Automated Damage Inspection in Vehicle Headlights Using U-Net and Resnet50

Kevila Cezario de Morais [1], Karin Satie Komati [2], Kelly Assis de Souza Gazolli[3]
[123]*Programa de Pós-graduação em Computação Aplicada (PPComp)*
Instituto Federal do Espírito Santo (IFES) Campus Serra
kevila.morais@gmail.com [1], kkomati@ifes.edu.br [2], kasouza@ifes.edu.br [3]

***Abstract*—The image analysis of vehicle damage is a procedure performed by insurance companies to determine whether the policy covers the service or not. In the case of damages in the headlights, the company receives a picture of the vehicle and a specialist analyzes the damage. This article proposes a system for the detection and classification of vehicle headlight images to automate the inspection. The method is based first on the U-Net structure for detecting the headlight in the image and then on the Resnet50 structure for classifying the damage. The U-Net database is made up of 2,000 vehicle images and 2,000 masks with headlight detection. Resnet50's database is made up of 2,000 images divided into 4 classes: broken, blurred, infiltrated, or undamaged. The results obtained in the test had an IOU of 70 percent in the detection and an accuracy of 76 percent in the classification.**

***Index Terms*—Insurance, Headlight, detection, Classification, Resnet50, U-Net**

I. Introdução

Os faróis veiculares desempenham um papel importante na garantia da segurança no trânsito, sendo itens obrigatórios que desempenham duas funções essenciais: sinalizar a presença do veículo a outros condutores e proporcionar iluminação adequada das vias [1]. Avarias, como faróis embaçados, lâmpadas queimadas ou trincas, devem ser reparadas em oficinas especializadas, contratadas em particular ou indicadas por uma seguradora.

O seguro de um veículo é estabelecido por meio de uma apólice, um documento que contém informações cruciais, tais como, os tipos de sinistros cobertos, cláusulas, condições e riscos associados a esse contrato. O segurado conta com indenizações, reparos e trocas em seu veículo quando há necessidade. Para a seguradora garantir que o cliente e seus prepostos cumpram com suas responsabilidades, existe uma equipe de especialistas responsável por analisar os detalhes do sinistro ocorrido. Isso acontece porque qualquer descumprimento das disposições contratuais pode resultar em perdas financeiras e materiais para a seguradora.

No caso de danos em faróis, não há cobertura para serviços relacionados à manutenção do veículo, tais como, regulagem do farol, manchas superficiais, troca exclusiva de lâmpada e outras situações. A seguradora apenas cobrirá os faróis em caso de quebra ou trinca. Essa avaliação é feita a partir de imagens do veículo danificado, onde é possível observar se o objeto está quebrado, embaçado, infiltrado ou sem danos.

A análise dessas imagens é realizada com a colaboração de vários especialistas para validação e determinação do procedimento de conserto. As imagens recebidas possuem diversos padrões em relação ao posicionamento do veículo, à qualidade da iluminação, ao dimensionamento e ao ambiente. Em alguns casos, a imagem contém apenas o farol enquadrado completamente, facilitando a classificação do dano. Em outros, a imagem apresenta a frente inteira do veículo, tornando necessária uma operação de aproximação para exibir apenas o farol e seus detalhes.

O objetivo geral deste trabalho é propor um sistema automático para a determinação de danos em farol veicular por meio de análise imagens, aproveitando o avanço de técnicas especializadas em detecção e classificação de objetos, com destaque para as redes neurais convolucionais (CNNs). Uma abordagem de aprendizado profundo, amplamente utilizada para resolver problemas complexos [2], as CNNs vêm sendo utilizadas em diversas aplicações na área de visão computacional, tais como, segmentação, classificação, detecção de objetos, reconhecimento de face, dentre outras [3] [4] [5] [6].

O sistema proposto se baseia em duas estruturas de redes neurais profundas. Inicialmente, é utilizada uma rede neural convolucional chamada U-Net [7] para detecção do farol na imagem. Em seguida, a imagem detectada é isolada e submetida a um outra rede neural convolucional, conhecida como Resnet50 (*Residual neural network*) [8] para classificar o farol em quatro classes: quebrado, embaçado, infiltrado ou sem danos.

