UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Detecção de anomalias em veículos ferroviários através de Visão Computacional

Victor Henrique Mendes Pereira

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Victor Henrique Mendes Pereira

Detecção de anomalias em veículos ferroviários através de Visão Computacional

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Valdir Grassi Junior

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

M538d

Mendes Pereira, Victor Henrique
Detecção de anomalias em veículos ferroviários através de Visão Computacional / Victor Henrique Mendes Pereira; orientador Valdir Grassi Junior. -- São Carlos, 2023.
60 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Detecção de defeitos. 2. Aprendizado Profundo. 3. Visão Computacional. 4. Ferrovia. 5. Yolo. I. Grassi Junior, Valdir , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Victor Henrique Mendes Pereira

Detection of anomalies in railway vehicles through Computer Vision

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Valdir Grassi Junior

Original version

São Carlos

2023

Esta monografia é dedicada aos meus filhos Murilo e Beatriz e a minha esposa Juliana, que me apoiaram a todo tempo na busca do aprendizado.

AGRADECIMENTOS

Ao Orientador Prof. Dr. Valdir Grassi Junior, pela sua paciência, atenção e todo o direcionamento neste trabalho acadêmico.

Aos professores e professoras deste excelente MBA, pela dedicação, empatia e a beleza de compartilhar o conhecimento.

Aos amigos e profissionais da Ferrovia Rumo pelo apoio na obtenção de informações relevantes para este estudo aplicado.

"Não sabendo que era impossível, foi lá e fez"
Jean Cocteau

RESUMO

PEREIRA, V. **Detecção de anomalias em veículos ferroviários através de Visão Computacional**. 2023. 60p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A identificação visual de defeitos ou desvios de qualidade em equipamentos e produtos é um tema amplamente estudado na industria e da mesma forma nas ferrovias. Esta identificação de defeitos que geralmente é realizada por pessoas além de demandar altos recursos financeiros, exige muita atenção, sendo também muito complexo padronizar os critérios entre as pessoas. No Brasil as ferrovias transportaram em 2021 21,5% de todo o volume de cargas incluindo os demais modais. A qualidade dos ativos ferroviários pode afetar tanto a produtividade das ferrovias quanto a segurança ferroviária. No período de 2006 a 2023 houveram em média mais de hum mil acidentes por ano nas ferrovias Brasileiras, sendo 14,68% causados por falhas no material rodante. Com o objetivo de realizar a detecção automática de defeitos em material rodante de trens em movimento, este trabalho aborda a utilização de técnicas de detecção de objetos existentes em modelos estado da arte de aprendizado profundo. Foram capturados vídeos reais da parte inferior dos vagões que trafegam na maior ferrovia de carga do Brasil e posteriormente feito a rotulagem das imagens dos rodeiros e do detector de descarrilamento de vagão (DDV). Foram comparadas características das principais redes de aprendizado profundo quanto a precisão mAP e velocidade FPS, tendo sido escolhido para o estudo de caso o modelo YOLOV8 que até o início de 2023 era considerado o modelo estado da arte para detecção de objetos. A rede YOLOV8 foi treinada no banco de dados criado, tendo atingido uma acurácia mAP50-95 de 59,1% e uma velocidade de 88FPS durante a inferência no ambiente *Google Colab* e uma velocidade de 30FPS durante a inferência com o algoritmo *DeepStream* utilizando um dispositivo Jetson AGX Orin. Foi implementando também no ambiente *Google Colab* através do algoritmo *DeepSort* o rastreamento e identificação dos objetos, além de ter sido implantado contadores que identificam quantos objetos cruzaram a linha de referência definida no vídeo.

Palavras-chave: Monografia. Detecção de Defeitos. Aprendizado Profundo. Visão Computacional. Ferrovia. Material Rodante. Vagão. Manutenção. Yolo.

ABSTRACT

PEREIRA, V. **Detection of anomalies in railway vehicles through Computer Vision**. 2023. 60p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

The visual identification of defects or quality deviations in equipment and products is a subject widely studied in the industry and in the same way in the railways. This identification of defects, which is usually carried out by people, in addition to demanding high financial resources, requires a lot of attention, and it is also very complex to standardize the criteria among people. In Brazil, in 2021, railroads transported 21.5% of the entire volume of cargo, including other modes. The quality of rail assets can affect both rail productivity and rail safety. In the period from 2006 to 2023, there were on average more than one thousand accidents per year on Brazilian railways, 14.68% of which were caused by failures in the rolling stock. With the objective of automatically detecting defects in rolling stock of moving trains, this work addresses the use of existing object detection techniques in state-of-the-art deep learning models. Real videos of the lower part of the wagons that travel on the largest freight railroad in Brazil were captured and then the images of the wheel sets and the wagon derailment detector (DDV) were labeled. Characteristics of the main deep learning networks were compared in terms of accuracy mAP and FPS speed. YOLOV8 model, which until the beginning of 2023 was considered the state-of-the-art model for object detection, was chosen for this case study. The YOLOV8 network was trained on the created database, reaching a mAP50-95 accuracy of 59.1% and a speed of 88FPS during inference in the Google Colab environment and a speed of 30FPS during inference with the DeepStream algorithm using a Jetson AGX Orin device. The tracking and identification of objects was also implemented in the Google Colab environment using the DeepSort algorithm. In addition, counters that identify how many objects crossed the reference line defined in the video was also implemented.

Keywords: Monography. Defect Detection. Deep Learning. Computer vision. Railroad. Rolling Stock. Wagon. Maintenance. Yolo.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAR	*Association of American Railroads*
ABNT	Associação Brasileira de Normas Técnicas
ANTF	Associação Nacional dos Transportadores Ferroviários
ANTT	Agência Nacional de Transportes Terrestres
DDV	Detector de Descarrilamento de Vagão
ERA	*European Union Agency for Railwats*
IA	Inteligência Artificial
ICMC	Instituto de Ciências Matemáticas e de Computação
NBR	Norma Técnica Brasileira
OTIF	*Intergovernmental Organisation for International Carriage by Rail*
USP	Universidade de São Paulo
YOLO	*You Only Look at Once*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

O Transporte ferroviário de cargas tem um papel muito importante na sociedade. No Brasil este modal correspondeu em 2021 a 21,5% da participação de todos os meios de transporte, equivalente a 506,8 milhões de toneladas úteis (TU), enquanto em países como Rússia e Austrália esta participação é de 81% e 55% respectivamente (ANTF, 2022), demonstrando o potencial de crescimento ferroviário no Brasil e a importância deste modal globalmente.

Um fator muito importante para o transporte ferroviário é a produtividade, sendo uma das suas principais variáveis o tempo de trânsito entre a origem e o destino. O tempo de trânsito é afetado por diferentes fatores, entre eles pela confiabilidade do material rodante. A incidência de falhas ou defeitos de material rodante, seja nos vagões ou nas locomotivas gera a necessidade de parada do trem e consequentemente eleva o tempo de trânsito previsto, reduzindo a produtividade da ferrovia.

Como medida para reduzir o risco de acidentes ferroviários devido às falhas mais comuns que historicamente ocorreram em vagões, a Associação das Ferrovias Americanas (AAR - *Association of American Railroads*) recomenda que seja realizada uma inspeção pré-partida do trem (AAR, 2023), no Brasil conhecida como revistamento. Durante o revistamento, os vagões que estão em um trem já formado em um pátio ferroviário, passam por uma inspeção visual, onde mecânicos caminham de vagão em vagão inspecionando itens pré-definidos. O tempo para realização do revistamento por vagão depende diretamente da quantidade de pessoas envolvidas na atividade, sendo esta quantidade definida pelas ferrovias, de acordo com a produtividade esperada. De acordo com o levantamento de dados de uma grande ferrovia americana, a inspeção pré-partida, leva na média 2 minutos por vagão considerando a amostra realizada em 1.293 trens em três turnos diferentes de trabalho (EDWARDS, 2006). Considerando como exemplo uma ferrovia onde são carregados mil vagões por dia na origem e os trens são formados por 100 vagões cada, é necessário que diariamente seja planejado a partida de 10 trens. Para cada um destes 10 trens haveria a necessidade de 200 minutos para revistamento dos 100 vagões. Este tempo de revistamento também afeta diretamente a produtividade das ferrovias. Caso as ferrovias adotem estratégias de inserir equipes maiores para o revistamento das composições, colocando ao invés de 2 pessoas no processo (1 por lado do vagão), 6, ou até 8 pessoas por composição, seria possível reduzir substancialmente o tempo de revistamento dos 100 vagões para até 30 minutos, no entanto exigiria um maior custo operacional para a realização dos revistamentos.

Um segundo fator crítico de uma operação ferroviária é a segurança operacional. No Brasil as ferrovias tem apresentado uma redução na taxa de acidentes ferroviários, tendo atingido em 2021 a menor taxa histórica, com 10,07 acidentes por milhão de km (ANTF, 2022). A contínua evolução da segurança operacional nas ferrovias é fruto de muitos investimentos em processos, desenvolvimento e implantação de tecnologias, sendo que dois grandes pilares da segurança são tratados separadamente, a quantidade de acidentes ferroviários e a gravidade destes acidentes ferroviários.

No pilar da redução da quantidade de acidentes ferroviários, as causas raízes são estratificadas, sendo que investimentos em capacitação, processos e tecnologias são feitos para atuar em cada uma das causas. No período de 2006 a 2013 ocorreram no Brasil 8.738 acidentes em ferrovias de carga, equivalente a mais de hum mil acidentes por ano. Conforme apresentado na Figura 1, neste período as causas destes acidentes foram 37,94% relacionadas a via permanente, 15,94% com interferência de terceiros, 14,68% com material rodante, 9,21% com falha humana e 31,44% com outras causas. Desta forma ocorreram na média 160 acidentes por ano causados por falhas de material rodante (ANTT, 2014).

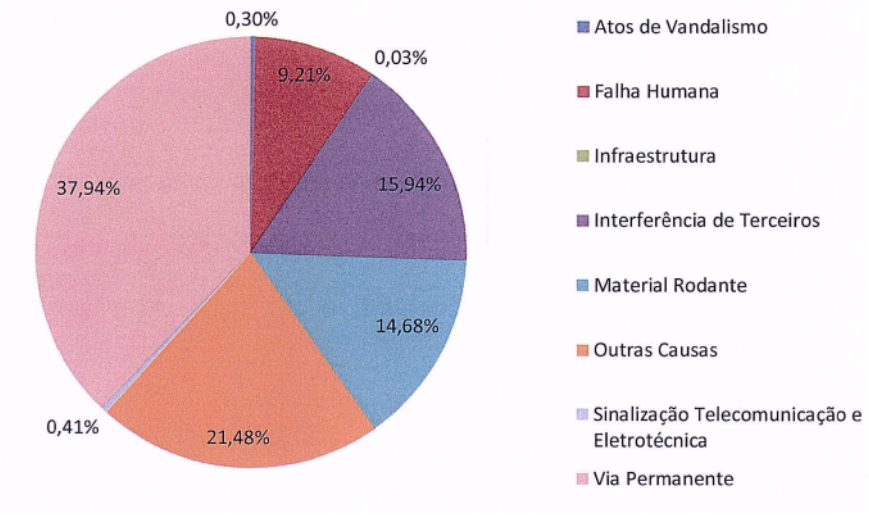


Figura 1 – Total de acidentes Ferroviários por causa entre 2006 e 2013 no Brasil (ANTT, 2014).

Como comparação, em 2022 nos Estados Unidos ocorreram 1.708 acidentes ferroviários, sendo que as causas destes acidentes foram 41,33% relacionadas a fatores humanos, 19,38% com via permanente, 12% com material rodante, 1,05% com sinalização e 26,23% com fatores diversos. O custo direto destes acidentes foi de USD254.876.149,00, sendo os acidentes onde a causa foi material rodante apresentaram um custo de USD33.774.791,00 (FRA, 2022).

No pilar da redução da gravidade dos acidentes ferroviários a classificação da gravidade não tem relação direta com a causa destes acidentes, mas sim com as suas consequências. Entre diversos fatores a gravidade de cada acidente ferroviário tem relação

com a existência de fatalidades ou lesões corporais graves, com o tipo de mercadoria transportada, com o impacto ambiental, com o impacto na sociedade, com o número de veículos ferroviários envolvidos no descarrilamento/tombamento, com a distância de via permanente danificada durante o acidente, com o tempo em que a ferrovia fica parada até liberação para circulação e com o custo total do acidente. Como exemplo, a princípio um acidente A onde 3 vagões foram tombados tem uma gravidade maior que um acidente B onde houve apenas um vagão descarrilado. Por outro lado se neste acidente B onde houve apenas um vagão descarrilado, o trem tenha continuado a trafegar com o vagão descarrilado, por 10 quilômetros de distância, danificando nesta região todos os dormentes da via permanente, até que alguém avise o maquinista para parar a composição, possivelmente os custos envolvidos na reparação dos 10 quilômetros de via permanente poderia ser muito superior ao custo da reparação dos 3 vagões tombados do acidente A.

