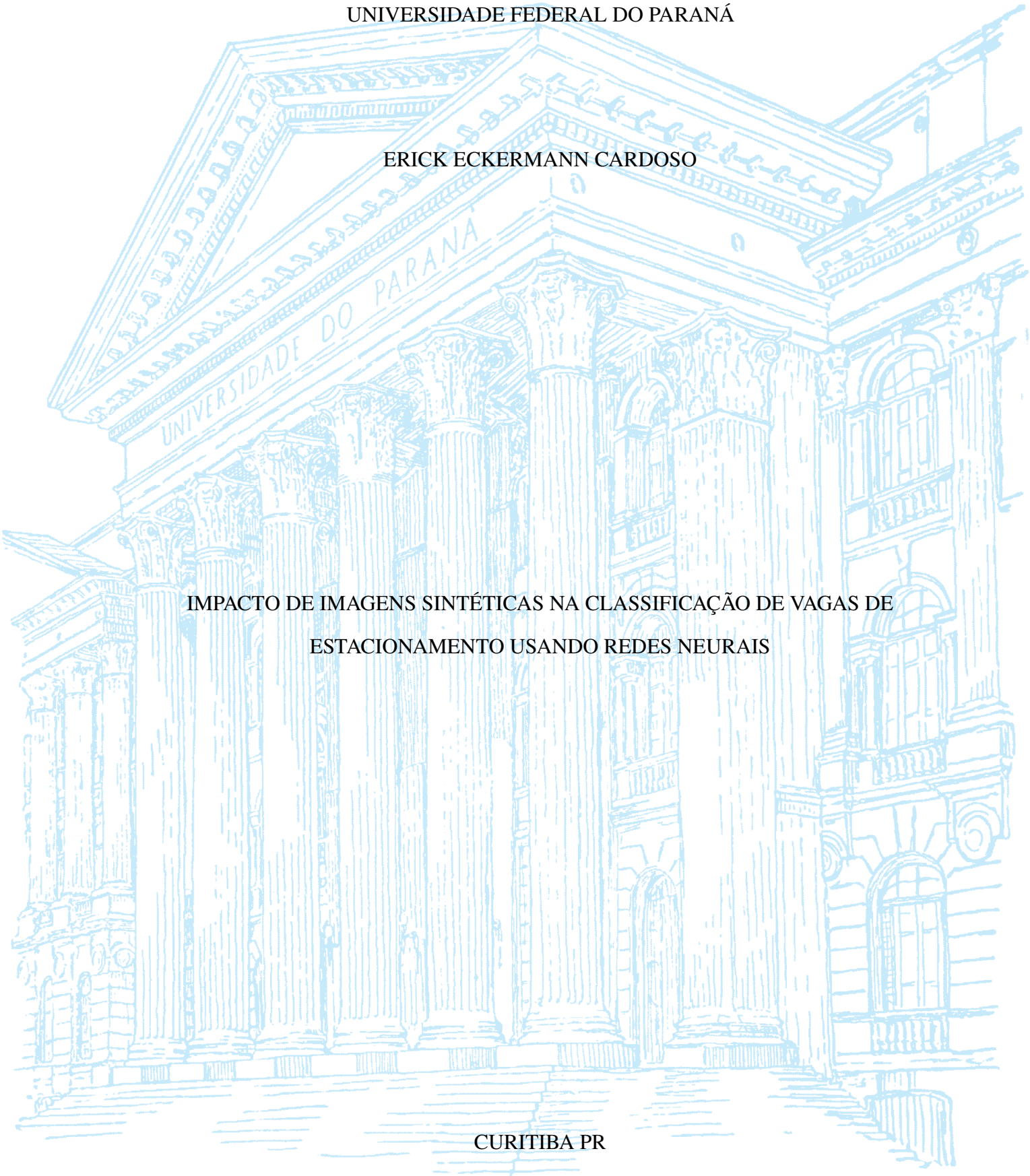
UNIVERSIDADE FEDERAL DO PARANÁ

ERICK ECKERMANN CARDOSO



IMPACTO DE IMAGENS SINTÉTICAS NA CLASSIFICAÇÃO DE VAGAS DE ESTACIONAMENTO USANDO REDES NEURAIS

CURITIBA PR

2024

ERICK ECKERMANN CARDOSO

IMPACTO DE IMAGENS SINTÉTICAS NA CLASSIFICAÇÃO DE VAGAS DE ESTACIONAMENTO USANDO REDES NEURAIS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Paulo R. Lisboa de Almeida.

CURITIBA PR

2024

A minha amada esposa, Isabelle, por todo o amor, apoio e carinho, e a meus pais por nunca terem medido esforços para me proporcionar um ensino de qualidade.

AGRADECIMENTOS

A Deus, pela minha vida, e por me ajudar a ultrapassar todos os obstáculos encontrados ao longo do cruso. À minha esposa, pelo amor, paciência e compreensão durante os momentos em que estive ausente, dedicando-me aos estudos. Sua presença e apoio constante foram fundamentais para que eu chegasse até aqui. Aos meus pais e irmãos, que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto me dedicava á realização deste trabalho, e ao longo de todo o curso. Aos professores, pelas correções e ensinamentos que me permitiram apresentar um melhor desempenhono meu processo de formação profissional. À Universidade Federal do Paraná, essencial no meu processo de formação profissional, pela sua história, e por tudo o que pude aprender aqui.

RESUMO

O crescimento urbano e o aumento da frota de veículos têm intensificado a demanda por soluções eficientes para a gestão de estacionamentos. Nesse cenário, sistemas de monitoramento por imagem baseados em aprendizado de máquina têm se destacado devido ao seu baixo custo e facilidade de instalação em comparação a métodos tradicionais, como sensores físicos. Esses sistemas já alcançam uma acurácia média de 95% em validações cruzadas, utilizando bases de dados conhecidas, como PKLot e CNRPark-EXT. No entanto, apesar da existência dessas bases extensas, ainda persistem desafios relacionados à disponibilidade e à diversidade dos dados necessários para treinamento, especialmente ao buscar melhorar a acurácia de modelos generalistas ou especializá-los para cenários específicos, onde em cada aplicação uma quantidade de imagens precisa ser coletada, segmentada e rotulada para ajustar o modelo se obter a melhor acurácia..

Este trabalho propõe a utilização de imagens sintéticas, geradas com o motor gráfico Unity 5 em conjunto com o pacote Unity Perception, para complementar ou substituir dados reais no treinamento de modelos de classificação de vagas de estacionamento. Um protocolo de geração de imagens foi desenvolvido, visando menor custo de criação comparado ao custo de se coletar, segmentar e rotular imagens reais. As imagens geradas por meio desse protocolo são denominadas de baixa fidelidade, devido à baixa qualidade das imagens e menor capacidade de simular um ambiente específico.

Utilizando a MobileNetV3 e aprendizado por transferência, foram realizados experimentos em três cenários: substituição total de dados reais, complemento de bases diversificadas e especialização em cenários específicos. Os resultados demonstraram que, em bases com poucos dados reais, o uso de imagens sintéticas pode aumentar a acurácia em até 2% (ex.: CNRPark-EXT), melhorando a generalização do modelo. Contudo, as imagens sintéticas não foram capazes de substituir completamente os dados reais devido à falta de fidelidade em replicar condições reais, reforçando a necessidade de combinações com dados reais ou dados mais realistas para melhores resultados.

Palavras-chave: Visão Computacional. Dados Sintéticos. Deep Learning. Classificação de Estacionamento.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná
PUCPR	Pontifícia Universidade Católica do Paraná
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
IoT	*Internet of Things* (Internet das coisas)
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)
HOG	*Histogram of Oriented Gradients*
LBP	*Local Binary Patterns*
SIFT	*Scale-Invariant Feature Transform*
NDDS	*NVIDIA Deep Learning Dataset Synthesizer*
EER	*Equal Error Rate*
ROC	*Receiver Operating Characteristic*

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o crescimento acelerado das cidades e a crescente frota de veículos têm levantado o problema de se estacionar carros de forma ágil em vias públicas e grandes espaços de estacionamento e destacado a necessidade de uma gestão eficiente desses espaços. A busca por soluções inovadoras que otimizem o seu uso, proporcionando comodidade aos motoristas e reduzindo congestionamentos, tem sido objeto de extensas pesquisas na última década (de Almeida et al., 2022; Paidi et al., 2018). Dentro desse contexto, soluções de monitoramento por imagem utilizando métodos baseados em visão computacional e aprendizado de máquina são comumente escolhidas devido ao seu baixo custo e facilidade de implantação (de Almeida et al., 2022; Almeida et al., 2013; Amato et al., 2017; Hochuli et al., 2023), comparado a outras técnicas de gerenciamento, como as baseadas em sensores.

Nos métodos baseados em visão computacional e aprendizado de máquina, imagens capturadas de uma alta e ampla perspectiva servem para monitorar uma grande área de estacionamento. Cada vaga é identificada e segmentada da imagem e modelos de classificação baseados em aprendizado de máquina são utilizados para classificar as vagas segmentadas como ocupadas ou livres.

Os modelos de classificação de vagas de estacionamento demandam uma boa quantidade de imagens com qualidade, diversidade e boa representatividade do mundo real para o treinamento. A coleta dessas imagens é uma tarefa complexa e dispendiosa, requerendo tempo, recursos financeiros e esforço. Nesse contexto, bases de dados como a PkLot (Almeida et al., 2015) e CNRPark-EXT (Amato et al., 2017) foram criadas, a fim de disponibilizar imagens para o treinamento e validação de modelos no problema de classificação de vagas de estacionamento. Experimentos com essas e outras bases de dados apresentaram resultados de, em média, 95% de acurácia na classificação com modelos treinados e testados em cenários cruzados (de Almeida et al., 2022; Hochuli et al., 2023) e 97% ao ajustar um modelo generalista apra um cenário específico com poucos dados rotulados do cenário (Hochuli et al., 2023). Ainda assim, a dificuldade e esforço em se obter e rotular dados persiste, sendo necessário realizar esse trabalho de coleta a cada novo cenário em que se desejar um modelo com a melhor acurácia possível. No entanto, dado que é um problema em evolução contínua, a necessidade de novas bases de dados para treinamento e validação de modelos ainda é um problema (de Almeida et al., 2022).

