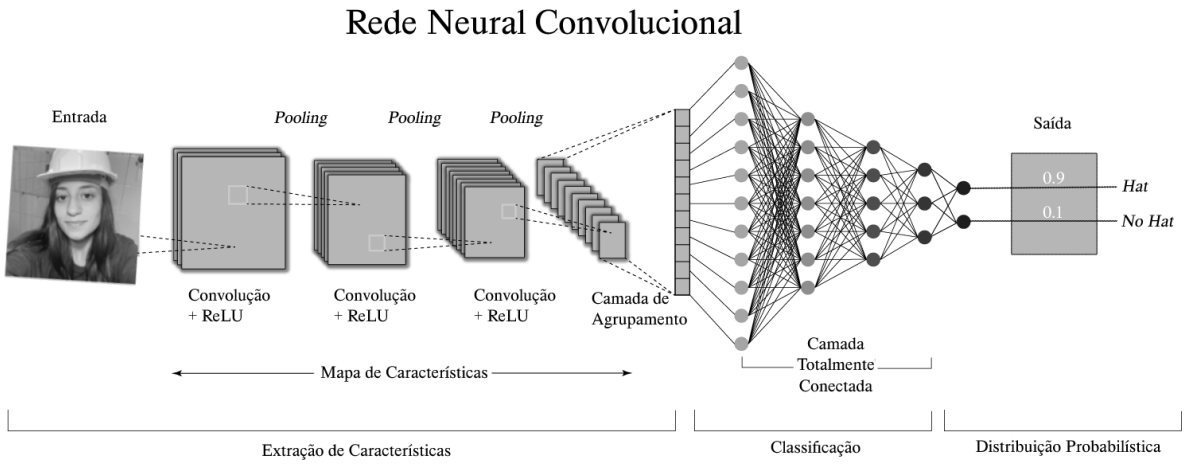
2.1.3 *Convolutional Neural Network* (CNN)

As redes neurais convolucionais, do inglês, CNNs, surgiram em 1998 por Fukushima (1980), e, hoje, após diversas readaptações e aprimoramentos, possuem um grande destaque na área de reconhecimento e classificação de imagens (Fukushima, 1980). Essas redes são feitas por neurônios artificiais, onde cada um deles tem um peso e um viés. A rede tem uma camada de entrada (*input layer*), uma de saída (*output layer*) e várias camadas internas (*hidden layers*), que, por sua vez, consistem em camadas convolucionais, camadas de sub-amostragem (*pooling*) e camadas totalmente conectadas (Dhillon; Verma, 2019). A Figura 2.2 representa a arquitetura de uma CNN genérica.

A continuação, são definidas as camadas componentes de uma CNN:

- **Camada Convolucional:** aplica uma operação de convolução para unir duas coleções de informações. Ela simula o *feedback* de um neurônio a um estímulo visual. Todos os *pixels* recebem o mesmo tratamento, independente de serem de *background* ou *foreground*.

Figura 2.2 – Esquematização de uma CNN.



Fonte: da Autora, adaptado de Shahriar (2023)

São aplicados filtros que reconhecem bordas, texturas e padrões e, ao final, são gerados os mapas de características (O'Shea1; Nash, 2015);

- **Camada de Agrupamento ou *Pooling*:** é usada para reduzir a dimensionalidade dos mapas recebidos da camada anterior, preservando as informações mais importantes e associando a saída de um *cluster* de neurônios a uma camada com um único neurônio (O'Shea1; Nash, 2015);
- **Camada totalmente conectada:** conecta cada neurônio de uma camada a cada neurônio de outra camada. Seu propósito é de classificar as imagens de entrada em diversas classes, baseadas no conjunto de treino. É localizada ao final da CNN pois é responsável por produzir a saída desejada (O'Shea1; Nash, 2015).

2.1.3.1 *You Only Look Once* (YOLO)

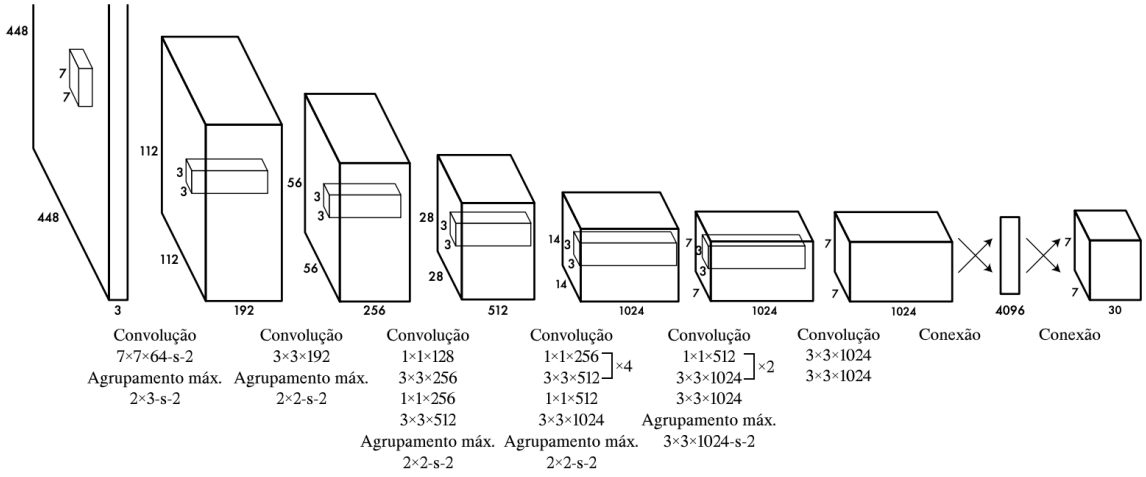
Através do uso de CNNs, uma nova abordagem de detecção de objetos foi proposta como YOLO, uma técnica que trata a detecção de objetos como um único problema de regressão. Isso é, uma única rede neural prediz as probabilidades de *bounding boxes* de objetos e as probabilidades das classes diretamente das imagens inteiras, apenas com uma avaliação. Dessa forma, como o nome diz, só se olha uma vez para os objetos para predizer o que são. Redes YOLO usam uma única rede convolucional para predizer múltiplas *bounding boxes* e suas respectivas classes prováveis para estas caixas. Esta técnica tem se mostrado extremamente rápida, se mostrando capaz de processar até mesmo vídeos em tempo real, com menos de 25 milissegundos de latência.

Diferentemente das técnicas que definem janelas deslizantes para analisar uma região da imagem passo a passo, as redes YOLO veem a imagem completa durante o treinamento, e, então,

implicitamente coletam informações sobre o contexto e outras classes de objetos presentes na imagem (Redmon et al., 2016).

As camadas iniciais (camadas de convolução) da YOLO são pré-treinadas usando o conjunto de dados *ImageNet* e extraem as características da imagem, enquanto as camadas totalmente conectadas (camadas de conexão) preveem as probabilidades de saída e as coordenadas, como ilustrado na Figura 2.3. A arquitetura da rede é inspirada no modelo *GoogLeNet* para classificação de imagem, com 24 camadas convolucionais seguidas por 2 camadas totalmente conectadas (Redmon et al., 2016).

Figura 2.3 – Estrutura de uma YOLO.



Fonte: da autora, adaptado de Redmon et al. (2016).

A saída final da rede, como pode ser visto na Figura 2.3, é um tensor de previsões de dimensões $7 \times 7 \times 30$ (Redmon et al., 2016). Isso é, a imagem final possui uma grade que a divide em 7 linhas e 7 colunas. Cada célula dessa grade é responsável por prever objetos em uma parte específica da imagem. Já o valor 30 refere-se à quantidade de valores por célula - eles representam as previsões associadas a cada célula e incluem informações como as coordenadas das *bounding boxes* dos objetos detectados, as probabilidades associadas a cada predição, e possivelmente outras informações relevantes para o modelo.

Nas camadas de *max-pool*, traduzidas na Figura 2.3 como "Agrupamento máximo", as notações numéricas indicam, respectivamente, as dimensões da janela de *pooling* (2×3, na primeira camada de convolução) e o passo (*stride*) do *pooling* (s-2, na primeira camada de convolução). Isso quer dizer que a janela de *pooling*, de tamanho 2×3, será movida de 2 em 2 unidades (Riad et al., 2022).

O *max-pooling* em cada sub-região retangular seleciona o valor máximo, preservando as características mais relevantes da entrada original. Essa técnica é usada para reduzir a resolução espacial e o número de parâmetros em uma rede neural, mantendo as características mais

importantes para a tarefa de detecção de objetos (Riad et al., 2022).

2.1.4 Redes *Transformers*

As redes *Transformers* surgiram como uma classe de arquitetura Redes Neurais que calcula as saídas usando mecanismos de atenção. Elas estão cada vez mais populares e têm sido muito exploradas para resolver tarefas de classificação de imagens, detecção de objetos e segmentação semântica (Yin et al., 2022) - elas são o primeiro modelo que utiliza somente camadas de transformação. A *transformer* divide cada imagem em uma sequência de 14×14 ou 16×16 *tokens* com uma largura fixa, aplica essas diversas camadas para conseguir criar uma relação global entre os *tokens* e então classificar a imagem (Yuan1 et al., 2021).

