



ANTONIO GAUDENCIO GUIMARÃES FILHO

Deep Convolutional Neural Network (DCNN) para
Extração automática de rede viária a partir da FUSÃO de
dados varredura LASER aerotransportado e imagens de
altíssima resolução em ambiente urbano



ANTONIO GAUDENCIO GUIMARÃES FILHO

***Deep Convolutional Neural Network (DCNN)* para
Extração automática de rede viária a partir da FUSÃO de
dados varredura LASER aerotransportado e imagens de
altíssima resolução em ambiente urbano**

Tese apresentada ao Programa de Pós-Graduação em Ciências Cartográficas da Universidade Estadual Paulista – Campus de Presidente Prudente, como parte dos requisitos para obtenção do título de doutor em Ciências Cartográficas.

Orientador: Prof. Dr. Aluir Porfírio Dal Poz

Co-orientador: Prof. Dr. Milton Hirokazu Shimabukuro

Presidente Prudente

2022

G963d	Guimarães Filho, Antonio Gaudencio Deep Convolutional Neural Network (DCNN) para extração automática de rede viária a partir da FUSÃO de dados varredura LASER aerotransportado e imagens de altíssima resolução em ambiente urbano / Antonio Gaudencio Guimarães Filho. -- Presidente Prudente, 2022 119 f. Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente Orientador: Aluir Porfírio Dal Poz Coorientador: Milton Hirokazu Shimabukuro 1. Extração de rede viária. 2. Fusão de dados em Aprendizado Profundo. 3. Fusão Multimodal. 4. Segmentação de imagens. 5. área urbana. I. Título.
-------	---

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: Deep Convolutional Neural Network (DCNN) para extração automática de rede viária a partir da FUSÃO de dados varredura laser aerotransportado e imagens de altíssima resolução em ambiente urbano

AUTOR: ANTONIO GAUDENCIO GUIMARÃES FILHO

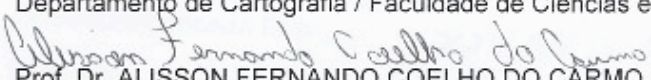
ORIENTADOR: ALUIR PORFIRIO DAL POZ

COORIENTADOR: MILTON HIROKAZU SHIMABUKURO

Aprovado como parte das exigências para obtenção do Título de Doutor em CIÊNCIAS CARTOGRÁFICAS, área: Aquisição, Análise e Representação de Informações Espaciais pela Comissão Examinadora:


Prof. Dr. ALUIR PORFIRIO DAL POZ (Participação Presencial)
Departamento de Cartografia / Faculdade de Ciências e Tecnologia de Presidente Prudente

Prof. Dr. NILTON NOBUHIRO IMAI (Participação Virtual)
Departamento de Cartografia / Faculdade de Ciências e Tecnologia de Presidente Prudente


Prof. Dr. ALISSON FERNANDO COELHO DO CARMO (Participação Presencial)
Inspectral


Prof. Dr. EDSON APARECIDO MITISHITA (Participação Presencial)
Departamento de Geomática / Universidade Federal do Paraná

Prof. Dr. RAUL QUEIROZ FEITOSA (Participação Virtual)
Departamento de Engenharia Elétrica / Pontifício Universidade Católica / Rio de Janeiro

Presidente Prudente, 26 de agosto de 2022

À minha querida avó Alice (in memoriam)

Por sua força, coragem, determinação, sabedoria, alegria e amor que demonstrou durante toda sua vida.

À minha esposa e ao meu filho

Que sofreram com os momentos de minha ausência mas que sempre me apoiaram e acreditaram que esse dia de conquista chegaria.

Agradecimentos

À Deus por guiar-me e sustentar-me durante todo o caminho da pesquisa.

À minha esposa e ao meu filho por todo o apoio, amor, paciência e dedicação durante esse período desafiador.

Aos meus pais e irmãos por todo amor, carinho e por fornecerem a base de sustentação do meu caráter.

Aos meus sogros por todo o carinho e alegria que sempre compartilham comigo.

Ao meu orientador Prof. Dr. Aluir Porfírio Dal Poz, pela confiança, sabedoria, pelos ensinamentos e pela atenção em todos os momentos.

Ao meu co-orientador Prof. Dr. Milton Hirokazu Shimabukuro, pela confiança desde o início da pesquisa e pelos ensinamentos.

À FAPESP pelo auxílio à pesquisa (processo nº 2021/03586-2, Fundação de Amparo à Pesquisa do Estado de São Paulo). Apoio fundamental para a realização da Tese.

À Companhia Imobiliária de Brasília (Terracap) e à empresa Topocart pela cessão dos dados utilizados na pesquisa.

Aos professores do Departamento de Cartografia e do PPGCC, que contribuíram para a minha formação acadêmica. Em especial ao Prof. Dr. Tommaselli pelo acolhimento e respeito desde a chegada ao Programa.

Ao Exército Brasileiro, à Diretoria de Serviço Geográfico (DSG) e ao 37º Batalhão de Infantaria Mecanizado (37º BI Mec) pela confiança depositada e suporte para a realização do curso.

Aos meus colegas de turma do PPGCC pelas discussões, apoio e incentivos. Em especial Fernando, Bruno, Michele, Érika e Marcela que me acolheram com muito carinho no programa.

Aos meus colegas Pacheco, Thiago e Thaisa do Laboratório de Integração de Sensores pelo apoio e suporte no início desafiador da pesquisa com aprendizado de máquina.

Ao Prof. Dr. Cassio Oishi do Departamento de Matemática e Ciência da Computação pelas discussões sobre *Deep Learning* que muito contribuíram para meu conhecimento.

Ao amigo Fernando do Laboratório *Cluster* pelo incansável apoio na manutenção, instalação e implementação do suporte de hardware/software da pesquisa.

Aos meus amigos de caserna Philipe Borba e Daniel Andrade pelas discussões, conversas, piadas, reflexões e parceria em todos os momentos.

Ao Dr. José Eduardo, ao Personal Trainer Mihagui e à Prof^{ra}. Pilates Cláudia Paccini por me ajudarem a voltar a andar. E às senhoras Helena e Rose que sempre me incentivaram e apoiaram nas sessões de Pilates.

Confie no Senhor de todo o seu coração e não se apoie em seu próprio entendimento; reconheça o Senhor em todos os seus caminhos, e ele endireitará as suas veredas.

Provérbios 3:5,6 - NVI

Resumo

A extração de rodovias é um importante tópico de pesquisa, dada a importância para a gestão de tráfego, desenho urbano, monitoramento de estradas, navegação por satélites e atualização de mapas. O principal problema na extração de vias, em Sensoriamento Remoto, é como descrever as suas características. Algoritmos de classificação de aprendizado profundo possuem a capacidade de extrair características a partir dos dados e não dependem de intervenção humana para construir regras. Atualmente, as Redes Neurais Convolucionais Profundas são o estado da arte da segmentação de rede viária e destacam-se pela capacidade de explorar informação de contexto multinível. Entretanto, com o aumento da resolução espacial, maior detalhamento das vias bem como mais interferências por prédios, sombras, obstruções (p.ex: carros, árvores) que ainda limitam a obtenção da rede viária em áreas urbanas. Uma das estratégias disponíveis é a utilização da fusão de dados de outros sensores/fontes para complementar a informação de extração de objetos. A metodologia proposta utilizou a arquitetura Unet adaptada para fusão de Ortoimagens de altíssima resolução espacial com dados de varredura LASER (LiDAR) do estado do Distrito Federal (DF). Duas abordagens de fusão foram testadas: inicial e multinível. Bem como dois tipos de entrada de dados: imagens intensidade e de modelos de superfície. A fusão inicial com a imagem intensidade alcançou o percentual de 74,92 IoU e 84,68 F1 no conjunto teste e demonstrou melhoria de 2,1% IoU e 1,2% F1 em relação ao modelo treinado sem a fusão na área urbana. Assim como, a fusão inicial com imagem de modelos acrescentou 2,3% IoU e 1,5% F1 em região com menor densidade de vias e menos oclusões. Os resultados demonstraram que a fusão aumenta a ativação da classe via e de outras classes diretamente relacionadas com oclusões, reduz a busca de contexto para inferência da rede em áreas oclusas e aumenta a precisão da segmentação. Implementação e resultados disponíveis em <https://github.com/tunofilho/thesis2022_data>.

Palavras-chave: Extração de rede viária. Fusão de dados em Aprendizado Profundo. Fusão Multimodal. Segmentação de imagens. área urbana.

Abstract

Road extraction is an important research topic, mainly traffic management, urban design, surveillance, satellite navigation, and map update. The main problem with road network extraction in Remote Sensing is describing its attributes. Deep learning algorithms can extract features from data without human intervention. Nowadays, deep convolution neural networks are state of the art in road network segmentation and are known for their remarkable ability to explore multi-level contexts. Nevertheless, with increased spatial resolution, the roads become more detailed, and the interferences (building, shadows, obstructions - e.g., cars, trees) still make extraction in urban areas challenging. One of the available strategies is data fusion from other sensors/sources to add information in object extraction. The proposed methodology uses the adapted Unet in the fusion of Highest resolution Orthoimages and LASER scanner data from the Federal District state (DF). Two fusion approaches were tested: early and multi-level. And two input types: intensity image and surface models. The early fusion with intensity image reaches 74.92 IoU and 84.68 F1 in test data, an improvement of 2.1% IoU and 1.2% F1 over the model without fusion in urban areas. Moreover, the early fusion with surface models increases 2.3% IoU and 1.5% in the region with few roads density and rare occlusions. The results showed that the fusion increases activation of road and other classes directly involved with occlusion, decreases the context dependence in occlusion areas, and enhances segmentation precision. Implementation and results are available at <https://github.com/tunofilho/thesis2022_data>.

Keywords: Road network extraction. Deep learning data fusion. Multi-modal fusion. Image segmentation. Urban area.

Lista de ilustrações

Figura 1 – Modelo Genérico de Rodovias	19
Figura 2 – Exemplos de reconhecimento de imagem, segmentação semântica, por instância e detecção de objetos	24
Figura 3 – Diferença entre a classificação CNN por fragmento e segmentação se- mântica.	24
Figura 4 – Linha do Tempo de ML	28
Figura 5 – Modelo de Neurônio Artificial - ANN	30
Figura 6 – Etapas do procedimento de BP de erro	31
Figura 7 – Arquitetura CNN - representação por unidade e imagens	32
Figura 8 – Funções de Ativação e Derivadas	33
Figura 9 – Modelo de Rede Neural com <i>Dropout</i>	37
Figura 10 – Esquema de otimização da Rede Neural com GC	38
Figura 11 – Representação simplificada da arquitetura VGG-16	39
Figura 12 – Bloco de aprendizagem residual	40
Figura 13 – Representação de <i>Transfer Learning</i> e <i>Fine Tuning</i>	42
Figura 14 – <i>Fully Convolutional Neural Network</i> (FCN)	44
Figura 15 – Arquitetura U-Net	45
Figura 16 – Níveis de processamento da fusão de imagens	48
Figura 17 – Padrões de amostragem dos Mecanismos de Varredura	50
Figura 18 – Principais Componentes do SVLA	51
Figura 19 – Casos de múltiplos retornos	52
Figura 20 – Componentes de SVLA e suas relações geométricas	53
Figura 21 – Diferentes arquiteturas de Fusão CNN para segmentação de imagens RGB-D	57
Figura 22 – Arquitetura FuseNet	58
Figura 23 – Estratégias de Fusão na FuseNet	58
Figura 24 – Comparação de mapas de característica de cada ramo na primeira camada convolucional	59
Figura 25 – Regiões urbanas pertencentes ao Distrito Federal	60
Figura 26 – Conjunto de dados ISPRS Potsdam	63
Figura 27 – Máscara binária da classe Superfícies Impermeáveis	64
Figura 28 – Problema máscara no conjunto de dados Potsdam	64
Figura 29 – Erros nas máscara de vias disponibilizadas por Kaiser et al. (2017) . .	65
Figura 30 – Exemplo de problemas existentes no conjunto de dados de Massachusetts	66
Figura 31 – Imagem de modelos MDT,MDS,MDSn	67
Figura 32 – Fluxograma de Geração das Imagens Intensidade	68

Figura 33 – Gráfico de dispersão de índices Diversidade x Densidade de vias	70
Figura 34 – Metodologia geral da pesquisa	71
Figura 35 – Arquiteturas UNet e ResUnet adaptadas	72
Figura 36 – Abordagens para fusão CNN	74
Figura 37 – Comparação das classes extraídas (predição) com os dados de teste . .	75
Figura 38 – Comparativo resultado modelo ResUnet_bce+dice com dados RGB x RGIR	78
Figura 39 – Influência canal IR na predição	79
Figura 40 – Gráfico <i>Precision x Recall</i> das predições dos modelos conjunto Massa- chusetts	81
Figura 41 – Resultados da predição imagens conjunto Massachusetts	83
Figura 42 – Casos de oclusão copa de árvores	84
Figura 43 – Exemplos de inferência de vias no conjunto Massachusetts	84
Figura 44 – Interpretação <i>encoder</i> da rede Unet da imagem 1	85
Figura 45 – Interpretação <i>decoder</i> da rede Unet da imagem 1	86
Figura 46 – Interpretação <i>encoder</i> da rede Unet da imagem 2	87
Figura 47 – Predições da rede ResUnet na área 10 - sem fusão	90
Figura 48 – Predições da rede Unet na área 1 - sem fusão	91
Figura 49 – Exemplos de Inferência de vias no conjunto DF - área 10	92
Figura 50 – Mapas de características extraídos do <i>encoder</i> do recorte 1	92
Figura 51 – Mapas de características extraídos do <i>encoder</i> do recorte 2	93
Figura 52 – Mapas de características extraídos do <i>decoder</i> do recorte 2	94
Figura 53 – Gráfico resultados das métricas dos modelos sem fusão e fusionados - Área 1	96
Figura 54 – Mapas de características extraídos do <i>encoder</i>	97
Figura 55 – Gráfico resultados métricas dos modelos sem fusão e fusionados - Área 10	99
Figura 56 – Mapas de características extraídos do <i>encoder</i> fusionado do recorte 1 .	100
Figura 57 – Mapas de características extraídos do <i>encoder</i> fusionado do recorte 2 .	100
Figura 58 – Comparativo Mapas de características extraídos do <i>encoder</i> fusionado x não fusionado do recorte 1	101
Figura 59 – Comparativo Predições das redes fusionadas x não fusionadas - área 10	102
Figura 60 – Comparativo Predições das redes fusionadas x não fusionadas - área 1 .	102
Figura 61 – Gráfico comparativo dos diferentes modelos na Área 10	104

Lista de tabelas

Tabela 1 – Arquiteturas criadas para classificadores ResNet.	41
Tabela 2 – Comparativo LiDAR x Sensor Passivo	54
Tabela 3 – Principais características do levantamento aéreo fotogramétrico	61
Tabela 4 – Especificações Técnicas dos conjuntos de dados.	61
Tabela 5 – Principais softwares/bibliotecas para auxílio à pesquisa.	62
Tabela 6 – Parâmetros de pré-processamento imagens intensidade	68
Tabela 7 – Resultado no conjunto teste de ISPRS Potsdam	77
Tabela 8 – Comparativo dos principais modelos (%) no conjunto teste de Massa- chusetts	81
Tabela 9 – Resultado nos conjuntos testes das Área 10 e 1 - sem fusão	88
Tabela 10 – Comparativo resultados da rede ResUnet no conjunto teste da Área 01 - DF fusão com Imagem Intensidade e com Modelos	95
Tabela 11 – Comparativo resultados Unet no conjunto teste da Área 01 - DF fusão com Imagem Intensidade e com Modelos	95
Tabela 12 – Comparativo resultados da rede ResUnet no conjunto teste da Área 10 - DF fusão com Imagem Intensidade e com Modelos	97
Tabela 13 – Comparativo resultados da rede Unet no conjunto teste da Área 10 - DF fusão com Imagem Intensidade e com Modelos	98

Lista de abreviaturas e siglas

ADAM	<i>Adaptive Moment estimation</i>
ALS	<i>Airborne LASER Scanning</i>
ANN	<i>Artificial Neural Network</i>
ASPP	<i>Atrous Spatial Pyramid Pooling</i>
ASPRS	<i>American Society for Photogrammetry and Remote Sensing</i>
BCE	<i>Binary cross entropy</i>
BN	<i>Batch Normalization</i>
BP	<i>Backpropagation</i>
CE	<i>Cross entropy</i>
CLAHE	<i>Contrast Limited Adaptive Histogram Equalization</i>
CNN	<i>Convolutional Neural Network</i>
CONCAR	Comissão Nacional de Cartografia
DCNN	<i>Deep Convolutional Neural Network</i>
DF	Distrito Federal
DL	<i>Deep Learning</i>
ET-EDGV	Especificação Técnica de Estruturação de Dados Geoespaciais Vetoriais
FCN	<i>Fully Convolutional Network</i>
FL	<i>Focal Loss</i>
FN	<i>False Negative</i>
FOV	<i>Field of View</i>
FP	<i>False Positive</i>
GAN	<i>Generative Adversarial Network</i>
GC	<i>Gradiente Centralization</i>
GCS	<i>Ground Coordinate System</i>

GD	<i>Gradient Descent</i>
GNSS	<i>Global Navigation Satellite System</i>
GPU	<i>Graphics Processing Unit</i>
GRSS	<i>Geoscience and Remote Sensing Society</i>
GSD	<i>Ground Sample Distance</i>
IADF TC	<i>Image Analysis and Data Fusion Technical Committee</i>
IDW	<i>Inverse Distance Weight</i>
IEEE	<i>Institute of Electrical and Electronic Engineers</i>
IFOV	<i>Istantaneous Field of View</i>
ILSVRC	<i>ImageNet Large-Scale Visual Recognition Challenge</i>
IMU	<i>Inertial Measurement Unit</i>
INDE	Infraestrutura Nacional de Dados Espaciais
IoU	<i>Intersection over Union</i>
IR	<i>Infra red</i>
ISPRS	<i>International Society of Photogrammetry and Remote Sensing</i>
LASER	<i>Light Amplification by Stimulated Emission of Radiation</i>
LADAR	<i>LASER Detection And Ranging</i>
LB	<i>LASER Beam</i>
LiDAR	<i>Light Detection And Ranging</i>
LU	<i>LASER Unit</i>
MDS	Modelo Digital de Superfície
MDSn	Modelo Digital de Superfície normalizado
MDT	Modelo Digital do Terreno
ML	<i>Machine Learning</i>
MLP	<i>Multilayer Perceptron</i>
NAS	<i>Neural Architecture Search</i>
NIR	<i>Near infrared</i>

OSM	<i>OpenStreetMap</i>
PA	<i>Pixel Accuracy</i>
ResNet	<i>deep Residual Network</i>
RGB	<i>Red, Green and Blue bands</i>
SC	Santa Catarina
SGD	<i>Stochastic Gradient Descent</i>
SICAD	Sistema Cartográfico do Distrito Federal
SIRGAS	Sistema de Referência Geodésico para as Américas
SMRF	<i>Simple Morphological Filter</i>
SVLA	Sistema de Varredura LASER Aerotransportado
SVM	<i>Support Vector Machine</i>
TN	<i>True Negative</i>
TP	<i>True Positive</i>
UTM	<i>Universal Transversal Mercator</i>
VANT	Veículo Aéreo Não Tripulado
VGG	<i>Vision Geometry Group</i>

Sumário

1	INTRODUÇÃO	18
1.1	Contextualização	18
1.2	Estado da Arte	22
1.3	Motivação	25
1.4	Objetivos	26
1.4.1	Objetivo Geral	26
1.4.2	Objetivos Específicos	26
1.5	Organização	27
2	APRENDIZADO PROFUNDO - <i>DEEP LEARNING</i>	28
2.1	Redes Neurais Artificiais (<i>Artificial Neural Network</i> - ANN)	28
2.2	Algoritmo <i>BackPropagation</i>	30
2.3	Redes Neurais Convolucionais - CNN	32
2.3.1	<i>Stochastic Gradient Descent</i> - SGD	34
2.3.2	Regularizadores	36
2.3.3	Classificadores - <i>Backbones</i>	39
2.3.3.1	Arquitetura - VGG	39
2.3.3.2	Arquitetura - ResNet	40
2.3.4	Transferência de aprendizado - <i>Transfer Learning</i>	41
2.4	Segmentação Semântica	43
2.4.1	U-Net	44
2.4.2	Função de perda - <i>Loss</i>	46
3	FUSÃO DE IMAGENS SENSORIAMENTO REMOTO	48
3.1	Princípios básicos do LiDAR	49
3.1.1	Sistema de Varredura a LASER Aerotransportado - SVLA	49
3.1.2	Integração LiDAR-Fotogrametria	54
3.1.3	Imagem Intensidade	55
3.2	Fusão por CNN	56
3.2.1	FuseNet	57
4	MATERIAIS E MÉTODOS	60
4.1	Área de Estudo	60
4.2	Coleta de dados e recursos computacionais	61
4.3	Pré-processamento dos dados	62
4.3.1	ISPRS Potsdam <i>Dataset</i>	62

4.3.2	Massachusetts <i>Dataset</i>	65
4.3.3	Distrito Federal <i>Dataset</i>	66
4.4	Método Proposto	71
4.5	Avaliação dos resultados	74
5	RESULTADOS E ANÁLISES	76
5.1	Experimento I	77
5.2	Experimento II	80
5.3	Experimento III	88
5.3.1	Discussão sobre segmentação semântica sem fusão	88
5.3.2	Discussão sobre segmentação semântica com dados fusionados	95
5.3.3	Comparação com métodos do Estado da Arte	103
6	CONCLUSÕES E RECOMENDAÇÕES FUTURAS	105
	REFERÊNCIAS	107

1 Introdução

1.1 Contextualização

O Brasil é um país que possui 1,7 milhões de quilômetros de rodovias (CNT, 2019) distribuídas de forma irregular pelo território nacional. Além disso, características destas vias como tipo, jurisdição, largura, comprimento, condições asfálticas, entre outras, são diversas e dinâmicas.