Este artigo está dividido da seguinte forma: na Seção II são apresentados os trabalhos correlatos. Na Seção III são apresentados os métodos utilizados para detecção (U-Net) e classificação (Resnet50) das imagens. A Seção IV contém os experimentos, onde são apresentadas as informações sobre base de dados, métricas de avaliação e resultados encontrados. E, por último, a Seção V apresenta a conclusão e os trabalhos futuros.

II. Trabalhos Relacionados

A visão computacional é uma área de estudo cada vez mais popular para incorporar em iniciativas de automação e

As autoras agradecem à FAPES e CAPES pelo PDPG (Programa de Desenvolvimento de Pós-Graduação - Parcerias Estratégicas nos Estados, processo 2021-2S6CD, FAPES nº132/2021). A professora Karin Komati agradece ao CNPq pela Bolsa de Produtividade DT-2 (308432/2020-7) e pelo projeto 407742/2022-0, também agradece à FAPES pelo Auxílio Taxa de Pesquisa (nº 293/2021) e pelo projeto nº1023/2022 P:2022-8TZV6.

transformação digital [9]. Sistemas de automação inteligente utilizando visão computacional podem reconhecer todo o conteúdo importante do dado, extraí-lo e processá-lo usando regras baseadas em IA para automação de ponta a ponta do que antes era uma tarefa manual [10].

Entre as principais técnicas de visão computacional estão detecção e classificação de imagens [11]. A detecção encontra a parte específica desejada e a separa do resto da imagem. Já a classificação é o processo de extração de informação em imagens para reconhecer padrões e objetos homogêneos [11]. Com essa informação extraída é possível separar as características em grupos. Entre os casos de uso dessas duas técnicas estão: inspeções de segurança de equipamentos [12], automatização no varejo [13], diagnóstico médico [14], reconhecimento de impressão digital e biometria [10].

Na medicina, [14] construiu um sistema de classificação e segmentação de imagens de ultrassom mamário usando redes neurais convolucionais, uma solução eficiente na análise e diagnóstico precoce. Nesse artigo, foi utilizado um conjunto de dados de ultrassom de mama (com 1.418 amostras normais e 1.182 cancerígenas), e proposto um sistema de diagnóstico auxiliado por computador de dois estágios para diagnosticar o câncer de mama automaticamente. Em primeiro lugar, o sistema utiliza uma rede neural do tipo ResNet pré-treinada para candidatos normais excluídos e, em seguida, usa um modelo Mask R-CNN aprimorado para a segmentação precisa do tumor.

Na agricultura esse tipo de sistema também já é estudado na identificação automática de doenças em plantas, onde uma análise rigorosa é necessária para a segurança alimentar, estimativa de perda de rendimento e gestão de doenças. Trabalhando em um conjunto de dados aberto [15], que inclui 15.200 imagens de folhas de culturas, uma Rede Residual foi treinada para realizar esta tarefa de classificação e atingiu uma acurácia de 99,40% no teste. Em outro estudo, cinco arquiteturas de aprendizagem profunda para classificação de imagens de pragas da soja foram testadas por [16] Inception-v3, Resnet-50, VGG-16, VGG-19 e Xception, todas se mostraram eficazes para apoiar especialistas e agricultores no manejo de controle de pragas em lavouras de soja, atingindo precisões de até 93,82%.