Cada rede *Transformer* é construída por vários codificadores (*transformer encoders*). Esses blocos são redes multicamadas que englobam camadas lineares simples, redes *feed-foward* e camadas de autoatenção; essas últimas que são a principal inovação das *transformers* - representadas pelo bloco *Multi-Head Attention* na Figura 2.4 (Poulinakis, 2023).

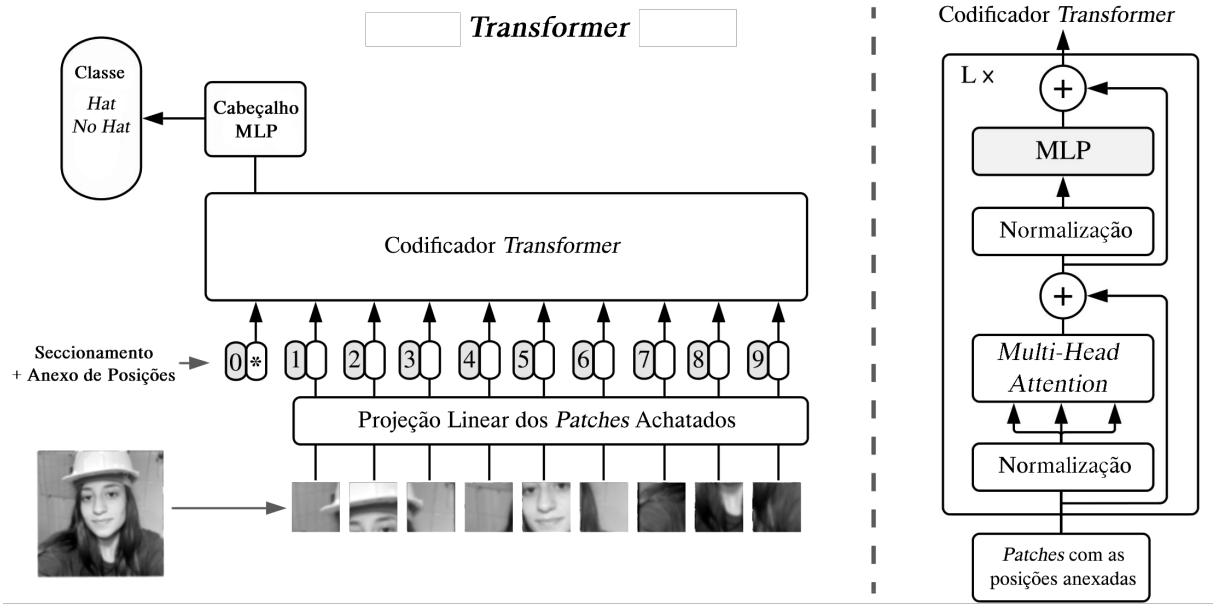
É essa camada que permite que as *Transformers* extraiam informações de grandes contextos, comparando um item de interesse a uma coleção de outros itens na sequência, revelando sua relevância no ambiente (Poulinakis, 2023). Por exemplo, comparar a presença de um capacete a outros elementos na sequência da imagem, como cabeça, luvas, pessoas, mãos, etc. Essa comparação é feita para descobrir a correlação semântica entre as diferentes seções de imagem calculando os índices de atenção entre elas, tarefa que é capaz de relacionar partes das imagens mesmo que estejam distantes (Beal et al., 2020).

Na Figura 2.4 é ilustrada a arquitetura genérica das redes *Transformers*. Primeiramente, a imagem é dividida em pequenas seções quadradas e de mesmo tamanho que serão achatadas, cada uma em um vetor, concatenando os canais de todos os *pixels* em um *patch*. Em seguida, serão anexados marcadores de posição para que a estrutura resultante seja usada como entrada para o codificador *transformer* (Dosovitskiy et al., 2020).

O codificador, estrutura ilustrada à direita da Figura 2.4, é composto de uma pilha de camadas idênticas. Cada camada é composta por duas subcamadas, a *MultiHead Attention* e os Blocos *Multi-Layer Perceptron* (MLP) - estes últimos são uma rede *feedfoward* totalmente conectada e consciente sobre posição, ou seja, são capazes de considerar a posição relativa dos píxeis durante o processamento (Vaswani et al., 2017). Entre cada subcamada, existe uma conexão residual e uma camada de normalização (Xiong et al., 2020).

Tanto durante o pré-treino quanto durante o ajuste fino, existe um cabeçalho de classificação anexado. Ele é implementado por uma MLP com uma camada oculta durante o pré-treino e por uma camada linear durante a etapa de ajuste-fino e é responsável pela classificação das imagens, ou seja, transforma as características aprendidas pela rede na resposta final da predição (Dosovitskiy et al., 2020).

Figura 2.4 – Esquematização da arquitetura de uma *transformer*.



Fonte: da Autora, adaptado de Dosovitskiy et al. (2020).

2.1.4.1 *DEtection TRansformer* (DETR)

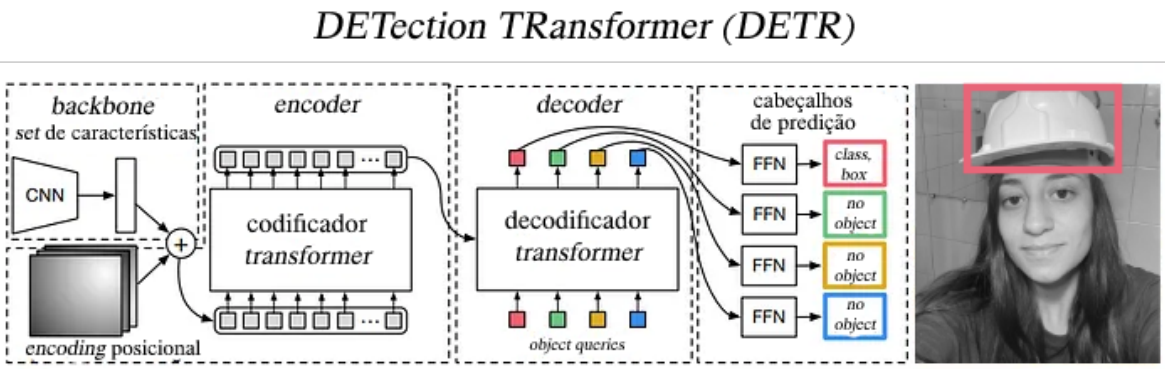
A DETR, modelo proposto pela equipe de IA do *Facebook*, também é uma abordagem inovadora para a detecção de objetos que, por sua vez, utiliza como foco a arquitetura das supracitadas redes *Transformers*. Propostas por Carion et al. (2020), as DETRs tratam a detecção de objetos como um problema de classificação multiclasse, onde cada objeto é associado a uma classe específica e também à sua posição na imagem. A arquitetura *transformer* permite que o modelo leve em consideração as relações espaciais globais entre os *pixels* da imagem, capturando eficientemente contextos complexos (Poulinakis, 2023), ver Figura 2.5.

As DETRs usam uma CNN convencional como *backbone*, que é como a porta de entrada da rede neural, responsável por extrair características importantes da imagem que chega (Carion et al., 2020). A ideia principal é que a CNN redimensione a imagem e aumente a quantidade de canais, como de costume.

Esse modelo achata a imagem em um vetor e a complementa com uma codificação posicional antes de enviá-la para o *encoder*. O *encoder* produz como saída um vetor de mesmo tamanho, que servirá como *side input* para o *decoder*.

Por sua vez, o *decoder* recebe como entrada principal um pequeno número fixo N de *object queries* aprendidos, comumente utilizado o padrão de $N = 100$, o que refere-se, também, ao número de tuplas de predição que serão retornadas. Por fim, cada *embedding* de saída do decodificador é enviado para o cabeçalho de predição. Nele, as *Feed-Foward Networks* (FFNs) (Rede Feed-Foward) irão realizar a classificação, retornando tuplas compostas pela classe (ou também *no object*) seguida das informações da *bounding box* (Carion et al., 2020).

Figura 2.5 – Estrutura de uma DETR.



Fonte: da autora, adaptado de Carion et al. (2020).

2.1.5 Bases de dados e *annotations*

Para realizar o treinamento de uma rede neural de forma efetiva, é necessária uma base de dados bem rotulada.

Trabalhos relacionados comumente citam as bases de dados PictorPPE-v3 (Nath; Behzadan; Paal, 2020; njvisionpower, 2023a), a base *Colored Hats, Vest, Glass* (CHVG) (Ferdous; Ahsan, 2022) e também a *Safety Helmet Wearing Dataset* (SHWD) (njvisionpower, 2023b). Na Tabela 2.1, sumarizam-se informações básicas sobre cada um destes bancos de imagens.