Os diferentes ambientes/contextos também influenciam as características das vias. Por exemplo, em termos de dinamismo, a área urbana possui taxas maiores do que a rural. Mudanças de geometria, novas conexões, eliminação de trechos, mudanças de direções são alguns exemplos de como as informações sobre rodovias são relevantes para planejamento e gestão urbana, bem como para a economia (p.ex. logística para empresas).

Tal dinamismo e complexidade tornam a obtenção de informações da rede viária, de forma ágil, um grande desafio. Além disso, o Brasil é um país continental com uma grande extensão rodoviária e várias organizações governamentais responsáveis pelo mapeamento sistemático brasileiro estão procurando estratégias e metodologias para enfrentar os desafios impostos.

O Decreto federal nº 6666/08 (BRASIL, 2008) criou a Infraestrutura Nacional de Dados Espaciais (INDE) e foi um marco para a cultura da geoinformação brasileira. Organizações governamentais começaram a produzir, extrair e disseminar geodados em uma mesma infraestrutura. O resultado foi a economia de recursos públicos, otimização no uso dos dados e rápida disseminação para os usuários.¹

De acordo com Warnest (2005), uma infraestrutura de dados espaciais tem 5 componentes: pessoas, dados, instituições, tecnologias, normas e especificações. Todos estão fortemente relacionados e existe intensa interação entre eles. Ressaltam-se as normas e especificações pois a partir destas os outros componentes são construídos, regulam todos os relacionamentos entre os componentes, especificam parâmetros, critérios e requerimentos da infraestrutura.

Em 2018, a Comissão Nacional de Cartografia (CONCAR) homologou a Especificação técnica para Estruturação de dados Espaciais Vetoriais (ET-EDGV v.3.0). Essa especificação define o modelo conceitual para dados vetoriais de forma a garantir a consistência lógica dos dados. Em outras palavras, ela especifica um modelo para o mapeamento do território brasileiro. Além disso, cria uma seção específica de grandes escalas (mai-

¹ Visualizador da INDE disponível em <<https://inde.gov.br/CatalogoMetadados>>

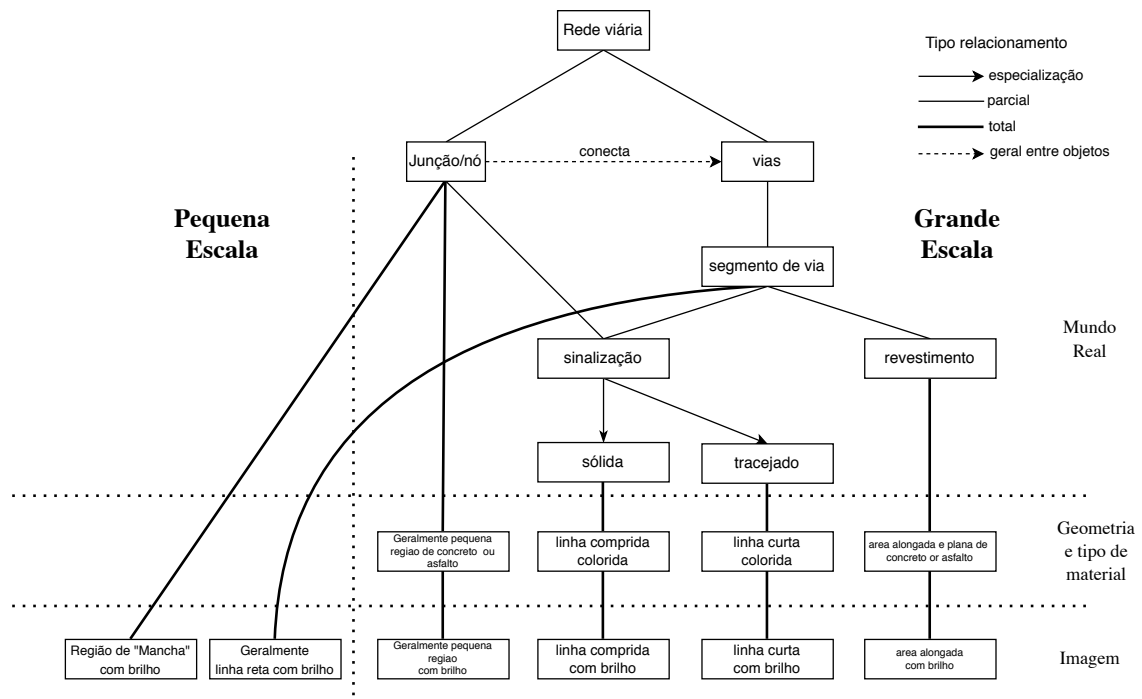
ores que 1:25.000), sendo a primeira especificação para estruturação dessas escalas no mapeamento sistemático brasileiro.

A partir da INDE, toda aquisição de informações para o mapeamento sistemático brasileiro está padronizada. Entretanto, o processo de extração é ainda demorado e custoso, isto torna difícil o mapeamento de feições do território, incluindo a rede viária.

A extração de rodovias é um importante tópico de pesquisa, dada a importância para a gestão de tráfego, desenho urbano, monitoramento de estradas, navegação por satélite e atualização de mapas (MENA, 2003; WANG et al., 2016). Shi, Miao e Debayle (2014) destacam que, após o surgimento dos computadores, muito trabalho e tempo foram reduzidos por diversos algoritmos propostos nas últimas décadas. Wei, Wang e Xu (2017) afirmam que a extração de rodovias visa detectar e segmentar a partir de imagens aéreas ou de satélite com processamento de imagens e algoritmos de visão computacional.

Nesse contexto, a extração de vias a partir de imagens de Sensoriamento Remoto é complexa devido a algumas características tais como: tipo de sensor, resolução espectral e espacial, clima, variação de luz e característica do solo. Além disso, encontrar um modelo único também é um desafio, dadas as muitas influências culturais (WANG et al., 2016). Apesar disso, Baumgartner e Hinz (2000) criaram um modelo geral de estradas (Figura 1) com alguns conceitos, características e relações.

Figura 1 – Modelo Genérico de Rodovias



Fonte: Adaptado de Baumgartner e Hinz (2000).

Este modelo esquematiza escalas e níveis de características que estão relacionados

às vias. Apresenta alguns conhecimentos importantes como geometria para a aquisição, influência radiométrica e relacionamento topológico. Os autores também destacam as relações com outras feições, como edifícios, árvores e veículos, o chamado modelo contextual (local ou global) que especifica as relações entre o objeto rodovia e os demais. De acordo com Dal Poz (2003) a aparência das rodovias, em geral, depende da resolução espacial (espaço-objeto) e espectral do sensor (espaço-imagem). Ademais, o fator espaço-escala é relevante na modelagem da malha viária.

O modelo apresentado por Baumgartner e Hinz (2000) é simples e não contempla os diversos tipos de redes viárias existentes. Porém, pode ser utilizado como base teórica para casos mais complexos e auxiliar a escolha do método de extração adequado. Especialmente na área urbana, requisitos como alta precisão, escalabilidade e flexibilidade para mudanças são relevantes para extração de informações. Nos últimos anos, imagens de Sensoriamento Remoto vem sendo disponibilizadas por várias fontes e muitas tecnologias surgiram, bem como a disponibilização de dados vetoriais, metadados e informações de contexto (p.ex. *Open Street Map* - OSM (OpenStreetMap contributors, 2017)). Assim, os algoritmos precisam operar neste novo cenário.

Dal Poz (2003) e Wang et al. (2016) afirmam que os métodos de extração de estradas começaram na década de 1970 devido ao acesso a imagens de satélite, sendo o estudo de Bajcsy e Tavakoli (1976) pioneiro na área. Desde então, vários outros algoritmos foram desenvolvidos. Wang et al. (2016) e Das, Mirnalinee e Varghese (2011) sugerem que, na literatura para extração de rodovias, os métodos mais utilizados são os baseados em classificação, a partir de conhecimento prévio, morfologia matemática, modelo de contorno ativo e programação dinâmica. Todos possuem vantagens e desvantagens, no entanto, a baixa flexibilidade para trabalhar com *big data* e a necessária intervenção do usuário no processo ainda afetam a precisão e o tempo de extração.

Além dos sensores passivos, os sensores *LASER Detection and Ranging* (LiDAR), mais especificamente utilizados em plataformas aéreas (Sistema de Varredura LASER aerotransportado - SVLA), podem fornecer informações complementares às obtidas na faixa espectral do visível. Há alguns anos, essa tecnologia era dispendiosa e com limitações técnicas. Entretanto, atualmente a nuvem de pontos LiDAR tem um papel relevante em muitas aplicações, como planejamento urbano e mapeamento de cobertura do solo. Em abordagens de aprendizado de máquina (*Support Vector Machine* (SVM), Redes Neurais, *Random Forest*, etc.), os dados de nuvem de pontos LiDAR são convertidos em imagens e concatenados com imagens brutas (SUN et al., 2018). O objetivo é aumentar a quantidade de informações aos classificadores, principalmente nos casos de oclusão nas cenas. Métodos que combinam dados LiDAR e fotogramétricos procuram vantagem na sinergia entre ambas as fontes. (Dal Poz, 2014)

Após o ano 2000, novas abordagens de aprendizado de máquina têm sido aplicadas

em diversas áreas do conhecimento, sendo uma delas o Sensoriamento Remoto (ZHU et al., 2017; MA et al., 2019). Essas ferramentas permitem extrair informações de um grande volume de dados a partir de diversas fontes. Existem muitos algoritmos de aprendizado de máquina. Contudo, desde o ano de 2014, as Redes Neurais Convolucionais têm obtido notáveis resultados nas áreas de Sensoriamento Remoto (LIU et al., 2018), principalmente devido à alta precisão alcançada, extração automática de características e flexibilidade de uso.

No tópico específico de extração de rodovias, a aplicação de Redes Neurais Convolucionais Profundas foi iniciado no ano de 2010 com Mnih e Hinton (2010). Devido às potencialidades descobertas no uso das arquiteturas de Rede ou algoritmos, como por exemplo: extração automática de características dos objetos; aplicação em grande volume de dados e enfoque prioritário no estudo de dados de múltiplos vetores como imagens, as Redes Convolucionais têm alcançado alta acurácia na extração de rede viária a partir de imagens de alta resolução espacial.

Na literatura existem diversos trabalhos com várias técnicas de extração por aprendizado profundo (*Deep Learning* - DL), entretanto as arquiteturas possuem limitação quanto aos casos de oclusão e obstruções que ocasionam descontinuidades e omissões na rede viária. Estas pesquisas tentam minimizar esses efeitos com estratégias de obtenção de contexto de cena (informações do entorno do objeto) e não exploram a complementariedade de informações advinda de outros sensores, como por exemplo o LiDAR.

Portanto, é oportunidade de pesquisa analisar a fusão de dados LiDAR e Ortoimagens por DL quanto às oclusões/obstruções na extração de rede viária. Com isso, as contribuições da pesquisa estão resumidas nos principais pontos:

1. Arquitetura de Rede Neural Convolucional ResUnet com fusão inicial de imagens multiespectrais/ intensidade LiDAR para extração de rodovias em áreas urbanas. A fusão inicial com dados de intensidade LiDAR adiciona informações para a separação de classes e diminui a dependência de contexto em áreas de oclusão/obstrução na segmentação de rede viária.
2. Metodologia para geração de imagem Modelos de Superfície e de Intensidade do sinal LASER. Ambas imagens acrescentam informação nas áreas de oclusão/ obstrução e aumentam a precisão da segmentação.
3. Discussão sobre a relevância da complementariedade de dados LiDAR e o contexto na fusão de dados para extração de rede viária urbana em regiões oclusas.
4. Criação de conjunto de dados (ortoimagens e pontos LiDAR) da região do Distrito Federal (DF) para pesquisa sobre extração de rodovias em altíssima resolução espacial com DL.

1.2 Estado da Arte

A extração de rodovias é assunto relevante nas áreas de visão computacional, Sensoriamento Remoto e Fotogrametria Digital. Dal Poz (2003) explica que o termo extração está relacionado com tarefas de reconhecimento (detecção de objetos) e de delineamento (contorno de objetos). Abdollahi et al. (2020) afirmam ainda que rodovias podem ser extraídas pelo eixo central ou pela representação da área. Algoritmos semi-automáticos realizam somente o delineamento e os automáticos realizam as duas tarefas. Os métodos de extração de área estão concentrados em segmentação ou classificação de rodovias enquanto os de eixo central a métodos de esqueletização. Atualmente, a quantidade de dados gerados ou adquiridos advindos de modernas tecnologias nas áreas citadas torna inviável, muitas vezes, a obtenção manual de informações. Principalmente quando a análise está relacionada a tríade custo-tempo-acurácia.

Desde a década de 70, vários métodos foram desenvolvidos baseados no tipo de imagem (baixa, média e alta resolução espacial) a ser utilizada, na primitiva ou modelo empregado, tipo de sensor (ativo/passivo) e no contexto da cena (rural, suburbano ou urbano). Wang et al. (2016) e Das, Mirnalinee e Varghese (2011) resumem métodos de extração a partir de imagens de satélite, especificamente para rodovias, ao longo de 30 anos. Técnicas baseadas em Conhecimento Prévio (modelo de via e uso de regras) (Dal Poz; ZANIN; Do Vale, 2006; PENG et al., 2008; SHI; ZHU, 2002), em Morfologia Matemática (operadores morfológicos) (NUNES, 2016; AMINI et al., 2002), Contornos Ativos (*Snakes*) (LAPTEV et al., 2000; MARTINS, 2017), Programação Dinâmica (VALE, 2003), Filtragem e Agrupamento ((BAUMGARTNER; HINZ, 2000; GAMBA; DELL'ACQUA; LISINI, 2006), bem como por Classificação (supervisionada e não-supervisionada) (MNIH; HINTON, 2010; DAS; MIRNALINEE; VARGHESE, 2011; ZHANG et al., 2018; MENDES; Dal Poz, 2019). Essas técnicas não possuem flexibilidade para trabalhar com muitos dados (ex. *Big Data*) e a intervenção do usuário no processo prejudicam, via de regra, a acurácia e tempo de obtenção das informações.

Sob o enfoque da acurácia da extração de rodovias a partir de imagens aéreas, os algoritmos de aprendizado de máquina têm destaque, principalmente as redes neurais (DAS; MIRNALINEE; VARGHESE, 2011). Isto porque o maior problema na extração de rodovias, em Sensoriamento Remoto, é como descrever as características das vias. Enfoque importante, tendo em vista que com o aumento da resolução espacial, maior detalhamento das vias, bem como maior interferência (prédios, sombras, obstruções - p.ex. carros, árvores) dificultam a elaboração de um modelo que possa extrair as características de forma rápida e acurada (WANG et al., 2016).

Algoritmos de classificação (supervisionada ou não-supervisionada) de aprendizado de máquina, especificamente de aprendizado profundo, possuem a capacidade de extrair características a partir dos dados, flexibilidade e escalabilidade, não dependem de intervenção

humana para construir regras. A performance dos algoritmos de aprendizado dependem da representação dos dados que são fornecidos como entrada. Cada parte de informação incluída na representação é conhecida como uma característica. Buscar automaticamente melhores representações (“aprender”) a partir de características relevantes para determinada tarefa de classificação é o problema que o aprendizado profundo (*Deep Learning* - DL) tem resolvido ao longo dos anos (GOODFELLOW; BENGIO; COURVILLE, 2016; CHOLLET, 2018).

A partir dos anos 2000, muitos algoritmos de aprendizado profundo foram desenvolvidos principalmente por avanços no potencial de processamento de dados dos computadores. As redes neurais convolucionais de aprendizado profundo (DCNN) têm obtido excelentes resultados nas tarefas de detecção, segmentação e reconhecimento de objetos/regiões em imagens (LECUN; BENGIO; HINTON, 2015; ABDOLLAHI et al., 2020). Competições com tarefas de visão computacional tornaram-se frequentes, sendo o trabalho de Krizhevsky, Sutskever e Hinton (2012) com a rede AlexNet considerado como o marco das DCNN por vencer a competição *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) (RUSSAKOVSKY et al., 2015) em 2012. Diversas arquiteturas foram desenvolvidas, tais como VGG (SIMONYAN; ZISSERMAN, 2015), ResNet (HE et al., 2016), GoogLeNet (SZEGEDY et al., 2015), entre outras.

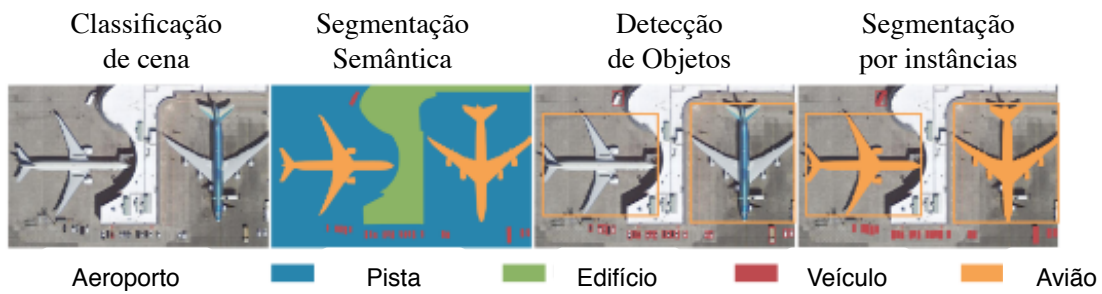
Ma et al. (2019) afirmam que, a partir de 2014, o interesse sobre DL aumentou na comunidade de Sensoriamento Remoto. O uso de redes neurais em aprendizado de máquina começou na década de 90, porém a falta de dados e a dificuldade de treinamento impossibilitaram um aprofundamento na área. Em Sensoriamento Remoto, o interesse foi renovado a partir de 2006. Além disso, a CNN é o tipo de DL mais utilizado, em imagens de áreas urbanas de alta resolução espacial (<10m) nas tarefas de classificação de uso e cobertura do solo, detecção de objetos, reconhecimento de cenas e segmentação.

É importante ressaltar as diferenças entre as nomenclaturas. Na área de visão computacional, reconhecimento de imagem (classificação de cena) é a predição da classe em toda imagem. Segmentação, segmentação semântica ou classificação ao nível do *pixel* (*pixel wise*) é a predição de cada *pixel* da imagem nas classes. Detecção de objetos é a predição da localização de objetos com polígonos envolventes e a classe correspondente. Segmentação por instâncias é a predição da localização e segmentação do objeto na imagem. A Figura 2 mostra exemplos destas definições e a Figura 3 destaca a diferença entre a classificação de *pixel* utilizando fragmento (“*patch*”) de contexto limitado para a classificação e a segmentação semântica² que utiliza da informação da cena para classificação de todos os *pixels*. (LIN et al., 2014; HOESER; KUENZER, 2020).

No caso específico da tarefa de segmentação, o estudo de Long, Shelhamer e Darrell (2015) sobre a arquitetura *Fully Convolutional Network* (FCN) foi o marco da segmentação

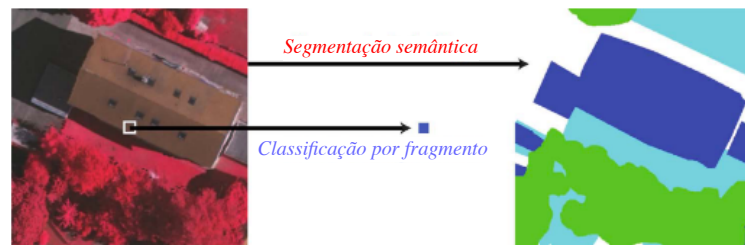
² mais detalhes na seção 2.4

Figura 2 – Exemplos de reconhecimento de imagem, segmentação semântica, por instância e detecção de objetos



Fonte: Adaptado de Hoeser e Kuenzer (2020)

Figura 3 – Diferença entre a classificação CNN por fragmento e segmentação semântica.



Fonte: Adaptado de Volpi e Tuia (2017)

por predição densa (todos os *pixels* da cena). A FCN proposta retira as camadas de neurônios totalmente conectadas da rede de classificação e define conexões diretas (*skip connections*) que combinam informações semânticas de níveis da rede para realizar a tarefa de segmentação. Posteriormente, novas arquiteturas foram desenvolvidas e aplicadas para a realização de segmentação semântica. Ma et al. (2019) afirmam que o estado da arte da segmentação semântica em Sensoriamento Remoto utilizam arquiteturas do tipo *encoder-decoder* e que a vantagem da DCNN para segmentação é sua capacidade de explorar informação de contexto multinível. Arquiteturas deste tipo possuem uma parte de codificação (*encoder*) ou classificador no qual são extraídas as características relevantes para a tarefa de segmentação em vários níveis, bem como uma parte de decodificação (*decoder*) ou expansão no qual o mapa de característica codificado é reconstruído (interpolação) para a geração da máscara final.