Com o intuito de atuar especificamente na redução da gravidade dos acidentes e afim de reduzir a distância que o trem percorre entre o acidente e a parada da composição, as ferrovias, associações, agências reguladoras e fabricantes discutem internacionalmente sobre o desenvolvimento de dispositivos para detecção do descarrilamento e parada imediata da composição. Entre 2014 e 2016 foram realizados grupos de trabalho técnicos com foco no tema Detecção de Descarrilamento de Vagões, onde foram discutidos diversos aspectos com o objetivo de padronizar e regulamentar o uso de detectores de descarrilamento de vagão, abrangendo a Europa, Asia e Africa (OTIF, 2016). Na Europa este tema vem sendo discutido desde 2007 após a Comissão Européia recomendar à Agência Ferroviária da Europa imposição do uso de detectores mecânicos de descarrilamento, no entanto a sua adoção é muito discutida e não foi regulamentada (ERA, 2014). No Brasil os Detectores de Descarrilamento de Vagão (DDV) são instalados nas frotas das principais ferrovias de carga, sendo que os seus requisitos mínimos de funcionalidade e desempenho são definidos na NBR 16865. De forma geral, estes requisitos funcionais da norma Brasileira exigem que os DDVs atuem somente no caso de um descarrilamento de um ou mais rodeiros do vagão, que funcionem tanto em vagões vazios quanto carregados e que atuem os freios da composição no máximo 1 segundo após o descarrilamento. A norma Brasileira permite o desenvolvimento dos DDVs com diferentes tecnologias, desde que atendam aos requisitos funcionais. Conforme apresentado na Figura 2, os DDVs mais comumente utilizados no Brasil são compostos principalmente por uma válvula pneumática de alívio de pressão de alta vazão, um fusível e um cabo de aço ou alça de metal. O DDV é ligado diretamente na tubulação do encanamento geral dos vagões, de forma que quando ocorre o descarrilamento do vagão e o eixo cai sobre o cabo de aço, provoca a quebra do fusível, liberando o ar pressurizado e acionando imediatamente os freios da composição (ABNT, 2020).

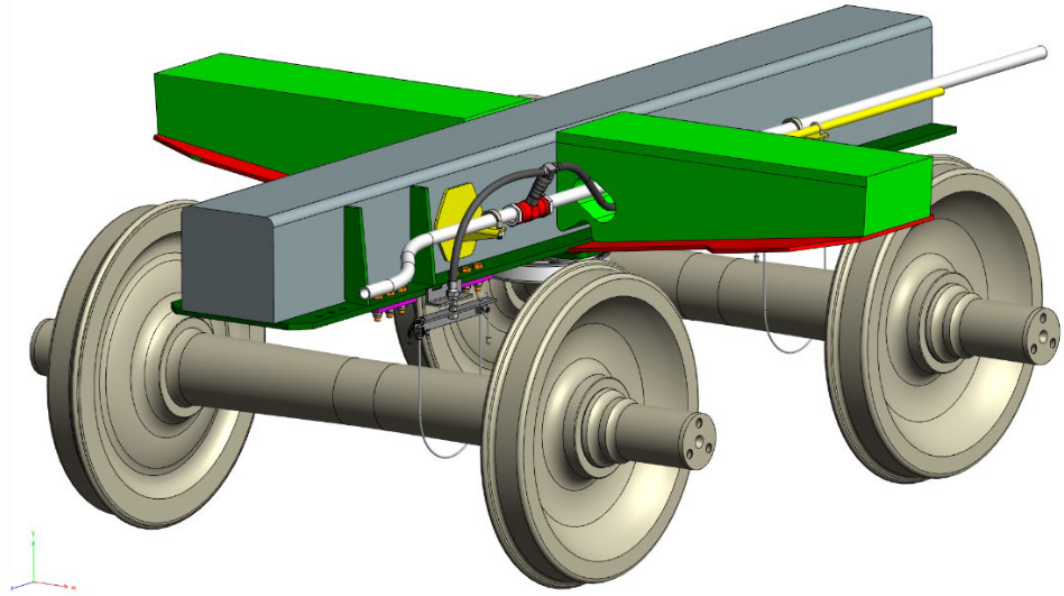


Figura 2 – Sistema de DDV comumente utilizado no Brasil(ABNT, 2020).

1.2 Justificativa e Motivação

A produtividade, a maximização da utilização dos ativos, a confiabilidade dos equipamentos, a segurança ferroviária e a segurança pessoal dos colaboradores são fatores extremamente importantes para as ferrovias.

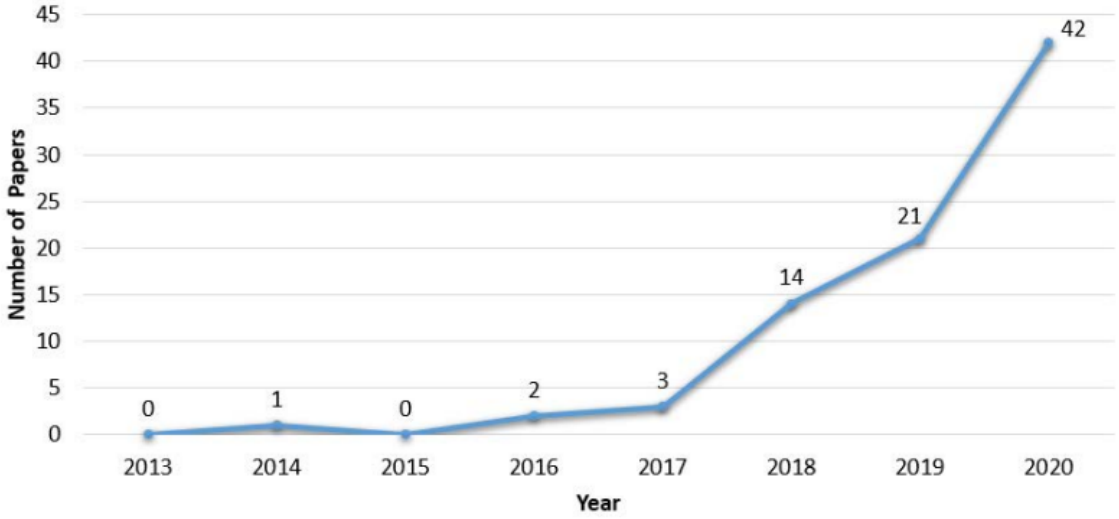
O uso de tecnologia para monitorar a condição do material rodante tem um potencial muito grande de elevar o desempenho dos ativos. Uma das técnicas de inteligência artificial (IA) mais investigadas e efetivas nas ferrovias tem sido a inspeção contínua baseada em vídeo, onde o Aprendizado Profundo tem apresentado desempenho inédito no processamento das imagens (DONATO *et al.*, 2022).

Conforme apresentado na Figura 3, entre 2013 e 2017 foram publicados apenas 6 artigos relevantes sobre o uso de Aprendizado Profundo para identificação de defeitos na manutenção ferroviária. No entanto a partir de 2018 houve um aumento expressivo de publicações, sendo 14 em 2018, 21 em 2019 e 42 em 2020 (DONATO *et al.*, 2022). Este crescimento exponencial das publicações relevantes demonstram a importância deste tema para as ferrovias internacionais, assim como relevância e alto potencial de identificação de defeitos com a utilização de técnicas de Aprendizado Profundo.

Dentre as áreas da manutenção ferroviária onde tem sido realizadas publicações relevantes entre 2013 e 2020, destacam-se conforme ilustrado na Figura 4 a inspeção de via permanente com 39 artigos, a inspeção de catenárias e pantógrafos, com 35 artigos, a inspeção de material rodante com 15 artigos e a inspeção de túneis e pontes com 8 artigos (DONATO *et al.*, 2022).

Entre diversos componentes dos vagões que podem ser monitorados foi escolhido

Figura 3 – Artigos relevantes publicados por ano sobre detecção de defeitos baseados em áudio/vídeo através de Aprendizado Profundo em manutenção ferroviária.



Fonte: Extraída de (DONATO *et al.*, 2022).

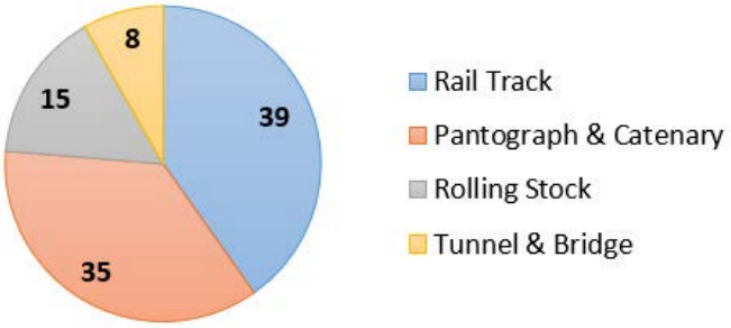


Figura 4 – Artigos relevantes publicados por tema sobre detecção de defeitos baseados em áudio/vídeo através de Aprendizado Profundo em manutenção ferroviária(DONATO *et al.*, 2022).

para este trabalho acadêmico o DDV, componente construtivamente simples mas que o seu correto funcionamento tem uma importância muito alta na redução da gravidade de acidentes ferroviários em geral, não causados especificamente por defeitos em vagões, tanto nas ferrovias nacionais quanto internacionais. Não foram encontrados nos artigos acadêmicos relevantes pesquisas relacionadas a inspeção de DDVs.

1.3 Questões de Pesquisa e Objetivos

Neste estudo, empregamos técnicas avançadas de Inteligência Artificial, utilizando especialmente o Aprendizado Profundo. Nosso objetivo foi realizar a detecção de anomalias e defeitos em vagões a partir de imagens reais, empregando um abordagem de aprendizado supervisionado. Diante dos desafios e problemas atualmente enfrentados em sistemas de

detecção de objetos por imagem foi elaborada a seguinte questão de pesquisa que norteou o projeto:

Q1 *"É possível desenvolver uma ferramenta de visão computacional de baixo custo capaz de detectar objetos e classificar defeitos em Detectores de Descarrilamento de Vagão, considerando a identificação em vídeo com o trem em movimento até 20km/h? "*

Diante desta questão de pesquisa, são definidos os seguintes objetivos para o desenvolvimento deste trabalho:

- Definir o hardware necessário para realizar as coletas dos vídeos de vagões ferroviários sendo movimentados a uma velocidade de até 20km/h.
- Realizar a aquisição de vídeos de vagões ferroviários com foco na parte inferior dos vagões.
- Realizar a rotulagem de imagens, identificando os rodeiros ferroviários, os detectores de descarrilamento sem defeito e os detectores de descarrilamento com defeito.
- Mapear os algoritmos CNN mais utilizados em artigos acadêmicos relevantes aplicados a detecção de anomalias em ferrovias e material rodante. Selecionar o algoritmo ideal para aplicação proposta com base nas características de velocidade de detecção, acurácia e nível de processamento exigido devido a necessidade de inferência em dispositivo embarcado.
- A partir do modelo proposto elaborar uma prova de conceito com um sistema que detecte a partir de vídeos de vagões ferroviários a existência/integridade dos detectores de descarrilamento de vagões.
- Incluir na prova de conceito o rastreamento dos objetos detectados a fim de contabilizar a quantidade de rodeiros, detectores de descarrilamento sem defeito e detectores de descarrilamento com defeito.

1.4 Metodologia

Para gerar registros de vídeo para se utilizar no desenvolvimento da aplicação de detecção de defeitos de DDVs, foi utilizada uma câmera previamente instalada entre os trilhos, sendo esta câmera posicionada para que haja o registro da parte inferior dos vagões que trafegam a uma velocidade de até 20km/h. A partir dos vídeos foram extraídas imagens com foco no rodeiro ferroviário, e estas imagens foram rotuladas de acordo com a existência e/ou integridade do detector de descarrilamento do vagão. Para extração de padrões, considerando os processos de Aprendizado Profundo, algoritmos de detecção de objetos utilizados em artigos relevantes de manutenção ferroviária foram mapeados, sendo escolhido o modelo mais adequado que apresente bons desempenhos de classificação e

velocidade. O modelo escolhido foi aplicado a uma prova de conceito de um sistema de detecção considerando a identificação em vídeo com o trem em movimento a 20km/h. O algoritmo foi avaliado considerando a medida mAP (*mean Average Precision*) e FPS (*Frames per second*), apropriadas para a detecção de objetos em tempo real utilizando uma única GPU (*Graphics Processing Unit*) convencional.

1.5 Principais Contribuições

As principais contribuições deste trabalho são:

1. Geração de um banco de dados com imagens rotuladas, através de imagens extraídas de vídeos de vagões ferroviários.
2. A implementação de uma rede de Aprendizado Profundo, incluindo treinamento, validação e testes, capaz de classificar eixos de vagões quanto a integridade/existência de DDVs.
3. Os resultados obtidos alcançaram desempenho de acurácia e velocidade satisfatórios para permitir a aplicação em uma única GPU convencional que estará processando os dados de vídeo de uma composição ferroviária com velocidade até 20km/h.

2 REVISÃO BIBLIOGRÁFICA

2.1 Aprendizado Profundo para Detecção de Objetos

O Aprendizado de Máquina se refere à detecção automatizada de padrões significativos em dados. As redes neurais artificiais são modelos computacionais baseados nas estruturas das redes neurais do cérebro humano, que consistem em um grande número de neurônios conectados entre eles através de redes complexas de comunicação (SHWARTZ; DAVID, 2014). O Aprendizado profundo, do inglês *Deep Learning* é um subcampo do Aprendizado de Máquina (*Machine Learning*) e inclui diversos modelos e algoritmos, entre eles as Redes Neurais Convolucionais (CNN - *Convolutional Neural Networks*) (DONATO *et al.*, 2022). As CNNs são atualmente o estado da arte para problemas de classificação e têm revolucionado o aprendizado de máquina, especialmente a área de visão computacional (PONTI; COSTA, 2018). As CNNs apesar de já terem sido apresentadas para o reconhecimento de padrões de dígitos manuscritos em imagens desde 1998 com a CNN LeNet (LECUN *et al.*, 1998), passaram a ser amplamente utilizadas para reconhecimento de imagens após a criação da CNN AlexNet que venceu a competição ILSVRC (*Imagenet classification with deep convolutional neural networks*) em 2012, classificando 1.2 milhões de imagens em 1000 classes e atingindo resultados recordes de desempenho (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Os dois principais fatores que fizeram com que as técnicas de aprendizado profundo passassem a se tornar populares são: i) A disponibilidade de grandes bancos de dados compostos por milhões de imagens, exemplificados na Figura 5; e ii) A disponibilidade de hardwares capazes de processar rapidamente estas gigantescas bases de dados, como as unidades de processamento gráfico (GPU - *Graphic Processor Unit*) (PONTI; COSTA, 2018).