Tabela 2.1 – Informações sobre bases de dados populares

Base de Dados	Tamanho	Rótulos	Formato padrão dos rótulos
PictorPPE v3	1472 imagens	*hat, vest, worker*	*YOLO txt*
CHVG	1699 imagens	*white, blue, red, and yellow hardhats, person head, vest, person, body and safety glas*	*Pascal VOC*
SHWD	7581 imagens	*person, hat*	*Pascal VOC*

Fonte: da Autora

A quantidade de imagens em um conjunto de dados não é, por si só, um indicador confiável da qualidade do conjunto de dados, e sua presença em grande número não assegura necessariamente um bom desempenho do modelo a ser treinado. Também é importante ponderar a qualidade das imagens, diversidade dos cenários, representatividade das classes e assertividade das anotações.

Os formatos dos rótulos têm um papel crucial nas fases pré-treinamento de modelos de detecção de objetos. A depender da rede neural selecionada, pode ser necessário fazer uma conversão de todas as anotações para o formato requerido - por exemplo, as redes YOLO não são capazes processar o formato VOC PASCAL, apresentado em formato de arquivo de extensão

.xml, (Mark; Luc, 2010) fazendo-se necessário realizar a conversão para o formato YOLO TXT, apresentado em formato de arquivo de extensão .txt, antes de iniciar o treinamento (Redmon; Farhadi, 2023).

A escolha bem ponderada da base de dados ideal influencia diretamente no resultado do treinamento da rede neural, visto que os conjuntos de treino serão tidos como verdade, sendo a base do conhecimento do modelo a ser construído.

2.1.6 *Region of Interest* (ROI)

A ROI é um conceito utilizado em processamento de imagens que se refere a partes específicas de uma imagem que são consideradas relevantes para determinado contexto. Ao aplicar essa técnica, em vez de processar a imagem inteira, os algoritmos se concentram apenas nessas áreas de interesse, o que pode melhorar a eficiência e a precisão de tarefas como detecção, classificação e recuperação de imagens (Charaya; Jindal; Kaur, 2016).

Quando aplicada no pré-processamento de imagens para o treinamento de redes neurais, a ROI pode ajudar no aprendizado, pois permite que o modelo extraia características importantes dessas áreas específicas. Isso reduz o tempo de processamento e pode aumentar a relevância dos resultados, já que a rede foca apenas nas partes mais significativas da imagem (Jan; Zainal; Jamaludin, 2020).

2.2 Trabalhos Relacionados

Nos últimos anos, diversos trabalhos têm sido realizados em busca de uma maior garantia de segurança dos trabalhadores em construções. Nas abordagens que utilizam a visão computacional, o uso das redes neurais são uma técnica muito comum para a detecção de objetos em imagens, e, em alguns experimentos, outras redes e algoritmos auxiliares são utilizados para aprimorar o desenvolvimento, seja detectando se o uso está sendo feito de forma correta (Xiong; Tang, 2021; Chen; Demachi, 2020) e, em alguns casos, fazendo a detecção em tempo real (Wu et al., 2023a; Chen et al., 2023; Lo; Lin; Hung, 2022; Ahmed et al., 2023).

2.2.1 Usando CNNs

Em 2020, Nath, Behzadan e Paal (2020) usaram a arquitetura YOLO juntamente com alguns classificadores baseados em CNN ou outros algoritmos, como, por exemplo, a rede *Visual Geometry Group* (VGG) com 16 camadas (VGG-16), uma *Residual Neural Network* (ResNet) com 50 camadas (ResNEt-50), Xception e árvores de decisão para realizar diferentes abordagens de experimentos, onde não somente a existência do EPI é identificada, mas também é feita a verificação de se o equipamento está sendo usado de forma correta por todos os trabalhadores. Nesse caso, a base de dados utilizada (Pictor-PPE-v3) foi criada por meio de mineração na web e

fotos enviadas por pessoas, contando com cerca de 1400 imagens anotadas e 4700 exemplos de trabalhadores usando diferentes combinações de EPIs.

Ainda em 2020, Delhi, Sankarlal e Thomas (2020) publicaram uma pesquisa onde um *framework* é desenvolvido de forma integral para identificar em tempo real o uso de EPIs, usando modelos de CNN juntamente à YOLOv3. Nesse caso, o modelo identifica o uso ou ausência de jaquetas de segurança e capacetes, e retorna 4 categorias possíveis para cada trabalhador: *SAFE, NOT SAFE, NO HARD HAT* e *NO JACKET*. Uma base coletada de gravações de câmeras de vigilâncias em diversas obras foi utilizado para treinar o modelo, usando cerca de 2500 imagens e esse modelo alcançou um f1-*score* de 96%.

No ano de 2022, Lo, Lin e Hung (2022) verificaram, em tempo real, o uso adequado de EPIs. Nesse estudo, foram usadas as redes YOLOv3, YOLOv4 e YOLOv7, com grande capacidade de detecção de objetos em tempo real. Um conjunto de dados foi coletado e elaborado especificamente para esta pesquisa, contendo cerca de 11.000 imagens e 88.000 rótulos. Os resultados obtiveram uma precisão de 97% com 25 *Frames Per Second* (FPS).

Em 2023, Wu et al. (2023b) desenvolveram um modelo baseado por conhecimento de atributo, que realiza a identificação semântica dos atributos a serem reconhecidos. A detecção inicial foi feita com coletes e capacetes, mas obteve resultados muito produtivos, mesmo submetendo o sistema à testes com diferentes poses, com oclusão e com imagens cheias de detalhes. O algoritmo proposto utiliza a rede YOLOv5 para localizar os trabalhadores, auxiliada pelo *Deep Sort*, um algoritmo usado como localizador para extrair o movimento e lidar com as poses em *frames* contínuos.

Ainda em 2023, Ahmed et al. (2023) também realizaram a detecção em tempo real do uso de EPIs. O dataset utilizado contava com oito classes diferentes de objetos, como capacetes, óculos, coletes e cabeças. Foi utilizada uma abordagem de detecção em duas etapas, usando a Rede Convolucional Rápida baseada em Regiões (Faster RCNN) e o valor médio de precisão obtido foi de 96%.

2.2.2 Usando Microsoft Azure Custom Vision AI

Em 2020, Balakreshnana et al. (2020) demonstraram o uso de IA para identificar EPIs, especificamente, óculos de proteção. Nesse caso, uma abordagem não muito comum foi utilizada: serviço da Microsoft, *Custom Vision AI* e alguns dispositivos de baixo custo promoveram uma plataforma treinada usando imagens públicas, na tentativa de criar um sistema capaz de identificar em tempo real o uso de EPIs. Esse modelo é mais simples, proposto apenas para identificar apenas o uso de óculos de proteção, e sua precisão final foi de 93.6% para a detecção de pessoas, 66.7% para a detecção de faces e apenas 50% para a detecção dos óculos, além de ser incapaz de identificar a diferença entre óculos de proteção e óculos de grau.

2.2.3 Usando *Transformers*

Em 2020, é introduzido o modelo *Vision Transformer* (ViT)-FRCNN (Beal et al., 2020), que combina um *backbone* baseado em *transformers* com uma rede de detecção. Nesse estudo, os modelos são capazes de transferir representações aprendidas entre as tarefas de detecção. Os autores fazem uma breve comparação entre *Transformers* e CNNs, mostrando como as *Transformers* podem ter desempenho igual ou superior em tarefas de detecção de objetos. A base de imagens utilizando em alguns dos experimentos foi o *OBJECTNET-D*, que possui um leque bem variado de objetos. Nesse caso, as pesquisas não foram focadas apenas em EPIs, mas trazem uma abordagem interessante no avanço da detecção de objetos em geral.

Em 2022, Li et al. (2022) realizaram um estudo do uso de *Transformers* para a detecção de objetos genéricos, onde essa rede é utilizada como um *backbone* (extrator de características de entrada) simples, ao invés de um modelo de classificação de imagens, sem o uso de redes hierárquicas muito complexas. Essa nova estrutura do modelo é chamada de ViTDet, que, nesse caso, atua mais coletando as características dos objetos ao invés de tentar necessariamente classificá-los - tarefa que será feita em outras camadas da aplicação. Nesse trabalho, os autores utilizam de janelas de atenção não sobrepostas, usando poucos blocos de propagação.

Chen et al. (2023), inseriram as *Transformers* com outro papel - nesse trabalho, destaca-se o uso das redes YOLO com uma abordagem diferente, usando as redes *Transformers* para suprir suas limitações. Neste estudo, foi introduzida a YOLOv5s-gnConv, que conta com um mecanismo adicional de modelagem espacial baseado em autoatenção *dot-product*, inspirado pelas *Transformers*, conseguindo, então, um melhor desempenho para capturar informações de longo alcance, além de tornar a YOLO mais veloz e eficaz, obtendo uma precisão de 97,32%, com FPS de 56,12.