Segundo Abdollahi et al. (2020), um dos primeiros trabalhos sobre extração de rodovias utilizando DL foi em 2010 (MNIH; HINTON, 2010) utilizando redes neurais não convolucionais. Outros trabalhos utilizam *patch CNN* (WANG et al., 2015; ALSHEHHI et al., 2017), a arquitetura FCN (WEI; WANG; XU, 2017), *encoder-decoder* (CHENG et al., 2017; SUN et al., 2018; XU et al., 2018; YANG et al., 2019; DING; BRUZZONE, 2021; ZHU et al., 2021; BANDARA; VALANARASU; PATEL, 2022) e *Generative Adversarial Network* (GAN) (COSTEA et al., 2017). Esses trabalhos utilizam técnicas de obtenção de

contexto para reduzir os efeitos de descontinuidades e de omissões causados por sombras de prédios e árvores, condições climáticas e similaridade espectral com outros objetos em imagens de alta resolução espacial. Poucos trabalhos utilizam a complementariedade das informações com outros sensores ou fontes para analisar problemas de oclusões e obstruções, mais comuns em pesquisas sobre classificação de cobertura de solo (SUN; WANG, 2018; AUDEBERT; SAUX; LEFÈVRE, 2018; SANG; MINH, 2018; SUN et al., 2018).

Paralelamente, avanços tecnológicos nas décadas de 80/90 impulsionaram pesquisas sobre técnicas de fusão de dados multisensores (MENA, 2003; SCHENK; CSATHÓ, 2002; LIU et al., 2018). Audebert, Saux e Lefèvre (2018) afirmam que dados de observação da Terra não estão limitados aos sensores ópticos que atuam na faixa do visível. Existem outros sensores e fontes que podem ser utilizadas para complementar a informação de extração de objetos. Ghanma (2006) diz que se as características complementares entre fotogrametria e LiDAR forem exploradas, maior será a compreensão da superfície imageada.

No imageamento óptico, as vias são os objetos mais baixos nas cenas, frequentemente obstruídos por sombras de objetos mais altos como vegetação e prédios. De forma contrária, com o SVLA, rodovias são menos afetadas porque o sensor tem alta penetrabilidade na vegetação e pequeno campo de visão (*Field of View* - FOV) (ZHANG, 2010). Os sensores LiDAR podem complementar informações das imagens de alta resolução, ou até mesmo ser a principal fonte de dados. Atualmente, a classificação de imagens e de nuvem de pontos tem papel relevante em muitas aplicações. Sun et al. (2018) afirmam que imagens podem ser geradas a partir de nuvem de pontos LiDAR e concatenadas com imagens multiespectrais, assim como Heydari e Mountrakis (2019) mostram várias estratégias para essa integração em redes de DL.

Pela revisão realizada na literatura, não existem muitos trabalhos de fusão com dados para extração de rodovias especificamente por DL. Recentemente foi explorada somente pelos autores: Parajuli, Choudhury e Kumar (2020) utilizando a rede TriSeg para fusão RGB/LiDAR e Wu et al. (2020), Liu et al. (2022) com trajetórias GNSS.

1.3 Motivação

A relevância científica e tecnológica da tese é evidenciada pela importância dada ao tema pela *International Society for Photogrammetry and Remote Sensing* (ISPRS), na Comissão II (*Photogrammetry*) e na Inter comissão II/III (análise de padrões em Sensoriamento Remoto) que possuem os seguintes temas: 1) identificação automática e aprendizado de padrões 2D e 3D em dados uni-modais ou multimodais de Sensoriamento Remoto; 2) Integração e fusão de dados de múltiplas fontes para detecção de objetos e entendimento de cena. Ademais, Ma et al. (2019) apresentam um estudo de como os

algoritmos de DL estão sendo aplicados em Sensoriamento Remoto e Zhu et al. (2017) encorajam os pesquisadores da área a contribuir para o aperfeiçoamento dos algoritmos de aprendizado automático.

Outro aspecto a ser considerado é a possibilidade de contribuir para o estado da arte em extração automática de feições, no caso para extração de rede viária em ambiente urbano. O uso de redes convolucionais profundas permite o ajuste automático aos dados, extraindo principais características para o objetivo de obtenção ágil de informações.

A utilização de conceitos e tecnologias pertencentes à área de Fotogrametria, Sensoriamento Remoto e Visão Computacional são imprescindíveis para que a extração ágil de informações possa aprimorar a gestão territorial urbana, principalmente em grandes escalas. E a fusão de dados é tema de destaque na comunidade científica. Por exemplo, a *Geoscience and Remote Sensing Society* (GRSS) por meio do *Image Analysis and Data Fusion Technical Committee* (IADF TC) realiza competições (*IEEE GRSS Data Fusion Contest*) anuais, desde o ano 2006, para divulgar novas ideias e avanços para a comunidade científica.

Por fim, a inspiração inicial deste projeto de pesquisa foi a continuidade da tese de Mendes (2012). Pesquisa posteriormente publicada em *International Journal of Image and Data Fusion* (MENDES; Dal Poz, 2019).

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver uma metodologia de extração automática de rede viária a partir de dados fusionados (imagens digitais de altíssima resolução espacial e dados de varredura LASER aerotransportado) por DL em ambiente urbano.

1.4.2 Objetivos Específicos

- Realizar a extração de vias em conjuntos de dados *benchmark* (Estado da Arte), bem como na região de estudo;
- Desenvolver método de fusão dos dados LiDAR/imagens ópticas como *input* de rede convolucional profunda;
- Avaliar a segmentação da rede viária urbana e a contribuição dos dados LiDAR na segmentação semântica das vias.

1.5 Organização

A tese possui seis capítulos. Os capítulos 2 e 3 apresentam a fundamentação teórica, assim como o capítulo 4 apresenta a metodologia adotada na pesquisa. O capítulo 5 contém os resultados e análises, o 6 as conclusões bem como recomendações futuras e por fim, as referências utilizadas.

2 Aprendizado Profundo - *Deep Learning*

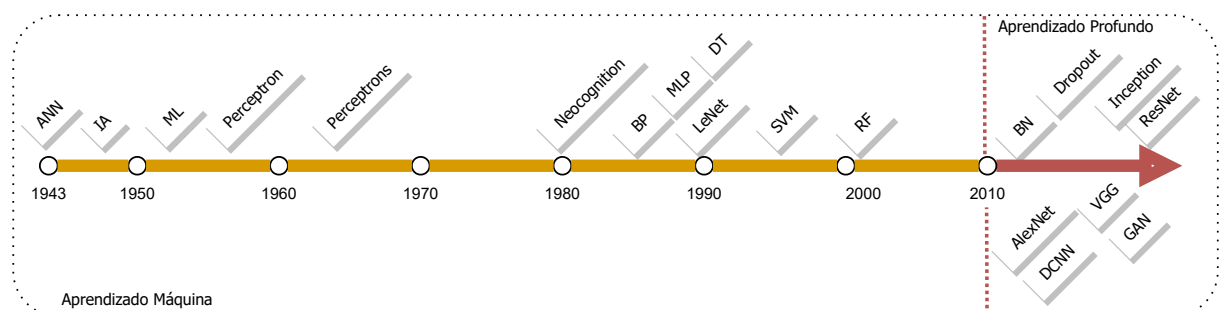
O aprendizado de máquina (*Machine Learning* - ML) é o campo da ciência dedicado ao estudo e desenvolvimento de algoritmos que permitam as máquinas aprender. Trask (2019) afirma que DL é um subconjunto de métodos de ML que utilizam principalmente Redes Neurais Artificiais. Hoeser e Kuenzer (2020) afirmam que o tema DL tem recebido muita atenção por pesquisadores e profissionais nos últimos anos. Esse crescimento decorre da disponibilidade de conjunto de dados para treinamentos de rede e do aumento do poder computacional para processamento. O DL evoluiu da ciência da computação, onde os métodos de DL são objetos de pesquisa, para outras áreas do conhecimento que fazem pesquisa utilizando técnicas de DL. Comparativamente ao ML, DL utiliza mais camadas de neurônios artificiais e maior profundidade dos modelos, conseqüentemente mais características complexas, abstratas e distintivas podem ser extraídas. Nas tarefas de Visão Computacional, as Redes Neurais Convolucionais são os algoritmos que concentram grande parte dos esforços na literatura de DL.

Antes de iniciar a fundamentação teórica sobre redes convolucionais, conceitos básicos sobre Redes Neurais Artificiais e *backpropagation* serão apresentados.

2.1 Redes Neurais Artificiais (*Artificial Neural Network* - ANN)

O aprendizado de máquina possui um longo histórico. Alguns pontos podem ser destacados como referenciais ao desenvolvimento atual. A Figura 4 ilustra resumidamente o histórico das ANN.

Figura 4 – Linha do Tempo de ML



Fonte: De autoria própria.

O estudo de Mcculloch e Pitts (1943) sobre redes neurais é considerado o início dessa era. Em 1950, Turing (1950) introduz a noção de inteligência artificial e discute

possibilidades para implementação de aprendizado de computadores. Anos posteriores, Samuel (1959) apresenta trabalho utilizando o termo “*machine learning*” e descreve princípios para serem utilizados no aprendizado de máquina por meio de jogo de xadrez.

Em 1958, Rosenblatt (1958) apresenta o Perceptron, modelo matemático de representação do neurônio. Após esta descoberta, Minsky e Papert (1969) lançam o livro no qual eles corrigem o modelo simples de Perceptron (problema do aprendizado da lógica XOR) de Rosenblatt e generalizam para classes de Perceptrons. Em 1980, Fukushima (1980) propõe um novo modelo de rede neural (Neocognitron) que reconhece padrões baseado na similaridade geométrica sem ser afetado pela posição ou pequenas distorções no formato do objeto. Trabalho este que inspirou a criação de redes convolucionais.

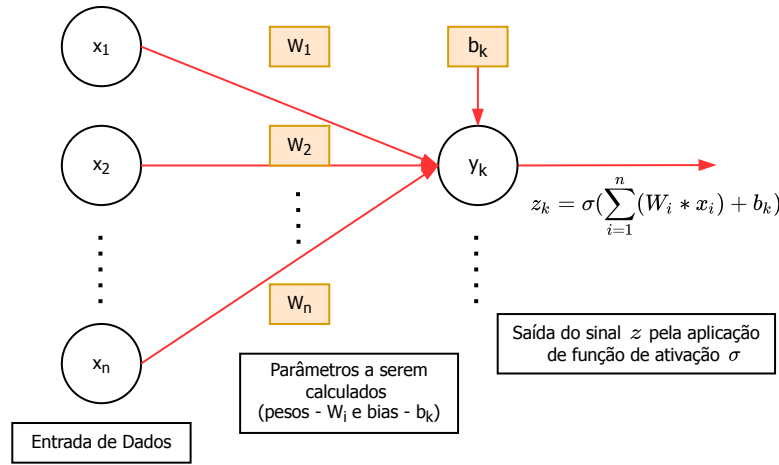
Outro trabalho relevante e correlacionado ao desenvolvimento das redes neurais foi o de Rumelhart, Hinton e Williams (1986). No qual foi utilizado o procedimento para aprendizagem em redes utilizando o algoritmo *BackPropagation* de erros da função custo (será descrito mais detalhadamente na Seção 2.2) com *Perceptrons* em camadas múltiplas (*Multi Layer Perceptron* - MLP).

No período dos anos 90, os estudos que empregaram redes neurais não obtiveram resultados relevantes. Exceção a este período foi a criação da LeNet de Lecun et al. (1989) para reconhecimento automático de números de código postal nas cartas. Novos algoritmos surgiram (por ex.: *Decision Trees* de Quinlan (1986), SVM por Boser, Guyon e Vapnik (1992), *Random Forests* por Breiman (2001)) e são amplamente utilizados.

A melhoria dos algoritmos de ML e o desenvolvimento de hardwares mais potentes possibilitaram o retorno da pesquisa com redes neurais. Principalmente, na sub-área de DL com as redes convolucionais, especificamente a partir de 2012 com a rede AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Assim como o surgimento de técnicas de regularização da rede para melhorar o aprendizado (por ex.: *Batch Normalization* de Ioffe e Szegedy (2015), *Dropout* de Srivastava et al. (2014)) e de outras arquiteturas (VGG de Simonyan e Zisserman (2015), GAN de Goodfellow et al. (2014), ResNet de He et al. (2016) e Inception de Szegedy et al. (2015)).

A Figura 5 mostra a representação do modelo de neurônio artificial (unidade ou nó). Basicamente, o modelo tem 3 partes: dados de entrada, parâmetros de ajuste e função de ativação. Sundelius (2017) afirma que os neurônios são representados por “nós ponderados” organizados em camadas, sendo a conexão dos neurônios realizada somente entre camadas diferentes. A função de ativação aplicada na soma ponderada das entradas do nó introduz a não-linearidade necessária para o aprendizado da rede.

Figura 5 – Modelo de Neurônio Artificial - ANN



Fonte: De autoria própria.

Ressalta-se que as redes neurais de múltiplas camadas são aproximadores universais de funções (HORNICK; STINCHCOMBE; WHITE, 1989; HORNICK, 1991). Ou seja, podem ser usadas para aproximar quaisquer funções mensuráveis desde que tenham unidades (nós) suficientes nas camadas e que a função de ativação seja contínua, limitadas no infinito (*squashing function*) e não-constante. Assim, dentre várias possibilidades, as redes neurais podem ser utilizadas como classificadores, ou seja, aproximar para uma função de separação de classes.

2.2 Algoritmo *BackPropagation*

Lecun, Bengio e Hinton (2015) afirmam que o procedimento de *backpropagation* (BP) utilizado para calcular o gradiente de uma função objetiva, custo ou perda (*loss function*) com respeito aos parâmetros de redes multicamadas é equivalente a utilizar a regra da cadeia no cálculo de derivadas. O algoritmo BP pode ser aplicado repetidamente para propagação do gradiente através de todas as camadas e nós.

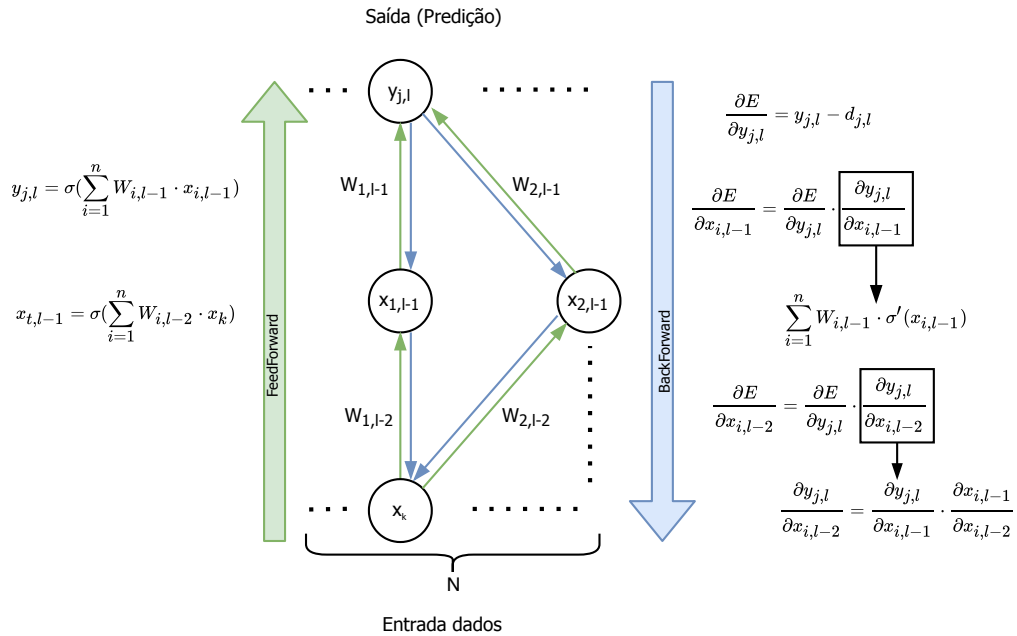
O trabalho pioneiro sobre a utilização do algoritmo em redes neurais multicamadas foi de Rumelhart, Hinton e Williams (1986). Eles explicam que o processo de aprendizado deve decidir em quais circunstâncias as unidades intermediárias (ocultas) devem ser ativadas para que a inferência se aproxime do valor correto desejado. O BP é descrito como procedimento que, de forma repetitiva, ajusta os parâmetros ("pesos") das conexões na rede para minimizar a diferença entre o vetor de saída e o desejado (correto). O erro total entre esses vetores é definido na Eq. 2.1.

$$E = \frac{1}{2} \sum_c \sum_j (y_{j,c} - d_{j,c})^2 \quad (2.1)$$

onde c é a quantidade de pares (entradas e saídas de dados), j é o índice das unidades (nós) da camada de saída da rede neural, y é a predição (inferência) da unidade e d é o valor correto.

O algoritmo possui 2 etapas: a primeira, chamada alimentação progressiva *feed-forward*, atualiza o “estado” de cada nó de cada camada pelo sinal recebido das unidades das camadas mais “baixas”. A segunda, chamada propagação regressiva *backward*, propaga o gradiente do erro calculado das camadas mais “altas” para as mais baixas. De forma a ilustrar as duas etapas, a Figura 6 apresenta um extrato de uma rede com as equações referentes ao cálculo das inferências e a propagação do gradiente do erro encontrado. Para maior clareza, o parâmetro *bias* não foi representado na figura.

Figura 6 – Etapas do procedimento de BP de erro



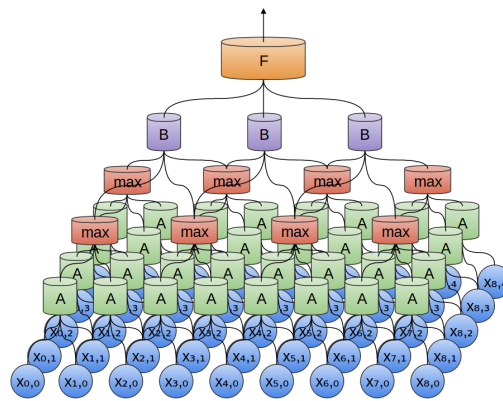
Fonte: De autoria própria

O nível mais “raso” da rede apresenta a entrada de dados e a ativação das unidades ponderadas resulta na predição no nível mais “alto” da rede neural multicamada. Após o cálculo do gradiente do erro encontrado, os parâmetros da rede são ajustados para que nova inferência possa ser realizada. Processo que se repete várias vezes com o conjunto de dados. O ajuste de pesos nesse processo foi otimizado pelo uso do algoritmo Gradiente Descendente (GD), o qual será explicado na Subseção 2.3.1.

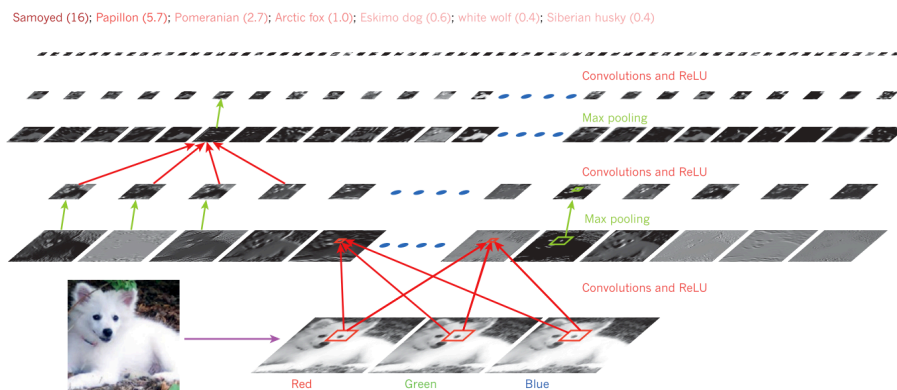
2.3 Redes Neurais Convolucionais - CNN

Lecun, Bengio e Hinton (2015) afirmam que as Redes Neurais Convolucionais (CNN) são próprias para processamento de dados na forma de múltiplos vetores. As CNN são uma categoria das redes neurais e seu uso mais comum tem sido com imagens, porém existem casos com linguagem, audio e vídeo. Desde o início dos anos 2000, as CNN têm obtido resultados relevantes no campo de visão computacional e possuem 4 características que diferem das redes neurais clássicas: conexões locais, compartilhamento de pesos, *pooling* e muitas camadas. Basicamente, a arquitetura das CNN está representada na Figura 7.

Figura 7 – Arquitetura CNN - representação por unidade e imagens



Fonte: Colah (<http://colah.github.io/posts/2014-07-Conv-Nets-Modular>). X representa unidades da camada de entrada; A e B representam unidades de mapas de característica gerados; max representa a operação de *pooling*; F representa a camada totalmente conectada.



Fonte: Lecun, Bengio e Hinton (2015). Representação dos mapas de característica como imagens e o classificador do objeto na camada mais alta.