O trabalho de [17] faz a classificação e detecção de imagens quanto à existência de rachaduras em construções, principalmente em fachadas feitas de paredes de alvenaria. O método proposto foi dividido em duas etapas: um modelo de rede neural convolucional (CNN) treinado em 26.177 imagens para criar a classificação (com trinca ou não), e outro modelo de rede neural U-Net treinado em 2870 imagens para detectar pixels de rachaduras dentro dos rótulos classificados como rachaduras. O trabalho de [18] propõe fazer a detecção de rachaduras em rodovias em duas etapas: a primeira etapa classifica se há ou não rachadura e a segunda etapa classifica a severidade da rachadura. Na primeira etapa, a rede neural VGG16 é utilizada, com algumas camadas da rede substituídas. Na segunda etapa, as imagens classificadas como rachadura passam por um *framework* integrado. Esse *framework* combina um modelo VGG16 [19] pré-treinado com uma rede neural recorrente, a *long short-term memory* (LSTM) [20], de maneira que as camadas escondidas da VGG16 são substituídas pela camada LSTM. A severidade de classificação de rachaduras é dividida em duas classes: leve (de 3 a 4,5 mm) e severa (10 a 12 mm).

O trabalho de [21] também propõe fazer a detecção de rachaduras em pavimentos de concreto em duas etapas, uma de classificação e uma de segmentação. Na primeira etapa, é utilizada transferência de aprendizado da rede neural LeNet [22] para se desenvolver um classificador. O modelo de classificação é treinado utilizando um conjunto de dados estabelecido pelo autor, o CCD1500 e esse conjunto contém cinco classes: rachadura falsa, rachadura, arranhão artificial, superfície intacta e planta. Na segunda etapa, a rede neural VGG16 teve seu modelo otimizado para a tarefa de segmentação e apresenta como saída imagens binárias, onde o preto representa píxeis do plano de fundo e o branco representa píxeis da rachadura.

Essa arquitetura foi utilizada também na detecção semântica de peças de automóveis 3D no trabalho de [23]. Dezesseis peças de automóveis foram detectadas a partir de uma rede neural convolucional baseada na arquitetura U-Net combinada com um codificador InceptionV3 [24] treinada em um conjunto de dados de peças automotivas disponível publicamente. O modelo foi capaz de detectar as seguintes partes: pára-choque traseiro, vidro traseiro, porta traseira esquerda, luz traseira esquerda, porta traseira direita, luz traseira direita, pára-choque dianteiro, vidro dianteiro, porta dianteira esquerda, luz dianteira esquerda, frente porta direita, farol direito, capô, espelho esquerdo, espelho direito e rodas. Os resultados indicam a capacidade da U-Net quanto à detecção semântica de imagens de automóveis.

Em outro trabalho, para classificar veículos em 11 categorias como carro, bicicleta, ônibus e motocicleta, [25] utilizou-se rede residual com 18 camadas, acrescentando *joint fine tuning* com um método de *dropout* e obteve uma acurácia de 97,95%. Já [8], classificou grupos de modelos de veículos com a arquitetura Resnet50, e em seguida modificada para usar o agrupamento espacialmente ponderado e com uma etapa de localização antes do processo de classificação.

Este trabalho agrega aos artigos existentes uma nova proposta para aplicação desses sistemas de detecção e classificação de imagens com aplicação de interesse econômico para empresas do ramo automotivo, que vai além dos casos de uso citados nos trabalhos relacionados.

III. MÉTODOS

As redes neurais de aprendizagem profundo são úteis em muitas tarefas que simulam o reconhecimento visual humano. Neste trabalho, as redes utilizadas para resolução do problema proposto foram U-Net, para detecção da trinca, e Resnet50 [26], para classificação quanto ao tipo de dano.

O sistema proposto neste artigo está dividido em duas etapas, Figura 1. Na primeira, o objetivo é detectar a presença do farol na imagem, ou seja, será feita uma detecção. Apenas

as imagens em que os faróis forem detectados corretamente passarão para a etapa seguinte. Na segunda etapa, os faróis detectados nas imagens serão classificados como quebrado, embaçado, infiltrado ou sem dano.