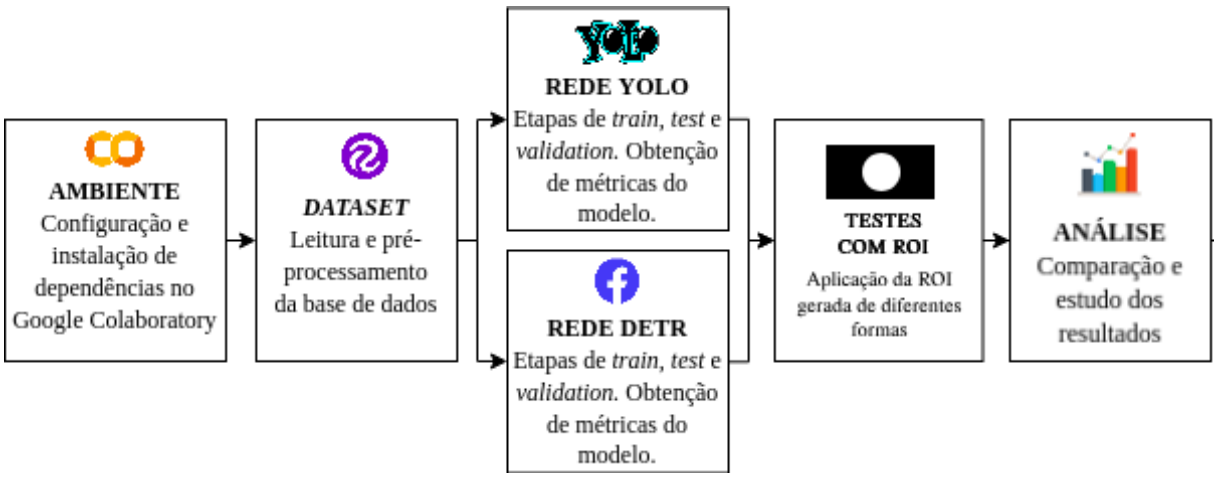
2.2.4 Síntese

Em resumo, os avanços recentes na aplicação de redes neurais para a detecção de EPIs na indústria da construção demonstram um progresso significativo na garantia da segurança dos trabalhadores, com uma variedade de abordagens e técnicas mostrando promessa em termos de precisão e eficiência em tempo real. Esse trabalho pretende verificar a possibilidade de aprimoramento da precisão das redes através do uso da ROI, dado que não é uma abordagem comumente utilizada nesse tipo de tarefa.

3 Experimentos e Resultados

Durante os experimentos, as redes YOLO e DETR serão postas à prova, treinadas e avaliadas sob as mesmas condições (bases de dados, parâmetros comuns, métricas de avaliação). Em seguida, as redes serão impostas à testes visando otimizar sua performance através do uso de ROIs geradas por diferentes técnicas. Por fim, os resultados obtidos serão analisados para entender o impacto das ROIs em suas diferentes abordagens na detecção de EPIs. A Figura 3.1 mostra o passo a passo das tarefas que serão realizadas ao longo deste trabalho.

Figura 3.1 – Mapa de etapas seguidas durante o desenvolvimento desse estudo.



Fonte: da autora.

Para dar início, foi selecionado um ambiente de desenvolvimento e um *dataset*. A partir dessas definições, as redes YOLO e DETR foram treinadas três vezes, sem nenhum tipo de pré-processamento, de moto a se obter três modelos com pesos a serem avaliados. Para cada um dos experimentos realizados neste trabalho, as métricas apresentadas são a média da avaliação destes três modelos.

Após treinadas as redes, as ROIs foram geradas de diferentes maneiras e aplicadas nos conjuntos de teste e de validação - para, em seguida, avaliar os três modelos obtidos na etapa anterior. Obtidos os resultados dos experimentos, foi feita uma análise comparativa entre as métricas obtidas em cada avaliação.

Nas subseções seguintes, cada uma das etapas citadas na Figura 3.1 será descrita, de modo a fornecer mais detalhes quanto às decisões, passos e implementações.

3.1 Ambiente de desenvolvimento

Os experimentos realizados neste trabalho foram feitos utilizando o *Google Colaboratory*, um serviço gratuito de nuvem do *Google Research*. Através dele é possível executar diretamente pelo navegador os algoritmos que aqui foram desenvolvidos em *python*, utilizando recursos de hardware mais potentes, que podem diminuir o tempo que seria gasto no treinamento das redes em uma máquina local.

Para acelerar o processo de treinamento dentro do *Colaboratory*, será necessário substituir o uso padrão de *Central Processing Unit* (CPU) pelo uso de *Graphics Processing Unit* (GPU), que é um acelerador de *hardware* muito usado em treinamentos usando IA, especializado em lidar com cálculos complexos e operações longas. Com sua arquitetura altamente paralela, as GPUs são capazes de realizar muitos cálculos de forma simultânea, sendo uma alternativa mais eficiente para casos como este, em que é necessário processar um grande volume de dados (Kimm; Paik; Kimm, 2021). Além dessa configuração, foi necessária a instalação do *framework ultralytics* para treinar e avaliar as redes.

3.2 Base de dados

O *dataset* selecionado para dar início aos experimentos foi uma versão do *Colored Hats, Vest, Glass* (CHVG) devido à sua organização, diversidade de imagens, precisão das *bouding boxes* e por ter sido utilizado em um trabalho relacionado (Ferdous; Ahsan, 2022). O conjunto de treino possui 1359 imagens, o de validação 170 e o de teste 170. A disposição das classes em cada *split* pode ser vista na Tabela 3.1.

Tabela 3.1 – Quantidades de objetos por classe no banco de dados CHVG

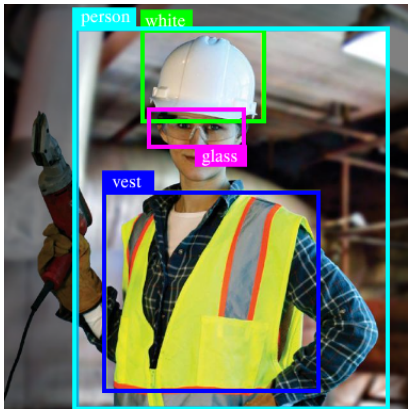
					Colored hats			
Split / Classe	***person***	***glasses***	***vest***	***head***	***blue***	***red***	***white***	***yellow***
Treino	3783	415	1758	603	460	370	1066	603
Teste	462	68	192	34	41	72	125	139
Validação	429	49	187	93	42	54	101	107

Fonte: da autora.

Os rótulos do dataset são armazenados em um arquivo .txt com o mesmo nome da imagem correspondente, conforme ilustrado na Tabela 3.2. Cada linha do arquivo contém a anotação da imagem, que inclui, no início, o ID da classe, seguido de quatro números que representam as coordenadas da *bounding box*: as posições x e y do centro da *bounding box*, além da largura e altura, respectivamente. Todos os valores são normalizados entre $[0, 1]$. No exemplo apresentado, é possível identificar um objeto de cada classe: 6 (*white hat*), 5 (*vest*), 3 (*person*) e 1 (*glasses*).

Conhecidas as especificidades do *dataset*, é necessário avaliar, segundo a implementação de cada rede, se será necessário pré-processar as anotações ou fazer alguma adaptação antes de

Figura 3.2 – Imagem do banco CHVG referente ao rótulo da Tabela 3.2.



Fonte: Ferdous e Ahsan (2022), acesso em 30 de agosto de 2024.

Tabela 3.2 – Exemplo de anotação da ubicação de objetos no formato YOLO TXT.

Classe	x_{centro}	y_{centro}	Altura	Largura
6	0.4968750	0.1812500	0.2906250	0.2250000
5	0.5203125	0.7101563	0.5218750	0.4890625
3	0.5687500	0.5328125	0.7562500	0.9218750
1	0.4820312	0.3125000	0.2328125	0.0875000

Fonte: Ferdous e Ahsan (2022), acesso em 30 de agosto de 2024.

iniciar a etapa de treinamento.

3.3 Treinamento

Na fase de treinamento, as redes terão seus pesos ajustados para que elas aprendam a classificar corretamente as imagens de entrada. Como explicado durante a fundamentação teórica deste trabalho, a YOLO e a DETR possuem arquiteturas diferentes, o que pode implicar no tempo, formato dessa fase e resultados obtidos com a aplicação do ROI.

Apesar das especificidades, o objetivo comum do treinamento é minimizar a diferença entre as previsões da rede e os resultados desejados, otimizando a performance de ambas. Para cada uma das redes, foram realizados três treinamentos, salvando o arquivo com os melhores pesos de cada um deles.

3.3.1 Experimentos

Inicialmente, ao instanciar o modelo, foram criadas redes limpas (*from scratch*), sem pré-treinamento, o que resultou em um desempenho insatisfatório. Como solução, foram empregados modelos oficial pré-treinados em uma ampla variedade de objetos do cotidiano, usando o banco de dados COCO (Redmon; Farhadi, 2023). Essa estratégia, denominada *fine-tuning*, provou ser mais eficaz, já que, dada uma rede apta a reconhecer padrões gerais, é mais eficiente simplesmente

ajustá-la para lidar com detecções mais específicas (Amisse; Jijón-Palma; Centeno, 2021), como é o caso dos EPIs.