Uma forma de contornar o problema da falta de dados é a utilização de dados sintéticos (Ekbatani et al., 2017; Tobin et al., 2017). Neste trabalho, propõe-se a utilização de imagens sintéticas no treinamento de modelos de classificação de vagas de estacionamento como forma de mitigar os desafios existentes no uso de dados reais. Dados sintéticos são dados gerados de forma artificial e algorítmica, assemelhando-se aos dados reais, embora não surjam de observações diretas ou coletas reais. Isso possibilita superar alguns desafios associados à coleta de dados reais, permitindo a criação de conjuntos de dados volumosos, com maior variedade e a um custo reduzido, em um período de tempo menor e com a capacidade de alcançar e até superar os resultados obtidos com dados reais (Tremblay et al., 2018).

Em Tobin et al. (2017), podemos ver que imagens sintéticas randomizadas, não realistas e de baixa proximidade com o cenário original podem alcançar resultados comparáveis a de imagens reais ao serem utilizadas em modelos de aprendizado profundo para reconhecimento de objetos, sendo necessário um esforço consideravelmente pequeno para serem geradas. Com isso, a proposta deste trabalho é utilizar deste método de geração de dados sintéticos para avaliar o impacto das imagens sintéticas nos modelos de classificação de estacionamento. Vamos

denominar as imagens geradas por meio desse método como imagens sintéticas de baixa fidelidade. As seguintes perguntas foram praparadas para orientação da pesquisa:

- P1 - De que forma a combinação de imagens reais e sintéticas de baixa fidelidade afeta a generalização de modelos de classificação de vagas de estacionamento?
- P2 - Como imagens sintéticas de baixa fidelidade se comportam na aplicação e especificação de modelos de classificação de vagas de estacionamento para um cenário alvo?
- P3 - É possível superar a acurácia dos modelos treinados com imagens reais treinando os modelos somente com imagens sintéticas de baixa fidelidade?

Para a geração das imagens sintéticas de baixa fidelidade, será utilizado o motor gráfico Unity 5 em conjunto com o pacote Unity-Perception (Borkman et al., 2017), escolhido por sua flexibilidade na criação de ambientes personalizados e pela capacidade de gerar grandes volumes de dados rotulados automaticamente, reduzindo significativamente o custo de coleta e anotação. No treinamento, será adotada uma Rede Neural Convolucional (CNN) utilizando aprendizado por transferência, uma técnica eficiente em cenários com dados limitados. A MobileNetv3 (Howard et al., 2019), em sua versão *large* pré-treinada na ImageNet (Deng et al., 2009), foi escolhida como modelo base devido aos resultados obtidos na classificação de vagas de estacionamento em trabalhos relacionados, demonstrando equilíbrio entre custo computacional e acurácia, tornando-a adequada para lidar com os diferentes cenários simulados. As bases de dados PKLot e CNRPark-EXT serão usadas tando como referência para geração dos dados sintéticos, quanto para o treinamento e validação dos modelos, servindo como base de comparação para avaliação do impacto das imagens sintéticas.

O restante deste trabalho é organizado da seguinte forma. No Capítulo 2, são apresentados os conceitos fundamentais para o desenvolvimento de modelos de classificação de ocupação de vagas de estacionamento e uso de dados sintéticos, com destaque para problemas de classificação, redes neurais convolucionais e geração de dados sintéticos. O Capítulo 3 discute o estado da arte, abordando as principais abordagens para classificação de vagas, as bases de dados disponíveis e o uso de imagens sintéticas em problemas de aprendizado de máquina.

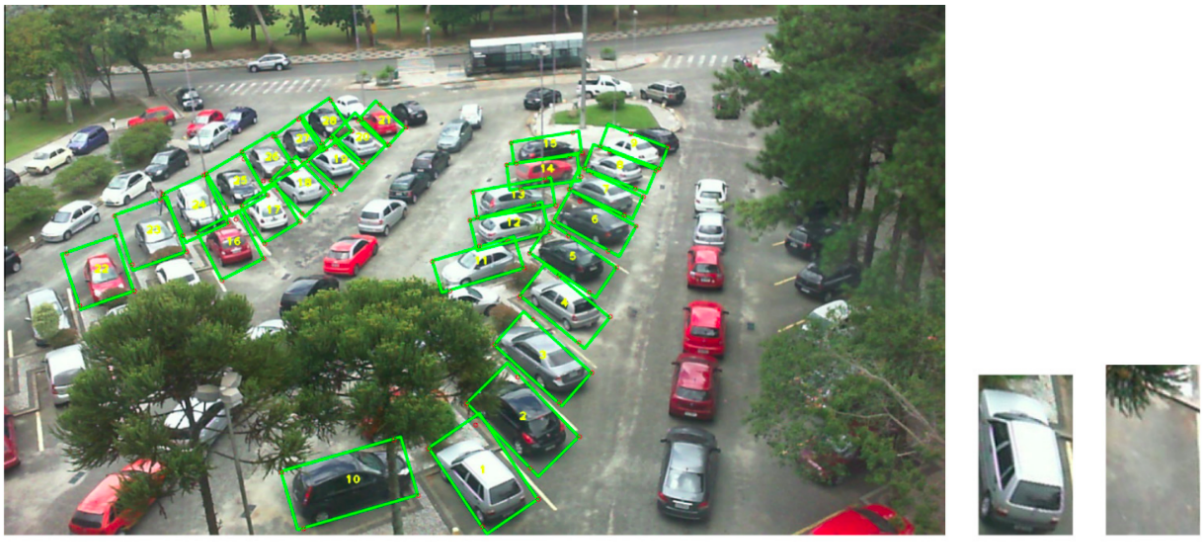
No Capítulo 4, é detalhada a proposta deste trabalho, incluindo o processo de geração de imagens sintéticas de baixa fidelidade utilizando o Unity Perception e o protocolo de treinamento dos modelos de classificação. O Capítulo 5 descreve os experimentos realizados, apresentando os cenários de treinamento e os resultados obtidos com diferentes configurações de dados. Por fim, no Capítulo 6, são apresentadas as conclusões, discutindo as limitações e sugerindo direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são abordados os conceitos e técnicas fundamentais para o desenvolvimento de modelos de classificação de ocupação de vagas de estacionamento, com foco no uso de dados sintéticos. Serão explorados tópicos como problemas de classificação, redes neurais convolucionais (CNNs), modelos pré treinados, e geração de dados sintéticos.

2.1 PROBLEMAS DE CLASSIFICAÇÃO

Os problemas de classificação são tarefas centrais em aprendizado de máquina e visão computacional, nos quais o objetivo é categorizar dados em uma ou mais classes predefinidas. Na classificação de vagas de estacionamento, por exemplo, os dados de entrada (imagens) devem ser classificados para determinar se uma vaga está ocupada ou livre. Esse tipo de problema é fundamentalmente binário (duas classes), mas também pode ser expandido para várias classes dependendo do contexto, como a detecção de tipos específicos de veículos (Suhao et al., 2018) ou detecção e classificação da nacionalidade de placas veiculares (Henry et al., 2020). A Figura 2.1 exemplifica um problema de classificação de vagas de estacionamento, um problema de natureza binária onde o objetivo é classificar cada vaga como ocupada ou livre.



(a) (b) (c)

Figura 2.1: Classificação de vagas de estacionamento: (a) Visão geral do estacionamento com a marcação de cada vaga, (b) exemplo de vaga classificada como ocupada, (c) exemplo de vaga classificada como livre. Retirado de Almeida et al. (2015).

Algoritmos como Support Vector Machine (SVM) e Redes Neurais Convolucionais (CNNs) são amplamente utilizados em tarefas de classificação. As SVMs, por exemplo, têm sido empregadas com sucesso para classificação de vagas utilizando descritores de textura como LBP e LPQ onde as amostras de treino e teste são originadas do mesmo cenário (Almeida et al., 2013), enquanto as CNNs destacam-se em cenários mais complexos, como a identificação de vagas ocupadas em imagens capturadas sob diferentes condições climáticas e de iluminação (Grbić e Koch, 2023) ou onde as amostras de treino e teste são de cenários diferentes, ou seja, validação cruzada (Hochuli et al., 2023). Esses cenários frequentemente envolvem variações significativas

no contexto visual, exigindo modelos capazes de capturar padrões detalhados e invariantes a essas mudanças. Por isso, as CNNs têm se mostrado mais eficazes, principalmente em problemas que demandam maior adaptabilidade a diferentes cenários operacionais e condições ambientais.

2.2 REDES NEURAIS CONVOLUCIONAIS (CNN)

Uma rede neural convolucional (CNN, do inglês *Convolutional Neural Network*) é um tipo específico de rede neural projetada para lidar com dados que possuem uma estrutura de grade, como imagens. Esse tipo de rede é amplamente usado em tarefas de visão computacional, como reconhecimento de objetos, classificação de imagens, detecção de objetos, entre outras. O grande diferencial das redes convolucionais é que elas conseguem extrair automaticamente características dos dados, capturando padrões de forma hierárquica.

A camada de convolução é a base das redes convolucionais. Nela, um filtro (ou *kernel*) percorre a imagem de entrada, realizando operações de convolução. Esse filtro é uma pequena matriz (por exemplo, 3×3 ou 5×5) com pesos treináveis. A convolução é a operação que calcula uma nova matriz de saída, chamada *feature map*, que representa certas características da imagem original.

Cada filtro tem a capacidade de identificar um padrão específico, como bordas, texturas ou formas. À medida que a rede aprende, esses filtros se especializam em detectar diferentes padrões, capturando informações da imagem em níveis variados de abstração. Após cada operação de convolução, uma função de ativação, como a ReLU (*Rectified Linear Unit*), geralmente é aplicada ao *feature map*. A função ReLU zera todos os valores negativos e mantém os positivos, aplicando uma não linearidade e facilitando a identificação de características complexas.

Depois da convolução e da ativação, as CNNs geralmente aplicam uma camada de *pooling*. A camada de *pooling* reduz a dimensionalidade dos dados, resumindo informações em blocos menores, mantendo apenas as características mais importantes. Uma técnica comum é o *max pooling*, que pega o valor máximo de uma pequena área, como 2×2, na imagem de entrada. A redução de dimensionalidade tem duas vantagens principais:

- Reduz a quantidade de parâmetros, acelerando o treinamento e a inferência.
- Introduz invariância à translação, o que ajuda a rede a reconhecer padrões, independentemente de pequenas variações na posição dentro da imagem.

As CNNs modernas geralmente têm várias camadas de convolução e *pooling*. As primeiras camadas detectam características de baixo nível, como bordas e texturas, enquanto camadas mais profundas capturam características de alto nível, como formas e objetos inteiros. Esse processo cria uma representação hierárquica da imagem, o que é crucial para entender e classificar padrões complexos.