Fig. 1. Sistema proposto para avaliação do dano

A. Detecção de Imagens com U-Net

Para a parte de detecção, a rede neural convolucional U-Net [7], no artigo *U-Net: Convolutional Networks for Biomedical Image detection* com o objetivo detectar estruturas neuronais de células em um vidro plano registradas por microscopia, foi adotada. U-Net é uma extensão da arquitetura FCN (Fully Convolutional Networks, em português rede totalmente convolucional) [27] que funciona com poucas imagens de treinamento [28] e atua em campos mais precisos de detecção. Uma FCN transforma a altura e a largura dos mapas de características intermediárias de volta às da imagem de entrada, como resultado, a saída da classificação e a imagem de entrada têm uma correspondência um-para-um no nível do pixel. Nela os operadores de *pooling* da FCN foram substituídos por operadores *upsampling* com um grande número de canais de recursos, melhorando a resolução das entradas de cada camada.

A estrutura da rede neural convolucional U-Net possui um caminho contraído e um caminho expandido, o que lhe confere a arquitetura em forma de U [7]. O caminho de contratação é uma rede convolucional típica que consiste na aplicação repetida de convoluções, cada uma seguida por uma unidade linear retificada (ReLU) e uma operação de agrupamento máximo. O caminho expansivo combina o recurso e as informações espaciais por meio de uma sequência de convoluções e concatenações com recursos de alta resolução do caminho de contração [7]. As operações realizadas nas camadas podem ser: convoluções, normalização, ativação, *pooling* ou deconvoluções [7].

B. Classificação de Imagens com Resnet50

A arquitetura de Rede Neural Residual é um modelo proposto por [8]. Originalmente desenvolvida a partir da ResNet-34, que compreendia 34 camadas ponderadas. Essa abordagem apresentou uma nova maneira de adicionar mais camadas convolucionais a uma CNN, sem enfrentar o problema do gradiente *vanishing* [29].

A arquitetura Resnet50 é composta pelos elementos:

- Uma convolução de *kernel* 7×7 ao lado de 64 outros *kernels* com *stride* tamanho 2.
- Uma camada de *max pooling* com *stride* tamanho 2.
- 9 camadas – convolução de *kernel* 3×3, 64, outra com 1×1, 64 *kernels* e uma terceira com 1×1,256 *kernels*. Estas 3 camadas são repetidas 3 vezes.
- Mais 12 camadas com 1×1.128 *kernels*, 3×3.128 *kernels* e 1×1.512 *kernels*, iteradas 4 vezes.
- Mais 18 camadas com 1×1.256 núcleos e 2 núcleos 3×3.256 e 1×1.1024, iterados 6 vezes.
- Mais 9 camadas com 1×1.512 núcleos, 3×3.512 núcleos e 1×1.2048 núcleos iterados 3 vezes.
- *Average pooling*, seguido por uma camada totalmente conectada com 1.000 nós, usando a função de ativação *softmax*

Segundo [8], redes residuais são mais fáceis de otimizar e podem ganhar acurácia com o aumento considerável da profundidade das camadas. No trabalho de [8] a rede rede neural residual foi implementada usando o conjunto de dados ImageNet e foram avaliadas diferentes redes com profundidade de até 152 camadas.

IV. Experimentos

Esta seção, descreve o conjunto de dados, método de avaliação, e resultados experimentais.

A. Datasets

Na etapa de detecção, a U-Net foi treinada com um conjunto de dados criado a partir de uma base de dados privada contendo 2000 imagens de veículos. Essas imagens foram rotuladas utilizando-se o *sofware* livre *Labelme* [30], que gera um arquivo com as coordenadas do retângulo demarcado na imagem, conforme Figura 2. Em seguida, foi gerado um conjunto de 2000 novas imagens com as dimensões da imagem original, porém com uma máscara em preto em branco indicando a região onde devia ser aprendida a detecção do farol pela rede (Figura 3).

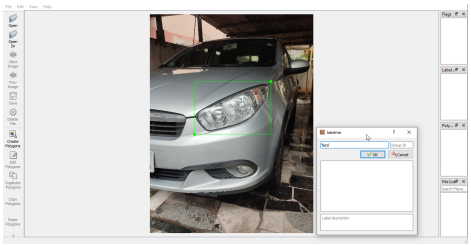

Fig. 2. Criação das máscaras com o software livre Labelme.

Fig. 3. Do lado esquerdo a imagem original e do lado direito a máscara criada para a detecção do farol.