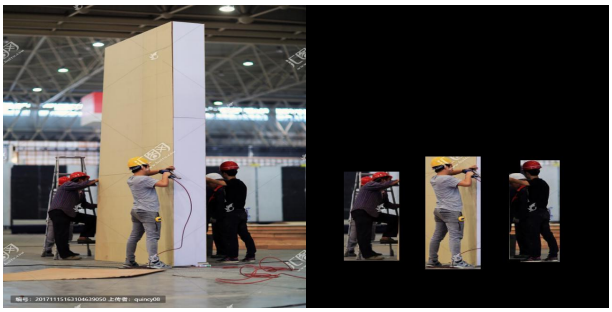
Os três treinos de cada rede foram executados com os parâmetros em suas configurações *default*, definindo apenas o número de épocas para 70, que se refere à quantidade de passagens completas de todas as imagens de treino pela arquitetura da rede neural. Esse valor foi determinado por ser próximo ao utilizado em trabalhos relacionados (comumente 100 épocas), mas ligeiramente reduzido de modo a não ultrapassar os limites de GPU do *Google Colaboratory* em sua versão gratuita.

3.4 Geração de ROI

Para avaliar o impacto da limitação da quantidade de *pixels* disponíveis no desempenho de detecção das redes, será utilizada a ROI para isolar regiões da imagem onde existem humanos, excluindo *pixels* que não sejam relevantes na detecção de EPIs, dado que, por exemplo, um capacete em cima de uma mesa não deve ser considerado como um verdadeiro positivo, já que não está sendo utilizado.

As redes neurais utilizadas operam com tamanhos de imagens padronizados e pré-definidos (Lv et al., 2023) (Wang; Liao, 2024), e, por esse motivo, recortar os humanos das imagens redimensionando-as seria muito custoso e improdutivo. Ao invés disso, a ROI foi aplicada de modo a manter a dimensão da imagem, apagando os *pixels* que não fossem de interesse, simplesmente colorindo-os com preto. A hipótese é que isso fará com que essa informação seja desconsiderada, poupando a rede de processar *pixels* que trariam informações irrelevantes, como outros objetos ou excesso de *background*. A Figura 3.3 ilustra, à esquerda, uma imagem original, e, à direita, um exemplo de pré-processamento com ROI, mantendo a dimensão original de 640 x 640 px.

Figura 3.3 – Exemplo de pré-processamento com ROI sem alterações na dimensão da imagem.



Fonte: da autora, adaptado de Ferdous e Ahsan (2022), acesso em 17 de setembro de 2024.

Em todos os experimentos que serão descritos, a ROI fora aplicada apenas nos conjuntos de teste e de validação, utilizados para avaliar cada um dos três modelos obtidos nos treinos. O

treinamento das redes foi feito num conjunto de imagens sem nenhum tipo de pré-processamento, com todas as informações disponíveis, para não prejudicar as redes nesta etapa de aprendizado.

A princípio, a ROI foi obtida e aplicada na imagem utilizando as dimensões originais da *bounding box* de humano, isso é, sem realizar nenhum tipo de alteração na *bounding box* que definiria a região de interesse. Porém, logo nos primeiros experimentos, notou-se uma queda expressiva na *performance* das redes, trazendo métricas extremamente menores do que as obtidas na abordagem sem pré-processamento.

No caso de ROI geradas pela própria rede neural, identificou-se que havia diversas discrepâncias entre as *bounding boxes* de humanos da região de interesse obtida e as fornecidas pela base de dados, levando, muitas vezes, à eliminação inadequada de *pixels* durante a criação da ROI. A Figura 3.4, ilustra um exemplo deste caso, onde a ROI remove *pixels* indevidos, invadindo ligeiramente a *bounding box* de humano definida pela base de dados. Isto é, existe uma coluna de *pixels* preta dentro da caixa desenhada em amarelo - que é a bounding box de humano definida pela base - ou seja, ao criar a ROI, a rede selecionou uma *bounding box* de humano menor e acabou apagando *pixels* que a base definia como verdadeiros positivos (faziam parte da classe humano).

Figura 3.4 – Exemplo de *pixels* removidos dentro da *bounding box* da classe de humano da base de dados.

Fonte: CHVG, acesso em 6 de Agosto de 2024.

Além disso, até mesmo quando as *bounding boxes* das regiões de interesse foram coletadas da própria base, se estas fossem aplicadas em suas dimensões originais, notou-se que alguns EPIs, majoritariamente os capacetes, possuíam *bounding boxes* que se estendiam para além das *bounding boxes* de humanos, resultando na exclusão de *pixels* que ainda continham informações relevantes, como pode ser visto na Figura 3.5.

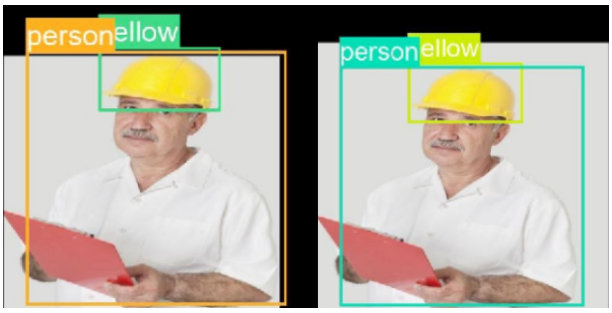
Por estas razões, optou-se por realizar a expansão da janela da região de interesse em todos os experimentos, de modo que a deleção errônea de *pixels* não mais acontecesse. Após a identificação de amostras de imagens afetadas pelos problemas supracitados, foi feita uma inspeção visual, de modo a definir uma taxa fixa de aumento da ROI que fosse capaz de suprimir o problema e, ainda, adicionar uma faixa de segurança para prevenir possíveis casos isolados onde seria necessária uma ROI ainda maior do que a calculada (*edge cases*). Dessa forma, a taxa

Figura 3.5 – Exemplo de *pixels* removidos na faixa onde a *bounding box* do capacete excede ligeiramente a *bounding box* de humano.

Fonte: CHVG, acesso em 6 de Agosto de 2024.

de expansão da *bounding box* foi definida para 20%, como mostra a Figura 3.6.

Figura 3.6 – Antes (esquerda) e depois (direita) da expansão da ROI para a correção do descarte indevido de pixels.



Fonte: da autora, adaptado de Ferdous e Ahsan (2022), acesso em 18 de setembro de 2024.

3.5 Avaliação

A seguir, são descritas as métricas que serão utilizadas para mensurar o desempenho final das redes. A seleção foi feita via inspeção dos trabalhos relacionados (Seção 2.2), priorizando métricas que apareciam com maior frequência. As avaliações foram feitas utilizando o valor *default* de 50% para interseção de *bounding boxes* considerada como acerto.

3.5.1 *Precision*

A precisão, ou *precision* em inglês, indica a proporção de instâncias positivas corretamente identificadas (TP) em relação ao total de instâncias classificadas como positivas ($TP+FP$). Essa métrica tem foco na redução de casos em que o modelo erroneamente classifica uma instância negativa (por exemplo, *no hat*) como positiva (*hat*) (Ting, 2017). Sua fórmula é dada por:

$$P = \frac{TP}{TP + FP} \tag{3.1}$$

Seu valor varia de 0 a 1, onde 1 indica uma precisão perfeita, enquanto valores mais baixos indicam uma maior taxa de falsos positivos (Ting, 2017).

3.5.2 *Recall*

O *recall* enfatiza a capacidade de um modelo em identificar corretamente todas as instâncias positivas (TP) em relação ao total de instâncias que são verdadeiramente positivas ($TP + FN$) (Ting, 2017). Sua fórmula é expressa como:

$$R = \frac{TP}{TP + FN} \tag{3.2}$$

Assim como o *precision*, o *recall* também varia de 0 a 1, onde 1 indica um *recall* perfeito, enquanto valores mais baixos indicam uma maior taxa de falsos negativos (Ting, 2017).

3.5.3 *F-score*

O *F-score* ou *F1-score* é uma métrica que combina *precision* e *recall* em um único valor, proporcionando uma medida geral do desempenho de um modelo de classificação (Ting, 2017). Ele é calculado pela fórmula harmônica entre *precision* (P) e *recall* (R):

$$F = 2 \cdot \frac{P \cdot R}{P + R} \tag{3.3}$$

Essa fórmula também fornece um valor que varia de 0 a 1. Um *F-score* de 1 indica um equilíbrio perfeito entre precisão e *recall* (Ting, 2017).

3.5.4 Matriz de Confusão

A Matriz de Confusão é uma tabela que organiza as previsões do modelo em quatro categorias diferentes com base nos resultados da classificação: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN) (Ting, 2017), ver Tabela 3.3.

Tabela 3.3 – Demonstração da disposição dos valores na matriz de confusão.

		Valor predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

3.6 Resultados

Após obtidos os resultados sem nenhum tipo de pré-processamento na base de dados, as redes tiveram selecionados os melhores pesos dos seus respectivos treinos, e estes foram avaliados em conjuntos de dados pré-processados com a aplicação de ROIs em suas divisões de teste e de validação, para testar a hipótese de que a definição de um ROI auxilia a rede na etapa de detecção, dado que existirão menos *pixels* com informações que possam confundi-la ou sobrecarregá-la.