Figura 5 – Exemplos de imagens de diferentes bancos de dados

(a) PASCAL VOC 12 (b) MS-COCO (c) ILSVRC (d) OpenImage

Fonte: Extraído de (ZAIDI *et al.*, 2021).

As tarefas de processamento de imagens podem ser divididas em três principais tipos: Classificação de imagens, Detecção de Objetos em imagens e Segmentação de objetos em imagens.

Para permitir a avaliação de quais técnicas de visão computacional são mais adotadas no contexto da detecção de defeitos em manutenção ferroviária, é importante que sejam avaliados os principais métodos de redes profundas utilizados em artigos e publicações relevantes para esta área. Como parte do projeto RAILS (*Roadmaps for A.I. Integration in the Rail Sector*) que faz parte da iniciativa *Shift2Rail* da União Européia, de acordo com (DONATO *et al.*, 2022) foi realizada uma pesquisa sobre os artigos mais relevantes publicados até agosto de 2021 que utilizaram Aprendizado Profundo para detecção de defeitos na manutenção ferroviária baseados em áudio ou vídeo. Dentre os 95 artigos mapeados, 90 eram baseados em modelos para detecção de defeitos a partir de dados de imagem ou vídeo e 5 eram baseados em modelos para detecção de defeitos a partir de dados de áudio, indicando uma forte tendência em detecção de defeitos na manutenção ferroviária utilizando dados de imagens e vídeos. Analisando os 90 artigos que utilizavam modelos para detecção de defeitos a partir de imagem ou vídeo, estes testaram e compararam o desempenho de 126 redes CNN. Conforme a Tabela 1, 68 modelos foram dedicados à tarefa de detecção de objetos em imagens, 41 dedicados à tarefa de classificação de imagens, e 17 dedicados à tarefa de segmentação de objetos em imagens. Desta forma, 54% de todos os modelos de detecção de defeitos a partir de dados de imagem utilizavam a tarefa de detecção de objetos.

Tabela 1 – Classificação dos modelos de CNN adotados em artigos relevantes de detecção de defeitos na manutenção ferroviária a partir de imagem de acordo com a tarefa de processamento de imagens

Tarefa de processamento de imagem	Quantidade de modelos adotados
Detecção de objetos em imagens	68
Classificação de objetos em imagens	41
Segmentação de objetos em imagens	17

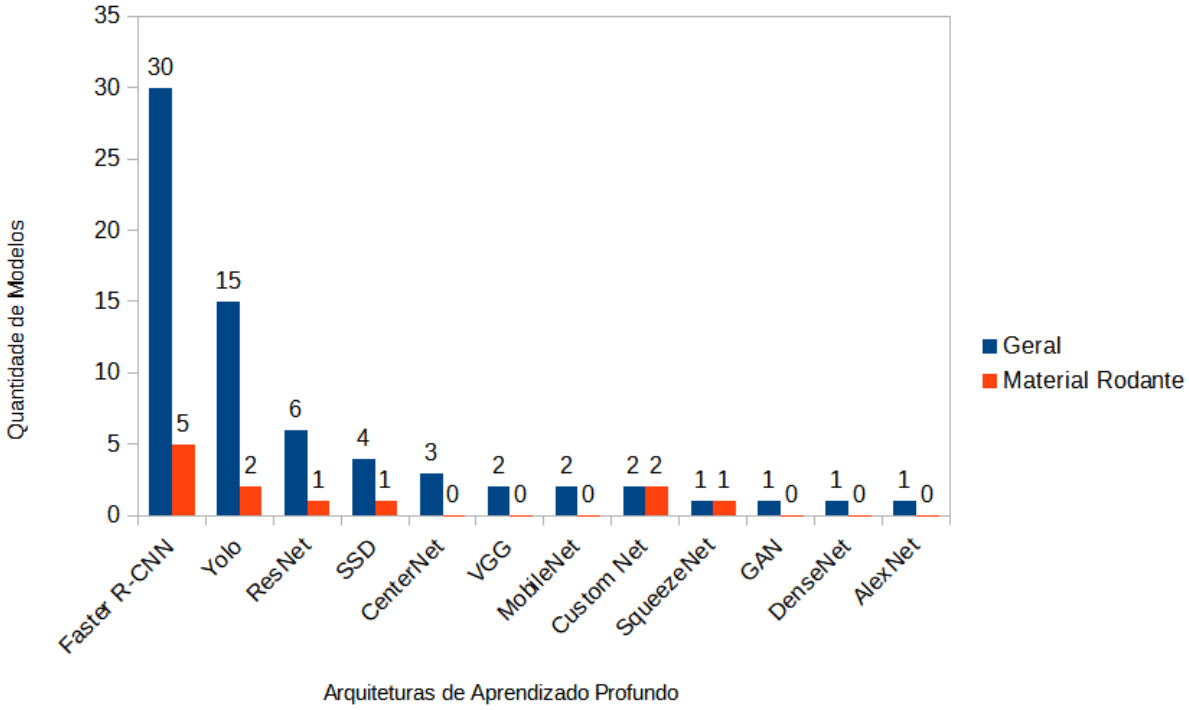
Fonte: Adaptado de (DONATO *et al.*, 2022)

A detecção de objetos é uma das mais importantes áreas de aplicação da visão computacional e permite a identificação e a descrição de conteúdos ou objetos de uma imagem, destacando a localização destes respectivos objetos através de regiões delimitadoras (BBox - *Bounding Box*). Ou seja, são tarefas que permitem tanto i) identificar objetos quanto ii) definir a região onde os objetos se encontram na imagem.

Conforme apresentado na Figura 6, pode-se avaliar os principais modelos de CNN utilizadas até agosto de 2021 nos 68 modelos que utilizaram a tarefa de detecção de

objetos. Tanto para a manutenção de material rodante, quanto para as demais áreas de manutenção ferroviária (onde inclui-se a detecção de defeitos em via permanente, túneis, pontes e catenárias), as abordagens mais utilizadas para detecção de objetos foram as CNN que pertencem a família baseada em região, na sua maioria Faster R-CNN que tem demonstrado uma maior precisão, e também as CNN que pertencem a família Yolo (*You only look at once*) por estarem dentre as redes CNN mais rápidas (DONATO *et al.*, 2022).

Figura 6 – Modelos CNN para detecção de objetos mais utilizadas nos artigos sobre identificação de defeitos na manutenção ferroviária até agosto de 2021



Fonte: Adaptado de (DONATO *et al.*, 2022).

2.1.1 Métricas de avaliação de desempenho

Os modelos de detecção de objetos são comparados utilizando múltiplos critérios para avaliação de seu desempenho, sendo *mean Average Precision* (mAP) o mais comumente adotado, ou simplesmente AP, quando se lida com apenas uma classe. A Precisão mede o percentual das predições corretas e é calculada através da derivação da Intersecção sobre a União (IoU - *Intersection over Union*), sendo a proporção de sobreposição entre as BBox verdade básica (*ground truth*) e preditas. Caso a IoU seja superior ao limite parametrizado para detecções corretas, é classificado como Verdadeiro Positivo e caso o IoU seja inferior é classificado como Falso Positivo. A classificação é Falso Negativo quando o modelo falha em detectar um objeto presente na verdade básica (ZAIDI *et al.*, 2021). Como exemplo, $AP_{0.5}$ é a precisão média de todas as classes quando uma BBox predita tem uma IoU > 0.5 com verdade básica.

$$Precisão = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo + Falso Positivo}} = \frac{\text{Verdadeiro Positivo}}{\text{Todas observações}} \tag{2.1}$$

Para detecção em tempo real de objetos que estão em movimento, uma métrica também muito importante é a quantidade de quadros por segundo (FPS - *Frames Per Second*), que mede a velocidade de processamento durante a fase de inferência, que é quando o modelo é submetido para uso em ambiente real, neste caso, quanto maior a quantidade de FPS, melhor. Em alguns casos é avaliado o tempo de inferência (*inference time*), medido em milissegundos (ms) por imagem, que é o inverso de FPS, neste caso, quanto menor o tempo de inferência, melhor.

$$\text{Tempo de Inferência (ms)} = \frac{1000}{FPS} \tag{2.2}$$

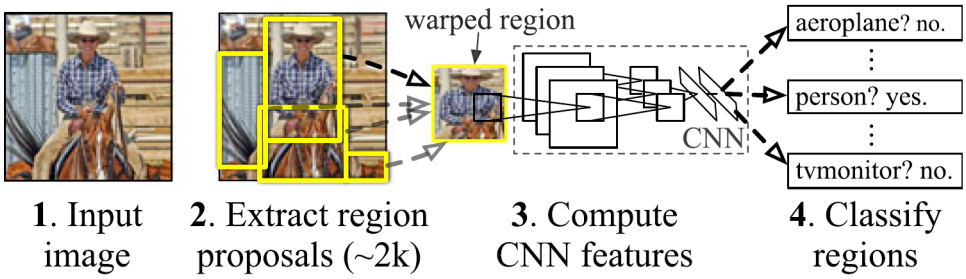
De forma geral os artigos relevantes desenvolvidos até 2021 que utilizam técnicas de aprendizado profundo em ferrovias abordam dois principais modelos de detecção de objeto estado da arte, Faster R-CNN e Yolo. Os modelos de detecção de objetos baseados em aprendizado profundo são classificados em dois tipos: i) detectores de dois estágios, que será representado pelo modelo Faster R-CNN; e ii) detectores de um estágio, que será representado pelo modelo Yolo.

2.1.2 Faster R-CNN

Um dos desafios fundamentais da visão computacional é o reconhecimento de objetos e localização destes nas imagens. No desafio VOC 2012 foi apresentado o modelo R-CNN (*Region-based Convolutional Network*) que elevou a mAP em mais de 50% em relação ao segundo melhor lugar, atingindo uma mAP de 62%. A R-CNN foi desenvolvida combinando duas idéias: i) aplicar CNNs de alta capacidade para propor regiões de imagens, que podem ter como saída BBox detectadas a fim de localizar objetos; e ii) quando houver poucos dados rotulados para treinamento, utilizar uma tarefa auxiliar de pré-treinamento supervisionado em um grande banco de dados auxiliar, seguido por uma calibração fina em um pequeno banco de dados específico, elevando consideravelmente o desempenho. Conforme Figura 7, a R-CNN possui como principais etapas: i) a imagem de entrada; ii) a extração de cerca de 2.000 propostas de região ou BBoxes; iii) o cálculo das características de cada proposta utilizando uma CNN; e iv) a classificação de cada região (GIRSHICK *et al.*, 2016).

Em 2015 foi apresentado um novo modelo baseado em região, denominado *Fast R-CNN*, devido a ser mais rápido para treinar (8,75 hrs vs 84 hrs) e testar em relação ao R-CNN. A Fast R-CNN apresentou para o banco de dados PASCAL VOC 2012 uma mAP de 66%, superando a R-CNN em 4%. Além disto algumas das principais características são: i) Treinamento em um único estágio; ii) Treinamento pode atualizar todas as camadas da rede. Conforme Figura 8, a Fast R-CNN possui as seguintes etapas: i) Uma imagem de entrada e múltiplas regiões de interesse (RoI - *Region of Interest*) são colocadas em uma rede totalmente convolucional (*fully convolutional network*). ii) Cada RoI é simplificada (*pooling*) em um mapa de características e então mapeado para um vetor de características

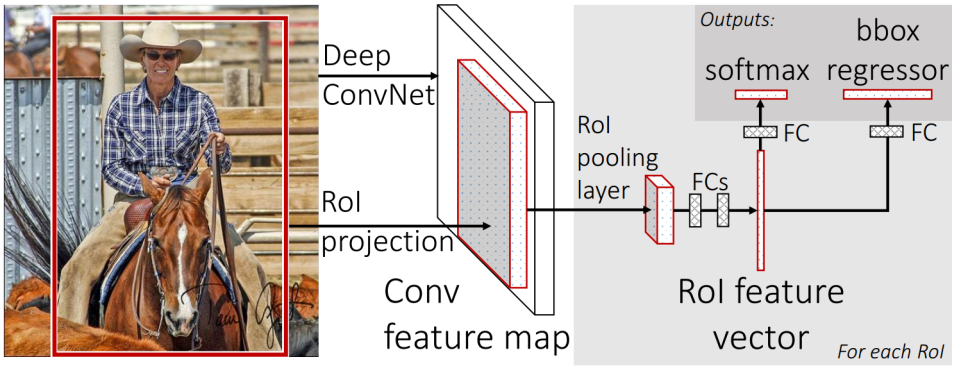
Figura 7 – Modelo R-CNN - Rede convolucional baseada em região



Fonte: Extraído de (GIRSHICK *et al.*, 2016).

por camadas totalmente conectadas (*fully connected layers*). iii) A rede possui dois vetores de saída por RoI: probabilidades softmax e regressores BBox por classe (GIRSHICK, 2015).