As unidades ou nós da camada convolucional (Fig. 7) estão organizados em mapas de característica (*feature maps*), no qual cada unidade da dimensão espacial dos mapas está conectada a uma região local padrão (*patch*), também chamado de campo receptivo (*receptive field*), dos mapas de característica de camadas prévias por meio de conjunto de pesos chamados de filtros (*kernel*), que desloca-se pela dimensão espacial dos mapas por passos (*strides*). Todas as unidades pertencentes a um mapa de característica compartilham

o mesmo filtro e diferentes mapas possuem outros filtros. Desta forma, a operação de convolução (Eq. 2.2) pode ser efetuada seguida por uma função de ativação (normalmente a ReLU (GLOROT; BORDES; BENGIO, 2011)) e *pooling*. (LECUN; BENGIO; HINTON, 2015)

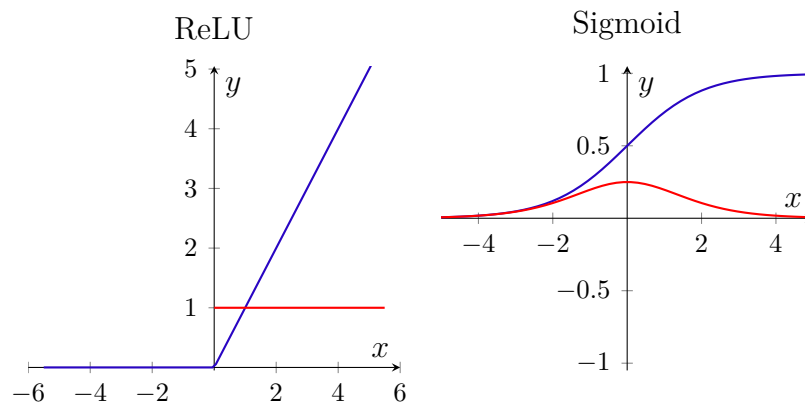
$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n) K(i - m, j - n) \quad (2.2)$$

onde S representa o resultado da operação de convolução em dados bidimensionais discretos, I representa uma imagem e K o *Kernel* (ou filtro) na operação de convolução.

A Figura 8 apresenta as diferentes funções de ativação utilizadas nas redes convolucionais e suas derivadas. A função ReLU é a mais utilizada nas camadas intermediárias nas diferentes arquiteturas por resolver o problema de desaparecimento do gradiente (*vanish gradient*) ao longo da rede e ser computacionalmente simples (GLOROT; BORDES; BENGIO, 2011), assim como a função *Sigmoid* para a classificação final.

Alshehhi et al. (2017) afirmam que a função *pooling* realiza uma subamostragem (função máximo ou média das unidades em uma região local) nos mapas de característica, reduzindo sua dimensão espacial. Lecun, Bengio e Hinton (2015) explicam ainda que o *pooling* realiza a fusão de características similares e cria a invariância a pequenos deslocamentos e distorções nas representações. Ou seja, o *pooling* amplia o campo receptivo para a convolução extrair padrões mesmo com a variação da posição relativa entre as características principais da representação (p.ex: deslocamento da posição do objeto na cena ou de alguma parte característica).

Figura 8 – Funções de Ativação e Derivadas



Fonte: De autoria própria. Funções de ativação representadas em azul e as derivadas em vermelho

Lecun, Bengio e Hinton (2015) esclarecem que a arquitetura da CNN foi criada desta forma por 2 motivos: dados de múltiplos vetores como imagens possuem regiões locais altamente correlacionadas, formando padrões locais distintos que podem ser detectados e as estatísticas locais das imagens são invariantes à localização. Hoese e Kuenzer (2020)

diferem as CNN das ANN pela extração de características de sinais sem alteração de sua estrutura, sensibilidade a conectividade local, os filtros são usados para conectar camadas em vez de conexões aos pares de neurônios (diminuição de parâmetros e compartilhamento de pesos) e a composição modular do classificador e da arquitetura permitem o uso em várias outras tarefas. Além disso, as CNN extraem características de forma hierárquica e imitam o comportamento do córtex visual de mamíferos.

Hoeser e Kuenzer (2020) dividem a arquitetura das CNN em 3 partes: entrada de dados, classificador (*Encoder*) e classificador principal. Ao longo do tempo, os classificadores foram evoluindo e este desenvolvimento acarretou melhor acurácia nas demais tarefas como localização de objetos e segmentação de objetos. Arquiteturas clássicas como AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e VGG (SIMONYAN; ZISSERMAN, 2015) iniciaram discussões sobre profundidade das redes, método de redimensionamento e redução de resolução (*pooling*), além de nova função de ativação (ReLU) para agilizar o treinamento e alcançar a convergência no processo de otimização.

Outras arquiteturas de CNN foram desenvolvidas ao longo dos anos, como por exemplo: conjunto das Inception que utilizam camadas convolucionais em ramos paralelos, redução da dimensionalidade para auxiliar no processamento da convolução (*bottleneck*) e fatorização nas dimensões espacial e de profundidade; conjunto ResNet utilizam blocos residuais, os quais são menos complexos e permitem arquiteturas de rede mais profundas (mais camadas convolucionais); conjunto das redes eficientes como MobileNets (uso em sistemas móveis - complexidade *versus* acurácia) e as construídas de forma automática por *Neural Architecture Search* (NAS). Estas últimas são arquiteturas que são modeladas por outras redes neurais, baseadas nas características mais relevantes das redes antecessoras. (HOESER; KUENZER, 2020)

2.3.1 Stochastic Gradient Descent - SGD

Segundo Goodfellow, Bengio e Courville (2016) a otimização de uma função $f(\theta)$ envolve a minimização ou maximização da mesma pela alteração de argumentos ou variáveis de projeto (θ). A função a ser otimizada é chamada função objetivo, no caso da minimização também pode ser chamada de custo ou perda (loss) ou erro. Bottou (1998) salienta que o objetivo de sistemas de aprendizado é minimizar a função de risco esperado (*expected risk function*) ou erro de generalização esperado, o qual difere-se da função objetivo comum utilizada em otimização pura devido à distribuição do conjunto de dados ser desconhecida. Desta forma, os algoritmos de ML realizam otimização de forma indireta. Utiliza-se uma distribuição empírica (conjunto de treinamento) e o objetivo é minimizar o **risco empírico** (eq. 2.3), onde não é garantida a determinação de mínimo local em tempo razoável, mas que permite encontrar um valor pequeno que atenda o problema de otimização da função no conjunto de dados de teste.

$$E_{x,y \sim \hat{p}_{\text{dados}}(x,y)}[\mathcal{L}(f(x; \theta), y)] = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(x^{(i)}; \theta), y^{(i)}) \quad (2.3)$$

onde x é a variável de entrada, y o valor conhecido do conjunto no aprendizado supervisionado, $\hat{p}$ a distribuição de probabilidade empírica do conjunto de treinamento, $\mathcal{L}$ a função perda (*loss*), $f(x, \theta)$ a predição de saída e m o número de amostras de treinamento.

É possível computar uma aproximação do risco usando um conjunto de treino finito de observações independentes. A minimização do **risco empírico** permite obter uma boa estimativa do risco esperado quando o conjunto de treinamento é suficientemente grande. A **generalização** do modelo reside na habilidade do sistema aprender a partir de um conjunto finito e, posteriormente, o modelo fornecer resultados válidos em geral. (BOTTOU, 1998)

Uma das técnicas utilizada para otimização é o GD de Cauchy (CAUCHY; 1847). A otimização baseia-se em reduzir $f(\theta)$ alterando θ em pequenos “passos” com sinal oposto a derivada da função ou ao gradiente. Pontos onde a derivada ou gradiente da função sejam nulos são chamados pontos críticos ou estacionários e não fornecem direção para otimização. Além disso, esses pontos críticos podem ser mínimos/máximos locais/globais e de sela (*saddle*). (GOODFELLOW; BENGIO; COURVILLE, 2016)

Goodfellow, Bengio e Courville (2016) ainda afirmam que, em aprendizado profundo, existem funções multidimensionais com mínimos locais que não são globalmente ótimos, um ou vários pontos de mínimo locais e muitos pontos de sela com regiões planas (*plateaus*). Geralmente, a otimização busca o valor da função custo pequeno, mas não necessariamente o mínimo global. Portanto, a Equação 2.4 do método GD propõe o cálculo do novo ponto a ser encontrado para otimização.

$$\theta' = \theta - \epsilon \nabla_{\theta} f(\theta) \quad (2.4)$$

onde ϵ é a **taxa de aprendizado**, esse escalar positivo determina o tamanho do “passo” a ser percorrido na função na direção contrária ao gradiente. A convergência é atingida quando todo elemento do gradiente é igual a zero ou muito próximo de zero. O tamanho do “passo” deve ser pequeno suficiente para não ultrapassar o ponto de mínimo, consequentemente evitando encontrar pontos onde a curvatura da função é muito acentuada.

O algoritmo estocástico é uma extensão do GD projetado para ser usado em grandes conjuntos de dados, que necessitam de potencial computacional para realização de treinamentos. Goodfellow, Bengio e Courville (2016) afirmam que o gradiente, no método SGD, é um valor esperado e pode ser estimado, sem viés (*bias*), utilizando pequenos conjuntos de amostras (*minibatch*). A Equação 2.5 representa a estimativa do valor esperado do gradiente na atualização de parâmetros da rede.

$$\theta' = \theta - \epsilon \left(\frac{1}{m} \right) \nabla_{\theta} \sum_1^m f(x^i, y^i, \theta) \quad (2.5)$$

onde m representa a quantidade de pequenas amostras do conjunto $B = \{x^{(1)}, \dots, x^{(m)}\}$, i variando de 1 a poucas centenas. Sendo x os valores de entrada, y valores de saída conhecidos e θ parâmetros a serem ajustados. Goodfellow, Bengio e Courville (2016) apresentam o Algoritmo 1 com as etapas para o cálculo do gradiente e busca da minimização da função.

Algoritmo 1: SGD na iteração k do treinamento da rede

Entrada: Taxa de aprendizado ϵ_k

Entrada: Parâmetro inicial θ

while não chegar no critério de parada **do**

selecione *minibatch* do conjunto de treinamento $\{x^{(1)}, \dots, x^{(m)}\}$ com
correspondente rótulo $y^{(i)}$;
estime o gradiente $\hat{g} \leftarrow + \left(\frac{1}{m} \right) \nabla_{\theta} \sum_1^m f(x^{(i)}, y^{(i)}, \theta)$;
atualizar $\theta \leftarrow \theta - \epsilon \hat{g}$;

end

Embora o algoritmo do SGD seja o mais utilizado, o aprendizado pode ainda ser lento. Várias estratégias foram criadas para acelerar o aprendizado e, conseqüentemente, a convergência do modelo. Os otimizadores *Momentum* (QIAN, 1999), *Adagrad* (DUCHI; BARTLETT; WAINWRIGHT, 2012), *RMSProp*, *ADAM* (KINGMA; BA, 2015) são alguns algoritmos utilizados.

2.3.2 Regularizadores

De acordo com Goodfellow, Bengio e Courville (2016), regularização é definida como qualquer alteração realizada no algoritmo de aprendizado com o objetivo de reduzir o erro de generalização, mas não o de treinamento. Ou seja, são estratégias para a rede tenha boa performance em dados novos para uma tarefa específica. Não existe o melhor algoritmo de aprendizado e nem a melhor forma de regularização. Em cada tarefa (p.ex: classificação, segmentação, reconhecimento, etc.) deve-se optar pela solução que seja mais adequada.

Algumas estratégias para regularização de modelos restringem os valores de parâmetros no modelo, outros adicionam restrições ou penalidades na função objetivo e ainda existem os comitês (*Ensemble*) que realizam combinações de métodos.

Ao diminuir o erro de generalização, conseqüentemente a regularização evita o ajuste excessivo aos dados (*overfitting*). Um exemplo que utiliza a função objetiva é a penalidade pela norma dos parâmetros. Basicamente, uma nova função custo é criada com a adição de uma penalidade $\Omega(\theta)$ (norma dos parâmetros) com um fator de regularização

$\alpha \in [0, \infty)$. O propósito é otimizar a função custo bem como a magnitude dos parâmetros a serem ajustados. (GOODFELLOW; BENGIO; COURVILLE, 2016)

As Equações 2.6 e 2.7 apresentam as normas mais utilizadas. A norma L^2 também é chamada de decaimento de peso (*Weight Decay*).

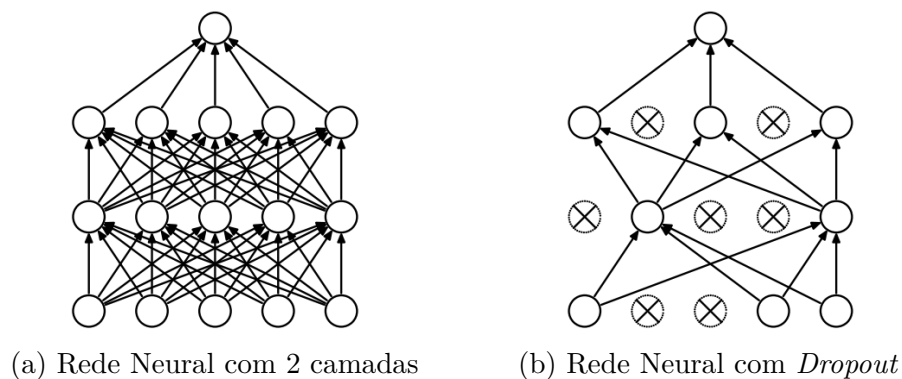
$$\Omega(\theta) = \frac{1}{2} \|w\|_2^2 \quad (2.6)$$

$$\Omega(\theta) = \|w\|_1 \quad (2.7)$$

Um exemplo de regularização implícita que não incide diretamente nos parâmetros é o aumento sintético de dados (***data augmentation***). É uma técnica de aumento artificial do conjunto de dados por transformações nas imagens (rotação, espelhamento, aumento de contraste, brilho, etc.). Ronneberger, Fischer e Brox (2015) afirmam que o uso do aumento sintético de dados ensina propriedades de invariância e robustez para a rede. Garcia-Garcia et al. (2018) explicam que além de regularizador, evitando ajuste excessivo aos dados, o aumento artificial de dados auxilia na convergência e aumenta a capacidade de generalização da rede. O objetivo destas transformações é criar mais amostras e aumentar o conjunto de dados artificialmente. Consequentemente, regulariza o modelo, realiza o balanceamento de classes, além da possibilidade de produzir amostras mais representativas para a objetivo de visão computacional. O artigo de (SHORTEN; KHOSHGOFTAAR, 2019) apresenta diversas técnicas utilizadas.

Srivastava et al. (2014) criaram a técnica *Dropout*. É um regularizador que, temporariamente, seleciona nós em diferentes camadas da rede que serão "desligados" ou "ligados" durante o treinamento. Uma rede neural com n unidades (nós) pode ser entendida como uma coleção de 2^n possíveis novas redes. A Figura 9 apresenta uma representação da técnica.

Figura 9 – Modelo de Rede Neural com *Dropout*



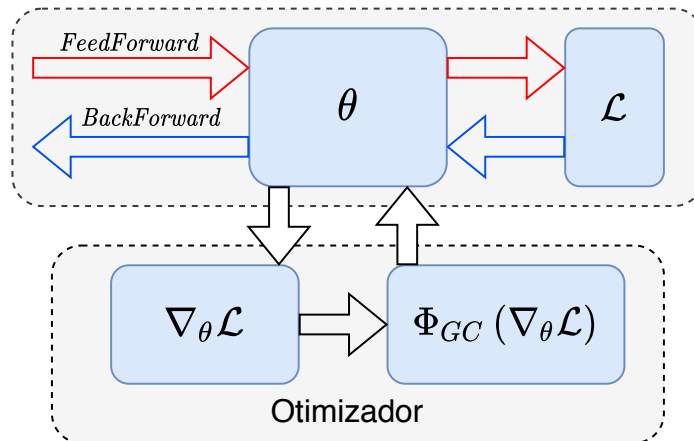
Fonte: Adaptado de Srivastava et al. (2014)

Sob o enfoque de restrição dos valores de parâmetros no modelo, pode-se citar a normalização por *mini-batch* (*Batch Normalization* - BN). Segundo Ioffe e Szegedy (2015), BN visa a normalização da distribuição dos valores obtidos nos filtros a cada *mini-batch* (subamostra) antes da função de ativação. Isso permite que o treinamento seja mais rápido (convergência) e se obtenha maior acurácia na otimização. Essa alteração na distribuição dos valores devido à atualização de parâmetros durante o treinamento é chamada de *internal covariance shift*.

Hazirbas et al. (2017) explicam que a camada de BN normaliza os mapas de características para distribuição com média igual a zero e variância unitária, posteriormente aplica fator de escala (γ) e deslocamento (β). Ambos os parâmetros são aprendidos no treinamento da rede convolucional para que a normalização não acarrete mudança na representação da camada, inclusive esses parâmetros podem restaurar os valores dos mapas sem a normalização. Ioffe e Szegedy (2015) ainda afirmam que BN auxilia no fluxo do gradiente na rede, possibilita utilizar taxas maiores de aprendizado, bem como reduz a necessidade de *dropout*.

O estudo de Yong et al. (2020) apresenta outra alternativa chamada centralização do gradiente (*Gradient Centralization* - GC) para a regularização dos parâmetros da rede. Essa técnica introduz uma restrição que a média dos vetores gradiente dos parâmetros (∇_{θ}) seja igual a zero, ou seja, o método (Φ_{GC}) calcula e subtrai a média do gradiente dos parâmetros em cada camada que é atualizada na rede. A Figura 10 ilustra o processo de treinamento com a inserção do módulo (Φ_{GC}) na atualização dos parâmetros durante a propagação reversa (*Backforward*).

Figura 10 – Esquema de otimização da Rede Neural com GC



Fonte: Adaptado de Yong et al. (2020). $\mathcal{L}$ representa a função de perda.

Yong et al. (2020) afirmam que o método combinado com BN em tarefas de segmentação e detecção de objetos pode melhorar a performance da generalização e acelerar o processo de treinamento.

2.3.3 Classificadores - *Backbones*

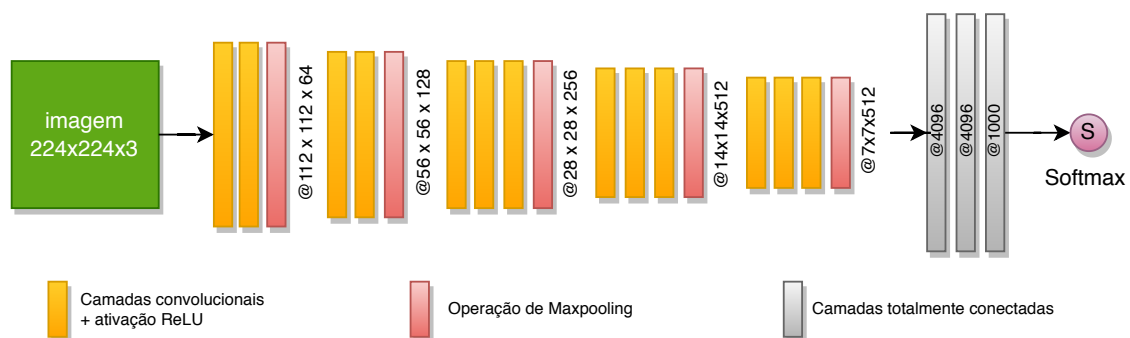
2.3.3.1 Arquitetura - VGG

Desde o sucesso da rede AlexNet em 2012 (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) na tarefa de classificação de cenas ImageNet, as redes convolucionais tem obtido grande destaque. Dois fatores são importantes neste processo: a disponibilização de grandes conjuntos de dados para treinamento e inferência, bem como o desenvolvimento de *hardwares* mais potentes (*Cluster* e GPU).

Várias arquiteturas de redes convolucionais foram criadas. Segundo Simonyan e Zisserman (2015), a utilização de pequenos filtros, passos (*strides*) com valores pequenos e que percorrem toda a imagem em várias escalas foram melhorias obtidas na linha de visão computacional. Porém, a arquitetura VGG foi a primeira a explorar a profundidade (quantidade de camadas convolucionais). VGG é uma sigla para nomear a rede convolucional muito profunda usada em classificação/reconhecimento de imagens de grandes escalas, além disso também é abreviatura do grupo de pesquisa da Universidade de Oxford - Visual Geometry Group (VGG).

Simonyan e Zisserman (2015) criaram e testaram 5 variantes de arquiteturas. Testes com 11, 13, 16 (ilustrada na Figura 11) e 19 camadas de parâmetros. A arquitetura final de 16 camadas, sendo 13 convolucionais agrupados em 5 blocos convolucionais intercalados pela camada de *maxpooling*, uso de normalização nos dados de entrada e 3 camadas iniciais totalmente conectadas para classificação final da cena. O total de mapas gerados nas camadas não foi alterado, sendo utilizado os valores de 64, 128, 256, 512 conforme a variação de escala da imagem ao longo da rede. Ao final, embora tenham utilizado um filtro convolucional pequeno (3x3), o empilhamento das camadas convolucionais resultou em funções mais discriminantes na separação das classes, diminuindo a quantidade de parâmetros na rede (133 a 144 milhões de parâmetros)(BIANCO et al., 2018).

Figura 11 – Representação simplificada da arquitetura VGG-16



Fonte: De autoria própria. Arquitetura com 13 camadas convolucionais e 3 camadas totalmente conectadas. @ - dimensão dos mapas de características ou das camadas de neurônios, Softmax - função de ativação (BRIDLE, 1990).

A arquitetura VGG, utilizando 16 e 19 camadas de parâmetros, conseguiu o primeiro e segundo lugares na competição ImageNet 2014 nas tarefas de localização de objetos e classificação de cenas, respectivamente. Além disso, esta arquitetura foi a base para a geração de redes convolucionais para segmentação semântica, detalhadas em na Seção 2.4.