Para a etapa de classificação do farol foi elaborada também uma base com imagens 1.000 imagens contendo apenas faróis veiculares. Para complemento da base e melhor contexto do sistema foram acrescentadas 1.000 imagens em que o farol já havia sido detectado corretamente pela U-Net e recortado para ser incluído no dataset, totalizando 2.000 imagens divididas igualmente em 500 imagens para cada uma das 4 classes: quebrado, embaçado, infiltrado ou sem danos. O recorte foi feito a partir de um código que automatizou o processo, com as coordenadas da localização do farol dadas pelo teste da U-net foi possível recortar exatamente a região esperada do farol. A Figura 4 mostra um exemplo de cada classe respectivamente.

Fig. 4. Da esquerda para a direita: farol quebrado, farol embaçado, farol infiltrado e farol sem dano.

Ambos os datasets da detecção e da classificação foram divididos para avaliação da seguinte maneira: 1400 imagens para treinamento, 400 imagens para validação e 200 imagens para teste. Nesta fase foram utilizadas as imagens segmentadas pela u-net tanto para treino quanto para validação. Já a base de teste da classificação foi composta apenas por imagens que foram detectadas inicialmente pela U-net e em seguida recortadas apenas para manter o farol.

As imagens de ambos os datasets estão em RGB, e foram redimensionadas para 256×256 pixeis e formato jpg. As imagens foram obtidas a partir de câmeras diversas, que podem ser de um telefone celular a máquinas profissionais. O posicionamento dos veículos nas imagens também é variado, alguns estão distantes e mostram a borda inteira, outros estão próximos e mostram apenas o farol. O tamanho das imagens também é variado entre 30KB e 5MB. Não é possível determinar a área física real da imagem representada, exatamente porque a distância de foco não foi pré-determinada, ou seja, não existe uma escala de comparação.

B. Avaliação

As bases de dados foram divididas em 70% para treinamento, 20% para validação e 10% pra teste. Para avaliar os resultados do modelo de detecção com U-Net, foram utilizadas a função de perda de entropia cruzada binária (loss) [31], que é apropriada para problemas de classificação entre duas categorias; e a interseção sobre união (IoU) [32], que mede a sobreposição entre duas caixas ou máscaras delimitadoras. Para avaliar o desempenho do modelo de detecção deste trabalho, foi utilizada a IoU, que mede o número de pixels comuns entre as máscaras criadas e da máscaras de predição dividido pelo número total de pixels presentes em ambas as máscaras.

A Equação (1) mostra o cálculo da IoU, onde TP (True Positive) é a quantidade de pixels corretamente prevista de acordo com a máscara de destino, enquanto TN (True Negative) é um verdadeiro negativo, ou seja, representa a quantidade de pixels que é corretamente identificado como não pertencente à determinada máscara. Um falso positivo (FP) indica que uma máscara de objeto prevista não tinha nenhuma máscara de objeto de verdade associada. Um falso negativo (FN) indica que uma máscara de objeto de verdade não tinha nenhuma máscara de objeto prevista associada.

$$IOU_c = \frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FP_c} + \mathrm{FN_c}} \tag{1}$$

Para avaliar os resultados do modelo de classificação com Resnet50, foi calculada a função de perda *sparse categorical cross entropy*. Essa função é apropriadas para problemas de classificação padrão onde o desempenho é medido pela acurácia geral da classificação multi classe. As saídas brutas da rede neural passam pela ativação softmax, que então gera um vetor de probabilidades previstas sobre as classes de entrada.

Para avaliar o modelo proposto também foi calculada a métrica de acurácia, apresentada na Equação (2). A acurácia refere-se ao número total de previsões corretas feitas, dividido pelo número total de todas as previsões das classes.

$$Accuracy_c = \frac{\mathrm{TP_c} + \mathrm{TN_c}}{\mathrm{TP_c} + \mathrm{FP_c} + \mathrm{FN_c} + \mathrm{TN_c}} \tag{2}$$

C. Resultados

O treinamento da U-Net para detecção do farol foi feito com 150 épocas (Figura 5 e Figura 6) e dessas, a que demonstrou o melhor resultado para média IoU da validação foi o treinamento da época 125. Nela a função loss atingiu 0,10 no treinamento, 0,75 na validação e 0,11 no teste. A função IoU resultou em 87% no treinamento, 72% na validação e 70% no teste.