Figura 8 – Modelo Fast R-CNN

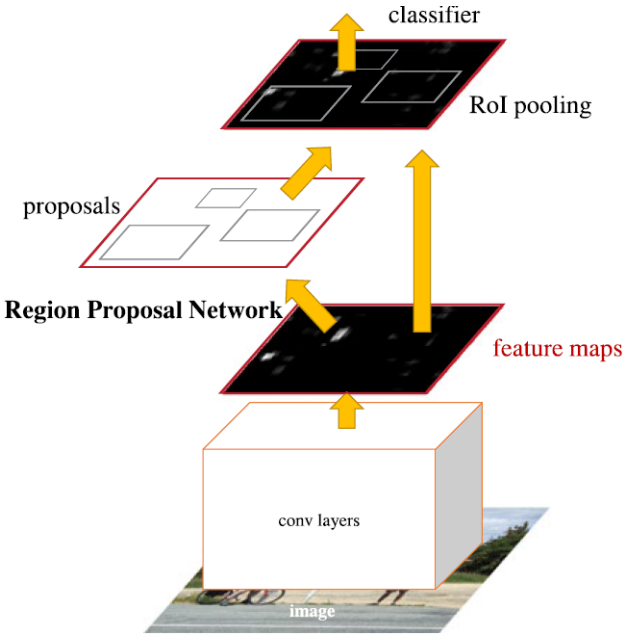


Fonte: Extraído de (GIRSHICK, 2015).

Ainda em 2015, vencendo diversas categorias nas competições ILSVRC e COCO 2015, os mesmos autores apresentaram a *Faster R-CNN*, que foi baseada na combinação de duas redes, a RPN (*Region Proposal Network*) e a *Fast R-CNN*. A RPN é uma rede completamente convolucional que para cada posição faz ao mesmo tempo a predição dos limites de região dos objetos e suas respectivas probabilidades. A RPN é uma rede que gera eficientemente predição de propostas de regiões e pode ser treinada de ponta a ponta. A fusão da RPN com a *Fast R-CNN* considera um esquema de treinamento que alterna entre um ajuste fino para a tarefa de proposta de região e um ajuste fino para a detecção de objetos, produzindo uma rede unificada e compartilhada entre ambas tarefas (REN *et al.*, 2017).

Conforme a Figura 9, o modelo Faster R-CNN é uma rede de detecção de objetos unificada, composta por dois módulos. O primeiro módulo RPN é composto por uma rede completamente convolucional que realiza as propostas de região e indica para o segundo módulo através do mecanismo de atenção para onde olhar. O segundo módulo é composto pelo detector Fast R-CNN que utiliza as propostas de região (REN *et al.*, 2017).

Figura 9 – Modelo Faster R-CNN

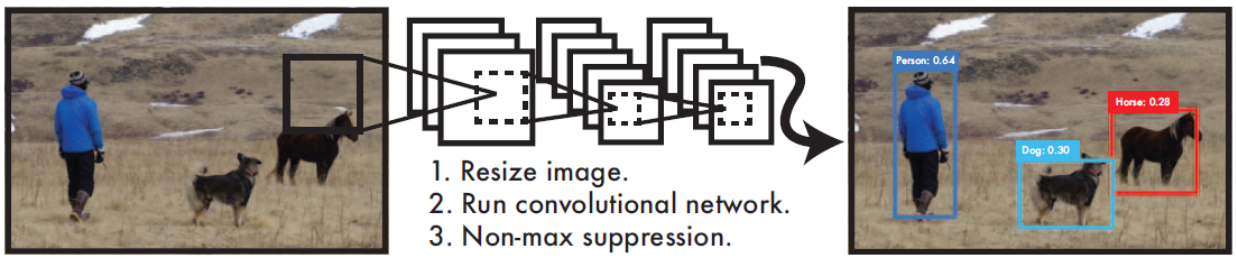


Fonte: Extraído de (REN *et al.*, 2017).

2.1.3 YOLO Original

Em 2016 foi apresentada a primeira versão da família de detectores em tempo real YOLO (*You Only Look Once*). Diferentemente dos modelos de detecção baseados em região, que possuem estágios separados para gerar as propostas de região e posteriormente executar um classificador para estas propostas de região, o modelo YOLO realiza a detecção em uma única rede neural. O YOLO trata a detecção de objetos como um único problema de regressão, realizando a predição das BBoxes e realizando a classificação das probabilidades de cada classe em uma única avaliação. Conforme a Figura 10, o processo é simples e de uma única direção, considerando as seguintes etapas: i) redução da imagem de entrada para 448x448; ii) execução de uma única rede convolucional na imagem; e iii) definição dos resultados de detecção (REDMON *et al.*, 2016).

Figura 10 – Sistema de Detecção YOLO

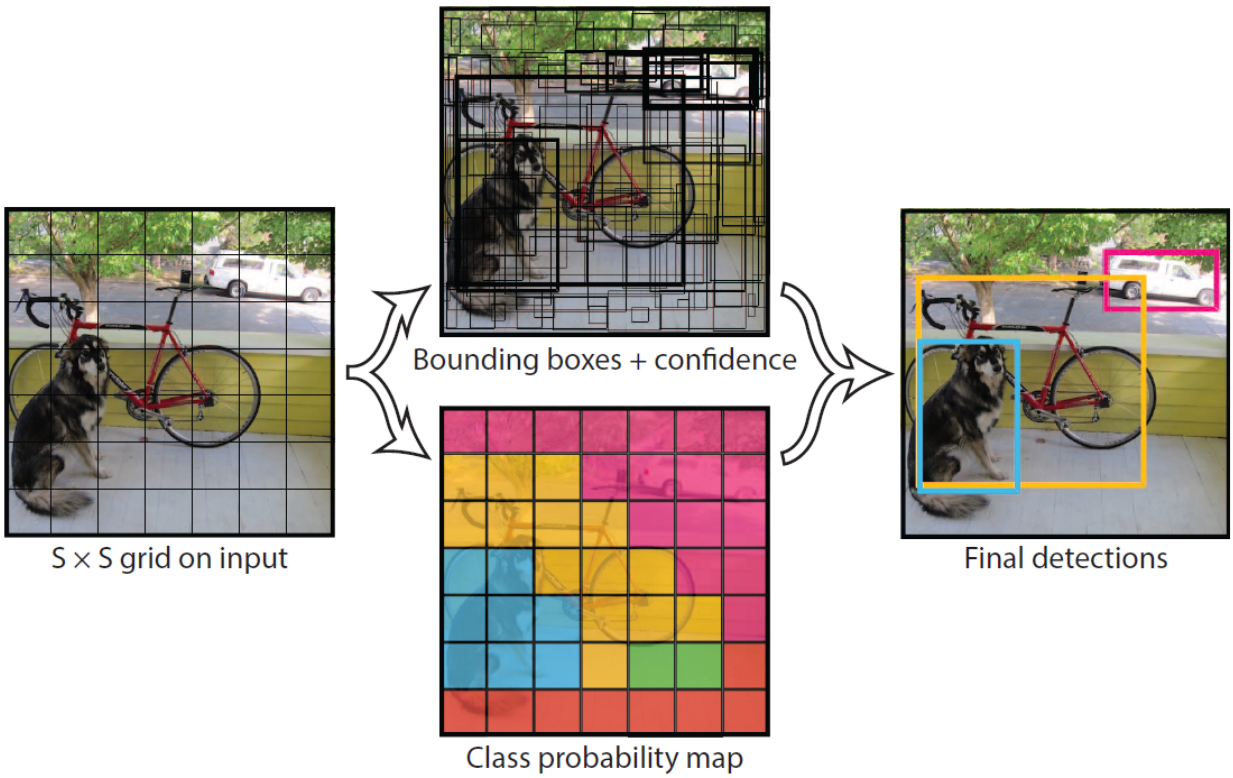


Fonte: Extraído de (REDMON *et al.*, 2016).

O modelo YOLO conforme Figura 11 divide a imagem em uma grade $S \times S$ e para cada célula da grade realiza a predição de B BBoxes, e a probabilidade de cada classe

estar presente na imagem C, sendo a predição representada pelo tensor $S \times S \times (B * 5 + C)$ (REDMON *et al.*, 2016).

Figura 11 – Modelo YOLO



Fonte: Extraído de (REDMON *et al.*, 2016).

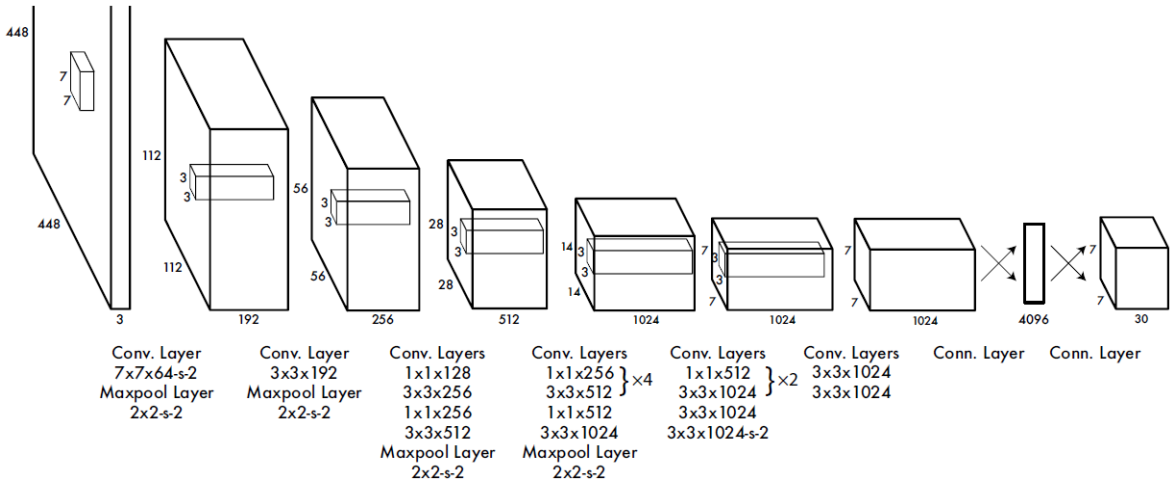
Na arquitetura do modelo YOLO, conforme Figura 12, são contempladas 24 camadas convolucionais seguidas por 2 camadas totalmente conectadas. O pré-treinamento das 20 primeiras camadas convolucionais seguidos por uma camada *average-pooling* e uma camada totalmente conectada é realizado na tarefa de classificação do banco *Imagenet* que possui 1.000 classes utilizando metade da resolução de entrada e então o dobro da resolução de entrada para a detecção. Posteriormente este modelo pré-treinado é convertido para executar a detecção adicionando quatro camadas de convolução e duas camadas totalmente conectadas, elevando o desempenho. A camada final realiza a predição tanto das coordenadas das BBoxes quanto das probabilidades de cada classe.

A partir do lançamento da arquitetura YOLO, diversas versões aprimoradas de modelos YOLO foram lançadas, incluindo as versões YOLOv2/YOLO9000 em 2016, YOLOv3 em 2018, YOLOv4 em 2020, YOLOv5 em 2020, YOLOv6 em 2022, YOLOv7 em 2022 e YOLOv8 em 2023.

2.1.4 YOLO-V8

De acordo com (HUSSAIN, 2023) a rede Yolo-V8 foi lançada em Janeiro de 2023 pela Ultralytics, atingindo resultados ainda melhores que as versões anteriores da família

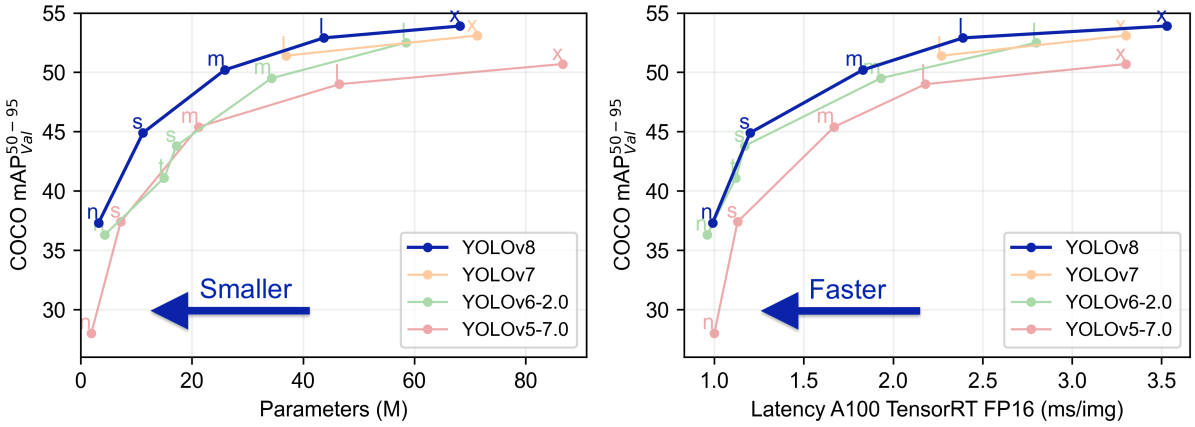
Figura 12 – Arquitetura YOLO



Fonte: Extraído de (REDMON *et al.*, 2016).