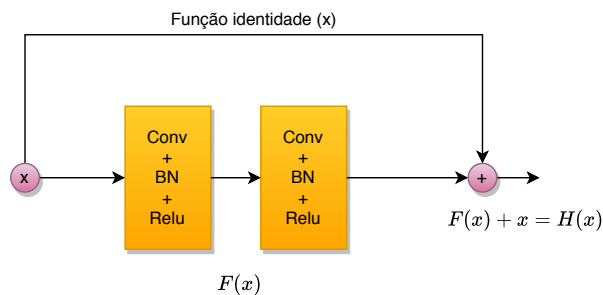
2.3.3.2 Arquitetura - ResNet

O termo ResNet representa redes muito profundas que possuem “aprendizado residual” (*deep residual learning*). He et al. (2016) apresentam essa arquitetura com o objetivo de encontrar uma solução para os problemas de acurácia de redes convolucionais profundas.

As redes convolucionais profundas, como por exemplo a VGG, apresentam resultados de acurácia nas tarefas de classificação superiores nas competições ImageNet. Antigos problemas no cálculo de gradiente (*vanishing/exploding gradient*), consequentemente na convergência da otimização da função custo durante o treinamento da rede, foram resolvidos com a inicialização (GLOROT; BENGIO, 2010; HE et al., 2015) dos pesos, normalização das camadas (IOFFE; SZEGEDY, 2015) e função de ativação adequada (HOESER; KUENZER, 2020). Porém, He et al. (2016) destacam que conforme a rede cresce, a acurácia no treinamento “satura” e um processo de “degradação” aumenta a taxa de erro. Tal processo não está relacionado ao ajuste excessivo aos dados e sim, à dificuldade de otimização da função custo.

O aprendizado residual foi desenvolvido para solucionar o problema da degradação da acurácia ao aumentar a profundidade da rede, bem como facilitar a otimização pela rede profunda. A hipótese utilizada é que se múltiplas camadas não lineares podem aproximar funções complexas, então da mesma forma essas camadas podem aproximar funções residuais. A Figura 12 representa um bloco de aprendizado residual onde as camadas devem otimizar $F(x)$. Função que pode ser representada por uma subtração de função subjacente $H(x)$ com a entrada x obtida por uma função identidade representada por uma conexão direta (*shortcut*).

Figura 12 – Bloco de aprendizagem residual



Fonte: Adaptado de He et al. (2016)

Formalmente, o bloco residual é definido pela Eq. 2.8. Se a entrada e saída tiverem dimensões diferentes na rede, é utilizada uma projeção linear W_s Eq. 2.9.

$$y = F(x, W_i) + x \quad (2.8)$$

$$y = F(x, W_i) + W_s x \quad (2.9)$$

onde x e y são a entrada e saída das camadas, $W_s x$ é a projeção linear escolhida (convolução 1x1) para dimensões diferentes e a função $F(x, W_i)$ representa o mapeamento residual a ser otimizado. A introdução da conexão direta não introduz parâmetros extras e nem aumenta a complexidade computacional. A ResNet-50, por exemplo, possui 50 milhões de parâmetros, significativamente menor que a VGG-16 (BIANCO et al., 2018).

A Tabela 1 apresenta as arquiteturas do classificador ResNet-XX em diferentes níveis de profundidade. Importante destacar que o número ao lado do nome representa a quantidade de camadas que possuem parâmetros a serem ajustados. Além disso, o tamanho da imagem original (224 x 224) refere-se aos modelos gerados na competição ImageNet.

Tabela 1 – Arquiteturas criadas para classificadores ResNet.

Nome Camada	Tamanho imagem	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-152
conv1	112x112	Conv 7x7, 64 filtros, stride 2				
conv2_x	56x56	Conv 3x3, maxpooling 2x2, stride 2				
		$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28x28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 8$
conv4_x	14x14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
conv5_x	7x7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$

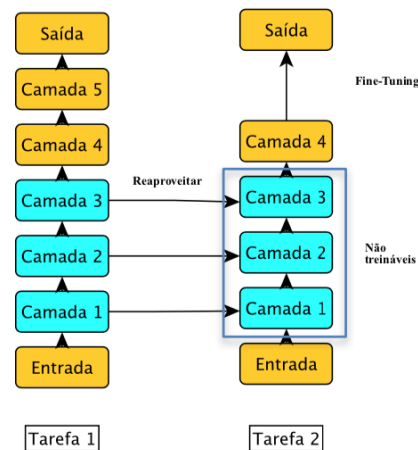
Fonte: Adaptado de (HE et al., 2016). Os blocos das Resnet50/101/152 são chamados de **bottleneck** - técnica adotada para manter a conexão identidade e projetar a entrada e saída do bloco residual (forma mais eficiente).

2.3.4 Transferência de aprendizado - *Transfer Learning*

Garcia-Garcia et al. (2018) explicam que o treinamento de uma rede neural profunda com pesos iniciais aleatórios (ou a partir de uma distribuição estatística), normalmente não é prática devido à falta de conjunto de dados extenso e o tempo necessário para alcançar a convergência na otimização. Mesmo que um conjunto de dados de grande extensão esteja disponível e a convergência seja alcançada em tempo razoável, é frequente o uso de pesos “pré-treinados” em vez de inicialização aleatória.

Goodfellow, Bengio e Courville (2016) definem *Transfer Learning*, termo utilizado anteriormente em (PRATT; MOSTOW; KAMM, 1991; CARUANA, 1995; BENGIO, 2011), como o aprendizado em um conjunto de dados que é utilizado para melhorar a generalização em outro conjunto. Yosinski et al. (2014) explicam que usualmente é um procedimento que utiliza n camadas mais "rasas" treinadas de uma rede base para uma nova. As camadas restantes da nova rede são iniciadas aleatoriamente e treinadas para uma nova tarefa. Esse fenômeno ocorre não somente em diferentes conjuntos de dados, mas também em objetivos de treinamento distintos (supervisionado e não-supervisionado), devido à generalidade das camadas mais baixas. E as camadas mais altas, por serem mais específicas na extração de características, são treinadas de forma a refinar (*fine-tuning*) os parâmetros do modelo para a nova tarefa (Fig. 13).

Figura 13 – Representação de *Transfer Learning* e *Fine Tuning*



Fonte: Adaptado de Géron (2019)

O estudo de Yosinski et al. (2014) no conjunto de dados da competição ImageNet tem sido utilizado como base para a realização de transferência de aprendizado em tarefas de classificação de cenas, localização de objetos e segmentação. Os autores mostraram que 2 fatores afetam negativamente a transferência de aprendizado: o problema de otimização nas camadas intermediárias que contêm *fragile co-adapted features*, características que interagem com outras de modo complexo ou frágil que não poderiam ser novamente ensinadas pelas camadas mais altas sozinhas; e o grau de especialização das camadas mais altas. Além disso, a inicialização com parâmetros pré-treinados pode melhorar a generalização do modelo mesmo após o refinamento na nova tarefa, principalmente se refinar a camada mais alta ou as duas últimas camadas.

O uso de transferência de aprendizado com pesos de redes pré-treinadas na competição ImageNet é frequente na segmentação semântica (GARCIA-GARCIA et al., 2018). As camadas do ramo de expansão são treinados aleatoriamente e todas as camadas refinadas para novo conjunto de dados. A inicialização com parâmetros pré-treinados tende

a aumentar a acurácia nos mapas de segmentação, evitar ajuste excessivo aos dados em pequenos conjuntos e permitir a convergência mais rápida do processo de otimização.

2.4 Segmentação Semântica

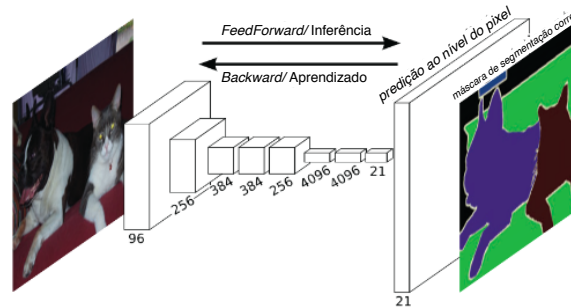
Segundo Ghosh et al. (2019) a segmentação de imagem é uma técnica de processamento digital que divide uma cena em 2 ou mais regiões (classes). Assim como pode ser entendida como um processo de determinação de bordas entre objetos (feições). Para cada *pixel* da imagem é atribuído um rótulo (*label*), e os *pixels* de mesmo rótulo estão conectados em alguma propriedade visual ou semântica. Além disso, possui diversas aplicações nas áreas médica, segurança, Sensoriamento Remoto, reconhecimento facial, direção autônoma, realidade aumentada entre outras.(GARCIA-GARCIA et al., 2018)

Como já explicado na Seção 1.2, na área de Visão Computacional existem 4 tarefas principais aplicadas em processamento de imagens (classificação de cena, detecção de objetos, segmentação semântica, segmentação por instâncias). Especificamente a segmentação semântica, Garcia-Garcia et al. (2018) apresentam a seguinte formulação para o problema de rotular cada *pixel*: encontrar uma forma de atribuir um estado do espaço de rótulos $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$ para cada um dos elementos de um conjunto de variáveis aleatórias $\chi = \{x_1, x_2, \dots, x_n\}$. Cada rótulo l representa uma classe ou objeto (sendo que existe a classe fundo de cena - *background* - ou classe vazia - *void*) . Normalmente, χ é uma imagem 2D de dimensões $W \times H = N$ pixels de x , mas pode ser estendida para multi/hiper dimensões.

A área de segmentação de imagens foi influenciada pelo desenvolvimento de DL. Principalmente pela efetividade no uso de convolução (produção de mapas de características - ativações internas) para segmentar imagens, bem como a diversidade de conjuntos de dados, desafios e competições sobre o assunto. (GHOSH et al., 2019)

Em aprendizado profundo, os algoritmos extraem características a partir de conjuntos de treinamento para predição com dados desconhecidos. Na predição de classe ao nível do *pixel*, o contexto é relevante. A segmentação de imagem é um problema de contexto multiescala. Ou seja, a informação contextual depende do tamanho e continuidade do segmento semântico a que o *pixel* pertence e da quantidade e densidade dos segmentos vizinhos de outras classes (HOESER; KUENZER, 2020).

O marco das CNN na segmentação semântica foi o estudo de Long, Shelhamer e Darrell (2015). A FCN, ilustrada na Figura 14, foi a primeira a substituir as camadas de neurônios totalmente conectados (classificador final), que perde a informação espacial da rede pela redução das dimensões espaciais dos mapas de características, para realizar a classificação direta de cada pixel(GHOSH et al., 2019).

Figura 14 – *Fully Convolutional Neural Network* (FCN)

Fonte: Adaptado de Long, Shelhamer e Darrell (2015)

Após o surgimento da FCN, diversos algoritmos foram desenvolvidos. Particularmente os modelos do tipo *Encoder-Decoder* têm obtido melhores resultados. Dentro do ramo de expansão os mapas de características, reamostrados (interpolação) ou reconstruídos por convolução, são combinados a mapas de maior resolução espacial por meio das conexões diretas (*skip connections*) da camada correspondente pertencente ao classificador. Ao final, predições de *pixels* são realizadas para a geração da máscara de segmentação (HOESER; KUENZER, 2020).

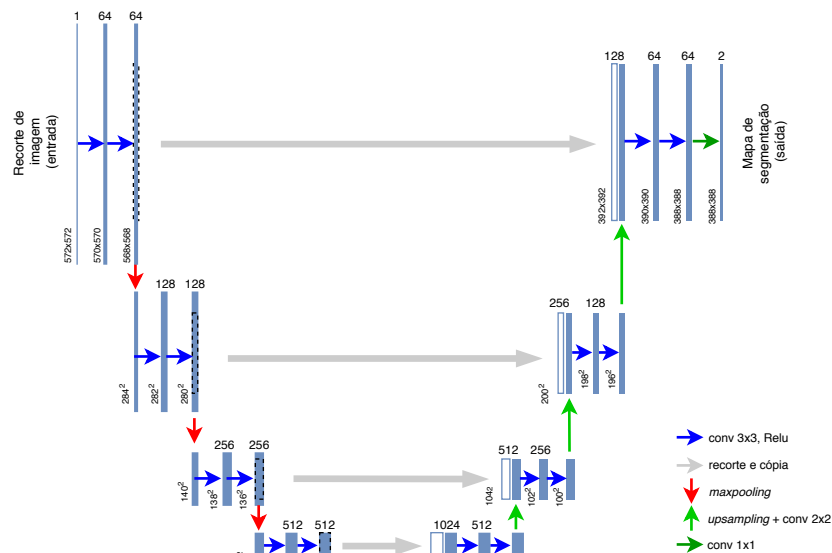
2.4.1 U-Net

A rede de segmentação semântica U-Net (RONNEBERGER; FISCHER; BROX, 2015) foi criada em 2015 para realizar a segmentação de neurônios em imagens microscópicas no ramo da Biomedicina. Foi inspirada na arquitetura FCN (LONG; SHELHAMER; DARRELL, 2015) e é a principal arquitetura do tipo *Encoder-Decoder* para segmentação Semântica. A parte de expansão possui muitos canais para que o contexto seja propagado às camadas de alta resolução. Por sua forma simétrica, tem desenho similar a um “U”, por isso o nome U-Net. Os autores utilizaram recortes (*tiles*) sobrepostos na imagem para manter o contexto na segmentação e também espelhamento de borda para solucionar o problema ao aplicar a convolução nas imagens. Outra estratégia foi o uso de aumento artificial dos dados para solucionar o problema de um conjunto de dados reduzido para treinamento do classificador. Essa estratégia permite que o aprendizado seja invariante às transformações geométricas realizadas no conjunto de dados. A escolha de uma função de perda entropia cruzada ponderada, no qual os *pixels* de borda próximos de auto-intersecção são penalizados, deve-se ao objetivo de separar as bordas das células que quase se interceptam na segmentação. Então os *pixels* de fundo recebem um peso alto na função.

A arquitetura U-Net (representada na Figura 15) é de fácil implementação e alteração, criada para segmentação binária e pode ser utilizada em pequenos ou grandes conjunto de dados. A rede é constituída por conjuntos de camadas convolucionais (*kernel*

3x3), ativadas por ReLU e operações de *max pooling* (2x2) com *stride* igual a 2 para reduzir a dimensionalidade espacial (impacto na resolução das imagens) no ramo de *encoder* em 5 níveis. Este ramo possui a arquitetura clássica da CNN para tarefas de classificação de cena. O classificador é o VGG-16 (SIMONYAN; ZISSERMAN, 2015), o mesmo utilizado pela arquitetura FCN. No ramo de expansão, cada camada possui processamento de reamostragem para aumentar a resolução espacial (*upsampling*) e convolução (2x2) seguido por concatenação de mapas de características oriundos do classificador (conexões diretas), além de operações convolucionais (3x3) ativadas pela Relu. Por fim, na última camada do ramo de expansão é realizada operação de convolução (1x1) ativada pela função *softmax* para mapear os 64 canais para a quantidade de classes desejada.

Figura 15 – Arquitetura U-Net



Fonte: Adaptado de Ronneberger, Fischer e Brox (2015)

A rede é organizada de forma a dobrar a quantidade de filtros convolucionais após cada operação de *pooling*. O recorte dos mapas de característica era executado para compensar o efeito de borda nas operações convolucionais e utilizavam imagens de entrada com tamanho 572x572 *pixels* e 1 banda espectral.

Por fim, importante destacar o uso de função de custo entropia cruzada binária (*binary cross entropy*) com penalização nos *pixels* de contorno ou borda, a inicialização dos pesos por distribuição gaussiana, regularização por *dropout* e aumento artificial de dados. A U-Net foi um marco para os tipos de arquitetura de segmentação em visão computacional.

2.4.2 Função de perda - Loss

A escolha da função de perda ou custo é importante para o processo de otimização em DL. Existem algumas mais utilizadas na literatura. A função *Binary Cross Entropy* (BCE) (Eq. 2.11) é uma das mais importantes pois mede a diferença entre duas distribuições de probabilidade. É o caso binário da função de *Cross Entropy* (CE) (Eq. 2.10).

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C g_i^c \log(s_i^c) \quad (2.10)$$

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C g_i^c \log(s_i^c) = -\frac{1}{N} \sum_{i=1}^N g_i \log(s_i) - (1 - g_i) \log(1 - s_i) \quad (2.11)$$

onde g representa o pixel de referência (mascara binária $[0;1]$), s a predição da rede para cada classe, C as classes e N o número de *pixels*.

Entretanto, a função BCE não possui a previsão de casos onde existe desbalanceamento de classes. Entre as opções para solucionar essa influência estão as funções custo que penalizam a classe mais frequente. No caso de detecção de objetos, Lin et al. (2020) criaram a *Focal Loss* (FL) (Eq. 2.12) caracterizada por penalizar as amostras que são “facilmente” classificadas pela rede.

$$FL = -\sum_{i=1}^N \sum_{c=1}^C (1 - s_i)^{\gamma} g_i^c \log(s_i^c) \quad (2.12)$$

Em segmentação binária de rodovias, os *pixels* da classe não-via predominam na cena. É um caso típico de desbalanceamento de classes, a classe mais frequente predomina no cálculo da função de perda e na acurácia final do modelo. Milletari, Navab e Ahmadi (2016) criaram a *DICE loss* (Eq. 2.13) e que se tornou, segundo Ma (2020), representante de um conjunto de funções que buscam maximizar a sobreposição de regiões de referência e a predição da rede.

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^N \sum_{c=1}^C g_i^c s_i^c}{\sum_{i=1}^N \sum_{c=1}^C g_i^{c2} + \sum_{i=1}^N \sum_{c=1}^C s_i^{c2}} \quad (2.13)$$

Salehi, Erdogmus e Gholipour (2017) e Abraham e Khan (2018) estenderam o conceito da Dice Loss. Eles criaram a *Tversky Loss* (Eq. 2.14) e a *Focal Tversky Loss* (Eq. 2.15), respectivamente. A função *Tversky* trabalha com balanceamento de falsos negativos e falsos positivos com hiperparâmetros α e β . Assim como a *Focal Tversky* utiliza o conceito da *Focal Loss* combinado com a *Tversky Loss*.

$$L_{tversky}(\alpha, \beta) = \frac{\sum_{i=1}^N \sum_{c=1}^C g_i^c s_i^c}{\sum_{i=1}^N \sum_{c=1}^C g_i^c s_i^c + \alpha \sum_{i=1}^N \sum_{c=1}^C (1 - g_i^c) s_i^c + \beta \sum_{i=1}^N \sum_{c=1}^C g_i^c (1 - s_i^c)} \quad (2.14)$$

$$L_{FocalTversky} = (1 - L_{Tversky})^{\frac{1}{\gamma}} \quad (2.15)$$

onde γ pertence ao intervalo de $[1, 3]$

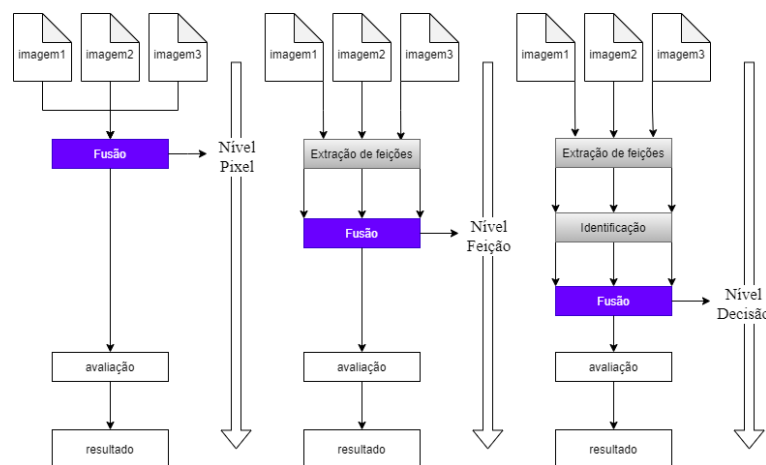
Existe ainda a possibilidade de combinação de funções. Por exemplo, a combinação (ou soma) das funções BCE e Dice Loss é frequentemente utilizada. Esta combinação busca alinhar distribuições e maximizar sobreposição de regiões de mesma classe.

3 Fusão de Imagens Sensoriamento Remoto

De acordo com Zhang (2010), a fusão combina dados de múltiplas fontes para agregar valor e melhorar a interpretação, produzindo uma representação visível de alta qualidade. Além disso, as técnicas de fusão são úteis para diversas tarefas, como detecção, identificação e classificação de objetos, detecção de mudanças e outros. Em sensoriamento remoto, especificamente, tem sido utilizado para integrar a informação adquirida com diferentes resoluções espaciais e espectrais de sensores embarcados em satélites, aeronaves e plataformas terrestres.

Pohl e Van Genderen (1998) explicam que a fusão de imagens é realizada em 3 diferentes níveis de processamento de acordo com o estágio da fusão: pixel, feição e decisão. Zhang (2010) esclarece que, ao nível do pixel, a fusão combina dados brutos de diferentes fontes em uma mesma resolução para ser mais informativo ou revelar mudanças nas cenas adquiridas em épocas distintas. Ao nível da feição, várias feições/características são extraídas (bordas, linhas e texturas) e combinadas em mapas que podem ser utilizados substituindo os dados originais. Ao nível da decisão, o resultado de vários algoritmos são combinados e decisões são realizadas por votação, estatística ou lógica fuzzy. A figura 16 ilustra esses diferentes níveis.