Fig. 5. Média IoU por época no treinamento e validação da U-Net.

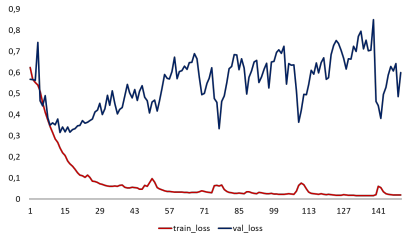

Fig. 6. Perda por época no treinamento e validação da U-Net.

O treinamento da Resnet50 para classificação do farol também foi feito com 150 épocas (Figura 7 e Figura 8) e dessas, a que demonstrou o melhor resultado de acurácia na

validação foi a época 137. Nela a função loss atingiu 0,24 no treinamento, 1,44 na validação e 1,49 no teste. A função de acurácia resultou em 92% no treinamento, 78% na validação e 76% no teste. Devido à especificidade do tema de tipos de danos em faróis, a Resnet50 usada na implementações foi pré-treinada no conjunto de dados ImageNet, porém as camadas ficaram congeladas com trainable = false . Uma vantagem disso é que o tempo de treinamento é menor.

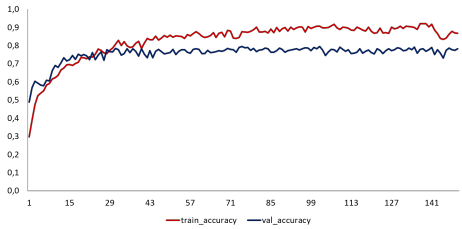

Fig. 7. Acurácia por época no treinamento e validação da Resnet.

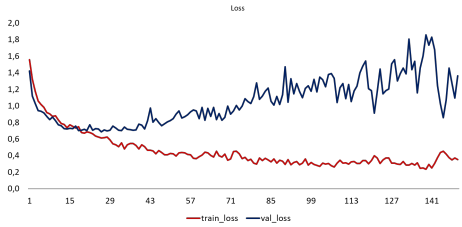

Fig. 8. Perda por época no treinamento e validação da Resnet.

O teste da Resnet50 indica um bom desempenho do modelo de arquitetura ResNet50, sendo o valor geral de acurácia no teste é obtido a partir de 76%. O farol embaçado foi classificado com uma acurácia de 94%, enquanto o farol infiltrado e neutro obtiveram acurácia de 76% e 96%, respectivamente. Já o farol quebrado obteve uma acurácia baixa de 40%. Pelo material do farol ser um vidro transparente, a trinca pode passar despercebida até numa verificação humana. Na matriz de confusão, apresentada na Figura 10, é possível identificar que 17 imagens consideradas farol neutro foram classificadas como farol quebrado.

Ao analisar as imagens, é possível compreender esse resultado, já que na maioria das imagens de farol quebrado e infiltrado, eles também estavam embaçados, o que indica a confusão para está classe. A Figura 9 mostra um exemplo onde o farol está embaçado e infiltrado ao mesmo tempo.

Para evidenciar o desempenho de um algoritmo de classificação em termos de quantidade de decisões por classe, a Figura 10 mostra a matriz de confusão que permite a visualização da frequência de classificação para cada classe do modelo.

Como pode ser visto, a taxa de reconhecimento deste método é diferente entre as diferentes classes, atingindo um nível superior a 76% no geral, entre os quais a acurácia deste método é relativamente alta na classificação de imagens claramente definidas, como farol embaçado. Isso pode ser devido ao

Fig. 9. Farol embaçada e infiltrado ao mesmo tempo.