Yolo. O autor indica ainda que a rede YOLO-V8 é aprimorada para o uso em dispositivos embarcados em altas velocidades de inferência. Na Figura 13 são comparadas as redes Yolo-V5, Yolo-V6, Yolo-V7 e Yolo-V8 quanto ao seu desempenho considerando imagens de 640 pixels. É possível observar que a rede Yolo-V8 é superior em relação as demais redes considerando a mesma quantidade de parâmetros. A rede YoloV8 suporta múltiplas

Figura 13 – Desempenho redes Yolo

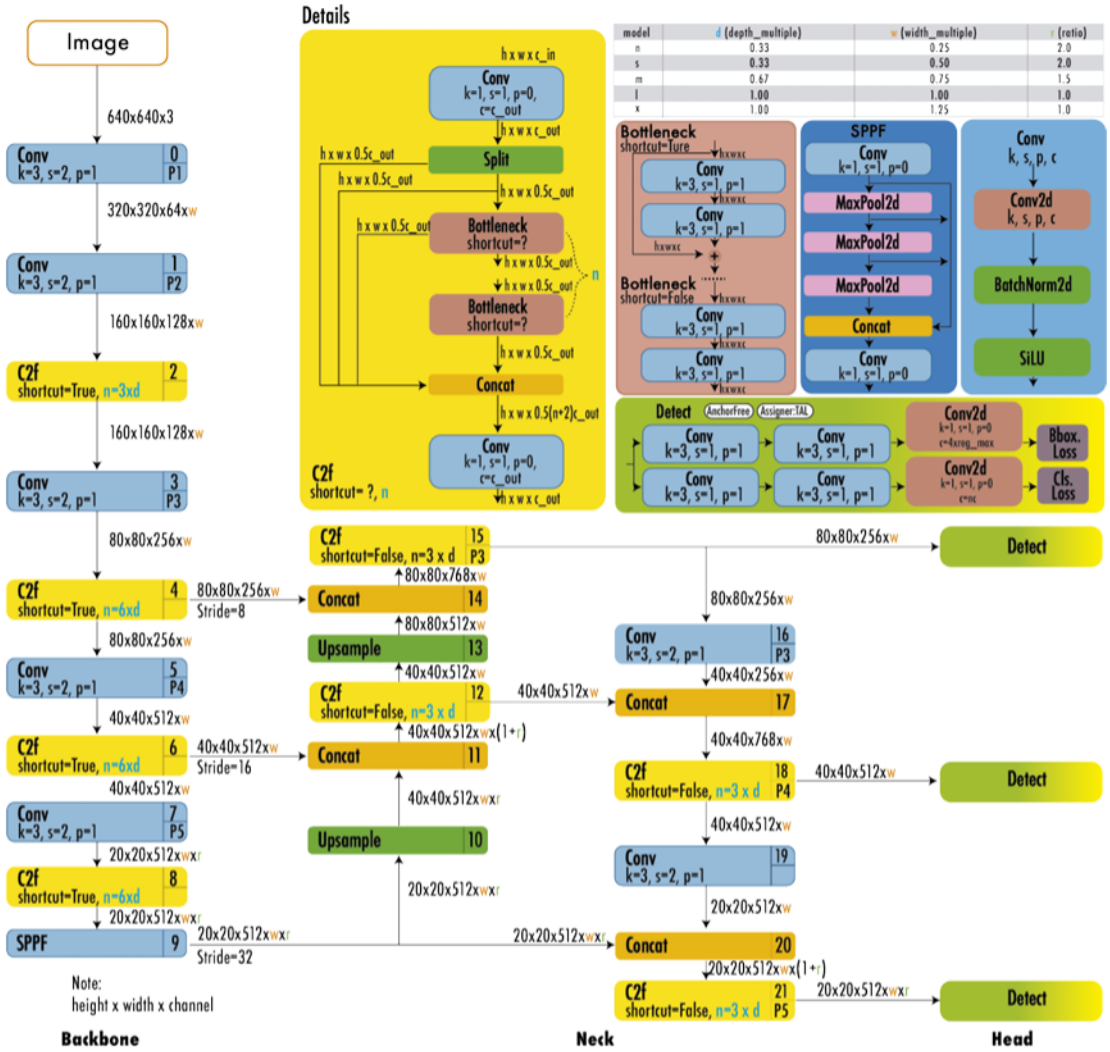


Fonte: Extraído de (JOCHER G.; CHAURASIA, 2023).

tarefas de visão computacional como a classificação, a detecção de objetos e a segmentação de objetos, além de tarefas mais avançadas como a estimativa de postura e o rastreamento e contagem dos objetos. De acordo com (TERVEN; CORDOVA-ESPARZA, 2023) a rede Yolo V8 utiliza um *backbone* com algumas alterações em relação a rede Yolo V5, agora denominado módulo C2f. A rede YoloV8 possui um design que permite que cada ramificação se concentre em sua tarefa, elevando o desempenho geral do modelo. Conforme a Figura 14, na camada de saída é utilizada como a função de ativação a função *sigmoid*. Para representar as probabilidades dos objetos pertencentes a cada classe possível é

utilizada a função *softmax*. O processamento é acelerado por uma camada *Spatial Pyramid Pooling Fast* (SPPF) através do *pooling* dos recursos em um mapa de tamanho fixo. A normalização e a ativação através da função *SiLU* são aplicadas em cada convolução. Foram realizados testes por (TERVEN; CORDOVA-ESPARZA, 2023) considerando o banco de dados MS COCO 2017, sendo que a YoloV8X atingiu uma precisão mAP de 53,9% para imagens com resolução de 640 *pixels*, sendo a velocidade de 280 FPS em uma GPU NVIDIA A100 utilizando TensorRT.

Figura 14 – Arquitetura rede YoloV8



Fonte: Extraído de (TERVEN; CORDOVA-ESPARZA, 2023).

Conforme a Tabela 2 a rede YoloV8 possui 5 modelos que variam quanto a precisão mAP, a velocidade de inferência, assim como a necessidade de processamento, devido ao aumento exponencial da quantidade de parâmetros da rede de acordo com o modelo selecionado.

Tabela 2 – Comparação da precisão mAP e velocidade FPS de acordo com o modelo da rede YoloV8 utilizada

Modelo	Tamanho (pixels)	mAP50-95	Velocidade A100 (ms)	parâmetros (M)	Flops (B)
YOLOV8n	640	37.3	0.99	3.2	8.7
YOLOV8s	640	44.9	1.20	11.2	28.6
YOLOV8m	640	50.2	1.83	25.9	78.9
YOLOV8l	640	52.9	2.39	43.7	165.2
YOLOV8x	640	53.9	3.53	68.2	257.8

Fonte: Adaptado de (JOCHER G.; CHAURASIA, 2023)

2.2 Trabalhos Relacionados

As seguintes referências são relacionadas a publicações que adotaram técnicas de aprendizado profundo para detecção de defeitos em material rodante, através das tarefas de detecção de objetos.

No trabalho de Zhan *et al.* (2018), é apresentado um modelo MRPN (*Multi region proposal generation*) proprietário para detecção de falhas nos sistemas de freio e de amortecimento de vagões de uma ferrovia de carga. O modelo MRPN é inspirado na Faster R-CNN e tem como objetivo a criação de uma plataforma unificada para geração de propostas de região de falha e a classificação das imagens de trens de carga. A base de dados consiste em imagens de 4 classes diferentes de componentes de vagões com e sem defeito, sendo as classes: 1) punho da torneira de isolamento, 2) coletor de pó, 3) parafusos fixados e 4) bloco limitador do rodeiro. O modelo MRPN apresenta resultados de até 100% de taxa de detecções corretas e uma velocidade de detecção de até 4 FPS em uma GPU K40.

Zhang *et al.* (2020) propuseram o modelo proprietário FTI-FDNet (*Fault Detection Network for Freight Train Images*) para detecção de falhas nos sistemas de freio e de amortecimento de vagões e de sua versão Light. O FTI-FDNet, inspirado nas redes Faster R-CNN e R-FCN, utiliza uma rede de proposta de multi-regiões a qual realiza a extração de BBoxes. Posteriormente a classificação e detecção é realizada pelo pooling de múltiplas regiões de interesse (RoI). Para redução do tamanho do modelo e aumento de sua velocidade é aplicado o método MRS (*Model Reduction Scheme*) no final. A mesma base de dados utilizada por (ZHAN *et al.*, 2018) foi adotada, com imagens de 4 classes. O modelo FTI-FDNet apresenta 99.28% de taxa média de detecções corretas (mCDN) e uma velocidade de detecção de 0,071 ou 14,08 FPS. Já o modelo Light FTI-FDNet apresenta 99,13% de taxa média de detecções corretas e uma velocidade de detecção de 0,058 ou 17,24FPS. Ambos modelos proprietários foram comparados com o modelo YOLOv3 que

atingiu velocidade superior na detecção (0,026 ou 38,46FPS) no entanto atingiu uma inferior taxa média de detecções (87,08%). Os experimentos foram conduzidos em um PC Intel I7 com GPU GTX1080Ti.

Yang *et al.* (2020) fazem a proposta de um modelo desenvolvido para detectar a fixação de parafusos da capa do rolamento dos rodeiros de vagões (ABCDFBs - *Axle Box Cover Device Fixing Bolts*). O modelo é baseado na combinação da Faster R-CNN com a OC-CNN (*One Class Convolutional Neural Network*). Inicialmente a Faster R-CNN é utilizada para localizar e detectar os ABCDFBs e então o modelo de classificação é treinado apenas com exemplos positivos de parafusos. O problema de poucos exemplos negativos é resolvido com o algoritmo. As imagens são obtidas a partir de uma *linear array camera* de alta precisão. 260 imagens rotuladas são dividas em treinamento e teste em uma proporção de 8:2, sendo 3 diferentes classes de parafusos. Os resultados do experimento demonstraram que este modelo atingiu uma taxa de acuracidade média de 96.55% e um tempo de teste para cada imagem de 200ms ou 5 FPS.

Já Chen *et al.* (2022) apresentam o modelo proprietário CDDF (*Component Defect Detection Framework*) de dois estágios para a detecção de defeitos em trens que estão em movimento. O primeiro estágio realiza a detecção dos componentes do trem utilizando um esquema proprietário de detecção de objetos (HOD - *Hierarchical Object Detection*) e o segundo estágio realiza a detecção dos defeitos dos componentes através de múltiplos métodos de tratamento de imagens e CNNs. O esquema hierárquico do HOD é motivado pelo módulo de atenção da Faster R-CNN. Três diferentes tipos de defeitos são classificados: i) Posicionamento incorreto do componente; ii) Componente defeituoso; e iii) Componente Faltante. O modelo CDFF busca resolver alguns desafios como a detecção de objetos muito pequenos, o grande número de tipos de defeitos de componentes e o limitado número de imagens com defeito para treinamento. 1.156 imagens com resolução de 1.400x1.024 pixels divididas em 15 classes de componentes (*b-plate, l-plate, bearing, collector, flange, spring, group, fixator, valve, nut-s, screw-s, nut-f, screw-f, bolt, and plug*) foram utilizadas para realizar a rotulagem manual de 12.207 BBoxes. Os resultados dos experimentos demonstraram um melhor desempenho para objetos pequenos em relação aos demais métodos, no entanto um pior desempenho para objetos grandes.

No modelo proprietário MPDD (*Multi-stage Pipeline for Defect Detection*) apresentado por Zhao *et al.* (2020), são incluídos dois estágios como uma proposta de método de detecção automática. O primeiro estágio de detecção de componentes, baseado em uma versão modificada da Faster R-CNN, e o segundo para a classificação de defeitos. 3.000 imagens com resolução de 2048x2000 pixels, divididas em 6 diferentes classes (*brake disc, brake caliper, tractor, side suspension, under suspension, plate bolt*) são coletadas de trens a 150km/h, sendo transformadas em 15.000 objetos rotulados no formato VOC. Os experimentos realizados baseados no API de detecção de objetos TensorFlow, executados

em um CPU Xeon 5600 e 4 GTX1080Ti indicaram que a MPDD conseguiu alcançar um *mAP* 0,792 com uma velocidade de 203ms por imagem ou 4,92 FPS, e a Faster R-CNN apresentou um *mAP* de 0,651 com uma velocidade de 114ms por imagem ou 8,77 FPS. A velocidade foi considerada baixa para o MPDD, ficando para trabalhos futuros o desenvolvimento de um novo modelo.

Para Yu *et al.* (2021), o baixo desempenho em métodos de detecção em tempo real é um problema a ser resolvido, além da baixa acuracidade em detecção de defeitos pequenos. Como solução é apresentada uma versão aprimorada do modelo YOLOv3, que adota uma amostra da FPN (*Pyramidal Feature Networks*) para aprimorar a acurácia de detecção de trincas e riscos na superfície do disco da roda ferroviária. 583 imagens são coletadas de 4 classes de defeitos de rodas (*cracks, dents, inclusions, and scratches*) por câmera industrial durante a fabricação de rodas de trens de alta velocidade, sendo posteriormente ajustadas para uma resolução de 416x416 e transformadas em 1.132 imagens após o processo de aumentação de dados. A plataforma Pytorch foi utilizada e testada em uma GPU Nvidia GTX2060. A versão aprimorada da YOLOv3 apresentou um *mAP* de 88,3 e uma velocidade de 28,9ms ou 34,60FPS, a versão original da YOLOv3 apresentou um *mAP* de 84,1 e uma velocidade de 28,2ms ou 35,46FPS e a Faster R-CNN apresentou um *mAP* de 86,1 e uma velocidade de 121,1 ou 8,25FPS.