Figura 16 – Níveis de processamento da fusão de imagens



Fonte: Adaptado de Pohl e Van Genderen (1998)

Segundo Zhang (2010), a classificação feita por Pohl e Van Genderen (1998) não abrange todos os tipos de métodos. Na prática, a fusão é, frequentemente, uma combinação dos três níveis mencionados.

Quanto a aplicação, Pohl e Van Genderen (1998) indicam que a fusão é utilizada para melhorar a resolução espacial (*image sharpening*), em correções geométricas (p.ex. registro

imagens), complementar informações por visão estereoscópica, ressaltar características não visíveis nos dados individualmente, complementar conjunto de dados para melhorar a classificação, detecção de mudanças, além de substituir áreas de falta de informação.

Entre as várias fontes de dados que podem ser fusionadas estão os sensores passivos na faixa do visível e os sensores ativos LiDAR. Zhang (2010) afirma que esta fusão tem sido explorada por várias aplicações, desde geração de Modelos Digitais de Superfície e Terreno (MDS/MDT), modelagem e extração tridimensional de objetos até mapeamento de cobertura do solo.

3.1 Princípios básicos do LiDAR

Vosselman e Hans-Gerd (2010) afirmam que desde o início dos anos 70, sistemas de sensoriamento remoto com LiDAR tem sido utilizados em várias áreas. Ashraf, Hur e Park (2017) citam, como exemplo, as áreas de planejamento de transporte, imageamento e visualização.

Sob o enfoque medição de distâncias, os termos de *LAser Detection And Ranging* (LADAR) e LiDAR se confundem. Wehr e Lohr (1999) explicam que o uso de LADAR garante o uso de LASER como fonte de energia e LiDAR pode utilizar outra fonte (p.ex: xenon). Outro ponto relevante é a diferença dos termos LASER e LiDAR, onde o primeiro refere-se ao dispositivo emissor de pulso e o segundo à técnica de sensoriamento remoto que utiliza a luz como fonte para medir distâncias.

As subseções seguintes apresentam conceitos básicos sobre levantamento aéreo com LiDAR, a integração com sensores passivos e a resposta radiométrica do sensor ativo (intensidade do sinal).

3.1.1 Sistema de Varredura a LASER Aerotransportado - SVLA

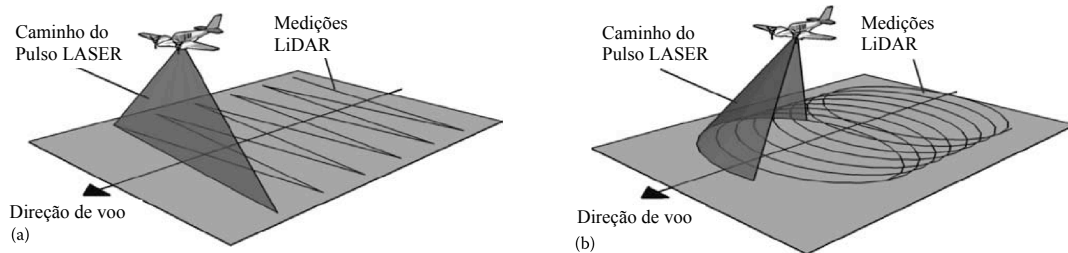
O Sistema de Varredura a LASER Aerotransportado (SVLA) ou *Airborne Laser Scanner* (ALS) é um sistema de aquisição de dados terrestres que utiliza o método de varredura a LASER sob uma plataforma aérea. O sistema de varredura pode ser utilizado em plataformas terrestres (veículos), aérea (aeronaves tripuladas ou não tripuladas - Veículo Aéreo Não-Tripulado (VANT)) ou orbital (satélites).

Wehr e Lohr (1999) afirmam que a varredura LASER utiliza um conjunto óptico-mecânico para direcionar os pulsos LASER e gerar um padrão de varredura no terreno. O sensor ativo emite pulso na direção da superfície terrestre e recebe parte deste no retorno após determinado tempo. O sensor mede a intensidade do sinal e o intervalo de tempo entre a emissão do pulso e a captação do retorno.

O mecanismo de varredura óptica utiliza um conjunto de lentes e espelhos oscilantes

(ou fibra óptica) que direcionam os pulsos LASER com ângulos conhecidos ao longo da direção transversal à linha de voo. Cada mecanismo de varredura possui um padrão de amostragem¹ no terreno. A Figura 17 apresenta dois tipos de amostragem existentes. (PETRIE; TOTH, 2018)

Figura 17 – Padrões de amostragem dos Mecanismos de Varredura

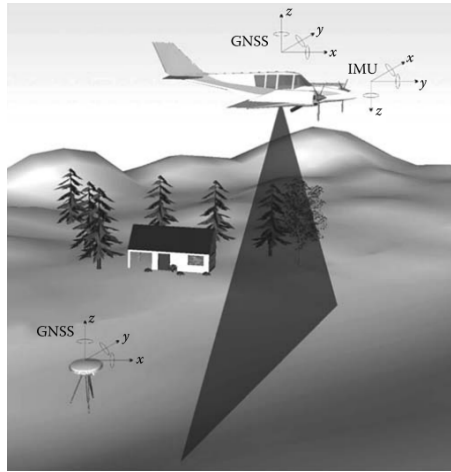


Fonte: Adaptado de BOSSLER e HABIB (2010). a) padrão “zig-zag” oriundo do mecanismo espelho oscilante ; b) padrão elíptico oriundo do espelho nutante - sistema Palmer

Normalmente, o SVLA é constituído pelos seguintes componentes: unidade LASER (transmissor, receptor, amplificador, detector de sinal, contador de tempo e demais componentes), mecanismo de varredura óptica, unidade eletrônica para processamento e controle do sistema, subsistema de posicionamento *Global Navigation Satellite System* (GNSS) e orientação *Inertial Measurement Unit* (IMU), estação de referência terrestre, *storage* para armazenamento dos dados medidos, *software* para planejamento, coordenação/control operacional e processamento preliminar dos dados em voo. De forma simplificada, o mecanismo de varredura direciona o pulso LASER na direção do terreno, a unidade LASER mede a distância percorrida pelo pulso (transmitido e recebido), os subsistemas de posicionamento GNSS e orientação IMU obtêm as posições e orientações da tomada de distância LASER em determinado instante. Além disso, outros sensores de imageamento podem ser utilizados conjuntamente (PETRIE; TOTH, 2018; BALTSAVIAS, 1999). A Figura 18 ilustra os principais componentes do sistema.

¹ Detalhes sobre mecanismos de varredura podem ser encontrados em (WEHR; LOHR, 1999; PETRIE; TOTH, 2018; WEHR, 2018)

Figura 18 – Principais Componentes do SVLA



Fonte: BOSSLER e HABIB (2010)

Os valores de distâncias com LASER podem ser realizadas com medidas diferença de tempo entre pulsos ou diferença de fase do sinal contínuo. No caso específico da pesquisa, os dados foram obtidos a partir de medidas por diferença de tempo entre pulsos. O sensor mede a intensidade do sinal de retorno e o intervalo de tempo entre a emissão e recepção do pulso refletido. No caso do SVLA, o sensor óptico encontra-se embarcado na aeronave e o objeto pertence a superfície terrestre. Wehr e Lohr (1999) apresentam a Equação 3.1 utilizada no cálculo da distância.

$$R = \frac{1}{2} \cdot t \cdot c \quad (3.1)$$

onde t é o tempo medido a partir da emissão até a recepção do sinal, R é a distância (*Range*) e c é a velocidade da luz.

Wehr (2018) salienta que existem alguns parâmetros importantes, pertencentes à geometria de aquisição de dados, no planejamento de voo com o sistema de varredura LASER. Os pontos iluminados no terreno (emitidos e recebidos) necessitam estar em um mesmo campo de visada (θ) ou ângulo de abertura máximo (FOV) do receptor a uma altura média de voo (H). A divergência do feixe emitido/recebido (ângulo γ) define o campo de visada instantâneo (*Instantaneous Field Of View* – IFOV) que depende do diâmetro (d) da abertura do emissor do pulso e o comprimento de onda do feixe (λ). Quanto maior a distância entre o sensor/objeto e o ângulo divergente (γ), maior será o diâmetro (D) da projeção do feixe no solo (*footprint*). As Equações 3.2 a 3.4 mostram os valores aproximados do diâmetro (D) do *footprint* e da faixa de varredura (S_w) sob um terreno plano horizontal e ângulo (γ) pequeno. (WEHR; LOHR, 1999; MACHADO, 2018; SHAN; TOTH, 2018; SANTOS, 2019)

$$\gamma = 2.44 \frac{\lambda}{d} \quad (3.2)$$

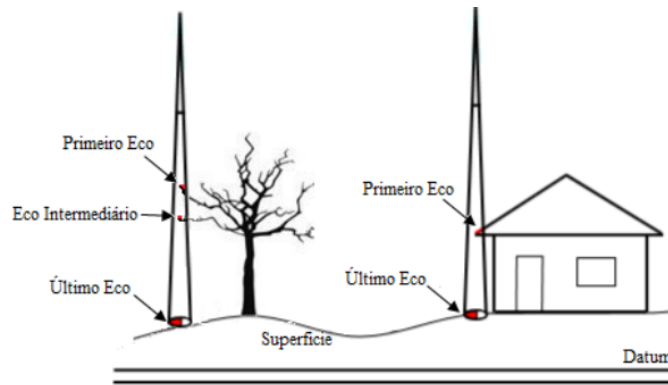
$$D = \frac{H}{\cos^2(\beta)} \gamma \quad (3.3)$$

$$S_w = 2H \tan\left(\frac{\theta}{2}\right) \quad (3.4)$$

onde β é o ângulo de varredura instantâneo, $\beta \in [0, \pm \frac{\theta}{2}]$

Santos (2019) afirma que devido ao ângulo de divergência, o pulso emitido pode ser refletido por outros objetos gerando múltiplos retornos (“ecos”). Frequentemente, esse efeito ocorre com vegetação e com edificações. A figura 19 ilustra múltiplos retornos de um mesmo pulso.

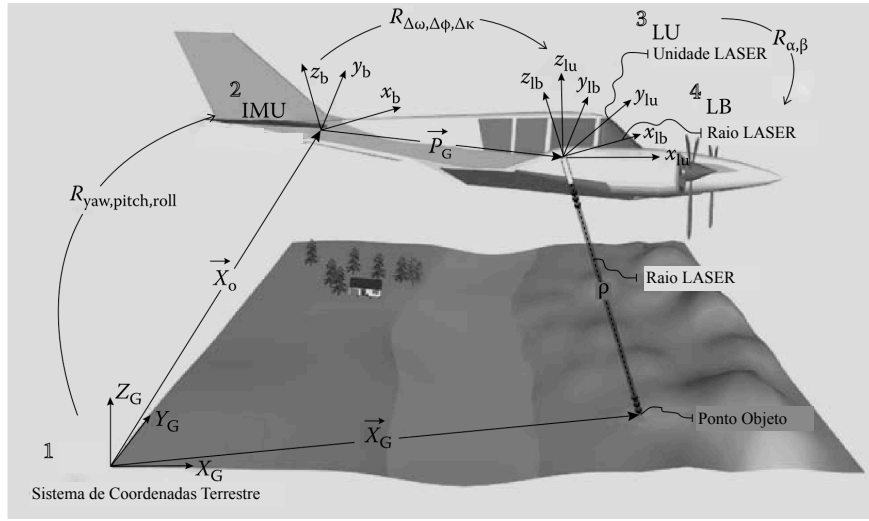
Figura 19 – Casos de múltiplos retornos



Fonte: Galo (2018)

Quanto modelo matemático utilizado, a Equação 3.5 (EL-SHEIMY; VALEO; HABIB, 2005) modela o relacionamento entre quatro sistemas para obtenção das coordenadas de terreno do ponto iluminado pelo pulso: 1) Sistema de coordenadas terrestre; 2) Sistema de coordenadas GNSS/IMU - origem no centro do sistema de coordenadas do IMU; 3) Sistema de coordenadas da unidade LASER - origem no ponto de disparo do LASER; 4) Sistema de coordenadas do pulso LASER - origem no ponto de disparo do LASER e eixo z coincidente com a direção do feixe. A Figura 20 ilustra o relacionamento geométrico entre os sistemas.

Figura 20 – Componentes de SVLA e suas relações geométricas



Fonte: Adaptado de BOSSLER e HABIB (2010).

$$\vec{X}_G = \vec{X}_o + R_{yaw,pitch,roll} \vec{P}_G + R_{yaw,pitch,roll} R_{\Delta\omega,\Delta\phi,\Delta\kappa} R_{\alpha,\beta} \begin{bmatrix} 0 \\ 0 \\ -\rho \end{bmatrix} \quad (3.5)$$

onde o primeiro vetor $\vec{X}_o$ é o vetor da origem no sistema de coordenadas terrestre (*Ground Coordinate System* - GCS) até a origem do sistema IMU, $\vec{P}_G$ é o vetor de translação entre o sistema da unidade LASER (*LASER Unit* - LU) e o do IMU, o terceiro $\vec{p}$ é o deslocamento entre o ponto de disparo do LASER (*LASER Beam* - LB) e o ponto objeto. A soma dos três vetores, após a aplicação das rotações ($R_{yaw,pitch,roll}$, $R_{\Delta\omega,\Delta\phi,\Delta\kappa}$, $R_{\alpha,\beta}$), determina o vetor $\vec{X}_G$ que é a coordenada do ponto objeto. ((BOSSLER; HABIB, 2010)

El-Sheimy, Valeo e Habib (2005), BOSSLER e HABIB (2010) definem os componentes da Equação 3.5:

- $\vec{X}_G$ - vetor de coordenadas no sistema terrestre do ponto objeto iluminado pelo pulso LASER;
- $\vec{X}_o$ - vetor de coordenadas no sistema terrestre da origem do sistema de coordenadas GNSS/IMU;
- $R_{yaw,pitch,roll}$ - matriz de rotação relacionada aos sistemas de coordenadas terrestre e do GNSS/IMU;

- $\vec{P}_G$ - vetor de translação da origem dos sistemas de coordenadas GNSS/IMU e da unidade LASER. Esse vetor é chamado de *lever arm offset*;
- $R_{\Delta\omega, \Delta\varphi, \Delta\kappa}$ - matriz de rotação relacionada aos sistemas de coordenadas GNSS/IMU e da unidade LASER. Essa matriz é chamada de *boresight matrix*;
- $R_{\alpha, \beta}$ - matriz de rotação relacionada aos sistemas de coordenadas da unidade e do pulso LASER;
- $\vec{p}$ - vetor de coordenadas no sistema do pulso LASER do ponto objeto.

Em relação ao armazenamento das informações, a *American Society for Photogrammetry and Remote Sensing* (ASPRS) criou um padrão de estrutura de dados chamado *LiDAR Exchange Format* (formato .LAS). Esse formato foi criado para intercâmbio dos dados entre os diferentes sistemas e usuários. Atualmente a versão é a 1.4. (ASPRS, 2019)

3.1.2 Integração LiDAR-Fotogrametria

Nos últimos anos, a tecnologia LiDAR está mais acessível, principalmente pelo custo menor e pela rapidez e custo-benefício de aquisição de dados 3D. A alta acurácia, o acesso a sistemas GNSS / IMU e a versatilidade dos sensores LiDAR são razões para difusão do seu uso em atividades de mapeamento.

Como exemplo de aplicação, os sistemas LiDAR aerotransportados podem ser utilizados de forma complementar aos sistemas passivos de forma a aumentar a acurácia e melhor descrever o espaço objeto. A Tabela 2 mostra limitações e potencialidades de cada uma das tecnologias.

Tabela 2 – Comparativo LiDAR x Sensor Passivo

LiDAR		Sensor Passivo	
Potencial	Limitação	Potencial	Limitação
Densa informação em superfícies homogêneas	Sem redundância	Alta redundância	Quase sem informação em superfícies homogêneas
Coleta dia e noite	Somente informação posicional e dificuldade em extrair semântica	Rica informação semântica	Coleta somente durante o dia
Aquisição direta de coordenadas tridimensionais	Quase sem informação em <i>break lines</i>	Informação espacial densa em <i>breaklines</i>	Complicados métodos de restituição 3D dos pontos
Acurácia altimétrica maior que planimétrica		Acurácia planimétrica maior que altimétrica	

Fonte: Adaptado de Shan e Toth (2018), Baltsavias (1999)

Por ser um sensor ativo e possuir um ângulo de abertura menor que as câmeras aéreas, o LiDAR possibilita a minimização dos efeitos de sombras e oclusões. Enquanto no sensor passivo, as bordas dos objetos são frequentemente melhor definidas. Isto ocorre devido ao LiDAR obter uma representação amostral do terreno, consequentemente não garante que os pontos obtidos coincidam com as bordas. Sob o aspecto semântico, as

câmeras aéreas permitem o registro da energia em diferentes bandas (p.ex: RGB) e o sensor ativo, normalmente monoespectral, registra somente a informação de intensidade de retorno. (SHAN; TOTH, 2018; SANTOS, 2019)

BOSSLER e HABIB (2010) destacam ainda que sistemas LiDAR são ferramentas de aquisição de dados de alta acurácia planialtimétrica e de forma ágil. Possuem capacidade de penetração em copas de árvores e possibilidade de mapeamento ao nível do solo. E permitem mapeamento em superfícies de pouca ou nenhuma textura (como p.ex: gelo, neve ou deserto).

O aproveitamento entre as potencialidades dos dois sensores depende da integração entre os sistemas. Os dados necessitam estar **co-registrados**² em um mesmo sistema de coordenadas. Após isto, a imagem RGB e a informação posicional do LiDAR (MDS e MDT) podem ser utilizadas no processo de ortorretificação (retificação das imagens para a projeção ortogonal). A ortorretificação é necessária pois existe distorção na imagem devido à variabilidade do relevo e altura dos objetos, além da inclinação do sensor no instante da aquisição (desalinhamento entre o objeto e a lente do sensor). Ortofotos ou Ortoimagens tem escala uniforme e representam todos os objetos em correta posição planimétrica (GHARIBI; HABIB, 2018).

A projeção ortogonal é essencial para ser utilizada em mapeamentos topográficos. Principalmente pela recuperação da correta posição planimétrica. Uma breve revisão sobre métodos de ortorretificação, detecção de oclusão e reconstrução de terreno pode ser encontrada no artigo de Gharibi e Habib (2018).

3.1.3 Imagem Intensidade

Os sensores LiDAR fornecem informações altimétricas e radiométricas dos pontos levantados. Esta última obtida a partir da intensidade do pulso laser recebido. A intensidade é definida como a razão da magnitude do sinal da luz refletida pela emitida sendo influenciada principalmente pela reflectância da superfície imageada. O uso da intensidade do LiDAR tem se destacado no registro de dados, extração de feições, classificação, segmentação, análise de superfícies e detecção/reconhecimento de objetos. (SONG et al., 2002; KASHANI et al., 2015)

O sinal refletido é registrado na faixa do Infravermelho Próximo (*Near Infrared - NIR*), normalmente no comprimento de onda igual a $0,9 \mu\text{m}$. A partir das informações registradas, é possível gerar a imagem intensidade da área. É considerada imagem de alta resolução espacial, não sofre influência de sombra (sensor ativo) e tem sido utilizada em métodos de classificação para distinguir diferentes alvos, principalmente de forma

² Ghanma (2006), Kumar Mishra (2012) apresentam vários métodos de co-registro imagem-pontos LiDAR

complementar aos dados multiespectrais no espectro do visível. (EL-ASHMAWY; SHAKER; YAN, 2012; SONG et al., 2002)

De forma a exemplificar a distinção dos materiais pela intensidade refletida, Song et al. (2002) apresentaram os diferentes percentuais de reflectância em quatro classes. Classe Asfalto possui taxas que variam entre 10-20%, Grama em torno de 50%, Árvore entre 30-60% e Telhado 20%-30%. Brennan e Webster (2006) ainda ressaltam que essa diferença na reflectância é útil quando objetos possuem alturas próximas, assim como os dados altimétricos mais efetivos nas diferentes alturas entre classes. Kashani et al. (2015) afirmam que a adição da informação de intensidade aos dados das superfícies imageadas pode acarretar um aumento de 10 a 20% na acurácia da classificação orientada a objeto de cobertura do solo.

Alguns fatores influenciam a qualidade das imagens intensidade. Kashani et al. (2015) agrupa em 4 categorias: característica da superfície do alvo, geometria de aquisição dos dados, efeitos de instrumentos ou do ambiente. Sem a influência dos outros fatores, valores de intensidade aumentam quanto maior for a reflectância da superfície dos objetos, exceto em superfícies altamente reflectivas (ex: espelho, vidro e água), estes podem saturar os valores de intensidade. Sobre a geometria, quanto maior a distância e o ângulo em relação ao alvo, maior ruído no sinal refletido, consequência do retroespalhamento do feixe de LASER e do ângulo de reflexão. Song et al. (2002) destacam que o ângulo é a principal fonte do ruído e influencia os valores de intensidade dos materiais. Além disso, diferentes sensores, com diferentes configurações também influenciam os valores de intensidade, bem como a atmosfera da região a ser mapeada e a umidade. Por fim, ainda é possível citar que a esparsidade dos pontos, consequência da densidade dos pontos coletados, influencia na resolução espacial da imagem intensidade final.