No final de uma CNN, as camadas convolucionais e de *pooling* são seguidas por uma ou mais camadas completamente conectadas (*Fully Connected*). Essas camadas funcionam como redes neurais tradicionais e são usadas para tomar decisões finais com base nas características extraídas nas camadas anteriores. Cada neurônio em uma camada totalmente conectada recebe a informação de todos os neurônios da camada anterior.

O treinamento de uma CNN envolve ajustar os pesos dos filtros em cada camada convolucional e das conexões na camada totalmente conectada. A rede aprende esses pesos por meio de um processo chamado retropropagação, em que o erro de saída é retropropagado para ajustar os pesos de forma a melhorar o desempenho da rede.

A Figura 2.2 apresenta um diagrama genérico da arquitetura de uma CNN com os componentes mencionados. O número de camadas de convolução, ativação e pooling pode variar

de acordo com a arquitetura escolhida, mas, em geral, todas elas seguem a ordem apresentada no diagrama.

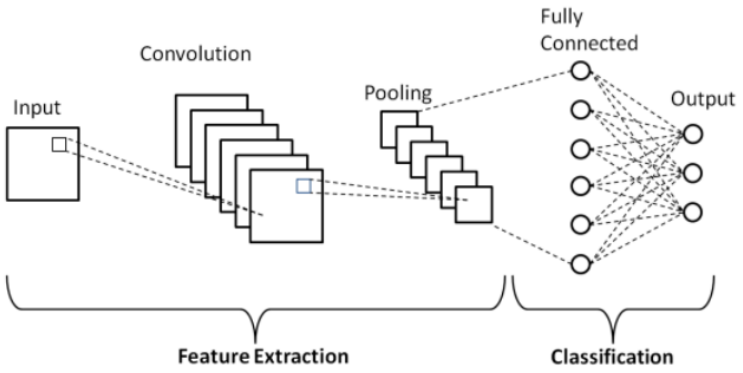


Figura 2.2: Diagrama genérico da arquitetura de uma Rede Neural Convolucional (CNN) (Phung e Rhee, 2019).

Arquiteturas populares de CNN incluem:

- AlexNet: Um dos primeiros modelos de CNN a demonstrar a eficácia de redes profundas para a classificação de imagens. Foi o vencedor do ImageNet Large Scale Visual Recognition Challenge (ILSVRC) em 2012, revolucionando a área de visão computacional (Krizhevsky et al., 2012).
- VGGNet: Caracterizada por sua simplicidade e profundidade, a VGGNet utiliza pequenas camadas convolucionais de 3x3 pixels empilhadas umas sobre as outras, permitindo a criação de redes muito profundas (até 19 camadas)(Simonyan e Zisserman, 2014).
- ResNet: Introduziu o conceito de conexões residuais (skip connections), permitindo a criação de redes extremamente profundas (com mais de 100 camadas) sem os problemas de degradação de desempenho que ocorrem em redes profundas tradicionais (He et al., 2015).
- MobileNetV3: Projetada para ser eficiente em termos de recursos computacionais, a MobileNet é ideal para aplicações móveis e embarcadas. Utiliza convoluções separáveis em profundidade para reduzir a complexidade computacional (Howard et al., 2019).

No problema de classificação de vagas de estacionamento, modelos como AlexNet e MobileNet têm demonstrado alta acurácia e têm sido amplamente utilizados (Hochuli et al., 2023; Amato et al., 2017; de Almeida et al., 2022). Neste trabalho foi utilizada a arquitetura MobileNetV3 (Howard et al., 2019) para a construção dos modelos de classificação.

2.2.1 MobileNetV3

A MobileNetV3 (Howard et al., 2019) é uma arquitetura de rede neural convolucional projetada especificamente para realizar tarefas de visão computacional (como classificação de imagens) e para ser leve e eficiente, visando aplicações em dispositivos móveis ou com baixa capacidade de processamento, como celulares e dispositivos IoT (Internet das Coisas, do inglês *Internet of Things*). Ela é parte da família MobileNet, que inclui versões anteriores como a MobileNetV1(Howard et al., 2017) e MobileNetV2(Sandler et al., 2019), cada uma consumindo menos recursos computacionais e aprimorando a acurácia em relação à anterior. A MobileNetV3

combina várias inovações e técnicas para reduzir o consumo de energia e melhorar a velocidade, sem perder a acurácia.

Entre as principais características da MobileNetV3 estão os blocos residuais invertidos, que aumentam a eficiência da rede ao expandir e comprimir as informações durante o processamento, e as convoluções com profundidade separável, que reduzem significativamente o número de cálculos necessários. A rede também utiliza os blocos *Squeeze-and-Excitation* (SE), que ajustam o peso de cada canal com base em sua relevância, e a função de ativação *Hard-Swish*, que é mais rápida e eficiente que funções tradicionais como ReLU e Swish.

A rede possui duas versões principais, a MobileNetV3-Large e a MobileNetV3-Small. A principal diferença entre elas está no equilíbrio entre acurácia e eficiência. A Large, com cerca de 5,4 milhões de parâmetros, é projetada para tarefas que exigem maior acurácia e podem tolerar um maior consumo de recursos computacionais, utilizando mais camadas e filtros, o que a torna ideal para dispositivos móveis modernos e aplicações mais complexas. Já a Small, com aproximadamente 2,9 milhões de parâmetros, prioriza eficiência e baixo consumo de energia, sendo mais leve e rápida, adequada para dispositivos com hardware limitado, como sensores IoT e drones, onde a velocidade e a economia de recursos são mais importantes que a acurácia absoluta (Howard et al., 2019; Kolosov et al., 2022).

2.2.2 Aprendizado por Transferência

O aprendizado por transferência (Zhuang et al., 2020) é uma técnica de aprendizado de máquina que reutiliza o conhecimento adquirido por um modelo treinado em uma tarefa para resolver uma segunda tarefa relacionada. Essa abordagem elimina a necessidade de treinar um modelo do zero, economizando tempo e recursos computacionais, além de facilitar a obtenção de bons resultados, especialmente em cenários com dados limitados. Por exemplo, uma rede neural treinada em um grande conjunto de dados genéricos, como o ImageNet(Deng et al., 2009), pode ser aproveitada para extrair características úteis em tarefas específicas, como a classificação de ocupação de vagas de estacionamento.

Esse processo ocorre em duas etapas principais. Na primeira, utiliza-se um modelo pré-treinado, que já aprendeu a extrair características gerais, como bordas, texturas e formas. Essas características podem ser aplicadas diretamente em novos problemas, mantendo todas as camadas do modelo inalteradas. Isso é chamado de aprendizado por transferência puro, que se limita ao uso do modelo como um extrator de características.

Na segunda etapa, caso a tarefa alvo exija mais especialização, pode-se realizar o ajuste fino (*fine-tuning*). Esse processo consiste em descongelar parte das camadas do modelo pré-treinado e ajustá-las aos dados da nova tarefa. Enquanto as primeiras camadas da rede, responsáveis por detectar padrões básicos, geralmente são mantidas inalteradas, as camadas finais, mais especializadas na tarefa original, podem ser adaptadas ou substituídas para ajustar o modelo à tarefa específica.

Entre as principais vantagens do aprendizado por transferência estão a economia de tempo e recursos computacionais, já que o modelo pré-treinado fornece uma base sólida para tarefas relacionadas. O ajuste fino, por sua vez, oferece maior flexibilidade e permite alcançar desempenhos superiores em aplicações específicas, mesmo quando a quantidade de dados disponíveis é limitada.

2.3 GERAÇÃO DE DADOS SINTÉTICOS

Dados sintéticos são gerados utilizando algoritmos e simulações para criar dados que imitam características dos dados reais. Ferramentas como Unity 5(Borkman et al., 2017) e Unreal

Engine (To et al., 2018) são frequentemente usadas para gerar imagens sintéticas, que podem ser utilizadas para treinar modelos de visão computacional. Esses ambientes simulados permitem a criação de cenários variados e controlados, facilitando a geração de grandes volumes de dados rotulados rapidamente. Ambientes simulados são criados para gerar dados sintéticos que podem replicar condições do mundo real, como iluminação, texturas e condições climáticas (Tschentscher et al., 2017). No caso de classificação de vagas, o uso de dados sintéticos pode ajudar a mitigar a falta de dados reais (de Almeida et al., 2022) e simulações realistas podem melhorar a acurácia dos modelos ao fornecer uma grande diversidade de exemplos para treinamento. Por exemplo, a Unity Perception, permite a criação e captura de ambientes simulados, gerando imagens rotuladas automaticamente com variáveis controladas (Borkman et al., 2017). A Figura 2.3 mostra um exemplo de um ambiente 3D simulado criado para a geração de imagens no problema de classificação de vagas de estacionamento.

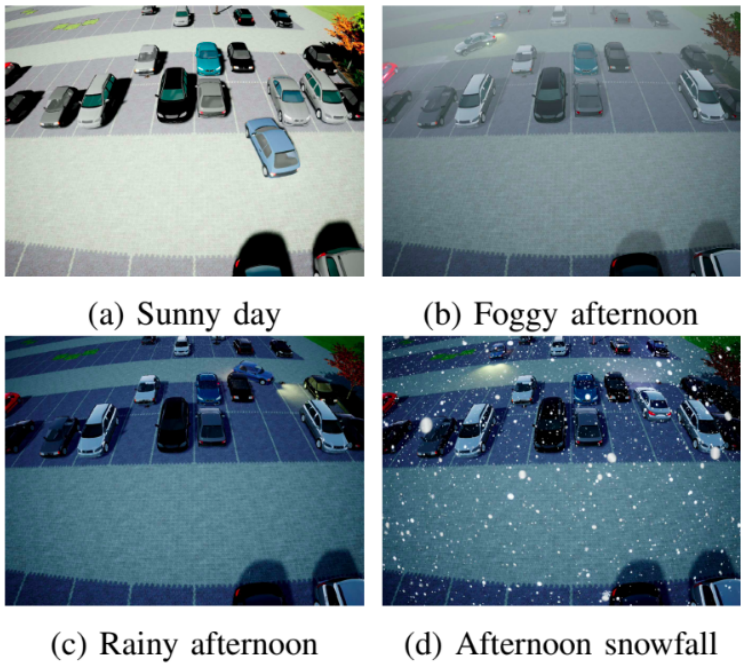


Figura 2.3: Ambiente 3D simulando um estacionamento em diferentes condições climáticas e de iluminação. Retirado de Tschentscher et al. (2017).