3.6.1 Técnicas de geração da ROI

As ROIs foram geradas seguindo duas abordagens: a primeira utiliza os *labels* da base de dados para realizar o recorte, enquanto a segunda usa a própria rede neural para realizar a pré-detecção. Se o uso do ROI for realmente positivo, acredita-se que utilizar os *bounding boxes* fornecidos pela própria base para gerar o ROI deverá trazer resultados melhores de desempenho da rede, já que utiliza do *ground truth*. Apesar disso, espera-se que a utilização de ROIs gerados pela detecção da própria rede neural também aumente a precisão dos resultados, já que também diminuirá a região de busca e a quantidade de *pixels* com informações desnecessárias.

Ainda, reforça-se que a validação desta segunda hipótese é indispensável para a viabilidade da aplicação da ROI em ambientes onde as redes treinadas seriam utilizadas na prática, já que o *ground truth* da base de dados pode apenas ser obtido nesse cenário de experimentação.

3.6.1.1 Sem pré-processamento

Este primeiro experimento serviu como linha de base para comparar e avaliar a performance dos modelos, já que fora realizado sem nenhum tipo de pré-processamento na base de dados.

As redes YOLO e a DETR obtiveram resultados satisfatórios, trazendo seu *F1-score* médio com valores de 0,8969 e 0,9114 (Tabela 3.4), respectivamente. Esses valores demonstram que elas são capazes, por si só, de aprender a identificar objetos com alta taxa de acerto.

3.6.1.2 Utilizando a própria rede para criar ROIs

Nesse experimento, utilizou-se das próprias redes para realizar um pré-processamento nas seções de teste e de validação da base, fazendo recortes nas regiões de interesse - onde eram detectados humanos. Esses *sets* foram utilizados na validação dos três treinos previamente executados, de modo a fazer com que as condições de cada experimento sejam as mesmas.

As redes apresentaram um *F1-score* médio de 0,8832 para a YOLO e 0,9135 para a DETR (Tabela 3.4). Isso demonstra que a YOLO, a princípio, não lidou bem com as ROIs geradas

Tabela 3.4 – Métricas obtidas em cada experimento

		YOLO	DETR
Sem ROI	precision	91,25	93,8
	recall	88,2	88,63
	f-score	89,69	91,14
ROI detectada	precision	87,41	88,59
	recall	89,25	94,29
	f-score	88,32	91,35
ROI da base	precision	82,56	88,66
	recall	90,46	93,54
	f-score	86,32	91,03

por sua própria detecção. Por outro lado, a DETR, teve um avanço sutil em seu resultado, que será discutido e melhor analisado na Seção 3.6.2.

3.6.1.3 Utilizando a própria base para criar ROIs

No terceiro experimento, a ROI foi implementada utilizando as próprias *bounding boxes* de humanos da base de dados, eliminando a necessidade de uma detecção preliminar pelas redes. Essa abordagem foi capaz reduzir a complexidade do processo e avaliar o desempenho das redes na detecção de EPIs de forma isolada, com a garantia de que não haveria nenhuma região selecionada erroneamente.

Nesse experimento, o *F1-score* médio obtido foi de 0,8632 pela YOLO e 0,9103 pela DETR (Tabela 3.4). Isso reforça a percepção de que a YOLO performa melhor sem a aplicação do ROI. Quanto à DETR, esse valor foi curiosamente o mais baixo dos experimentos, ainda que apresente uma diferença ínfima.

3.6.2 Análise

Durante a avaliação das redes, notou-se que a DETR trouxe resultados muito estáveis, com números muito próximos entre si, enquanto a YOLO foi notadamente prejudicada com a aplicação das regiões de interesse. Na Tabela 3.4, são comparados os resultados de precisão, *recall* e *F1-score* médios obtidos nos experimentos descritos na seção anterior.

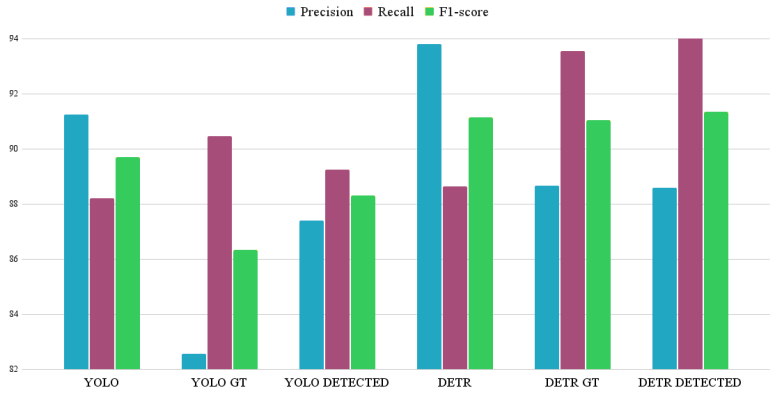
As diferenças de desempenho podem ser atribuída à maneira como cada rede processa os dados. Durante o treino, a YOLO coleta informações sobre o contexto e outras classes de objetos presentes para correlacionar com os objetos de interesse, portanto, ao aplicar a ROI, ela pode ter trazido um efeito contrário ao desejado: atuado como um oclusor do contexto ao invés de realmente ser produtiva no objetivo de poupar a rede de processar *pixels* inúteis.

Por outro lado, a DETR processa os dados de maneira diferente, utilizando uma abordagem baseada em tokens (Poulinakis, 2023) para representar a imagem, como citado na Seção 2.1.4.1. Seu mecanismo de atenção permite que a rede considere diferentes partes da imagem de maneira

mais global e menos local. Isso permitiu que ela mantivesse uma performance mais estável, pois, apesar de acabar gerando alguns *tokens* completa ou parcialmente pretos, a rede ainda foi capaz de identificar as classes de interesse de forma isolada.

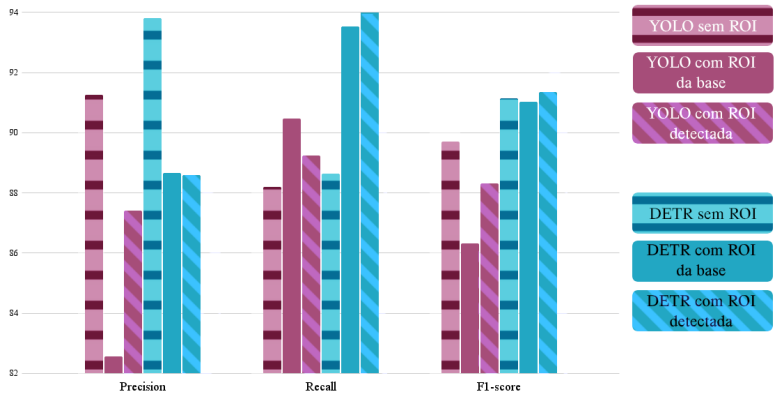
As Figuras 3.7 e 3.8 trazem uma visualização das métricas obtidas para cada um dos experimentos, onde YOLO-GT e DETR-GT representam os experimentos cuja ROI foi obtida através dos *bounding boxes* da própria base de dados (*ground truth*); YOLO *DETECTED* e DETR *DETECTED* representam os experimentos onde as próprias redes foram responsáveis pela pré-detecção.

Figura 3.7 – Comparação das métricas por experimento.



Fonte: da autora.

Figura 3.8 – Comparação dos experimentos agrupados por métrica.



Fonte: da autora.

O fato de a DETR ter obtido um melhor desempenho ao usar os *bounding boxes* de sua própria detecção contraria as expectativas, dado que a hipótese era que ambas as redes performariam melhor utilizando as *bounding boxes* da base, do *ground truth*. Acredita-se que esse efeito aconteceu, pois ao detectar os humanos no pré-processamento, a rede considerava uma área maior do que a definida como verdadeira pela base de dados. Essa diferença fez com

que a ROI se expandisse além do esperado, e, consequentemente, a DETR, se beneficiou do seu próprio erro, já que possuia uma área maior da imagem para construir os seus *tokens* com maior contexto.

A Figura 3.9 mostra um exemplo deste caso, onde a ROI é maior do que o esperado, devido a uma detecção errônea da rede DETR durante o pré-processamento. Pode-se notar que a ROI se excede à esquerda, já que o modelo detectou o humano considerando todo o braço - enquanto a base o ensina a considerar somente a porção do tronco e cabeça.

Figura 3.9 – Exemplo de caso onde a DETR se beneficia do seu erro cometido durante a construção da ROIs.

Fonte: da autora, adaptado de Ferdous e Ahsan (2022).

Além disso, pode-se notar que, apesar dos experimentos terem refletido de forma diferente em cada arquitetura, a DETR superou a YOLO em todos os cenários. Isso pode ser um indício de uma superioridade intrínseca às redes *transformers* se comparadas à uma CNN. Ainda, apesar de não validar integralmente a hipótese de que a ROI seja benéfica, foi possível observar que a DETR possui uma alta capacidade de adaptação a diferentes configurações de entrada.