2.3 Considerações finais

Conforme apresentado há poucas publicações sobre o uso de técnicas de aprendizado profundo para detecção de defeitos em material rodante, sendo que dentre os trabalhos relacionados não foram encontrados modelos que buscassem defeitos em sistemas de detecção de descarrilamento de vagões. O adequado desempenho de sistemas de detecção em tempo real para trens em movimento, considerando a combinação de uma elevada precisão mAP e uma alta velocidade de detecção (FPS) é um grande objeto de estudo das ferrovias e pesquisadores. Nas publicações listadas pode-se verificar que há uma grande diferença quanto a velocidade de detecção entre as redes Faster R-RCNN e Yolo. As publicações que adotaram a rede Faster R-CNN atingiram velocidades entre 4 e 8 FPS, tendo sido considerada baixa para detecção de objetos em movimento, enquanto as publicações que adotaram a rede YoloV3 atingiram velocidade de 34 a 38 FPS. Conforme descrito anteriormente, foram lançados após a rede YoloV3 diversos outros modelos no período de 2020 a 2023, tendo sido lançado em Janeiro de 2023 a rede YoloV8 que é considerada o estado da arte e permite velocidades ainda maiores de inferência. Desta forma, neste trabalho, foi escolhido o modelo YoloV8 para implementar o sistema de detecção proposto. Com base na literatura analisada, o modelo YoloV8 permite atender os requisitos do sistema de acurácia, velocidade de detecção, e implementação em hardware embarcado.

3 MATERIAIS E MÉTODOS

3.1 Projeto de Estudo

O experimento que será apresentado a seguir foi realizado para responder a principal questão de pesquisa do projeto: "É possível desenvolver uma ferramenta de visão computacional de baixo custo capaz de detectar objetos e classificar defeitos em Detectores de Descarrilamento de Vagão, considerando a identificação em vídeo com o trem em movimento até 20km/h?"

A fim de estruturar as principais etapas necessárias para responder esta questão de pesquisa, o experimento foi abordado conforme os objetivos previamente apresentados, que contemplam desde a aquisição dos vídeos reais de vagões até a avaliação do desempenho da detecção de objetos e defeitos a partir do modelo de Aprendizado Profundo YoloV8.

3.2 Aquisição de vídeos e extração de imagens

3.2.1 Características de Hardware para Coleta de Vídeos

A definição das características do hardware para coleta dos vídeos é uma etapa muito importante para a tarefa de detecção de objetos. As características devem ser amplamente avaliadas de acordo com o tipo de objeto a ser detectado, seu tamanho, a distância deste objeto em relação à câmera, a velocidade em que este objeto passa pela câmera, as condições de visibilidade que podem envolver tanto condições diurnas, noturnas, com ou sem chuva.

Os vídeos com a filmagem da parte inferior dos vagões foram obtidos com o apoio da Especialista de Vagões da Rumo Laura Braz, que coletou as imagens em um portal de câmeras existente na cidade de Paranaguá-PR. No local onde são registradas as imagens os trens circulam na faixa de velocidade de até 20 km/h. A câmera que registrou as imagens é fabricada pela Intelbras, modelo VIP5550. Como características principais a câmera possui resolução de até 5MP, Lente Varifocal 2.7 a 13.5mm motorizada, Sensor 1/2.7"Progressivo CMOS, IP-67, conexão IP e taxa de frames de 20 até 60 FPS.

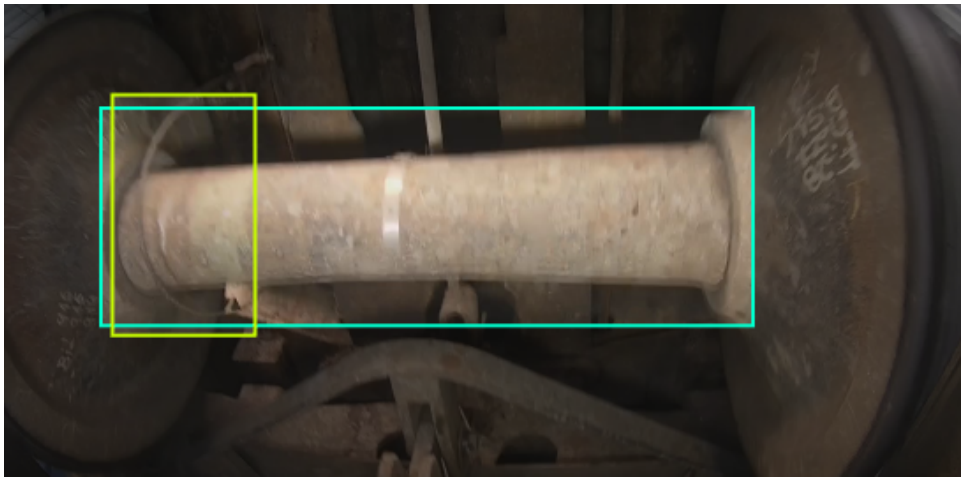
As características da câmera utilizada atendem de forma geral o objetivo deste trabalho, sendo 60 FPS adequados para o registro dos trens trafegando até 20 km/h. No entanto nota-se o efeito de *Motion Blur* em várias imagens capturadas por esta câmera, devido ao tipo de obturador utilizado, o que acaba afetando a precisão da rede.

Os vídeos foram extraídos em uma resolução de 1280x720, equivalente a 0.92 megapixels, resolução suficiente para treinamento de redes de aprendizado profundo.

3.3 Rotulagem das imagens e preparação da base de dados

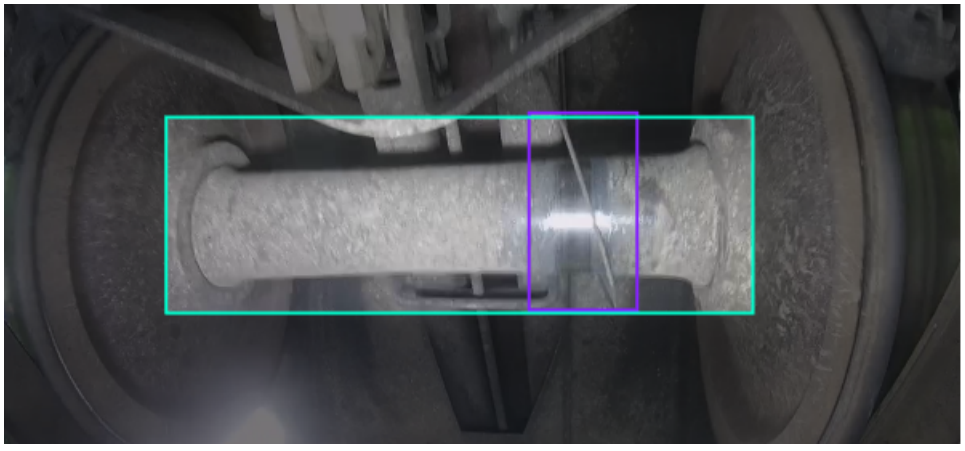
A primeira etapa adotada para a rotulagem das imagens foi transformar um terço de todos os frames dos vídeos em imagens individuais, tendo sido utilizado o pacote *Open CV* para capturar os frames e gravar cada imagem no formato PNG. Neste caso foram convertidos 31 vídeos em imagens, gerando 217 mil imagens. A segunda etapa adotada foi realizar a seleção das imagens que contém os objetos a serem rotulados, neste caso imagens com o rodeiro ferroviário, sendo as demais imagens excluídas. Nesta etapa foram mantidas 2.400 imagens que continham o rodeiro ferroviário. A terceira etapa adotada foi a realização da rotulagem das 2.400 imagens. Nesta etapa após a comparação de diversas opções de sites e softwares para anotação de imagens, foi utilizado o aplicativo CVAT (*Computer Vision Annotation Tool*) que possui interface amigável e boa velocidade para a tarefa, sendo criadas as seguintes classes: 1) Rodeiro (classe existente em todas as imagens), 2) DDV OK (conforme Figura 15), 3) DDV Com defeito (conforme Figura 16).

Figura 15 – Classe Rodeiro e Classe DDV OK



Fonte: O Autor.

Figura 16 – Classe Rodeiro e Classe DDV com Defeito



Fonte: O Autor.

Nesta etapa foram realizadas 4.620 anotações com *bounding boxes*, sendo 2.302 da Classe Rodeiro, 1.983 da classe DDV OK e 319 da classe DDV com Defeito. Haviam 98 imagens que não continham o rodeiro ferroviário e foram mantidas na base sem a rotulagem de classes. A classe DDV com Defeito predominantemente foi composta por imagens de eixos que estavam marcados devido ao cabo do DDV estar encostando no eixo do rodeiro ferroviário.

3.3.1 Divisão do conjunto, pré-processamento e aumentação dos dados

Nesta etapa foi utilizado o aplicativo Roboflow que possui diversas funções relacionadas a tarefas de visão computacional. Após a rotulagem das imagens, foi utilizado o aplicativo Roboflow para dividir a base de dados em 70% para treinamento, 20% para validação e 10% para o teste. Desta forma as 2.400 imagens foram divididas em 1.674 imagens para treinamento, 484 imagens para validação e 242 imagens para teste. Foi realizado na sequência um pré-processamento das imagens, reduzindo a resolução de 1280x720 para 640x640 pixels, resolução muito utilizada pelos modelos de detecção de objetos abordados, reduzindo o tempo para treinamento das redes, assim como maximizando a velocidade de inferência. Após o pré-processamento foi realizada a aumentação dos dados, onde foram geradas novas versões de cada imagem do conjunto de treinamento, considerando dois principais filtros, sendo o primeiro a Rotação das imagens em -15° e +15° e o segundo a utilização de ruídos em até 5% dos pixels. A base de dados para treinamento foi ampliada para 5.058 imagens, permanecendo a base de validação com 484 imagens e a base de teste com 242 imagens.

3.4 Treinamento e Validação

Neste trabalho foi utilizada a versão mais atual da família Yolo, o YoloV8 que atualmente tem sido considerada como o estado da arte para detecção de objetos em movimento. O treinamento foi realizado no ambiente do Google Colab, utilizando uma GPU Tesla A100 com 40GB de memória RAM. A rede neural YoloV8m foi treinada utilizando inicialmente 143 épocas. No entanto, como não foi observado melhoria nos resultados a partir da época 93, a própria rede Yolo considerou os melhores resultados como da época 93. No treinamento foram utilizadas imagens com resolução de 640 *pixels*, tendo o treinamento levado 4 horas. O arquivo best.pt com os pesos gerados na rede durante a etapa de treinamento ficou com um tamanho de 58,8 *Megabytes*. Conforme detalhado na Tabela 3, a precisão mAP50-95 considerando todas as classes foi de 59,1% resultado ainda melhor que a referência apresentada na etapa de revisão bibliográfica, onde foi atingido uma mAP de 53,9% para o banco de dados MS COCO 2017.

Tabela 3 – Treinamento da rede YoloV8M

Classe	mAP50-95
Todas	0.591
DDV OK	0.487
DDV com Defeito	0.493
Rodeiro	0.792

Fonte: O Autor

3.4.1 Resultados do Treinamento

Após a realização do treinamento da rede na base de dados rotulada, foram gerados gráficos para permitir o entendimento de sua precisão e desempenho. Na Figura 17 é apresentada a Matriz de Confusão que compara no eixo X a classificação verdadeira das classes em relação ao eixo Y a classificação predita das classes. Desta forma é possível observar que a classe Rodeiro foi predita corretamente em 100% dos casos, a classe DDV OK em 97% dos casos e a classe DDV com defeito em 78% dos casos. A classe *background* representa imagens sem objetos DDV ou rodeiro, que foram incluídas na base de dados para reduzir os Falsos Positivos (FP). Na base de dados utilizada, foram incluidas 97 imagens sem objetos. Na matriz confusão é possível visualizar que as imagens da classe *background* foram preditas em 77% dos casos como DDV OK e 22% dos casos como DDV com defeito, indicando uma quantidade insuficiente de imagens da classe *background* na base de dados.

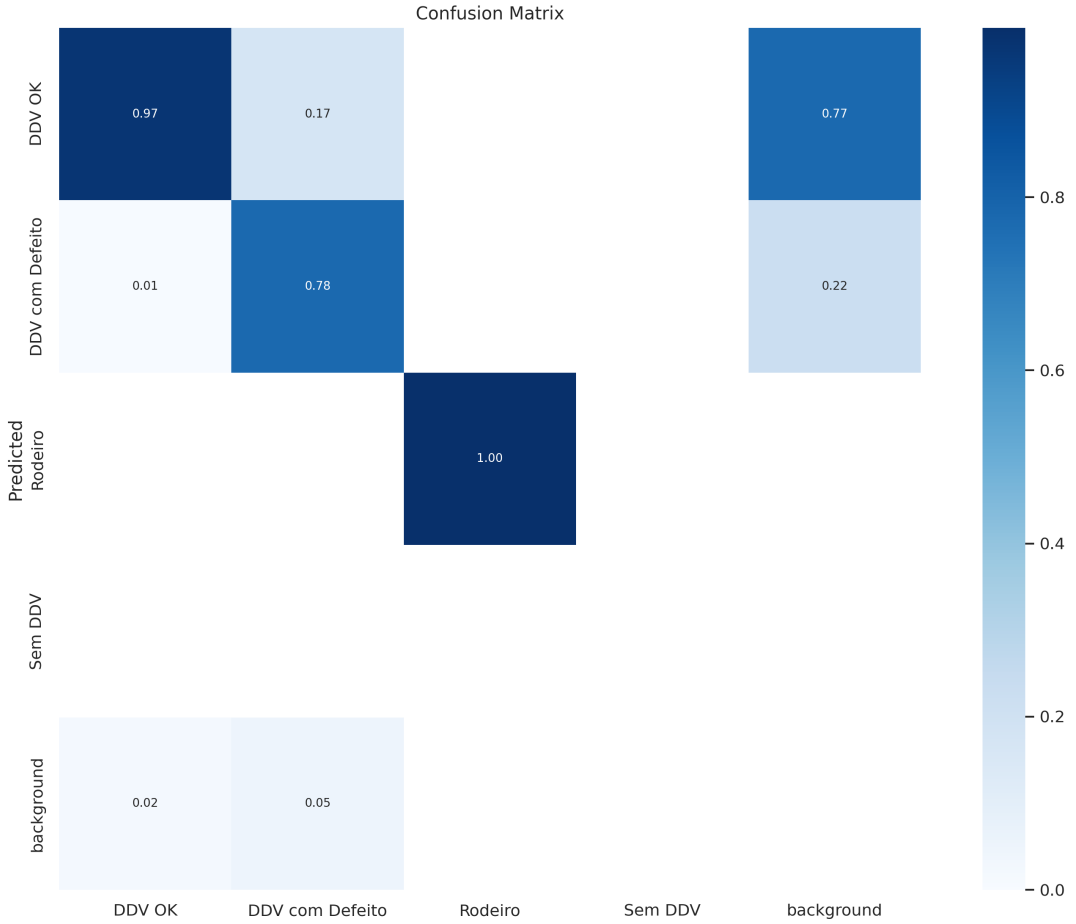
Na Figura 18 são apresentadas imagens do lote de validação que foram rotuladas utilizando os parâmetros da rede treinada. Todas as classes foram adequadamente identificadas nestes exemplos.