Entretanto, existem algumas barreiras quando se trata de utilizar dados sintéticos para treinamento de modelos de classificação. O *Reality Gap* (Tobin et al., 2017), conhecido como a lacuna entre ambientes sintéticos e a complexidade do mundo real, resulta na dificuldade dos modelos treinados apenas com dados sintéticos em se adaptarem adequadamente a situações reais. Essa discrepância surge devido à complexidade em simular fielmente todas as características visuais, físicas e dinâmicas do mundo real nos dados sintéticos. Elementos como iluminação, texturas, variações climáticas e interações complexas são desafios para a reprodução precisa. Consequentemente, modelos treinados exclusivamente com dados sintéticos podem ter dificuldade em generalizar para situações reais. Uma estratégia para superar essa limitação é combinar dados sintéticos e reais durante o treinamento (Tremblay et al., 2018), oferecendo ao modelo uma exposição mais diversificada e possibilitando uma melhor adaptação e desempenho em cenários do mundo real.

2.4 CONCLUSÃO

Nesta seção, foram apresentados os conceitos fundamentais para o desenvolvimento de modelos de classificação de vagas de estacionamento, explorando desde os problemas de classificação até a geração de dados sintéticos. Inicialmente, destacou-se a importância da classificação como uma tarefa central em aprendizado de máquina. Em seguida, o uso de Redes Neurais Convolucionais (CNNs) foi abordado como um avanço significativo, proporcionando a capacidade de identificar padrões complexos em imagens de forma hierárquica e automatizada. Arquiteturas modernas, como MobileNetV3, foram destacadas por seu uso eficiente de recursos computacionais e acurácia, especialmente em dispositivos com recursos computacionais limitados.

O aprendizado por transferência, por sua vez, foi discutido como uma solução prática para adaptar modelos pré-treinados a tarefas específicas, como a classificação de ocupação de vagas. Complementarmente, foi discutida a importância da geração de dados sintéticos como alternativa para superar desafios associados à coleta de dados reais. Apesar das vantagens de rapidez e controle proporcionadas por ferramentas como Unity Perception, foi evidenciado o impacto do *Reality Gap* na generalização dos modelos e a necessidade de combinar dados sintéticos e reais para melhorar a robustez das soluções.

Esses conceitos estabelecem a base teórica para a proposta deste trabalho, que busca explorar o potencial e as limitações dos dados sintéticos de baixa fidelidade no treinamento de modelos de classificação de vagas. No próximo capítulo, será apresentado o estado da arte, contextualizando as contribuições recentes e os desafios atuais dessa área de pesquisa.

3 ESTADO DA ARTE

Uma das maiores limitações dos modelos de classificação baseados em aprendizado de máquina é lidar com a falta de dados. Uma forma de contornar o problema da falta de dados é a utilização de dados sintéticos (Ekbatani et al., 2017) (Tobin et al., 2017). Neste capítulo serão discutidos os métodos atuais de aprendizado de máquina para classificação de vagas de estacionamento, as bases de dados disponíveis e também como são gerados os dados sintéticos hoje em dia e quais os resultados obtidos.

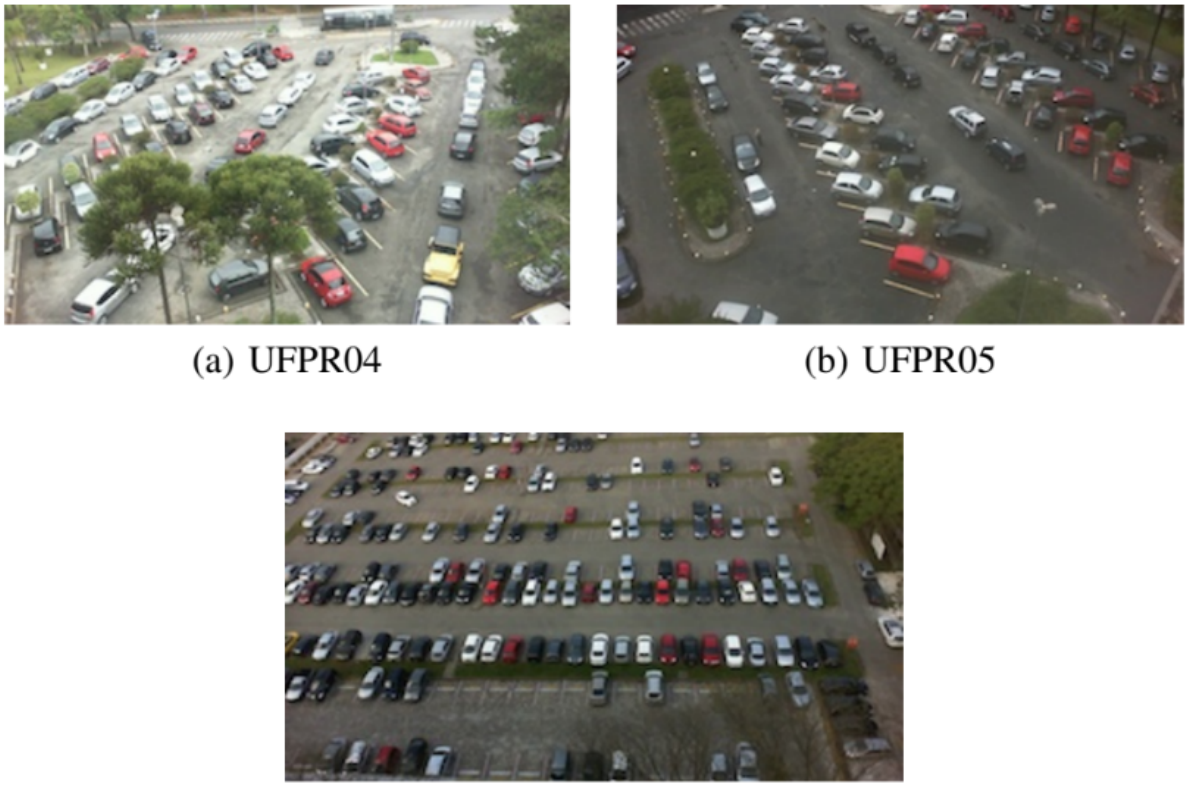
3.1 BASES DE DADOS

A base de dados PKLot (Almeida et al., 2015) é comumente usada para pesquisas em detecção e classificação de vagas de estacionamento por meio de visão computacional (de Almeida et al., 2023)(Hochuli et al., 2023)(de Almeida et al., 2022). Contém 12.417 imagens de tamanho 1280×720 capturadas de dois estacionamentos diferentes (UFPR e PUCPR) em dias ensolarados, nublados e chuvosos. O primeiro estacionamento possui dois ângulos de captura diferentes (UFPR04 e UFPR05). Cada imagem possui anotações das posições das vagas e cada vaga está associada a um rótulo indicando se está ocupada ou vazia. Usando essas anotações e segmentando as imagens, é possível ter cerca de 695.900 imagens de vagas de estacionamento nas mais diversas condições. Essa diversidade é valiosa para treinar e avaliar modelos de aprendizado de máquina e visão computacional, permitindo o desenvolvimento de sistemas robustos capazes de identificar automaticamente a ocupação de vagas em estacionamentos com base nas imagens fornecidas (Almeida et al., 2015)(de Almeida et al., 2022)(Hochuli et al., 2023).

Outra base de dados comumente usada em pesquisas é a CNRPark-EXT Amato et al. (2017). Especificamente, é uma expansão do conjunto de dados CNRPark, que foi desenvolvido para avaliar e treinar algoritmos de reconhecimento de ocupação de vagas em estacionamentos. É composta por 4.287 imagens capturadas por câmeras instaladas em 9 diferentes ambientes de estacionamento, cada imagem com anotações de segmentação das vagas com seus respectivos rótulos, entre ocupada ou livre, gerando cerca de 150.000 imagens rotuladas. As imagens apresentam variações em iluminação, condições climáticas e diferentes ângulos de visualização, tornando o conjunto de dados desafiador e representativo de situações do mundo real.

3.2 CLASSIFICAÇÃO DE VAGAS DE ESTACIONAMENTO

O trabalho de Almeida et al. (2015) propõe o uso de características LPQ e LBP extraídas das imagens como vetores de características e SVMs como classificadores. Conjuntos de SVMs treinados utilizando diversas variações dos métodos LPQ/LBP como características foram utilizados para classificação. Este método resultou em altas acurácias, contudo revelou limitações em termos de generalização. Em geral, ao treinar um classificador com um subconjunto de imagens de um determinado estacionamento e testá-lo com outro subconjunto do mesmo estacionamento, é alcançado consistentemente uma acurácia média de aproximadamente 99,5%. No entanto, no cenário de validação cruzada entre os subconjuntos, a acurácia média caiu para cerca de 85%. Essa redução na acurácia ao lidar com dados de estacionamentos distintos sugere que o método pode ser eficaz dentro do mesmo contexto de estacionamento, mas enfrenta dificuldades na generalização para diferentes ambientes ou cenários.



(a) UFPR04 (b) UFPR05

(c) PUCPR

Figura 3.1: A base de dados PKLot contém três cenários nomeados a) UFPR04, b) UFPR05 e c) PUCPR.

(a) (b)

Figura 3.2: A base de dados CNRPark-EXT abrange diferentes câmeras instaladas no mesmo ambiente de estacionamento.

Em Amato et al. (2017) além da criação da base de dados CNRPark-EXT, vemos o uso de modelos de aprendizado profundo. Os autores propuseram a mAlexNet, uma Rede Neural Convolucional (CNN) leve projetada para a classificação de vagas de estacionamento, e reportaram resultados entre 93% e 98% de acurácia em cenários de validação cruzada e no mesmo dataset, respectivamente.

No mesmo contexto, os autores de Grbić e Koch (2023); Dhuri et al. (2021); Nurullayev e Lee (2019); Hochuli et al. (2022, 2023); Alves et al. (2023) também propuseram abordagens baseadas em aprendizado profundo para resolver o problema de classificação de vagas de estacionamento. Em Nurullayev e Lee (2019), foi apresentada a rede CarNet, um método baseado em Redes Neurais Convolucionais Dilatadas (Dilated CNNs) que ignora certos pixels no núcleo de convolução. A *Dilated CNN* mostrou maior robustez e generalização em comparação com

abordagens anteriores, como mAlexNet (Amato et al., 2017), alcançando acurácias entre 94% e 98%.