As Tabelas 3.5a, 3.5c, 3.5b, 3.6a, 3.6c e 3.6b ilustram as matrizes de confusão em cada um dos experimentos. Através delas, é possível visualizar que, o erro que mais se destaca em ambas as redes é quando elas detectam um humano onde não existe (*predicted Human* x *label Background*). Uma percepção interessante é que, apesar de a YOLO não ter melhorado sua precisão média com a aplicação da ROI, a quantidade de vezes em que ela confundiu o *background* com uma pessoa foi reduzida com os experimentos que aplicavam a ROI (de 0.33 no experimento sem pré-processamento para 0.26 e 0.27 nos experimentos com ROI da detecção e extraída da base, respectivamente). Isso ocorreu, pois possivelmente os *pixels* responsáveis pelos falsos positivos de humanos foram removidos com o recorte. No entanto, a rede ainda teve sua precisão prejudicada nos experimentos com ROI, pois, apesar de confundir o *background* com um humano com menos frequência, a YOLO apresentou uma diminuição dos acertos em verdadeiros positivos (detecções de um objeto de interesse onde realmente existe).

Já a DETR apresenta um aumento na taxa desse mesmo erro, o que explica sua queda de precisão nos experimentos com ROI, já que teve um aumento expressivo em suas detecções falso positivo (detecta um objeto onde não há, como confundir o background com um humano).

Tabela 3.5 – Matrizes de confusão da YOLO

(a) Sem ROI

	blue	glass	head	person	red	vest	white	yellow	background
blue	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
glass	0.00	0.78	0.00	0.00	0.00	0.01	0.00	0.00	0.12
head	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.10
person	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.33
red	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.02
vest	0.00	0.00	0.00	0.00	0.00	0.93	0.05	0.00	0.14
white	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.10
yellow	0.00	0.00	0.00	0.00	0.02	0.00	0.03	0.97	0.15
background	0.00	0.22	0.10	0.05	0.07	0.07	0.06	0.03	0.00

(b) ROI detectada pela rede

	blue	glass	head	person	red	vest	white	yellow	background
blue	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
glass	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.14
head	0.00	0.00	0.95	0.00	0.00	0.00	0.01	0.00	0.14
person	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.26
red	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.02
vest	0.00	0.00	0.00	0.00	0.00	0.90	0.05	0.00	0.16
white	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.10
yellow	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.90	0.13
background	0.02	0.24	0.05	0.07	0.11	0.10	0.08	0.10	0.00

(c) ROI da base

	blue	glass	head	person	red	vest	white	yellow	background
blue	1.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03
glass	0.00	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.09
head	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.18
person	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.17
red	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.01
vest	0.00	0.00	0.00	0.00	0.00	0.91	0.06	0.00	0.22
white	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.01	0.17
yellow	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.92	0.12
background	0.00	0.22	0.12	0.04	0.07	0.09	0.05	0.07	0.00

Tabela 3.6 – Matrizes de confusão da DETR

(a) Sem ROI

	blue	glass	head	person	red	vest	white	yellow	background
blue	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
glass	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.08
head	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.08
person	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.39
red	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.02
vest	0.00	0.00	0.00	0.00	0.00	0.96	0.06	0.01	0.18
white	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.01	0.12
yellow	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.09
background	0.00	0.14	0.06	0.04	0.07	0.04	0.05	0.01	0.00

(b) ROI detectada pela rede

	blue	glass	head	person	red	vest	white	yellow	background
blue	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
glass	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.11
head	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.10
person	0.00	0.00	0.00	0.97	0.00	0.01	0.00	0.00	0.46
red	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.01
vest	0.00	0.00	0.00	0.00	0.00	0.93	0.06	0.00	0.20
white	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.01	0.05
yellow	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.99	0.06
background	0.02	0.14	0.03	0.03	0.06	0.06	0.04	0.00	0.00

(c) ROI da base

	blue	glass	head	person	red	vest	white	yellow	background
blue	0.98	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.04
glass	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.12
head	0.00	0.00	0.96	0.00	0.00	0.00	0.01	0.01	0.11
person	0.00	0.00	0.00	0.96	0.00	0.01	0.00	0.00	0.52
red	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.01
vest	0.00	0.00	0.00	0.00	0.00	0.92	0.06	0.00	0.10
white	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.05
yellow	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.05
background	0.02	0.10	0.04	0.04	0.06	0.07	0.06	0.03	0.00

Isso pode estar ocorrendo se o cálculo de intersecção dos *bounding boxes* para validar um acerto estiver resultando em valores menores do que o limiar para considerá-lo um verdadeiro positivo. Nesse cenário, a DETR pode ter seus *bounding boxes* sobre os humanos com valores baixos de sobreposição com o *ground truth* e ser penalizada, tendo sua detecção classificada como falso positivo.

Ainda, vale destacar que a rede DETR teve um aumento relevante em seu *recall*, apresentando mais verdadeiros positivos em outras classes (onde realmente existiam os objetos de interesse e a rede os identificou corretamente). Isso é, com a aplicação da ROI, a rede *transformer* deixou com que menos objetos passassem despercebidos.

4 Considerações Finais

4.1 Conclusão

Neste trabalho, investigamos a eficácia de diferentes abordagens de pré-processamento na detecção de EPIs utilizando as redes neurais artificiais YOLO e DETR. Os resultados obtidos refletem como diferentes arquiteturas de rede podem ser afetadas pela maneira como as imagens são processadas. Esses resultados reforçam que a escolha da arquitetura de rede neural e da forma de pré-processamento são decisões que devem ser tomadas de maneira mútua.

A introdução de ROIs não demonstrou melhora significativa como esperado - acabou prejudicando a detecção da rede YOLO, enquanto a DETR se mostrou adaptável e robusta, mantendo resultados bons e estáveis em todos os experimentos.

Apesar das diferenças nas métricas dos experimentos realizados, nota-se que ambas as redes obtiveram uma boa taxa de acerto e poderiam ser utilizadas na automação da detecção de EPI em construções sem a necessidade de pré-processamento das imagens, já que o ROI parece ter atuado como um dificultador para a YOLO e um fator não muito relevante para a performance da DETR.

Analisando as precisões obtidas nas avaliações das redes, pode-se dizer que é viável implementar o uso delas para a detecção de EPI na indústria em tempo real. Em um contexto de vídeo, nas câmeras de vigilância, a captura de múltiplos *frames* proporcionaria diversas oportunidades para que as redes identifiquem corretamente os EPIs em questão de segundos.

Na tomada de decisão de qual rede será utilizada e se a ROI deve ser aplicada, é importante levar em consideração a capacidade computacional e o tempo de resposta, visto que as redes *transformers*, apesar de apresentarem uma precisão maior, podem exigir mais poder computacional para operar - o que pode ser agravado caso seja implementada o pré-processamento de detecção da região de interesse.

4.2 Trabalhos Futuros

Para trabalhos futuros, várias direções podem ser exploradas para aprimorar ainda mais a detecção de EPIs. Uma possibilidade é investigar outras técnicas de pré-processamento que possam ser mais adequadas para as redes, como a suavização dos *pixels* fora da *bounding box* ao invés de colori-los em preto: essa estratégia poderia ajudar a preservar informações contextuais.

Além disso, é importante avaliar o benefício de se utilizar cada uma das redes, levando em consideração o custo computacional e tempo de detecção. Isso poderia ser feito aplicando as

redes treinadas à câmeras de segurança em tempo real e registrando o momento no tempo em que cada rede detecta uma classe que surge na imagem capturada.

Outra possibilidade é validar o uso do EPI de forma mais precisa - apesar de a base utilizada conter apenas imagens de EPIs em uso, para realizar essa abordagem, seria interessante fazer a interseção entre as classes de pessoa e os EPIs (capacete, óculos e colete), verificando se eles estão de fato sendo utilizados pela pessoa. Para isso a base de dados precisaria conter exemplos de EPIs que não estão em uso (exemplos negativos), onde esses itens apareceriam sem estar sendo usados por uma pessoa, e assim não seriam considerados verdadeiros positivos.

Referências

AHMED, M. I. B.; SARAIREH, A. R. L.; AL-QARAWI, S.; MHRAN, A.; AL-JALAOUD, J.; AL-MUDAIFER, D.; AL-HAIDAR, F.; ALKHULAIFI, D.; YOULDASH, M.; GOLLAPALLI, M. Personal protective equipment detection: A deep learning-based sustainable approach. *MDPI*, v. 1, 2023.

AMISSE, C.; JIJóN-PALMA, M. E.; CENTENO, J. A. S. Fine-tuning deep learning models for pedestrian detection. *Boletim de Ciências Geodésicas*, Universidade Federal do Paraná, v. 27, n. 2, p. e2021013, 2021. ISSN 1982-2170. Disponível em: <https://doi.org/10.1590/s1982-21702021000200013>.