3.5 Testes

3.5.1 Inferência no ambiente Google Colab

Na inferência realizada no ambiente Google Colab foi utilizada uma GPU Tesla A100 com 40Gb de memória RAM. Em uma primeira etapa foi realizada a inferência em arquivos de imagens. As 242 imagens do lote de teste, que não foram utilizadas na etapa de treinamento e de validação da rede, foram submetidas à etapa de teste utilizando os pesos rede Yolov8 treinada previamente, utilizando a tarefa "*detect*", o modo "*predict*"e utilizando como fonte a pasta onde 242 as imagens estava armazenadas. Para esta etapa foram necessários cerca de 11,3ms por imagem, equivalente a 88FPS. As imagens com as *bboxes* foram exportadas para uma pasta. Em uma segunda etapa foi realizada a inferência utilizando como fonte um arquivo de vídeo de um trem em movimento. Da mesma forma,

Figura 17 – Matriz de Confusão Experimento



Fonte: O Autor.

este vídeo foi submetido à etapa de teste utilizando os pesos da rede YoloV8 treinada previamente. Conforme ilustrado na figura 19, o vídeo foi adequadamente exportado com as *bboxes* indicando as classes de objetos em cada um dos *frames*.Para esta etapa foram necessários cerca de 11,7ms por imagem, equivalente a 85FPS.

3.5.2 Inferência em um dispositivo Jetson AGX Orin

A realização da etapa de inferência para aplicações reais, em operações que rodam 24 horas por dia e em alguns casos em regiões onde não há comunicação de internet exige o uso de uma GPU local. Para este trabalho, conforme mostra a Figura 20, foi utilizado um dispositivo Jetson AGX Orin, que possui 64Gb de memória RAM.

Para realizar a inferência no dispositivo *Jetson* foram instalados no ambiente *Ubuntu* as bibliotecas *DeepStream, Ultralytics, Pytorch* e *Torchvision*. De forma geral foram realizadas configurações para permitir realizar a inferência com os pesos da rede previamente treinada. Além disto a inferência pode ser realizada com diferentes configurações da precisão de bits de números de ponto flutuante. Na Figura 21 é possível observar para um dispositivo Jetson similar, que para a rede YoloV8m que foi utilizada durante a etapa de treinamento,

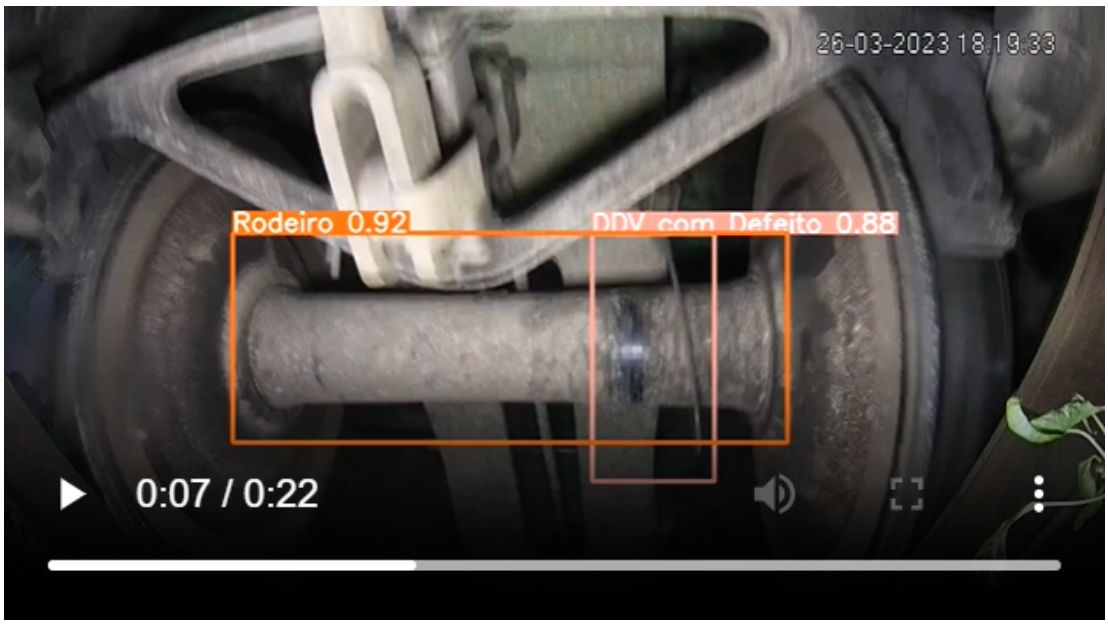
Figura 18 – Exemplos Imagens Rotuladas lote Validação

Fonte: O Autor.

realizando a inferência com precisão FP32 a velocidade esperada seria de 56 FPS. Seria possível atingir até 385 FPS caso o treinamento fosse realizado com base na rede YoloV8n, realizando a inferência com precisão INT8, no entanto o impacto da redução na precisão mAP deveria ser melhor analisada de acordo com cada estudo de caso.

Neste trabalho foi utilizada a FP32 como a precisão de bits de números de ponto flutuante, formato que possui a maior precisão e ocupa uma maior memória de processamento quando comparado as configurações FP16 e INT8, o que reduz a velocidade de inferência. Na Figura 22 é apresentado a inferência realizada pelo algoritmo *DeepStream* utilizando um vídeo com o trem em movimento. As classes ao longo do vídeo são adequadamente identificadas, tendo sido atingida uma velocidade de inferência de 30 FPS.

Figura 19 – Vídeo após etapa de inferência no ambiente Colab



Fonte: O Autor.

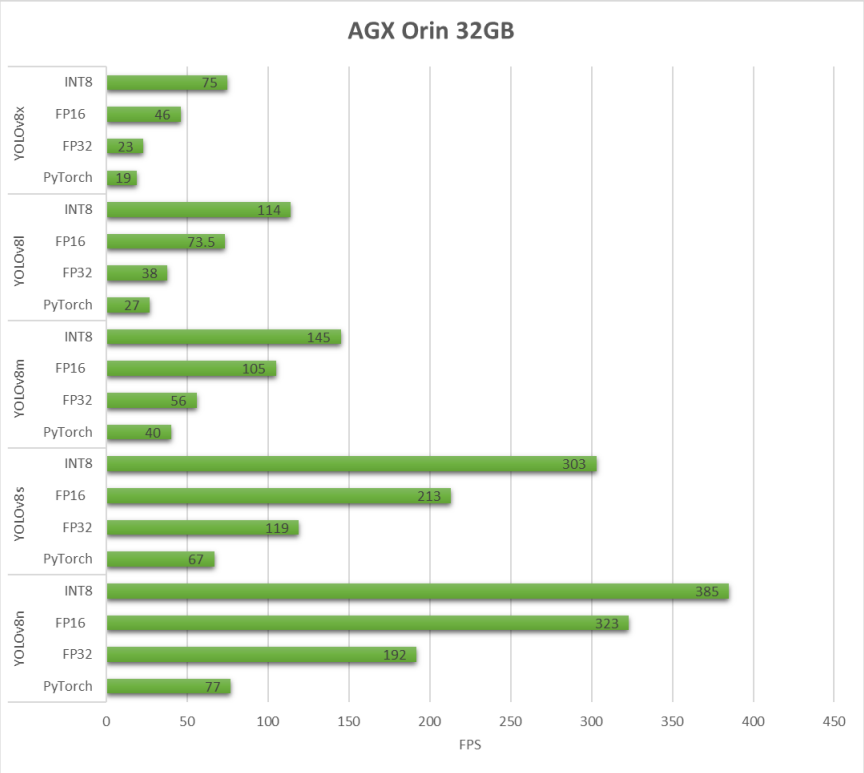
Figura 20 – Dispositivo Jetson AGX Orin

Fonte: O Autor.

3.6 Rastreamento e contagem dos Objetos

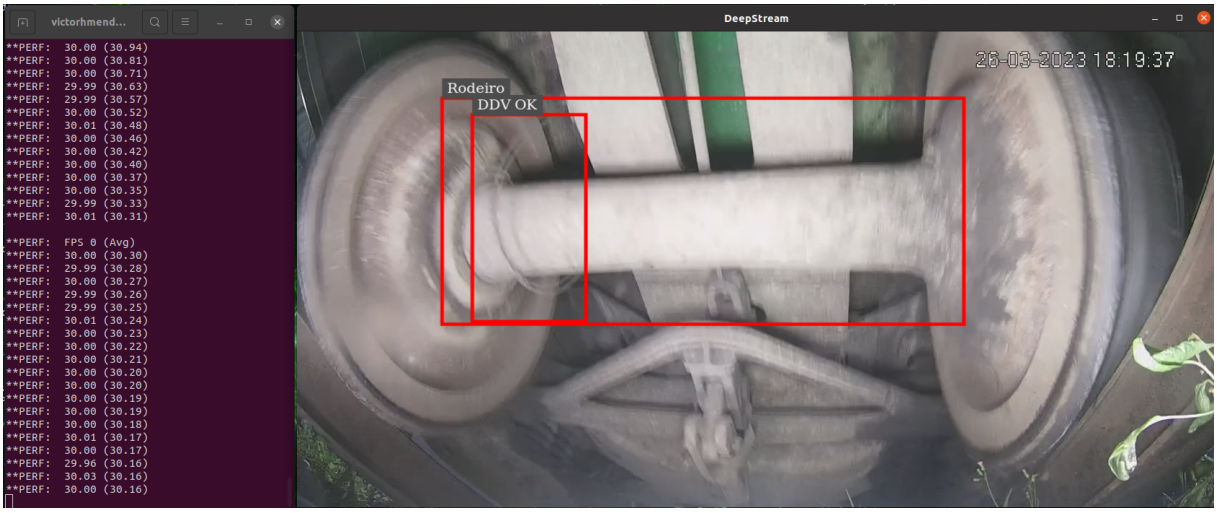
Conforme destacado no Capítulo 2, o modelo YoloV8 possui tarefas avançadas de visão computacional relacionadas ao rastreamento dos objetos e a contagem destes objetos. O rasteramento de objetos é uma técnica para detectar objetos entre quadros de vídeos, utilizando suas características. Para cada objeto é atribuído um numero identificador aleatório distinto, que é mantido ao longo dos frames. Neste trabalho foi utilizado no ambiente *Google Colab* o algoritmo de rastreamento *Deep Sort (Simple Online Real Tracking)* para permitir tanto o rastreamento e distinção entre os objetos, quanto a contagem do número de objetos que passaram pelo detector. Na Figura 23 é apresentado

Figura 21 – Comparação de velocidade de acordo com a rede YoloV8 e precisão de bits de números de ponto flutuante



Fonte: O Autor.