Os autores de Dhuri et al. (2021) propuseram um sistema de detecção de ocupação de vagas de estacionamento em tempo real baseado na rede neural VGG16 e treinado com a CNRPark-EXT Amato et al. (2017) em conjunto com um dataset privado. O modelo alcançou uma acurácia média de 93,4%. Em Grbić e Koch (2023) foi utilizada a ResNet34 para a classificação das vagas, previamente localizadas e segmentadas em quadrados por um método que identifica veículos em uma série de imagens e aplica um algoritmo de agrupamento em uma visão aérea (*bird's-eye view*). O modelo foi extensivamente avaliado nos datasets públicos PKLot e CNRPark-EXT e alcançou acurácias entre 92% e 99%.

Em Hochuli et al. (2022) foi analisado o impacto do tipo de segmentação (Retângulos rotacionados, bounding-boxes e polígonos) das vagas nos resultados e também medir a quantidade de imagens necessárias para aperfeiçoar um modelo para cenários específicos. Para isso, foi definida uma CNN pré-treinada composta por 3 camadas convolucionais. Os experimentos revelaram acurácias superiores a 99% para os cenários específicos (modelo treinado e testado no mesmo subset), atingindo os melhores resultados para o tipo de segmentação com retângulos rotacionados. Além disso, os resultados mostraram que com apenas 1.000 imagens é possível realizar a afinação do modelo pré-treinado e obter resultados satisfatórios na mudança de cenários, atingindo em média 97% de acurácia.

Considerando o tempo e esforço necessários para a coleta de imagens, segmentação das vagas e anotação dos rótulos, em Hochuli et al. (2023) diversos modelos e técnicas são testados a fim de validar qual modelo se comporta melhor num cenário de validação cruzada. É então concluído que um modelo global com a arquitetura da MobileNetV3 (Howard et al., 2019) é adequado para a classificação de vagas de estacionamento em diferentes cenários e é capaz de atingir uma taxa de acerto de, em média, 95% nos cenários de validação cruzada, dispensando assim a necessidade de fine-tuning em cenários em que a coleta de imagens é restrita.

A Tabela 3.1 reúne uma síntese dos resultados obtidos nos trabalhos citados nessa seção, com a informação do modelo de aprendizado de máquina utilizado, a menor e a maior acurácia reportada nos diferentes cenários de teste dos trabalhos e as bases de dados utilizadas no trabalho, bem como a indicação de avaliação em cenário cruzado. É difícil comparar os resultados diretamente, tendo em vista que os autores utilizaram métodos e formas de validação diferentes.

Tabela 3.1: Sínteze dos resultados obtidos pelos trabalhos relacionados

Trabalho	Modelo utilizado	Faixa de acurácia	Tipo de Avaliação	Bases de Dados
Almeida et al. (2015)	SVM	84% a 99,5%	Mesma Base	PKLot
Amato et al. (2017)	mAlexNet	93% a 98%	Cenários Cruzados & Mesma Base	PKLot + CNRPark-EXT
Nurullayev e Lee (2019)	CarNet	94% a 98%	Cenários Cruzados & Mesma Base	PKLot + CNRPark-EXT
Dhuri et al. (2021)	VGG16	87% a 95%	Mesma Base	CNRPark + Base de dados privada
Hochuli et al. (2022)	CNN customizada	89% a 97%	Mesma Base	PKLot
Hochuli et al. (2023)	MobileNetV3	82% a 95%	Cenários Cruzados	PKLot + CNRPark-EXT
Grbić e Koch (2023)	ResNet34	92% a 99%	Cenários Cruzados & Mesma Base	PKLot + CNRPark-EXT

Mesmo com esses resultados, o problema de classificação de vagas de estacionamento ainda carece de mais imagens para validação e treinamento de modelos, devido às limitações na generalização dos modelos e coleta de dados para cenários específicos. Nesse contexto, técnicas de geração de imagens sintéticas podem ser usadas tanto no treinamento como na validação de modelos, para tentar melhorar a acurácia de um modelo para um cenário global ou específico.

3.3 GERAÇÃO DE DADOS SINTÉTICOS

Os dados sintéticos são gerados artificialmente e algoritmicamente, sendo empregados no treinamento de modelos de aprendizado de máquina para complementar conjuntos de dados existentes ou compensar a falta de dados reais. Eles desempenham um papel crucial ao aumentar a diversidade e a quantidade de amostras disponíveis, especialmente em cenários onde os conjuntos de dados são limitados ou insuficientes. Além disso, esses dados podem ser úteis na representação de situações raras, na preservação da privacidade dos dados reais, na redução de vieses nos conjuntos de dados e na criação de cenários de teste e validação para garantir a robustez dos modelos de aprendizado de máquina.

Diversas ferramentas podem ser utilizadas para a geração de dados sintéticos. Com relação a imagens sintéticas, as ferramentas mais utilizadas são motores gráficos de jogos como Unity e Unreal Engine (Tremblay et al., 2018; Tschentscher et al., 2017; Jaipuria et al., 2020; Reutov et al., 2022). Alguns pacotes personalizados que auxiliam na aleatorização, captura e rotulação dos dados, são disponibilizados pela comunidade acadêmica, como é o caso do *Unity Perception* (Borkman et al., 2017), para o Unity5, e o NDDS (To et al., 2018) para a Unreal Engine 4.

Uma das técnicas para geração de dados sintéticos é a *domain randomization* (Tobin et al., 2017), que envolve introduzir aleatoriedade deliberada nos ambientes de treinamento sintéticos usados para ensinar modelos de aprendizado de máquina, principalmente com imagens. Essa técnica visa criar uma variedade maior de cenários, ajustando aleatoriamente parâmetros como texturas, iluminação e formas geométricas. Ao variar aleatoriamente as características do ambiente virtual os modelos são expostos a uma ampla gama de condições durante o treinamento. Essa diversidade ajuda os modelos a se adaptarem a diferentes variações que podem ser encontradas no mundo real, capacitando-os a generalizar de forma mais eficaz para situações reais. Em essência, ao simular uma maior variedade de cenários durante o treinamento, os modelos se tornam mais robustos e capazes de lidar com a complexidade e as variações do mundo real, diminuindo a diferença entre os ambientes virtuais e o mundo real.

O estudo de Tobin et al. (2017) demonstrou que um detector de objetos treinado exclusivamente em simulação utilizando a técnica de *domain randomization* pode atingir uma precisão suficientemente alta no mundo real, possibilitando a realização de agarramentos em ambientes com obstáculos.

O estudo de Tobin et al. (2017) propõe a técnica de *domain randomization* como uma solução para reduzir o *Reality Gap* entre dados sintéticos e cenários reais. Os autores demonstraram que um modelo treinado exclusivamente com dados gerados em simulação foi capaz de alcançar alta acurácia ao identificar objetos específicos no mundo real em um ambientes com obstáculos, guiando um braço robótico para agarrar esses objetos. A técnica mostrou-se eficaz ao ampliar a capacidade de generalização dos modelos, permitindo sua aplicação prática em contextos reais sem a necessidade de ajuste fino com dados coletados do ambiente operacional. A Figura 3.3 mostra um exemplo de imagens geradas no trabalho.

Em Tremblay et al. (2018) foi demonstrado que a combinação de imagens sintéticas não fotorealistas (*Domain Randomization*) com imagens sintéticas fotorealistas, para o treinamento

Figura 3.3: Exemplo de imagens geradas sinteticamente utilizando a técnica de *domain randomization*, extraída do trabalho de Tobin et al. (2017).

de redes neurais, pode superar com sucesso o problema de *reality gap* para aplicações no mundo real, alcançando acurácias comparáveis com redes de última geração treinadas em dados reais. A combinação de imagens fotorealísticas e *domain randomization* no treinamento de modelos de aprendizado aumenta a capacidade do modelo de classificar corretamente porque atenua a diferença entre ambientes sintéticos e o mundo real. Enquanto as imagens fotorealísticas replicam detalhes visuais precisos do mundo real, o *domain randomization* introduz variações controladas ou aleatórias nos ambientes virtuais. Essa abordagem amplia a diversidade dos dados de treinamento, permitindo que os modelos se adaptem a uma variedade de condições presentes na prática, fortalecendo sua capacidade de generalização para situações reais ao minimizar a lacuna da realidade durante o treinamento.

No contexto de administração e classificação de vagas de estacionamento em tempo real, os autores de Tschentscher et al. (2017) apresentam um ambiente virtual de simulação 3D de estacionamento desenvolvido no Unreal Engine para avaliar sistemas de orientação de vagas baseados em vídeo. O ambiente permite simular condições variadas de clima, iluminação e obstruções, como veículos em movimento, proporcionando dados altamente personalizáveis para treinamento e validação de modelos. Além disso, uma câmera virtual foi projetada para gerar imagens realistas, incorporando restrições físicas como desfoque, ruído e profundidade de campo.

Em complemento, no trabalho Horn e Houben (2018) o sistema proposto em Tschentscher et al. (2017) foi utilizado para treinar classificadores como SVM e kNN com dados exclusivamente sintéticos. Duas sequencias de imagens reais, sequencia A e B, foram extraídas de uma base de dados privada e utilizadas para a validação dos modelos. Acurácias médias de 79,24% e 91,66% foram obtidas na validação com a sequencia A e B, respectivamente, demonstrando que tarefas como classificação de vagas de estacionamento podem ser resolvidas sem a necessidade de dados reais, embora a precisão seja ligeiramente inferior em relação a modelos treinados com dados

reais. Essa abordagem destaca o potencial dos ambientes simulados para reduzir custos e superar desafios de coleta de dados aplicados ao problema de classificação de vagas de estacionamento.

3.4 CONCLUSÃO

Este capítulo abordou as soluções atuais para a classificação de vagas de estacionamento utilizando aprendizado de máquina, destacando as principais bases de dados utilizadas, como PKLot e CNRPark-EXT, e os métodos mais comuns aplicados a essa tarefa. As redes neurais convolucionais (CNNs), utilizadas em Grbić e Koch (2023); Dhuri et al. (2021); Nurullayev e Lee (2019); Hochuli et al. (2022, 2023); Alves et al. (2023), demonstraram ser eficazes, com acurácias entre 93% e 99%, mas ainda enfrentam dificuldades em termos de generalização para cenários diversos e custo de coleta e segmentação de datos para o ajuste fino em cenários específicos. O uso de dados sintéticos, particularmente por meio de técnicas como *domain randomization*, tem sido explorado como uma solução para ampliar a diversidade dos dados, mas sua aplicação ainda apresenta limitações no que diz respeito à fidelidade dos dados gerados e ao *reality gap* entre os dados simulados e os reais.