3.2 Fusão por CNN

Recentemente, o desenvolvimento e a popularidade de sensores que captam a informação de profundidade de cena cresceram. Isto devido ao surgimento de câmeras RGB-D (equipamentos de sensoriamento de profundidade que são integrados ao sensor RGB de câmeras) mais acessíveis e portáteis, bem como de conjuntos de dados *benchmark*. (HU; CHEN; LIN, 2018; GARCIA-GARCIA et al., 2018)

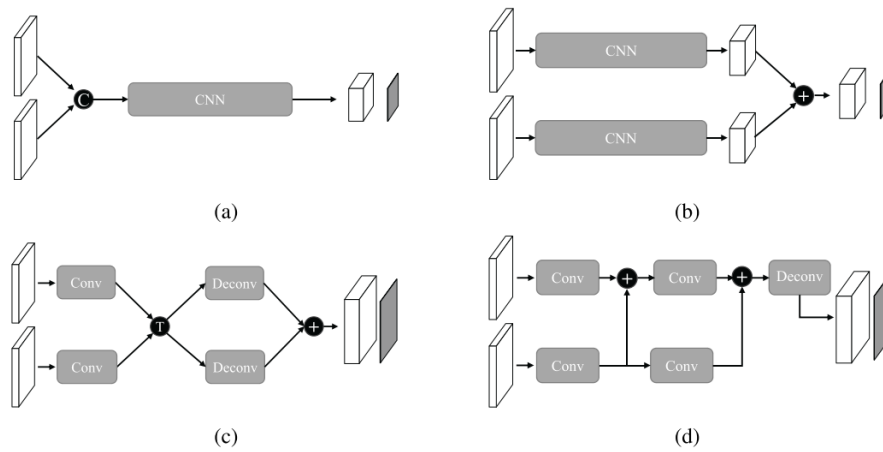
Hu, Chen e Lin (2018) afirmam que métodos clássicos necessitam de conhecimento prévio e muitas regras para extração de características principais. Os resultados de DL em tarefas de visão computacional motivaram pesquisadores a explorar redes convolucionais em segmentação semântica.

A adição de informação de profundidade a imagem RGB poderia melhorar a performance da segmentação semântica. O canal de profundidade tem informação complementar

aos canais RGB com informação estrutural da cena. Com isso, poderia diminuir a incerteza na segmentação de objetos que tenham aparência similar. (HAZIRBAS et al., 2017)

Uma das formas mais triviais na fusão de dados de profundidade é adicionar canais a imagens RGB, aumentando a dimensionalidade da entrada. Porém, Hazirbas et al. (2017) dizem que isso não permite explorar totalmente a informação estrutural presente na cena. Lee, Park e Hong (2017) categorizam as formas de fusão por CNN em 2 tipos: fusão inicial e multinível.

Figura 21 – Diferentes arquiteturas de Fusão CNN para segmentação de imagens RGB-D



Fonte: (LEE; PARK; HONG, 2017). a) Fusão inicial, b-c-d) variantes de fusão multinível. 'C', 'T', '+' representam Concatenação de camadas, Transformações e soma de nós.

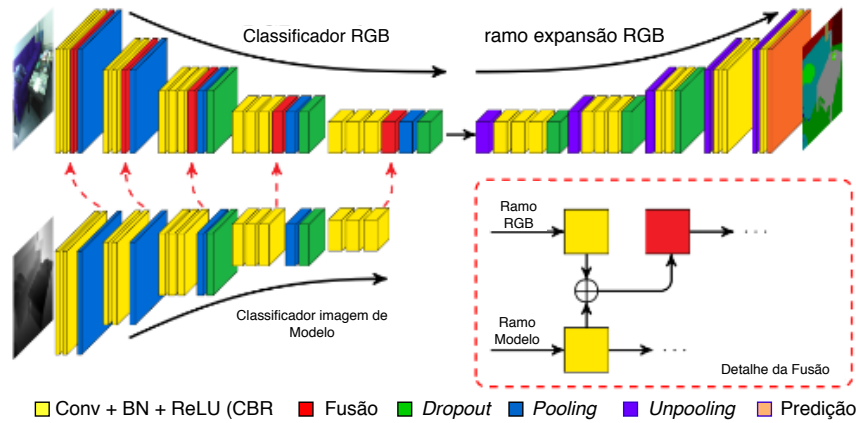
A Figura 21 ilustra as diferentes formas. O item a) apresenta a fusão inicial que realiza a fusão dos mapas de características ao nível de baixa abstração da rede (nível próximo da entrada de dados da rede). Os demais itens da figura apresentam tipos de fusão multinível que resultam em diferentes arquiteturas.

3.2.1 FuseNet

A rede convolucional FuseNet foi criada por Hazirbas et al. (2017). O objetivo da rede foi realizar um estudo de como utilizar a informação complementar de profundidade (dados altimétricos) em segmentação semântica. O estudo utilizou dados de cenas internas (*indoor*) de residências, as quais são mais complexas que cenas externas (*outdoor*), para segmentação de objetos.

A arquitetura escolhida para a FuseNet (Fig. 22) é do tipo *Encoder-Decoder*. Entre as várias arquiteturas deste tipo, a FuseNet foi inspirada na SegNet (BADRINARAYANAN; KENDALL; CIPOLLA, 2017), principalmente pelo ramo de expansão. A contribuição para a fusão das imagens consiste em utilizar um outro classificador para extrair as informações de profundidade paralelo ao classificador RGB e combinar os mapas de características extraídos.

Figura 22 – Arquitetura FuseNet

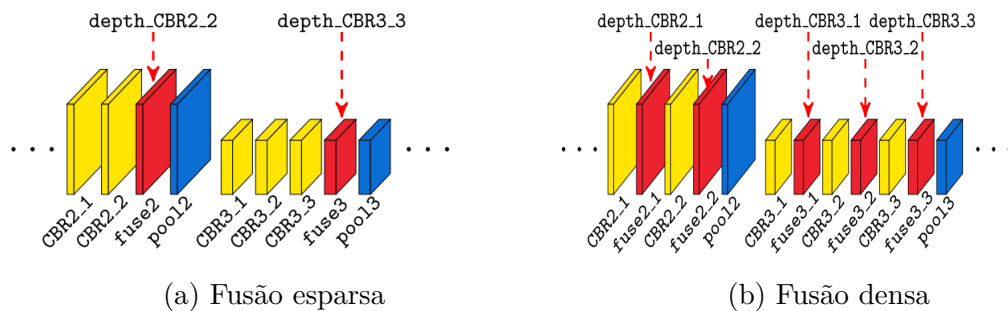


Fonte: Adaptado de Hazirbas et al. (2017)

Os ramos de codificação possuem o classificador VGG-16 com camadas de BN. Além disso, várias conexões diretas entre os ramos foram inseridas para a adição dos mapas de características. O ramo de expansão é o utilizado na SegNet. A principal característica é a utilização na expansão da técnica de recuperar as posições dos *pixels* amostrados no classificador e reamostrar na reconstrução do mapa de segmentação (*unpooling*).

Outro aspecto relevante no artigo de Hazirbas et al. (2017) é o estudo da quantidade de conexões diretas, bem como sua localização. Os autores analisam duas abordagens de fusão: conexões diretas a cada camada convolucional (Fusão densa - Figura 23b) ou somente antes da camada de *pooling* (Fusão esparsa - Figura 23a). Assim como se a adição dos mapas na rede deve ser realizada antes ou após a ativação ReLU.

Figura 23 – Estratégias de Fusão na FuseNet



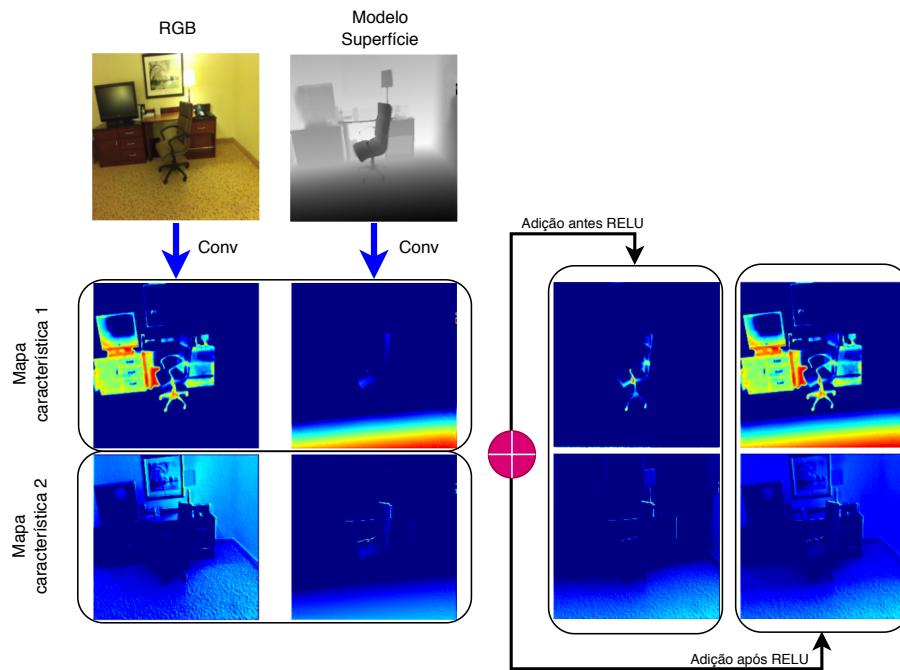
Fonte: Adaptado de Hazirbas et al. (2017).

Foram realizados 3 experimentos. O primeiro demonstrou que a abordagem da rede atingiu, e algumas vezes ultrapassou, a acurácia na segmentação multiclases do estado da arte. O segundo testou a quantidade de conexões diretas (1 a 5) entre os ramos e as abordagens de adicionar canais na imagem de entrada quanto a acurácia global. E o último experimento testou 4 canais (RGB-D) na imagem de entrada e as melhores abordagens

da FuseNet (testadas nos 2 experimentos anteriores) em 37 classes do *benchmark* SUN RGB-D (SONG; LICHTENBERG; XIAO, 2015).

A abordagem com 5 conexões diretas do tipo fusão esparsa, antes da ativação ReLU, apresentou os melhores resultados, principalmente com a métrica IoU. A Figura 24 apresenta a análise realizada com dois mapas de características do ramo RGB e de profundidade.

Figura 24 – Comparação de mapas de característica de cada ramo na primeira camada convolucional



Fonte: Adaptado de Hazirbas et al. (2017)

No treinamento da rede foram utilizadas técnicas de regularização BN, *dropout* e *L2_norm* dos pesos ou decaimento de pesos (*weight decay*). A função de perda escolhida foi a entropia cruzada e a função *softmax* destaca a probabilidade da classe em cada *pixel* no mapa da segmentação. Além de uso de pesos pré-treinados do conjunto de dados ImageNet. A arquitetura da FuseNet se tornou referência na fusão de imagens por CNN para segmentação semântica.

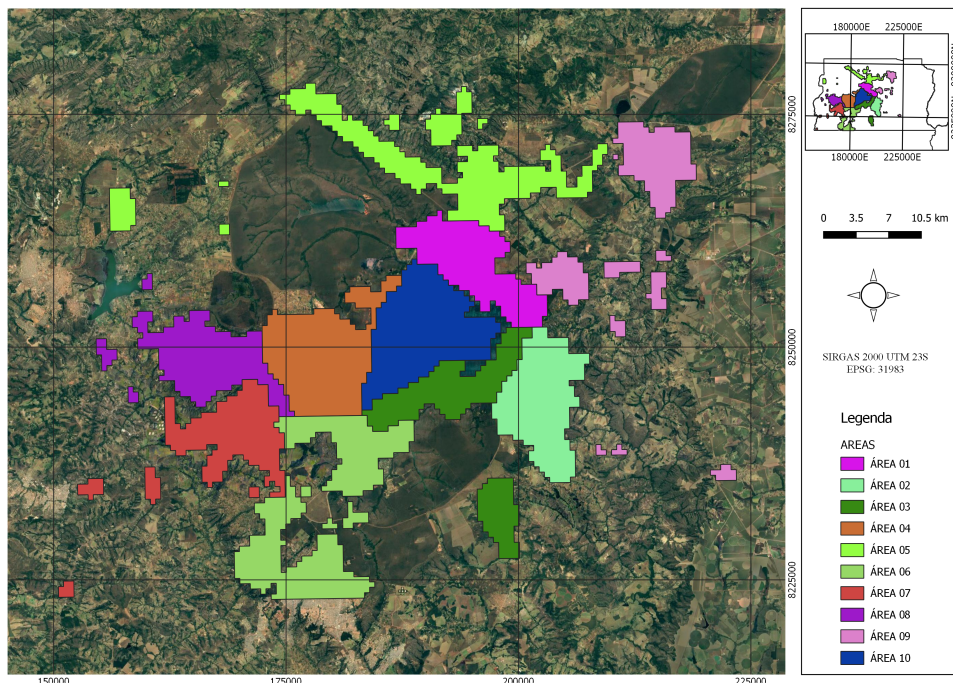
4 Materiais e Métodos

Este capítulo descreve a metodologia empregada na pesquisa. Etapas de pré-processamento, verificação dos dados, implementação das redes e o método proposto estão descritos nas seções e subseções.

4.1 Área de Estudo

A cidade de Brasília é uma grande metrópole com mais de 2 milhões de habitantes, pertencente ao estado do Distrito Federal (DF), capital do país. O último mapeamento da área urbana foi realizado no ano de 1997 na escala de 1:2.000. Em virtude da desatualização da base cartográfica, o governo estadual realizou a contratação de empresa técnica especializada para executar o levantamento aerofotogramétrico com sensores ópticos (passivo e ativo) em 10 áreas urbanas distribuídas pelo estado (Fig. 25). Cada área tem uma taxa de urbanização, densidade populacional e características específicas.

Figura 25 – Regiões urbanas pertencentes ao Distrito Federal



Fonte: De autoria própria

O levantamento aéreo corresponde a uma área total de 1200 km² e foi realizado no período de março a maio de 2016. O sistema de referência geodésica adotado é SIRGAS 2000, sistema de projeção *Universal Transversal of Mercator* (UTM) fuso 23S. Em relação

a nuvem de pontos, o referencial altimétrico é o Marégrafo de Imbituba - SC com o modelo geoidal local. As demais características técnicas estão contidas na Tabela 3.

Tabela 3 – Principais características do levantamento aéreo fotogramétrico

Característica	LiDAR	RGB
Sensor	ALS-60 Leica	Ultracam-XP
Altura média de voo	720 m	1400 m
Altitude média de voo	1820 m	2500 m
Densidade de pontos	4 pts/m ²	-
Superposição faixas/fotos	lateral 20%	Longitudinal 70±3% e lateral 40±3%
Direção de voo	Várias	Norte - Sul ou Leste - Oeste
Produtos	MDS e MDT (grade 50 cm)	Fotografias com <i>Ground Sample Distance</i> (GSD) de 8cm , resolução radiométrica de 8bits por banda. Ortofoto com GSD de 10cm

A região possui uma grande densidade de vias distribuídas nas áreas urbanas. Aspecto importante para a realização de treinamento do classificador, bem como testes e análises. Ademais, possui extensas áreas arborizadas com muitas oclusões de vias, bem como a altíssima resolução espacial das imagens permite visualizar diversas obstruções ocasionadas por veículos.

4.2 Coleta de dados e recursos computacionais

Os dados utilizados na tese são as ortoimagens e a nuvem de pontos LiDAR adquiridas no ano de 2016 por meio de voo fotogramétrico no Distrito Federal, como descrito na Seção 4.1. O banco de dados geoespacial adquirido a partir desses insumos também foi utilizado como a referência da extração das rodovias. Além disso, foram coletados os conjuntos de dados chamados “*Benchmark*”: ISPRS¹ Postdam (CRAMER, 2010), bem como o de Massachusets (MNIH, 2013). Esses conjuntos de dados são referência para diversos artigos e foram utilizados para que as redes convolucionais, implementadas nesta tese, tenham comparação com o estado da arte. As especificações de todos os conjuntos de dados utilizados estão na Tabela 4.

Tabela 4 – Especificações Técnicas dos conjuntos de dados.

Conjunto de dados	Quantidade Imagens	Tamanho Imagens(px)	Conjunto Teste	Resolução Espacial	Canais (espectrais ou auxiliares)	Classes
ISPRS Postdam	38	6000x6000	14	5cm	RGB+IR+MDS	6
Massachusets	1171	1500x1500	49	1m	RGB	2
Ortoimagem DF	>3700	6000x6000		10cm	RGB+MDS+MDT+I	2

Os recursos de hardware e software pertencem aos laboratórios de Fotogrametria do Departamento de Cartografia e do Cluster DMC (dois computadores com Processador Intel(R) Core(TM) i7-11700F @ 2.50GHz , Memória 128GB RAM DDR4 , Disco Rígido: SSD 1TB , um equipado com a placa de vídeo Nvidia RTX 3090 24GB DDR6x e outro com

¹ <https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/>

RTX 3080 10GB DDR6x) pertencente ao Departamento de Matemática e Computação. A Tabela 5 detalha os softwares ou bibliotecas utilizados com a linguagem de programação Python 3.7.

Tabela 5 – Principais softwares/bibliotecas para auxílio à pesquisa.

software/ biblioteca	Versão	Descrição	Referencia
Keras	2.x	Framework para DeepLearnig	(CHOLLET et al., 2015)
Tensorflow	2.x	Interface de Programação de Aplicação (API) para Machine Learning	(ABADI et al., 2015)
Docker Engine	20.x	Ambiente de desenvolvimento de aplicativos em containers e microserviços	(MERKEL, 2014)
CloudCompare	2.x	Software de Visualização e Processamento de Nuvem de pontos	(CLOUDCOMPARE, 2020)
QGIS	3.14.x	Sistema de Informações Geográficas	(QGIS Development Team, 2009)
PDAL	2.x	Biblioteca de conversão e manipulação de Nuvem de pontos	(CONTRIBUTORS, 2018)
GDAL	3.x	Biblioteca para conversão e processamento de dados raster e vetoriais geoespaciais	(GDAL/OGR contributors, 2021)

Ressalta-se que os computadores descritos, instalados no Cluster DMC, foram adquiridos em 2021 por meio de recursos do projeto FAPESP nº 2021/03586-2.

4.3 Pré-processamento dos dados

Todos os conjuntos de dados necessitam de processamento para serem utilizados na rede convolucional. A normalização das imagens é necessária para que a entrada tenha valores entre $[0,1]$ do tipo *float* para melhorar a convergência na otimização da rede convolucional. Outro processo comum no preparo, é a correção de *pixels nodata*. O preparo de cada conjunto de dados será explicado com mais detalhes nas subseções seguintes.

4.3.1 ISPRS Potsdam *Dataset*

O conjunto de dados da cidade de Potsdam, Alemanha, possui características de cidade histórica com grandes blocos de edificações, vias estreitas e densamente povoada. O conjunto é constituído por 38 recortes de ortofotos de mesmo tamanho gerados pela BSF Swissphoto² no ano de 2008. A Figura 26 mostra a distribuição dos recortes. Área total de 3,4km².

Cada recorte possui GSD de 5 cm e combinação RGB, RG-IR ou RGB-IR com resolução radiométrica de 8bits por banda. Assim como modelo digital de superfície (MDS), modelo de superfície normalizado (MDS-MDT = MDS_n) e máscaras de segmentação para treinamento e teste. (ROTTENSTEINER et al., 2013)

² <https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/>

Figura 26 – Conjunto de dados ISPRS Potsdam



Fonte: Rottensteiner et al. (2013)

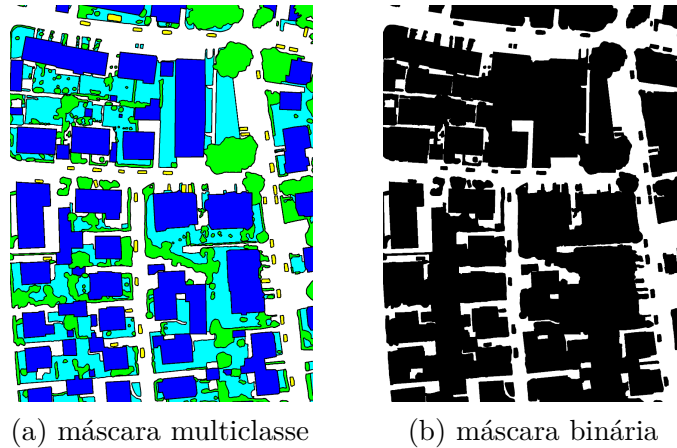
Ressalta-se que as máscaras possuem 6 classes de objetos:

- Superfícies Impermeáveis (*Impervious surfaces*) (RGB: 255, 255, 255)
- Prédio (*Building*) (RGB: 0, 0, 255)
- Vegetação baixa (*Low vegetation*) (RGB: 0, 255, 255)
- Árvore (*Tree*) (RGB: 0, 255, 0)
- Carro (*Car*) (RGB: 255, 255, 0)
- Fundo de cena (*Clutter/background*) (RGB: 255, 0, 0)

A classe via está incluída em superfícies impermeáveis juntamente com outras feições (p.ex: estacionamentos). Além disso, a classe superfícies impermeáveis apresenta “buracos”, tais como copa de vegetação e carros. Um procedimento de matemática de bandas foi realizado para separação da classe de superfícies e convertê-la em máscara binária. A figura 27 apresenta o resultado da classe isolada na cena.