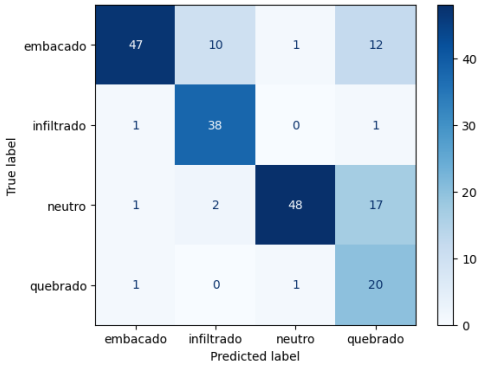

Fig. 10. Matriz de confusão da classificação com Resnet50.

fato de que imagens claramente definidas apresentam maiores vantagens na extração de características. Os resultados obtidos no teste do sistema obtiveram média de IOU de 70% na detecção e acurácia de 76% na classificação.

V. Conclusões e Trabalhos Futuros

A análise de imagem de danos veiculares é um procedimento realizado pelas seguradoras para determinar se o serviço é coberto pela apólice ou não. Em caso de danos nos faróis, a empresa recebe uma foto do veículo e uma equipe de especialistas analisam os danos. Como as imagens são recebidas com diferentes padrões é necessário primeiramente detectar o farol na imagem e depois determinar se existe algum dano e qual é este dano. Neste artigo, foi proposto um sistema de detecção e classificação de imagens de faróis de veículos para automatizar a inspeção. O método baseia-se primeiro na estrutura U-Net para detectar e recortar o farol na imagem e em seguida na estrutura Resnet50 para classificar o dano. A base de dados para treinamento da U-Net é composta por 2.000 imagens de veículos e 2.000 máscaras com detecção de faróis. A base de dados utilizada nos experimentos de classificação do farol com a Resnet50 é composta por 2.000 imagens divididas em 4 classes, que contém 500 imagens: quebrado, embaçado, infiltrado ou sem dano. Os resultados obtidos no teste tiveram média de IOU de 70% na detecção e acurácia de 76% na classificação como pode ser visto na Tabela I.

Para demonstrar uma melhor avaliação do modelo proposto é necessário aumentar o número de dados com faróis defeituosos. Além disso, é um desafio incluir imagens tiradas em várias posições no conjunto de dados, considerar padronizar

TABLE I
RESULTADOS DAS MÉTRICAS DE CADA EXPERIMENTO NOS TESTES.

Detecção com U-Net	
IOU	70%
Classificação com Resnet50	
ACURÁCIA	76%
TP	151 imagens
FP	8 imagens
FN	41 imagens

essas imagens e utilizar fotos com qualidade mais alta, adequando a iluminação com radiação ultravioleta (UV) e radiação infravermelha (IR) por exemplo, para destacar os faróis que precisem ser examinados, possivelmente melhoraria os resultados. No futuro, pretendemos melhorar a acurácia, aumentar e refinar as bases de dados de acordo com a qualidade das imagens. Também pretendemos aumentar o número de classes separando casos em que os faróis estão acesos pois entendemos que esta característica pode ter um grande impacto no treinamento da rede. Outra possível pesquisa futura seria comparar os resultados obtidos neste trabalho com classificações similares que usam outras estruturas de redes neurais e identificar soluções plausíveis para este problema que ainda não exploradas.

REFERÊNCIAS

[1] BRASIL, "Lei nº 9.503, de 23 de setembro de 1997 compilado. institui o código de trânsito brasileiro.," 2023.

[2] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual understanding of convolutional neural network-a deep learning approach," *Procedia computer science*, vol. 132, pp. 679–688, 2018.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *IEEE Access*, vol. 8, pp. 54564–54573, 2020.

[4] D. Dais, I. Bal, E. Smyrou, and V. Sarhosis, "Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning," *Automation in Construction, Elsevier: v. 125, n. 1*, 2021.

[5] S. Singh and S. Prasad, "Techniques and challenges of face recognition: A critical review," *Procedia Computer Science: v. 143, n. 1*, pp. 536–543, 2018.