Em Horn e Houben (2018) acurácias médias de 79,24% e 91,66% foram atingidas na classificação de vagas de um cenário específico utilizando de imagens sintéticas geradas a partir de um ambiente de simulação 3D altamente fiel ao cenário real, com condições de iluminação, câmera e clima realistas, replicando todos os pontos do cenário específico. Porém, construir esse ambiente simulado de alta fidelidade é mais trabalhoso que a coleta e rotulação de imagens reais de um cenário específico. Apesar dos avanços, apostar na combinação de dados reais e sintéticos parece ser a melhor abordagem para melhorar a acurácia dos modelos em diferentes cenários (Tremblay et al., 2018). O próximo capítulo irá apresentar uma proposta utilizando de dados sintéticos para superar essas limitações, focando no impacto das imagens sintéticas de baixa fidelidade na acurácia de modelos de classificação.

4 PROPOSTA

Como discutido na seção 3.2, os autores de Hochuli et al. (2023) demonstraram que é possível treinar um modelo de classificação de vagas de estacionamento generalista capaz de alcançar resultados de em média 95% de acurácia na validação cruzada entre as bases de dados, sem a necessidade de um ajuste fino para um cenário específico. Porém, resultados ainda melhores, em média 97% de acurácia, podem ser alcançados ao se realizar um ajuste fino em um modelo generalista com poucos dados rotulados de cenários específicos (Hochuli et al., 2022).

Ainda assim, a dificuldade e esforço em se obter e rotular dados persiste, sendo necessário realizar esse trabalho de coleta a cada novo cenário em que se desejar um modelo com a melhor acurácia possível. Uma forma de tentar contornar esse problema é a utilização de dados sintéticos no treinamento, mais especificamente imagens sintéticas, já que, além de ser possível gerar as mais diversas condições climáticas e de iluminação, milhões de imagens rotuladas podem ser geradas em minutos, podendo ser usadas para substituir imagens reais ou em conjunto com elas (Tremblay et al., 2018) (Tobin et al., 2017) (Hinterstoisser et al., 2019) (Ekbatani et al., 2017).

Na seção 3.3 foir discutido que em Horn e Houben (2018) acurácias médias de 79,24% e 91,66% foram atingidas na classificação de vagas de um cenário específico utilizando de imagens sintéticas geradas a partir de um ambiente de simulação 3D altamente fiel ao cenário real, com condições de iluminação, câmera e clima realistas, replicando todos os pontos do cenário específico. Porém, construir esse ambiente simulado de alta fidelidade é mais trabalhoso que a coleta e rotulação de imagens reais de um cenário específico, o que dificulta a aplicação desse método em larga escala. Também, não há ganhos na acurácia suficientes que justifiquem a utilização desse método em comparação a coleta e segmentação de imagens do cenário específico. Em Tobin et al. (2017), podemos ver que imagens sintéticas randomizadas e de baixa fidelidade à realidade podem alcançar resultados comparáveis a de imagens reais ao serem utilizadas em modelos de aprendizado profundo para reconhecimento de objetos, sendo necessário um esforço consideravelmente pequeno para serem geradas. Com isso, a proposta deste trabalho é utilizar deste método de geração de dados sintéticos para avaliar o impacto das imagens sintéticas nos modelos de classificação de estacionamento e responder as perguntas de pesquisa definidas na seção 1.

As próximas sub-seções fornecem uma descrição detalhada do processo de geração das imagens sintéticas e do treinamento e arquitetura dos modelos de classificação utilizados nesse trabalho.

4.1 GERAÇÃO DOS DADOS SINTÉTICOS

A ferramenta de simulação utilizada é o motor gráfico de jogos Unity5, em conjunto com o pacote Unity Perception (Borkman et al., 2017). Esse pacote oferece diversas ferramentas de aleatorização do ambiente e possibilita a construção de um ambiente de simulação tanto de alta como baixa fidelidade. As imagens são capturadas automaticamente e as informações de contexto para rotulação são geradas automaticamente por meio da visão de uma câmera que pode ser posicionada de qualquer maneira em qualquer angulação.

Para o ambiente simulado de baixa fidelidade neste trabalho, um script de aleatorização para as vagas do estacionamento e espaçamento entre elas foi criado e um conjunto de texturas para o chão e modelos 3D de carros e árvores foram selecionados. Primeiro seleciona-se a posição da câmera e o número de iterações da execução. Cada iteração gera uma imagem com

informações de contexto e as informações para segmentação das vagas e rotulação como ocupada ou livre são geradas automaticamente. O script segue os seguintes passos para cada iteração:

- Uma textura é escolhida e aplicada ao chão do ambiente.
- O script define as fileiras de vagas, suas posições, espaçamento entre as vagas e a rotação das mesmas. A cada 100 iterações, a posição de todas as vagas é rotacionada em 15 graus.
- Os carros são adicionados à cena. Cada vaga recebe ou não um carro, com uma probabilidade de 50% entre ocupada ou livre.
- A posição e quantidade de árvores é selecionada e as árvores são adicionadas à cena aleatoriamente.
- Ajusta-se a posição e intensidade da luz, mudando cor e clima do ambiente. Isso gera ambientes que simulam climas ensolarados, chuvosos e também possibilita capturas que simulam horários como tarde, noite e dia.
- A imagem é capturada pela câmera e as informações de contexto são registradas.

Foram geradas diversas bases de imagens, cada uma simulando a posição e parâmetros de câmera de cada sub-base disponível na CNRPark-EXT (Amato et al., 2017) e PKLot (Almeida et al., 2015). No total, 12 diferentes bases de imagens foram geradas. O número de iterações selecionado foi o de 1.500, cada iteração gerando uma imagem sintética, ou seja, cada base possui 1.500 imagens do estacionamento completo e em média 20.000 imagens das vagas segmentadas e rotuladas. As vagas são segmentadas no formato de retângulos rotacionados, pois é o formato que apresenta melhor acurácia na classificação de acordo com Hochuli et al. (2022). Um exemplo das vagas segmentadas está disponível na Figura 4.1. A Tabela 4.1 mostra a relação de quantidade de imagens segmentadas para cada base.

Figura 4.1: Exemplo das de vagas de estacionamento segmentadas em retângulos rotacionados. As vagas podem são rotuladas entre Ocupada ou Livre.

O programa open source Fspy (Stuffmatic, 2023) foi utilizado para coletar a posição e parâmetros da câmera de cada sub-base. A Figura 4.2 mostra uma comparação entre a primeira e a milésima iteração da base de dados com os parâmetros da UFPR04.

Tabela 4.1: Relação de quantidade de imagens sintéticas segmentadas

Câmera	Livre	Ocupada	Total
CNR-cam1	14.495	14.605	29.100
CNR-cam2	5.429	5.371	10.800
CNR-cam3	9.533	9.667	19.200
CNR-cam4	12.158	12.542	24.700
CNR-cam5	15.558	15.642	31.200
CNR-cam6	9.027	9.073	18.100
CNR-cam7	12.332	12.168	24.500
CNR-cam8	11.434	11.066	22.500
CNR-cam9	8.245	8.255	16.500
PKLot-UFPR04	18.713	18.787	37.500
PKLot-UFPR05	9.433	9.567	19.000
PKLot-PUCPR	35.994	36.106	72.100
Todas	162.351	162.849	325.200

Figura 4.2: Comparação entre a primeira e a milésima iteração da base de dados com os parâmetros de câmera simulando o subconjunto UFPR04

4.2 TREINAMENTO DOS MODELOS DE CLASSIFICAÇÃO

O processo de aprendizado por transferência foi a escolha para o treinamento dos modelos e realização dos experimentos. Esse processo utiliza de um modelo de classificação pré treinado em uma base de dados generalista, a camada de classificação é removida e então uma nova camada de classificação é incluída, então um ajuste fino é feito com as bases disponíveis. É um processo simples e reduz o custo computacional quando comparado a um treinamento completo, além de ter apresentado acurácias satisfatórias nas pesquisas recentes (Hochuli et al., 2023) (Alves et al., 2023) (de Almeida et al., 2022), e por isso foi o método de treinamento escolhido.

O modelo base utilizado neste trabalho é a MobileNetv3 (Howard et al., 2019) na sua versão *Large*, pré-treinada na base de dados ImageNet (Deng et al., 2009). Esse modelo foi escolhido por ser amplamente utilizado nesse tipo de exerimentos e pelo equilíbrio entre custo computacional e acurácia que essa arquitetura proporciona (de Almeida et al., 2022; Hochuli et al., 2023; Alves et al., 2023). A versão *Large* foi escolhida por experimentação, e apresentou os melhores resultados, sendo a acurácia próxima dos outros trabalhos que utilizam da MobileNetV3 para classificação de vagas de estacionamento (de Almeida et al., 2022; Hochuli et al., 2023; Alves et al., 2023). O formato de entrada das imagens é padronizado em 128x128 e uma camada de pré-processamento é incluída na entrada do modelo para fazer o redimensionamento das imagens. A camada de classificação (última camada) é trocada a fim de mudar o *output channel* para 2 classes e todas as camadas são descongeladas, com 4,6 milhões de parâmetros treináveis.

Foi utilizado o aumento dos dados (*data augmentation*) para aumentar a variabilidade dos dados de treinamento. As seguintes técnicas de aumento dos dados foram utilizadas:

- Rotação com taxa de 0.2.
- Direcionamento da imagem na horizontal.
- Alteração no contraste contraste com taxa de 0.4.

Cada modelo foi treinado com um *batch size* de tamanho 32 por 15 épocas, utilizando o otimizador Adam com uma taxa de aprendizado de 0.0001 e um *scheduler* para diminuir a taxa de aprendizado em 0.1 a cada 7 épocas. É escolhido o modelo que apresentou menor taxa de perda entre as épocas. Cada modelo foi treinado e validado 5 vezes, e a média aritimética das 5 execuções é o resultado final. Para a classificação, o limiar de decisão é calculado de acordo com o *Equal Error Rate* no espaço ROC. Esses parâmetros foram escolhidos por meio de experimentação e foram os que apresentaram melhores resultados, tendo como base de comparação a acurácia.

4.3 CONCLUSÃO

Neste capítulo, foi apresentada a proposta deste trabalho, que utiliza dados sintéticos de baixa fidelidade para treinar modelos de classificação de ocupação de vagas de estacionamento. A abordagem busca responder às perguntas de pesquisa propostas na seção 1, explorando o impacto das imagens sintéticas na generalização e acurácia de modelos de classificação, bem como sua viabilidade em superar o desempenho de modelos treinados exclusivamente com dados reais.