ARKIN, E.; YADIKAR, N.; XU, X.; AYSA, A.; UBUL, K. A survey: object detection methods from cnn to transformer. 2022.

BALAKRESHNANA, B.; RICHARDSB, G.; NANDAB, G.; MAOB, H. Ppe compliance detection using artificial intelligence in learning factories. *Elsevier*, v. 1, 2020.

BEAL, J.; KIM, E.; TZENG, E.; PARK, D. H.; ZHAI, A.; KISLYUK, D. Toward transformer-based object detection. 2020.

CARION, N.; MASSA, F.; SYNNAEVE, G.; USUNIER, N.; KIRILLOV, A.; ZAGORUYKO, S. End-to-end object detection with transformers. 2020.

CHARAYA, S.; JINDAL, S.; KAUR, B. Content based image retrieval using selective region matching with region of interest and svm. *International Journal of Computer Applications*, v. 137, p. 28–33, 03 2016.

CHEN, H.; LI, Y.; WEN, H.; HU, X. Yolov5s-gn conv: detecting personal protectiv equipment for workers at height. *frontiers*, v. 1, 2023.

CHEN, S.; DEMACHI, K. A vision-based approach for ensuring proper use of personal protective equipment (ppe) in decommissioning of fukushima daiichi nuclear power station. *MDPI*, v. 1, 2020.

DEEPAN, P.; SUDHA, L. *Chapter 8 - Object Classification of Remote Sensing Image Using Deep Convolutional Neural Network*. Academic Press, 2020. 107-120 p. (Intelligent Data-Centric Systems). ISBN 978-0-12-816385-6. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128163856000088>.

DELHI, V. S. K.; SANKARLAL, R.; THOMAS, A. Detection of personal protective equipment (ppe) compliance on construction site using computer vision based deep learning techniques. *Frontiers*, v. 1, 2020.

DHILLON, A.; VERMA, G. K. Convolutional neural network: a review of models, methodologies and applications to object detection. *Springer Nature*, v. 1, p. 28, 2019.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

FERDOUS, M.; AHSAN, S. M. M. Ppe detector: a yolo-based architecture to detect personal protective equipment (ppe) for construction sites. 2022. Disponível em: <https://figshare.com/articles/dataset/CHVG_Dataset/19625166>.

FILGUEIRAS, V. A.; SILVA, A. da; SOUZA, G. L. de; SOUZA, I. F. de; SCIENZA, L. A.; BRANCHTEIN, M. C.; CUNHA, S. F. da. *Saúde e segurança do trabalho na construção civil brasileira*. 1. ed.. ed. [S.l.]: UNICAMP, 2015.

FUKUSHIMA, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Buol Cybern*, v. 36, p. 193–202, 1980.

GEISEL, E. *DA SEGURANÇA E DA MEDICINA DO TRABALHO*. 1977. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l6514.html>.

GUIMARãES, Y. da S.; RAYMUNDO, P. M. J. C. A obrigatoriedade do uso de equipamento de proteÇÃo individual (epi). 2019.

JAN, M. M.; ZAINAL, N.; JAMALUDIN, S. Region of interest-based image retrieval techniques: a review. *IAES International Journal of Artificial Intelligence*, IAES Institute of Advanced Engineering and Science, v. 9, n. 3, p. 520, 2020.

JOSHI, N. *What is object detection?* 2022.

KELM, A.; MEINS-BECKER, A.; PLATZ, D.; KHAZAEE, M. J.; COSTIN, A. M.; HELMUS, M.; TEIZER, J. Mobile passive radio frequency identification (rfid) portal for automated and rapid control of personal protective equipment (ppe) on construction sites. 2013.

KIMM, H.; PAIK, I.; KIMM, H. Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning. p. 312–319, 2021.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *IEEE*, v. 1, p. 46, 1998.

LI, Y.; MAO, H.; GIRSHICK, R.; HE, K. Exploring plain vision transformer backbones for object detection. *Facebook AI Research*, v. 1, 2022.

LO, J.-H.; LIN, L.-K.; HUNG, C.-C. Real-time personal protective equipment compliance detection based on deep learning algorithm. *MDPI*, v. 1, 2022.

LV, W.; XU, S.; ZHAO, Y.; WANG, G.; WEI, J.; CUI, C.; DU, Y.; DANG, Q.; LIU, Y. *DETRs Beat YOLOs on Real-time Object Detection*. 2023.

MAIA, A. L. M. AnÁlise preliminar de riscos em uma obra de construÇÃo civil. 2014.

MARK, E.; LUC, V. G. The pascal visual object classes (voc) challenge. 2010.

NATH, N. D.; BEHZADAN, A. H.; PAAL, S. G. Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, v. 112, p. 103085, 2020.

NJVISIONPOWER. *Safety Helmet Wearing*. 2023. [Online; acesso em 6 de Novembro de 2023]. Disponível em: <https://github.com/ciber-lab/pictor-ppe>.

NJVISIONPOWER. *Safety Helmet Wearing*. 2023. [Online; acesso em 6 de Novembro de 2023]. Disponível em: <https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset>.

O'SHEA1, K.; NASH, R. An introduction to convolutional neural networks. 2015.

POULINAKIS, K. *Are Transformers replacing CNNs in Object Detection?* 2023. Disponível em: <https://www.picsellia.com/post/are-transformers-replacing-cnns-in-object-detection>.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You only look once: Unified, real-time object detection. *University of Washington and Allen Institute for AI and Facebook AI Research*, v. 5, p. 10, 2016.

REDMON, J.; FARHADI, A. *YOLO - Ultralytics*. 2023. Disponível em: <https://docs.ultralytics.com/>.

RIAD, R.; TEBOUL, O.; GRANGIER, D.; ZEGHIDOUR, N. Learning strides in convolutional neural networks. *CoRR*, abs/2202.01653, 2022. Disponível em: <https://arxiv.org/abs/2202.01653>.

ROJAS, A. P. K.; MAZUR, C. da S. A higienizaÇÃo correta das mÃos e a utilizaÇÃo de epi´s pela enfermagem em um setor pÓs-cirÚrgico. 2015.

SCHERER, F. *Fiscalização do uso de EPIs - Equipamentos de Proteção Individual*. 2017. Disponível em: <https://www.jusbrasil.com.br/artigos/fiscalizacao-do-uso-de-epis-equipamentos-de-protecao-individual/494235312>.

SEHSAH, R.; EL-GILANY, A.-H.; IBRAHIM, A. M. Personal protective equipment (ppe) use and its relation to accidents among construction workers. *Med Lav*, v. 1, p. 28, 2020.

SHAHRIAR, N. *What is Convolutional Neural Network — CNN (Deep Learning)*. 2023. Disponível em: <https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5>.

SILVEIRA, C. A. Acidentes de trabalho na construção civil identificados através de prontuários hospitalares. *Scielo*, v. 1, 2005.

TARLENGCO, J. *Everything You Need to Know About PPE Safety*. 2023. Disponível em: <https://safetyculture.com/topics/ppe-safety/>.

TING, K. M. Precision and recall. In: SAMMUT, C.; WEBB, G. I. (Ed.). *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017. p. 990–991. Disponível em: <https://doi.org/10.1007/978-1-4899-7687-1_659>.

TRT. *Aumento de mortes de trabalhador no país reforça campanha Abril Verde*. 2022. Disponível em: <https://www.trt21.jus.br/noticias/noticia/aumento-de-mortes-de-trabalhador-no-pais-reforca-campanha-abril-verde>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. 2017.

WANG, C.-Y.; LIAO, H.-Y. M. Yolov9: Learning what you want to learn using programmable gradient information. 2024.

WANG, Z.; CAI, Z.; WU1, Y. An improved yolox approach for low-light and small object detection: Ppe on tunnel construction sites. *Oxford*, v. 1, 2023.

WU, X.; LI, Y.; LONG, J.; ZHANG, S.; WAN, S.; MEI, S. A remote-vision-based safety helmet and harness monitoring system based on attribute knowledge modeling. *MDPI*, v. 1, 2023.

WU, Y. L. X.; LONG, J.; ZHANG, S.; WAN, S.; MEI, S. A remote-vision-based safety helmet and harness monitoring system based on attribute knowledge modeling. *MDPI*, v. 1, 2023.

XIONG, R.; TANG, P. Pose guided anchoring for detecting proper use of personal protective equipment. *Elsevier*, v. 1, 2021. Disponível em: <https://github.com/ruoxinx/PPE-Detection-Pose>.

XIONG, R.; YANG, Y.; HE, D.; ZHENG, K.; ZHENG, S.; XING, C.; ZHANG, H.; LAN, Y.; WANG, L.; LIU, T.-Y. On layer normalization in the transformer architecture. 2020.

YIN, H.; VAHDAT, A.; ALVAREZ, J. M.; MALLYA, A.; KAUTZ, J.; MOLCHANOV, P. A-vit: Adaptive tokens for efficient vision transformer. *CVF*, v. 1, 2022.

YUAN1, L.; CHEN, Y.; WANG, T.; YU, W.; SHI, Y. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *National University of Singapore*, v. 1, 2021.