[6] A. Pathak, M. Pandey, and S. Rautaray, "Application of deep learning for object detection," *Procedia Computer Science: v. 132,, n. 1*, pp. 1706–1717, 2018.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, (Las Vegas, NV, USA), pp. 770–778, 2016.

[9] J. Chai, H. Zeng, A. Li, and E. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Machine Learning with Applications: v. 6,, n. 1*, 2021.

[10] A. Karn, "Artificial intelligence in computer vision," *International Journal of Engineering Applied Sciences and Technology*, vol. 6, pp. 249–254, 07 2021.

[11] J. Wu, B. Peng, Z. Huang, and J. Xie, "Research on computer vision-based object detection and classification," in *Computer and Computing Technologies in Agriculture VI* (D. Li and Y. Chen, eds.), (Berlin, Heidelberg), pp. 183–188, Springer Berlin Heidelberg, 2013.

[12] L. Lu, "Improved yolov8 detection algorithm in security inspection image," 2023.

[13] A. Naumann, F. Hertlein, L. Dörr, S. Thoma, and K. Furmans, "Literature review: Computer vision applications in transportation logistics and warehousing," 2023.

[14] X. Xie, F. Shi, J. Niu, and X. Tang, "Breast ultrasound image classification and segmentation using convolutional neural networks," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19*, pp. 200–211, Springer, 2018.

[15] V. Kumar, H. Arora, J. Sisodia, *et al.*, "Resnet-based approach for detection and classification of plant leaf diseases," in *2020 international conference on electronics and sustainable communication systems (ICESC)*, pp. 495–502, IEEE, 2020.

[16] E. C. Tetila, B. B. Machado, G. Astolfi, N. A. de Souza Belete, W. P. Amorim, A. R. Roel, and H. Pistori, "Detection and classification of soybean pests using deep learning with uav images," *Computers and Electronics in Agriculture*, vol. 179, p. 105836, 2020.

[17] K. Chen, G. Reichard, X. Xu, and A. Akanmu, "Automated crack segmentation in close-range building façade inspection images using deep learning techniques," *Journal of Building Engineering*, vol. 43, p. 102913, 2021.

[18] T. U. Ahmed, M. S. Hossain, M. J. Alam, and K. Andersson, "An integrated cnn-rnn framework to assess road crack," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2019.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[20] R. C. Staudemeyer and E. R. Morris, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks," 2019.

[21] Z. Qu, J. Mei, L. Liu, and D.-Y. Zhou, "Crack detection of concrete pavement with cross-entropy loss function and improved vgg16 network model," *IEEE Access*, vol. 8, pp. 54564–54573, 2020.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] D. Jurado-Rodríguez, J. M. Jurado, L. Pádua, A. Neto, R. Munoz-Salinas, and J. J. Sousa, "Semantic segmentation of 3d car parts using uav-based images," *Computers & Graphics*, vol. 107, pp. 93–103, 2022.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

[25] H. Jung, M.-K. Choi, J. Jung, J.-H. Lee, S. Kwon, and W. Young Jung, "Resnet-based vehicle classification and localization in traffic surveillance systems," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, (Honolulu, HI, USA), pp. 61–67, 2017.

[26] R. Watkins, N. Pears, and S. Manandhar, "Vehicle classification using resnets, localisation and spatially-weighted pooling," *arXiv preprint arXiv:1810.10329*, 2018.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015.

[28] R. Galina, T. Melo, and K. Komati, "Pavement crack segmentation using a u-net based neural network," in *Anais do XVII Workshop de Visão Computacional*, (Porto Alegre, RS, Brasil), pp. 76–81, SBC, 2021.

[29] E. Coltri, G. Costa, K. Silva, P. Martim, and L. Bergamasco, "Automatic segmentation and roi detection in cardiac mri of cardiomyopathy using q-sigmoid as preprocessing step," in *Anais do XVII Workshop de Visão Computacional*, (Porto Alegre, RS, Brasil), pp. 143–147, SBC, 2021.

[30] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, pp. 157–173, 2008.

[31] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, oct 2020.

[32] M. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," vol. 10072, pp. 234–244, 12 2016.