A geração de imagens sintéticas, detalhada ao longo do capítulo, foi realizada utilizando o pacote Unity Perception da ferramenta Unity5, pacote que permite a criação de cenários altamente diversificados com um esforço computacional relativamente baixo. Essa escolha possibilitou a simulação de cenários adaptados para os diferentes ângulos e condições presentes nas bases reais PKLot e CNRPark-EXT, utilizando a tecnica de *domain-randomization* para a variação de texturas e condições de iluminação das imagens sintéticas.

Para o treinamento dos modelos, foi adotado o processo de aprendizado por transferência com a MobileNetv3, devido ao seu equilíbrio entre eficiência computacional e acurácia. Os parâmetros de treinamento foram expecificados e ajustados com base em experimentação prévia, visando maximizar a acurácia e a consistência dos resultados.

A proposta apresentada representa um esforço para contornar os desafios associados à coleta e rotulação de dados reais, oferecendo uma alternativa mais escalável e adaptável por meio de dados sintéticos. Nos próximos capítulos, os resultados dos experimentos serão analisados para validar a eficácia desta abordagem, respondendo às perguntas de pesquisa e contribuindo para o avanço do uso de dados sintéticos em visão computacional aplicada.

5 EXPERIMENTOS

5.1 PROTOCOLO EXPERIMENTAL

Este protocolo foi desenvolvido para avaliar a acurácia de modelos de classificação de ocupação de vagas de estacionamento treinados em diferentes cenários de dados e o impacto das imagens sintéticas de baixa fidelidade na acurácia dos modelos, comparando cenários em que os modelos são treinados apenas com imagens reais, apenas com imagens sintéticas, ou com uma combinação de ambas. Foram utilizadas as bases de dados CNRPark-EXT (Amato et al. (2017) e PKLot (Almeida et al., 2015) como base de imagens reais. As imagens sintéticas foram geradas como definido na seção 4. Os seguintes cenários de treinamento foram definidos, utilizando aprendizado por transferência:

- M1 - Modelos treinados somente com imagens reais: Representa a configuração convencional e serve como base de comparação para os demais experimentos.
- M2 - Modelos treinados somente com imagens sintéticas: Este cenário visa avaliar o potencial das imagens sintéticas de baixa fidelidade para substituir completamente as imagens reais no treinamento.
- M3 - Imagens Reais + Todas as imagens sintéticas: Modelo treinado com uma combinação de imagens reais e todas as imagens sintéticas geradas espelhando a base de dados que será usada para teste. Este cenário explora a influência do uso de uma grande quantidade de imagens sintéticas para complementar o conjunto de dados real, potencialmente aumentando a variabilidade e robustez do modelo.
- M4 - Imagens Reais + Sintéticas (Câmera Específica): Modelo treinado com imagens reais, acrescidas de imagens sintéticas geradas especificamente para um cenário (câmera) de teste. Cada câmera da base de teste recebe seu próprio modelo treinado, visando uma maior especialização. Esse tipo de modelo possibilita a comparação com o método proposto em Hochuli et al. (2022), onde vê-se que adicionar poucos dados de um cenário específico pode aumentar consideravelmente a acurácia do modelo.

As bases de imagens reais (PKLot e CNRPark-EXT) foram ordenadas por dia de captura e divididas na proporção de 70% para treino e 30% para validação, a fim de evitar vieses. No caso dos modelos treinados exclusivamente com imagens sintéticas, as imagens foram randomicamente divididas em 70% para treino e 30% para validação. Nos cenários que combinam imagens reais e sintéticas, as imagens sintéticas foram somadas ao subconjunto de treino da base real (70%), seja usando todas as imagens (M3) ou apenas as específicas de um cenário (M4). Ao todo, foram treinados e testados 17 modelos nos diferentes cenários experimentais.

Todos os modelos foram treinados com a arquitetura MobileNetv3 (pré-treinada na ImageNet) para garantir comparabilidade e eficácia. As seguintes configurações de treinamento foram adotadas:

- Taxa de aprendizado: 0.0001, com redução de 0.1 a cada 7 épocas.
- Número de épocas: 15, com escolha do modelo de menor taxa de perna na validação.
- Tamanho do lote: 32.

A métrica de avaliação utilizada é a acurácia média, com ênfase na comparação entre cenários para entender o impacto das imagens sintéticas sobre o desempenho e a generalização dos modelos, em comparação com os modelos de imagens reais. O limiar de decisão foi calculado com base no *Equal Error Rate* (EER) no espaço ROC (Receiver Operating Characteristic).

5.2 RESULTADOS

Tendo como base o protocolo experimental descrito, as tabelas 5.1 e 5.2 apresentam as acurácias obtidas com os modelos M1, M2 e M3. As melhores acurácias foram alcançadas pelo modelo treinado apenas com imagens reais (M1), com acurácia média de 94,57% no teste com a base PKLot e 95,29% no teste com a base CNRPark-EXT. Em contraste, o modelo treinado apenas com imagens sintéticas (M2) obteve uma acurácia significativamente menor em ambos os cenários. Já o modelo onde o treinamento usou como base a soma das imagens reais e um conjunto geral das imagens sintéticas (M3) não apresentou uma mudança significativa na acurácia no treinamento com a PKLot, porém apresentou uma melhora significativa na acurácia no cenário de treinamento CNRPark-EXT, com aproximadamente 2% de aumento.

Tabela 5.1: Acurácias obtidas testando com a CNRPark-EXT os tipos de modelos M1, M2 e M3

Conjunto de Treino	Teste CNRPark-EXT
PKLot (M1)	95,29%
Apenas imagens sintéticas (M2)	85,96%
PKLot + Todas as Imagens Sintéticas CNRPark-EXT (M3)	95,05%

Tabela 5.2: Acurácias obtidas testando com a PKLot os tipos de modelos M1, M2 e M3

Conjunto de Treino	Teste PKLot
CNRPark-EXT (M1)	94,57%
Apenas imagens sintéticas (M2)	85,44%
CNRPark-EXT + Imagens Sintéticas PKLot (M3)	96,49%

As tabelas 5.3 e 5.4 apresentam os resultados dos modelos M4, que foram treinados com a combinação do conjunto de treino de imagens reais e imagens sintéticas criadas para simular um cenário específico de câmera de teste. Para cada conjunto de teste baseado em uma câmera específica, foi utilizado um modelo M4 separado, treinado apenas com imagens sintéticas geradas para aquela câmera. Na maioria dos casos, os resultados dos modelos M4 não mostraram uma melhoria significativa em comparação com os modelos M1 (treinados apenas com imagens reais). Contudo, observamos um aumento considerável na acurácia do modelo PKLot ao utilizar dados sintéticos simulando a câmera 9, além de uma melhora no teste do modelo CNRPark-EXT adicionado das imagens PUCPR sintéticas.

Tabela 5.3: Acurácias obtidas com os modelos treinados com imagens sintéticas de cenários específicos da CNRPark-EXT.

Conjunto de Treino	cam1	cam2	cam3	cam4	cam5	cam6	cam7	cam8	cam9	Média
PKLot (M1)	94,35%	97,17%	92,71%	96,51%	95,23%	94,64%	95,2%	97%	95,38%	95,39%
PKLot + Cenário Sintético (M4)	93%	97,53%	91,8%	96,08%	94,84%	94,34%	94,2%	96,8%	97,05%	95,07%

5.3 ANÁLISE DOS RESULTADOS

O modelo base e o método de aprendizado por transferência escolhido evidenciaram a boa capacidade de generalização do modelo e apresentaram resultados próximos e comparáveis

Tabela 5.4: Acurácias obtidas com os modelos treinados com imagens sintéticas de cenários específicos da PKLot.

Conjunto de Treino	UFPR04	UFPR05	PUCPR	Média
CNRPark-EXT (M1)	91,88%	95,76%	94,83%	94,16%
CNRPark-EXT + Cenário Sintético (M4)	90,98%	94,81%	98,05%	94,61%

a outros trabalhos no estado da arte (Hochuli et al., 2023) (Paidi et al., 2018) (Grbić e Koch, 2023) que tratam desse mesmo problema. Assim, podemos ter uma boa base de comparação na avaliação do desempenho das imagens sintéticas de baixa fidelidade ao se utilizar os resultados obtidos com tipos de modelos M1 treinados. Pode-se concluir então:

P1 - P1 - De que forma a combinação de imagens reais e sintéticas de baixa fidelidade afeta a generalização de modelos de classificação de vagas de estacionamento? O modelo do tipo M3 da PKLot, em comparação com o M1 não apresentou melhora, mantendo-se próximo do resultado alcançado com o modelo M1. Já o modelo do tipo M3 da CNRPark-EXT teve um acréscimo considerável na acurácia em relação ao seu respectivo M1. Esse comportamento na PKLot pode ser atribuído à grande quantidade de imagens reais na base de treinamento, que facilita a convergência do modelo para uma solução estável, minimizando o impacto das imagens sintéticas adicionais. No caso da CNRPark-EXT, com menos imagens reais, a adição das imagens sintéticas contribuiu de forma mais evidente para melhorar a acurácia e a capacidade de generalização do modelo.

P2 - Como imagens sintéticas de baixa fidelidade se comportam na aplicação e especificação de modelos de classificação de vagas de estacionamento para um cenário alvo? Os resultados com os modelos do tipo M4 evidenciam que imagens sintéticas de baixa fidelidade talvez não sejam a melhor escolha para se especificar um modelo genérico para um cenário específico, tendo acurácia média inferior na maioria dos testes e uma melhora somente em 3 dos 12 casos. Comparando com os experimentos em Hochuli et al. (2022), o uso de imagens reais é uma melhor opção para a especificação, por mais que demandem mais trabalho para serem coletadas do que as sintéticas de baixa fidelidade para serem geradas.

Pode-se então concluir que o impacto das imagens sintéticas de baixa fidelidade na eficiência dos modelos de classificação de vagas de estacionamento é baixo, não sendo uma boa escolha para a tarefa de especificação de um modelo para um cenário específico e melhorando a generalização de um modelo somente quando a quantidade de imagens reais disponíveis para treinamento é relativamente pequena à quantidade de imagens sintéticas. Nesses casos a adição das imagens sintéticas proporciona maior variedade de características para treinamento, o que auxilia o modelo na classificação.