Figura 22 – Inferência realizada com o a biblioteca DeepStream no dispositivo Jetson

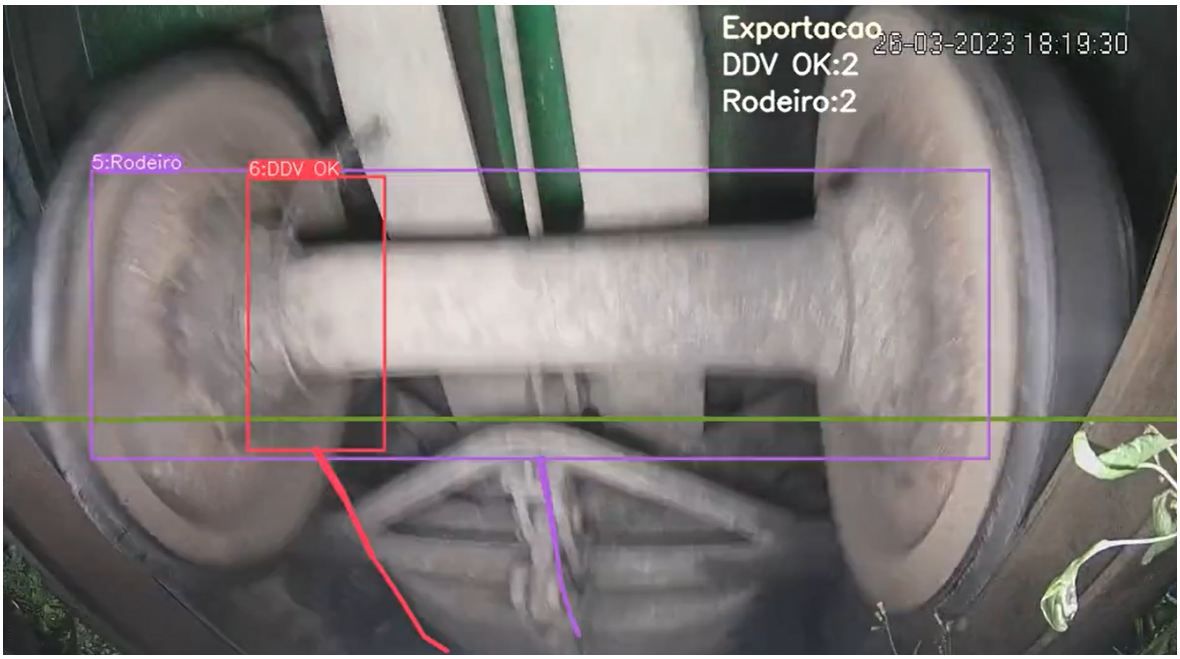


Fonte: O Autor.

um dos quadros durante o uso do *DeepSort*. É possível notar também que cada classe de objeto apresenta uma assinatura por onde passou nos quadros anteriores, sendo esta assinatura destacada na mesma cor da classe do objeto. Além disto é possível verificar uma linha na cor verde, onde foi definido uma região de referência para contabilizar a quantidade de objetos de cada classe identificada. É possível verificar que até o registro deste quadro haviam sido identificados 2 rodeiros e 2 DDVs OK que cruzaram a linha

verde estando o trem no sentido exportação para o porto.

Figura 23 – Rastreamento e Contagem dos Objetos



Fonte: O Autor.

3.7 Considerações Finais

No experimento realizado foi possível treinar um modelo de aprendizado profundo para realizar a detecção automática de anomalias em veículos ferroviários através de visão computacional. As imagens foram coletadas a partir de uma câmera que registrou vídeos da parte inferior dos vagões. O treinamento do modelo Yolov8 foi realizado no ambiente do Google Colab, sendo utilizadas 2.400 imagens dos rodeiros. Os resultados deste experimento foram satisfatórios, tendo havido uma assertividade de 100% para a classe Rodeiro, 97% para a classe DDV OK e 78% para a classe DDV com Defeito. Na inferência realizada no ambiente *Google Colab* foi possível atingir 88FPs de velocidade, sendo que na inferência realizada com o algoritmo *DeepStream* no dispositivo Jetson AGX Orin, foi atingida uma velocidade de 30FPS. Além disto utilizando no ambiente *Google Colab* o algoritmo *DeepSort* foi possível implementar o rastreamento e identificação dos objetos, além da contagem de quantos objetos passaram pelo ponto de referência no vídeo.

4 CONCLUSÃO

Neste trabalho foram avaliados através de pesquisa bibliográfica, trabalhos acadêmicos relevantes com a aplicação de redes de aprendizado profundo para detecção de anomalias na ferrovia. Nos trabalhos acadêmicos foi possível mapear as principais redes utilizadas, assim como as características destas redes quanto a precisão mAP e velocidade de inferência FPS. As principais redes utilizadas nos trabalhos acadêmicos até 2021 eram a Faster R-CNN e a YoloV3. Foi possível observar que as redes Faster R-CNN apresentaram bons resultados de precisão mAP, no entanto apresentaram uma baixa velocidade de inferência, entre 4 e 8 FPS, enquanto a rede YoloV3 apresentou velocidade de inferência de 34 a 38 FPS. Considerando a criação de outros modelos da familia Yolo nos últimos anos, foi proposta a utilização neste trabalho da rede YoloV8, considerada o estado da arte até o início de 2023. O banco de dados com imagens rotuladas utilizado neste trabalho foi criado pelo próprio autor, a partir de vídeos reais da parte inferior de vagões. Ao todo foram utilizados 31 vídeos, com 217 mil frames, tendo sido realizado o saneamento das imagens onde constavam as classes de interesse, sendo rotuladas 2.400 imagens. Posteriormente foram utilizadas técnicas de aumentação de dados, ficando a base de dados para treinamento com 5.058 imagens, a base de validação com 484 imagens e a base de teste com 242 imagens. As classes utilizadas foram Rodeiro, DDV OK e DDV com Defeito. Foi realizado no ambiente *Google Colab* o treinamento da rede YoloV8m utilizando 93 épocas e imagens de 640 *pixels*, tendo sido obtida uma precisão geral mAP50-95 de 59,1%, e precisão mAP50-95 para a classe rodeiro de 79,2%. Na matriz de confusão foi indicado que a classe Rodeiro foi predita corretamente em 100% dos casos, a classe DDV OK em 97% dos casos e a classe DDV com defeito em 78% dos casos. Na etapa de teste foi atingida uma velocidade de inferência de 88 FPS no ambiente *Google Colab*, sendo que no algoritmo *DeepStream* utilizado no dispositivo Jetson AGX Orin foi atingida uma velocidade de 30FPS. Como última etapa foi utilizado no ambiente *Google Colab* o algoritmo de rastreamento *Deep Sort* para permitir tanto o rastreamento e distinção entre os objetos, quanto a contagem do número de objetos que passaram pelo detector. De forma geral o trabalho permitiu identificar de forma adequada os defeitos dos vagões para as classes propostas utilizando a rede YoloV8, tendo o trabalho sido desenvolvido de ponta a ponta, desde o registro e rotulagem das imagens reais, até a inferência uma GPU local.

4.1 Trabalhos Futuros

Neste trabalho, além do dispositivo existente na cidade de Paranaguá-PR, de onde foram coletados os vídeos utilizados no treinamento da rede, também foi desenvolvido um segundo dispositivo em conjunto com o Especialista de Locomotivas da Rumo, Manoelino

Almeida e a empresa parceira RA. Conforme Figura 24, este dispositivo foi posicionado entre os trilhos para permitir a filmagem da parte inferior dos vagões durante a sua movimentação. O segundo dispositivo foi instalado em um pátio de cruzamento da ferrovia Rumo localizado na cidade de Araraquara-SP, onde a velocidade dos trens pode chegar a 20 km/h. O dispositivo é composto por uma caixa metálica, por refletores de led, 3 câmeras, além de uma proteção de vidro acima das câmeras. As características deste dispositivo foram definidas visando garantir a robustez necessária para esta aplicação ferroviária, a correta iluminação dos objetos a serem detectados, além de permitir a proteção das câmeras contra pequenos objetos que são projetados durante a passagem do trem, com um baixo custo.

Figura 24 – Dispositivo de Filmagem Inferior

Fonte: O Autor.

As câmeras utilizadas neste dispositivo que foi instalado na cidade de Araraquara-SP tem características similares a câmera utilizada na cidade de Paranagua-PR. As câmeras são fabricadas pela Intelbras modelo VIP5550 D Z IA, possuindo como principais

características técnicas relevantes: Resolução até 5MP, Lente Varifocal 2.7 a 13.5mm motorizada, Sensor 1/2.7"Progressivo CMOS, IP-67, conexão IP, taxa de frames de 20 até 60 FPS.

Como trabalho futuro, em conjunto com os times de engenharia ferroviária e inovação da empresa Rumo, será definido uma especificação técnica e criado um protótipo completo para aplicação na ferrovia, incluindo: a) Definição de matriz de priorização de defeitos a serem monitorados para locomotivas e vagões; b) Aprimoramento do dispositivo citado para registro das imagens da parte inferior dos ativos; c) Definição de câmeras mais adequadas para registro de objetos em movimento, dado que as câmeras utilizadas neste trabalho geraram em várias imagens um efeito de *Motion Blur*; d) Definição de servidor adequado para armazenamento dos vídeos e imagens, incluindo recursos para rotulagem das imagens; e) Definição da GPU mais adequada para inferência local; f) Definição de requisitos para desenvolvimento de interface das detecções com os usuários da ferrovia que acessarão os dados.

REFERÊNCIAS

AAR. **49 CFR PART 215 - Railroad freight car safety standards**: 215.13 pre-departure inspection. 2023. Available at: https://www.ecfr.gov/current/title-49/subtitle-B/chapter-II/part-215. Access at: 12 fevereiro 2023.

ABNT. **NBR 16865 - Freio ferroviário — Detector de descarrilamento de vagão — Requisitos de funcionalidade e desempenho**. 2020. Available at: https://www.abntcatalogo.com.br/pnm.aspx?Q=V1JWSHhmd1NsdURla0FRNFdDcU9VdXBpMG1XOCt5NkVNT0k0YmdFc3p1bz0=. Access at: 19 fevereiro 2023.

ANTF. **Informações Gerais**: O setor ferroviário de carga brasileiro. 2022. Available at: https://www.antf.org.br/informacoes-gerais/. Access at: 11 fevereiro 2023.

ANTT. **Relatório de Acidentes - 2006 a 2013**: O setor ferroviário de carga brasileiro. 2014. Available at: https://www.gov.br/antt/pt-br/assuntos/ferrovias/relatorios-e-plano-trienal-de-investimentos-pti-1/relatorio-de-acidentes. Access at: 11 fevereiro 2023.

CHEN, C. *et al.* A hybrid deep learning based framework for component defect detection of moving trains. **IEEE Transactions on Intelligent Transportation Systems**, v. 23, n. 4, p. 3268–3280, 2022.

DONATO, L. D. *et al.* A survey on audio-video based defect detection through deep learning in railway maintenance. **IEEE Access**, v. 10, p. 65376–65400, 2022.

EDWARDS, J. R. **Improving the efficiency and effectiveness of railcar safety appliance inspection using machine vision technology**. 2006. 169 p. Dissertação (Master of Science in Civil Engineering) — University of Illinois, Illinois, 2006.

ERA. **EU activities for reducing impacts of freight train derailments**. 2014. Available at: http://otif.org/fileadmin/user_upload/otif_verlinkte_files/05_gef_guet/03_AG-Entgleisungsdetektion/2014_10/CE_GTDD_2014-A_Annex_III_Presentation_EC-ERA_E.pdf. Access at: 21 fevereiro 2023.

FRA. **Train Accidents by Cause**. 2022. Available at: https://railroads.dot.gov/accident-and-incident-reporting/train-accident-reports/train-accidents-cause. Access at: 19 fevereiro 2023.

GIRSHICK, R. Fast r-cnn. *In*: **2015 IEEE International Conference on Computer Vision (ICCV)**. [*S.l.: s.n.*], 2015. p. 1440–1448.

GIRSHICK, R. *et al.* Region-based convolutional networks for accurate object detection and segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 38, n. 1, p. 142–158, 2016.

HUSSAIN, M. Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. **Machines**, v. 11, n. 7, 2023. ISSN 2075-1702. Available at: https://www.mdpi.com/2075-1702/11/7/677.

JOCHER G.; CHAURASIA, A. Q. J. **YOLO by Ultralytics. GitHub**. 2023. Available at: https://github.com/ultralytics/. Access at: 20 julho 2023.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. Imagenet classification with deep convolutional neural networks. **Neural Information Processing Systems**, v. 25, 01 2012.

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

OTIF. **WG Derailment Detection**. 2016. Available at: http://otif.org/en/?page_id=142. Access at: 21 fevereiro 2023.

PONTI, M. A.; COSTA, G. B. P. da. Como funciona o deep learning. arXiv, 2018. Available at: https://arxiv.org/abs/1806.07908.

REDMON, J. *et al.* You only look once: Unified, real-time object detection. *In*: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2016. p. 779–788.

REN, S. *et al.* Faster r-cnn: Towards real-time object detection with region proposal networks. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 39, n. 6, p. 1137–1149, 2017.

SHWARTZ, S. S.; DAVID, S. B. **Understanding Machine Learning: From Theory to Algorithms**. [*S.l.: s.n.*]: Cambridge University Press, 2014.

TERVEN, J.; CORDOVA-ESPARZA, D. **A Comprehensive Review of YOLO: From YOLOv1 and Beyond**. 2023.

YANG, Y. *et al.* Defect detection of axle box cover device fixing bolts in metro based on convolutional neural network. *In*: **2020 39th Chinese Control Conference (CCC)**. [*S.l.: s.n.*], 2020. p. 7504–7509.

YU, Y. *et al.* Surface defect detection of hight-speed railway hub based on improved yolov3 algorithm. *In*: **2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)**. [*S.l.: s.n.*], 2021. v. 4, p. 1386–1390.

ZAIDI, S. S. A. *et al.* A survey of modern deep learning based object detection models. **CoRR**, abs/2104.11892, 2021. Available at: https://arxiv.org/abs/2104.11892.

ZHAN, Y. *et al.* A unified framework for fault detection of freight train images under complex environment. *In*: **2018 25th IEEE International Conference on Image Processing (ICIP)**. [*S.l.: s.n.*], 2018. p. 1348–1352.

ZHANG, Y. *et al.* Real-time vision-based system of fault detection for freight trains. **IEEE Transactions on Instrumentation and Measurement**, v. 69, n. 7, p. 5274–5284, 2020.

ZHAO, B. *et al.* Defect detection method for electric multiple units key components based on deep learning. **IEEE Access**, v. 8, p. 136808–136818, 2020.