P3 - É possível superar a acurácia dos modelos treinados com imagens reais treinando os modelos somente com imagens sintéticas de baixa fidelidade? Não. As acurácias obtidas com os modelos de tipo M2 foram consideravelmente inferiores às obtidas com os modelos de tipo M1. Isso pode ter acontecido pois as imagens sintéticas de baixa fidelidade não apresentam características realistas o suficiente para se resolver o problema de classificação de vagas de estacionamento, como não reproduzir fielmente a qualidade da imagem e detalhes das condições climáticas e de iluminação. Talvez a melhor abordagem para a criação de um modelo somente com imagens sintéticas que possa competir com o estado da arte seja a descrita em Tremblay et al. (2018), que consiste em misturar imagens sintéticas altamente randomizadas com imagens sintéticas fotorealísticas.

Deve-se considerar as limitações do protocolo desenvolvido para geração das imagens sintéticas e também da qualidade das imagens geradas. A falta de variações climáticas, condições de iluminação texturas e contexto nas imagens certamente impactam o experimento. Apesar do

protocolo gerado ter sido pensado em diminuir o esforço humano na tarefa de coletar e gerar imagens de vagas de estacionamento, possivelmente um cenário mais realista pode melhorar os modelos tanto para o cenário generalista quanto para o ajustado a cenários.

5.4 CONCLUSÃO

Os experimentos mostraram que as imagens sintéticas de baixa fidelidade podem complementar bases de dados reais, especialmente em cenários com dados limitados, como na CNRPark-EXT, onde foi observada uma melhora na acurácia dos modelos M3. Contudo, em bases maiores, como a PKLot, as imagens sintéticas tiveram impacto reduzido, indicando que sua contribuição diminui quando há abundância de dados reais.

Para os modelos M4, os resultados foram em sua maioria inferiores aos dos modelos M1, sugerindo que as imagens sintéticas de baixa fidelidade são pouco eficazes para especialização dos modelos para cenários específicos. Além disso, os modelos M2, treinados exclusivamente com imagens sintéticas, não alcançaram acurácias competitivas, evidenciando as limitações na fidelidade e variabilidade dos dados gerados.

Pode-se concluir que, embora úteis em contextos específicos, as imagens sintéticas de baixa fidelidade não substituem os dados reais e são mais adequadas para ampliar a generalização quando há poucos dados disponíveis. Trabalhos futuros podem explorar métodos de geração mais realistas para aumentar seu impacto em aplicações práticas.

6 CONCLUSÃO

Este trabalho investigou o impacto do uso de imagens sintéticas de baixa fidelidade no treinamento de modelos de classificação de ocupação de vagas de estacionamento, explorando sua viabilidade tanto como complemento quanto como substituto de dados reais. A proposta abordou duas perguntas principais: o impacto das imagens sintéticas na acurácia dos modelos e a possibilidade de superação do desempenho de modelos treinados exclusivamente com dados reais.

Os experimentos demonstraram que as imagens sintéticas de baixa fidelidade são capazes de complementar bases reais em contextos com dados limitados, como evidenciado na base CNRPark-EXT, onde o modelo M3 apresentou aumento de acurácia em comparação ao modelo M1. No entanto, sua eficácia foi reduzida em cenários com maior disponibilidade de dados reais, como na base PKLot. Além disso, os modelos M4, que utilizaram dados sintéticos para especialização em cenários específicos, apresentaram desempenho inferior em relação aos modelos treinados apenas com imagens reais, reforçando que as imagens sintéticas de baixa fidelidade são mais adequadas para generalização do que para especialização.

Por outro lado, os modelos M2, treinados exclusivamente com imagens sintéticas, apresentaram acurácias consideravelmente inferiores aos modelos M1. Isso indica que, apesar do potencial de redução de custos e esforço na geração de dados, a baixa fidelidade dos dados sintéticos limita sua capacidade de representar adequadamente as condições do mundo real, como variações climáticas, texturas e iluminação.

Com base nos resultados obtidos, conclui-se que as imagens sintéticas de baixa fidelidade têm um papel importante como complemento em cenários com dados escassos, mas não são suficientes para substituir os dados reais. Trabalhos futuros podem explorar abordagens mais avançadas para geração de dados, como a combinação de imagens fotorealísticas e altamente randomizadas, além de métodos de domain randomization aprimorados.

Este estudo contribui para a área de visão computacional ao destacar os limites e as possibilidades do uso de dados sintéticos na classificação de vagas de estacionamento, abrindo caminho para novas investigações sobre a utilização de técnicas de geração mais sofisticadas para superar os desafios identificados.

REFERÊNCIAS

Almeida, P. R., Oliveira, L. S., Jr., A. S. B., Jr., E. J. S. e Koerich, A. L. (2015). Pklot - a robust dataset for parking lot classification. *Expert Systems with Applications*, 42(11):4937–4949.

Almeida, P. R., Oliveira, L. S., Silva, E., Britto, A. e Koerich, A. (2013). Parking space detection using textural descriptors. Em *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK.

Alves, P. L., Hochuli, A. G., de Oliveira, L. E. e de Almeida, P. R. L. (2023). Optimizing parking space classification: Distilling ensembles into lightweight classifiers. Em *International Conference on Machine Learning and Applications (ICMLA)*, páginas 1379–1384, Florida, USA.

Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C. e Vairo, C. (2017). Deep learning for decentralized parking lot occupancy detection. *Expert Systems with Applications*, 72:327–334.

Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., Romero, C., Smith, W., Thaman, A., Warren, S. e Yadav, N. (2017). Unity perception: Generate synthetic data for computer vision. Em *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, BC, Canada.

de Almeida, P. R. L., Alves, J. H., Oliveira, L. S., Hochuli, A. G., Frohlich, J. V. e Krauel, R. A. (2023). Vehicle occurrence-based parking space detection. Em *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oahu - Hawaii, USA.

de Almeida, P. R. L., Alves, J. H., Parpinelli, R. S. e Barddal, J. P. (2022). A systematic review on computer vision-based parking lot management applied on public datasets. *Expert Systems with Applications*, 198:116731.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. e Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. Em *2009 IEEE Conference on Computer Vision and Pattern Recognition*, páginas 248–255.

Dhuri, V., Khan, A., Kamtekar, Y., Patel, D. e Jaiswal, I. (2021). Real-time parking lot occupancy detection system with vgg16 deep neural network using decentralized processing for public, private parking facilities. *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, páginas 1–8.

Ekbatani, H. K., Pujol, O. e Segui, S. (2017). Synthetic data generation for deep learning in counting pedestrians. Em *International Conference on Pattern Recognition Applications and Methods (ICPRAM*, páginas 318–323, Porto, Portugal.

Grbić, R. e Koch, B. (2023). Automatic vision-based parking slot detection and occupancy classification. *Expert Systems with Applications*, 225:120147.

He, K., Zhang, X., Ren, S. e Sun, J. (2015). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 770–778.

Henry, C., Ahn, S. e Lee, S.-W. (2020). Multinational license plate recognition using generalized character sequence detection. *IEEE Access*, PP:1–1.

Hinterstoisser, S., Pauly, O., Heibel, H., Marek, M. e Bokeloh, M. (2019). An annotation saved is an annotation earned: Using fully synthetic training for object instance detection. *CoRR*, abs/1902.09967.

Hochuli, A. G., Barddal, A. P., de Almeida, P. R. L., Palhano, G. C. e Mendes, L. M. (2023). Deep single models vs. ensembles: Insights for a fast deployment of parking monitoring systems. Em *International Conference on Machine Learning and Applications (ICMLA)*, páginas 1379–1384, Florida, USA.

Hochuli, A. G., Jr, A. S. B., de Almeida, P. R. L. e Williams B. S. Alves, F. M. C. C. (2022). Evaluation of different annotation strategies for deployment of parking spaces classification systems. Em *International Joint Conference oN Neural Networks (IJCNN)*, páginas 1–8, Padua, Italy.

Horn, D. e Houben, S. (2018). Evaluation of synthetic video data in machine learning approaches for parking space classification. Em *2018 IEEE Intelligent Vehicles Symposium (IV)*, páginas 2157–2162, Changshu, China.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V. e Adam, H. (2019). Searching for mobilenetv3.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. e Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Jaipuria, N., Zhang, X., Bhasin, R., Arafa, M., Chakravarty, P., Shrivastava, S., Manglani, S. e Murali, V. N. (2020). Deflating dataset bias using synthetic data augmentation. Em *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Kolosov, D., Kelefouras, V., Kourtessis, P. e Mporas, I. (2022). Anatomy of deep learning image classification and object detection on commercial edge devices: A case study on face mask detection. *IEEE Access*, 10:109167–109186.

Krizhevsky, A., Sutskever, I. e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90.

Nurullayev, S. e Lee, S.-W. (2019). Generalized parking occupancy analysis based on dilated convolutional neural network. *Sensors*, 19(2).

Paidi, V., Fleyeh, H., Håkansson, J. e Nyberg, R. G. (2018). Smart parking sensors, technologies and applications for open parking lots: a review. *IET Intelligent Transport Systems*, 12(8):735–741.

Phung e Rhee (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9:4500.

Reutov, I., Moskvin, D., Voronova, A. e Venediktov, M. (2022). Generating synthetic data to solve industrial control problems by modeling a belt conveyor. *Procedia Computer Science*, 212:264–274. 11th International Young Scientist Conference on Computational Science.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. e Chen, L.-C. (2019). Mobilenetv2: Inverted residuals and linear bottlenecks.

Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Stuffmatic (2023). fspy - open source still image camera matching. `https://fspy.io/`. Acessado em 12/12/2023.

Suhao, L., Jinzhao, L., Guoquan, L., Tong, B., Huiqian, W. e Yu, P. (2018). Vehicle type detection based on deep learning in traffic scene. *Procedia Computer Science*, 131:564–572. Recent Advancement in Information and Communication Technology:.

To, T., Tremblay, J., McKay, D., Yamaguchi, Y., Leung, K., Balanon, A., Cheng, J., Hodge, W. e Birchfield, S. (2018). NDDS: NVIDIA deep learning dataset synthesizer. `https://github.com/NVIDIA/Dataset_Synthesizer` .

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. e Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world.

Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D. e Birchfield, S. (2018). Deep object pose estimation for semantic robotic grasping of household objects. Em *Conference on Robot Learning (CoRL)*, páginas 307–316, Zurich - Switzerland.

Tschentscher, M., Pruß, B. e Horn, D. (2017). A simulated car-park environment for the evaluation of video-based on-site parking guidance systems. Em *2017 IEEE Intelligent Vehicles Symposium (IV)*, páginas 1571–1576.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. e He, Q. (2020). A comprehensive survey on transfer learning.