As ortoimagens foram reamostradas para 10cm de modo a inserir contexto da via em recortes de tamanho 600x600 *pixels* com sobreposição vertical e horizontal de 50% e normalização para valores de pixel entre [0,1].

Figura 27 – Máscara binária da classe Superfícies Impermeáveis



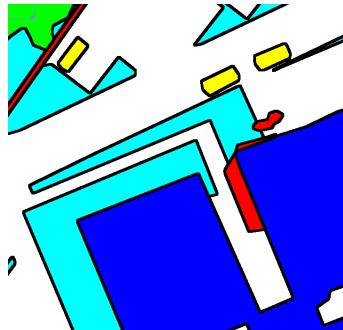
Fonte: De autoria própria.

As imagens disponibilizadas estão georreferenciadas, *pixels nodata* foram interpolados e não existem muitos problemas de distorção nas imagens devido a ortorretificação. (p.ex: Figura 28).

Figura 28 – Problema máscara no conjunto de dados Potsdam



(a) ortoimagem distorcida pelo processo de ortorretificação.



(b) atribuição de classe Fundo no erro de ortorretificação em vermelho.

Fonte: De autoria própria.

Inicialmente, a pesquisa buscou utilizar as máscaras de vias confeccionadas por Kaiser et al. (2017). Entretanto, problemas de georreferenciamento são significativos e frequentes nos recortes (Fig. 29). Consequentemente, o uso foi descartado.

Figura 29 – Erros nas máscara de vias disponibilizadas por Kaiser et al. (2017)



Fonte: De autoria própria.

4.3.2 Massachusetts *Dataset*

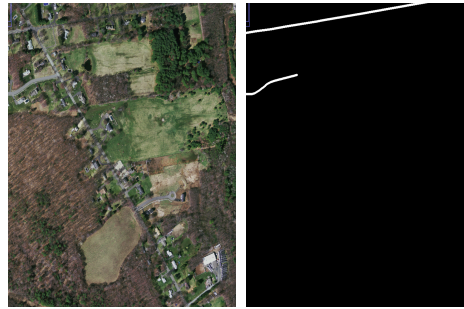
O conjunto de dados Massachusetts foi um dos primeiros a ser utilizado para extração de rodovias com o uso de aprendizado profundo. Mnih (2013) reuniu recortes de ortoimagens disponibilizadas pelo Estado de Massachusetts³ oriundas de imageamento aerofotogramétrico do ano de 2013.

Para a extração de rodovias, o autor utilizou 1171 recortes reamostrados para 1m de GSD, com bandas RGB e resolução radiométrica de 8 bits. Dados georreferenciados no sistema geodésico NAD83 e sistema de projeção UTM, recobrimento total de 2600km². Adicionalmente, foram geradas máscaras binárias de vias a partir dos dados OSM.

Mnih (2013) afirma que as imagens de teste e validação foram corrigidas manualmente. Entretanto, o conjunto de treinamento foi mantido com erros. Embora a rede convolucional seja robusta com alguns problemas de máscara, foram encontradas 215 cenas com erros grosseiros (p.ex: Fig. 30), aproximadamente 20% do conjunto de treino, valor significativo e que demandou a retirada dos recortes.

³ <https://www.mass.gov/orgs/massgis-bureau-of-geographic-information>

Figura 30 – Exemplo de problemas existentes no conjunto de dados de Massachusetts



Fonte: De autoria própria.

Como a reamostragem foi realizada pelo autor, somente os processos de recorte no tamanho 600x600 *pixels* sem sobreposição e normalização $[0,1]$ foram realizados.

4.3.3 Distrito Federal *Dataset*

A área do conjunto de dados do DF, descrito na Seção 4.1, compreende um total de aproximadamente 1200km². Além disso, foi objeto de rigoroso controle de qualidade nos dados (FILHO et al., 2020). As máscaras foram criadas a partir dos dados vetoriais oficiais disponibilizados pela Companhia Imobiliária de Brasília (TERRACAP)⁴, empresa pública distrital.

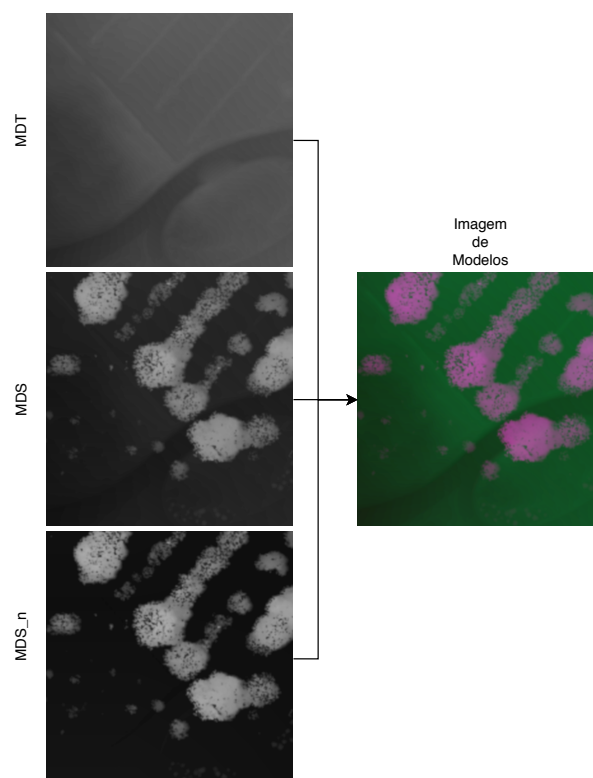
O conjunto original apresenta imagens que seguem o enquadramento do mapeamento sistemático do Sistema Cartográfico do Distrito Federal (SICAD). Devido à área do DF estar localizada entre os fusos UTM 22 e 23, o SICAD utiliza o fuso 23 estendido.

De forma semelhante aos outros conjuntos de dados, as imagens RGB foram reamostradas para GSD de 20cm, recortadas para tamanhos de 600x600 *pixels* (sem sobreposição) para a extração da rede viária sem fusão. Enquanto que os modelos MDS e MDT, originalmente com GSD de 50cm, foram escalados para 8 *bits* e reamostrados por *spline* cúbica para 20cm e recortados para tamanho de 600x600 *pixels* sem sobreposição.

Cada modelo foi adicionado a uma banda da imagem final. Tal procedimento visa fornecer à rede convolucional mais informações sobre as classes do terreno. A Figura 31 ilustra um exemplo de imagem de modelos oriundos do LiDAR.

⁴ <https://www.terracap.df.gov.br/>

Figura 31 – Imagem de modelos MDT,MDS,MDSn

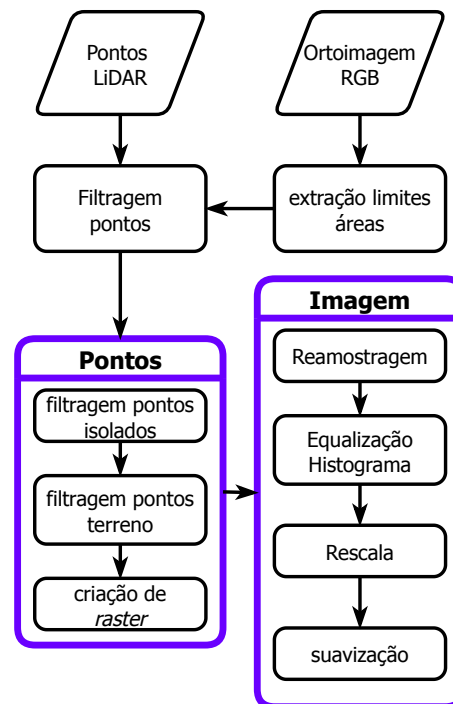


Fonte: De autoria própria.

As imagens intensidade possuem baixa resolução espacial devido ao ruído. Adicionalmente, no caso do DF com 4pts/m^2 , a característica de pontos esparsos pela baixa densidade não permite alta acurácia no contorno de objetos. Todos os pontos pertencentes ao conjunto de dados possuem atributo de informação de intensidade do sinal do LASER com resolução radiométrica de 16 bits e sem normalização. Informações técnicas como linha de voo utilizada e parâmetros para calibração/normalização não foram disponibilizados.

Com isso, alguns métodos foram aplicados para manipulação do conjunto de pontos LiDAR, geração das imagens intensidade e redução do ruído. O Fluxograma (Fig. 32) ilustra as etapas seguidas na pesquisa para obtenção das imagens.

Figura 32 – Fluxograma de Geração das Imagens Intensidade



Fonte: De autoria própria.

A etapa inicial compreende a filtragem dos pontos do conjunto que estão contidos no limite dos recortes das orthoimagens RGB. Posteriormente, dois algoritmos de filtragem foram aplicados: remoção de pontos por média da distância a 32 vizinhos mais próximos e *Simple Morphological Filter* (SMRF) (PINGEL; CLARKE; MCBRIDE, 2013) para classificação de pontos no terreno. Estes pontos são usados para criação de um *grid* com valores interpolados pelo algoritmo ponderação pelo inverso da distância (*Inverse Distance Weighting* - IDW). A aplicação de processos de reamostragem, equalização de histograma, escala e suavização foram utilizados para compatibilizar com a resolução espacial das orthoimagens, reduzir a discrepância entre os histogramas das cenas, transformar a resolução das imagens para 8 *bits* e reduzir ruído. Essa sequência foi aplicada em todas as 10 áreas do DF com os parâmetros descritos na Tabela 6.

Tabela 6 – Parâmetros de pré-processamento imagens intensidade

Parâmetro	Algoritmo SMRF				Imagem Intensidade		Reamostragem	Equalização	Filtro Gaussiano
	cell size (m)	slope tolerance (%)	maximum window (m)	elevation threshold (m)	scalar	resolution (m)	resolution (m)	limit	standard deviation
Valor	0,4	0,15	33	0,5	1,25	0,4	0,2	0,3	3

Pingel, Clarke e McBride (2013) apresentaram o algoritmo SMRF em 4 etapas principais: geração de superfície mínima; classificação de cada célula da superfície como terreno ou não ; criação de modelo digital de terreno temporário a partir destes pontos classificados e filtragem de pontos originais com base no relacionamento com o modelo temporário. Inicialmente, a superfície mínima é construída a partir do parâmetro *resolution* e de todos

os pontos de menor elevação no conjunto de dados. Ao mesmo tempo, uma operação morfológica de *opening* com elemento estruturante de disco é aplicada progressivamente na superfície mínima com raio que varia de 1 *pixel* ao parâmetro *maximum window*. De forma iterativa, a diferença entre a superfície mínima e a resultante de operação morfológica, em cada célula, são comparadas com um limiar calculado (multiplicação dos parâmetros *slope tolerance*, *windows radius* e *cell size*) e a célula classificada como terreno ou não em uma máscara binária. Ao final, a máscara resultante é aplicada na superfície mínima para a geração de um modelo digital de terreno provisório que é utilizado na medição de distância vertical entre os pontos originais e o modelo. A filtragem dos pontos é realizada entre a diferença e o limiar estabelecido (parâmetros *elevation threshold* + *scalar* multiplicado pela declividade calculada em cada ponto).

Após a classificação automática dos pontos, um *grid* de resolução radiométrica de 16 *bits* foi criado por interpolação a partir de valores do atributo intensidade do sinal dos pontos. Nesta etapa, a imagem intensidade tem baixa resolução espacial em relação às ortoimagens e pouco contraste. Processos de interpolação por spline cúbica, equalização adaptativa do histograma (*Contrast Limited Adaptive Histogram Equalization* - CLAHE), transformação de escala para 8 *bits* e suavização por filtro gaussiano foram utilizados para compatibilizar a resolução, aumentar o contraste, ajustar a escala radiométrica e reduzir o ruído da imagem resultante.

Recentemente, o artigo de Chen et al. (2021) realizou um estudo comparativo entre vários algoritmos de filtragem de pontos de terreno. Sendo o algoritmo SMRF apresentou melhor performance em áreas urbanas com baixo custo computacional.

Por fim, devido à grande quantidade de dados existente e a limitação de tempo na realização de experimentos, optou-se por utilizar as áreas 1 e 10. A primeira possui diversidade de vias e menor densidade, conseqüentemente menor oclusão e vias dispersas, enquanto a última possui maior diversidade e densidade de vias que todas as outras, com isso maior possibilidade de oclusão de vias (conforme Fig. 33). A escolha foi realizada com base nos índices de diversidade H' (índice de *Shannon–Wiener*) (Eq. 4.2) e densidade D (Eq. 4.1) de vias nos dados oriundos do OSM (ZHANG et al., 2015).

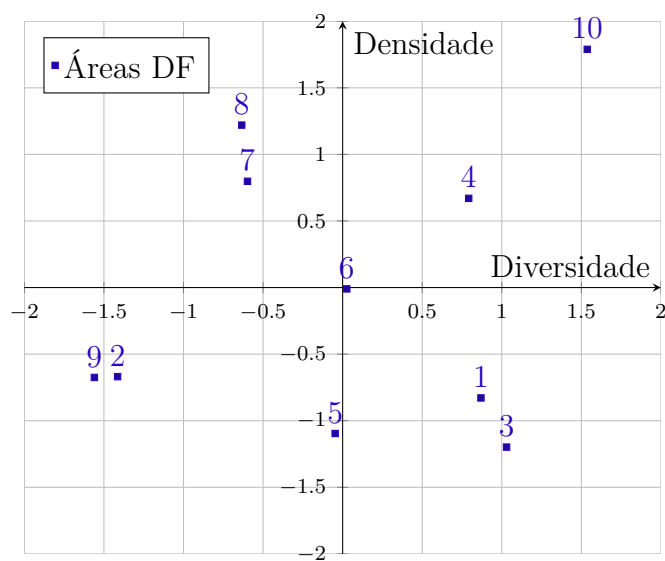
$$D_x = \frac{L_x}{A_x}, \quad x \in (1, 2, \dots, 10) \quad (4.1)$$

onde x representa cada uma das 10 áreas do DF, L_x é o comprimento total das vias OSM presentes na área x e A_x representa a área total da região x .

$$H' = - \sum_{i=1}^n P_i \ln P_i, \quad P_i = \frac{n_i}{N}, \quad i \in (1, 2, 3, \dots, n) \quad (4.2)$$

onde i representa cada tipo (classe) de via OSM, n os tipos de vias existentes na área, n_i

Figura 33 – Gráfico de dispersão de índices Diversidade x Densidade de vias



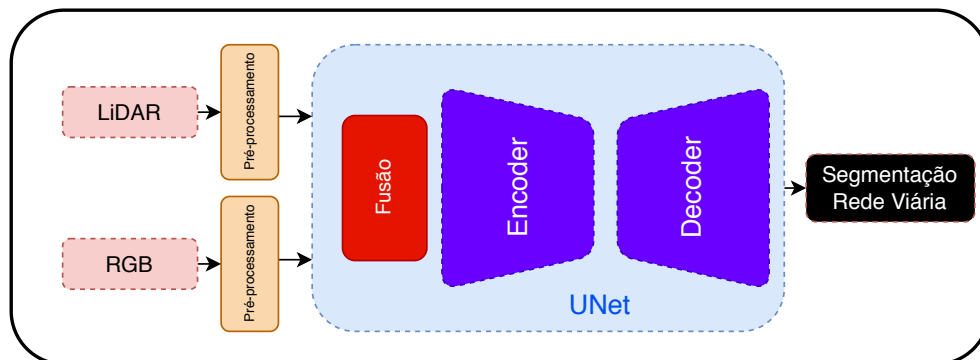
Fonte: De autoria própria.

o comprimento total das vias na classe i , N é o comprimento total de todas as vias na área e P_i é a importância relativa da classe de via i .

4.4 Método Proposto

A Figura 34 ilustra, de forma sumarizada, o método proposto na pesquisa para que a fusão de imagens (ortoimagens + LiDAR) por DCNN possa contribuir com informações complementares nas áreas de oclusões e obstruções na extração automática de rede viária em ambiente urbano. A metodologia envolve a realização de pré-processamento de dados, inspeção de hiperparâmetros e fusão de dados no classificador da arquitetura Unet.

Figura 34 – Metodologia geral da pesquisa



Fonte: De autoria própria. RGB - entrada de dados Ortoimagens sensor passivo; LiDAR - entrada de dados com imagens intensidade ou imagem modelos (MDS,MDT,MDSn) do sensor ativo ; Fusão - inicial ou multinível; UNet com os classificadores VGG-16 ou ResNet-50.

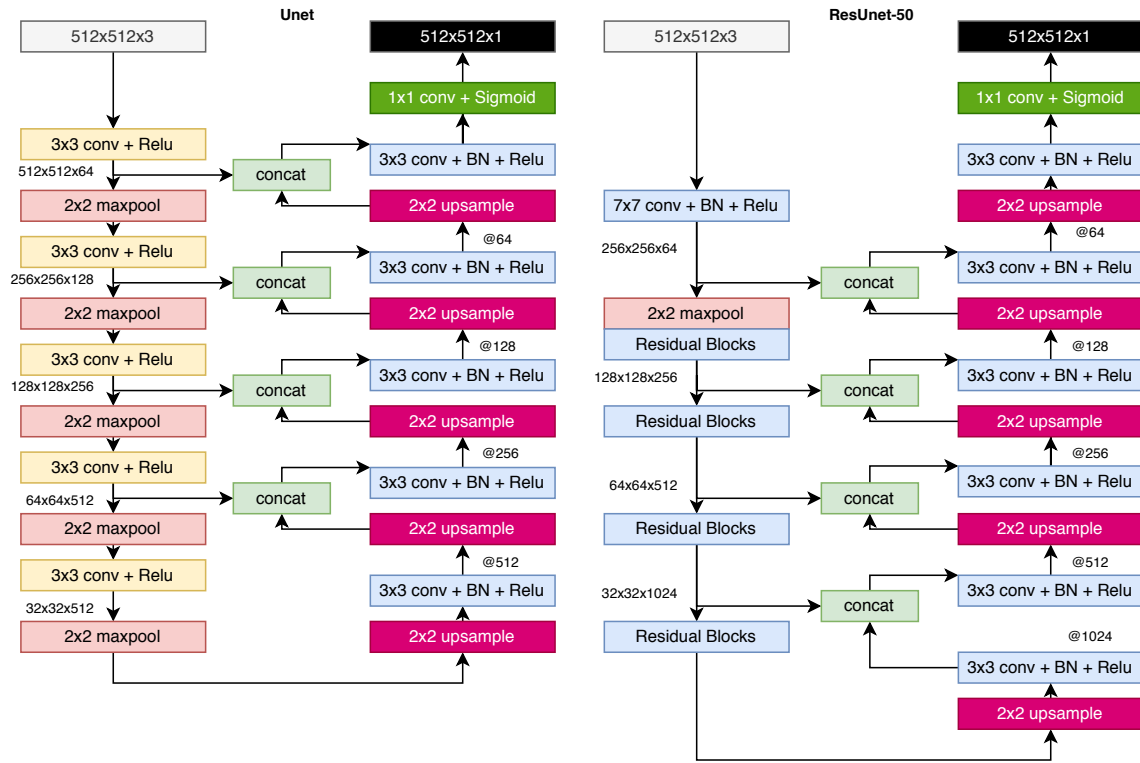
A Tese utiliza a arquitetura UNet com algumas modificações. Por exemplo, a utilização dos classificadores VGG-16 (original UNet) e ResNet-50 (também chamada de ResUnet), inserção de BN para regularização dos valores de mapas de característica, imagens de entrada com tamanho e canais diferentes da original, recortes aleatórios (*random crop*), outras transformações para aumento artificial de dados (espelhamento horizontal, vertical e rotação de 90°), utilização da *dice loss*, *Tversky* ou *Focal Tversky Loss* para o desbalanceamento com a classe não rodovia e utilização de transferência de aprendizado de redes treinadas no conjunto ImageNet.

Todas as adaptações foram realizadas de forma a aproveitar as melhores práticas de treinamento das redes convolucionais profundas. O uso de BN visa a normalização da distribuição dos valores obtidos nos filtros a cada sub amostra (*batch*) antes da função de ativação. Isso permite que o treinamento alcance a convergência de forma mais ágil e se obtenha maior acurácia na otimização. O uso de aumento artificial de dados permite aumentar a generalização do modelo e evita o superajuste aos dados de treinamento.

As redes UNet e ResUnet são apresentadas na Figura 35 com os classificadores VGG-16 e ResNet-50. A especificação detalhada da arquitetura encontra-se no repositório GitHub⁵.

⁵ https://github.com/tunofilho/thesis2022_data/tree/main/metadata

Figura 35 – Arquiteturas UNet e ResUnet adaptadas



Fonte: De autoria própria. Nas camadas do classificador são exibidas a resolução dos mapas de características; @ - representa dimensão (quantidade de mapas de característica) nas camadas de expansão; concat - concatenação; conv - convolução; maxpool - *maxpooling*.

Embora tenha mais camadas convolucionais, o classificador ResNet-50 possui arquitetura menos complexa (quantidade FLOPs⁶) para a otimização, além de criar conexões residuais entre as camadas convolucionais. Isso permitiu a criação de rede mais profunda com alta acurácia do classificador e auxiliou na propagação do gradiente da função custo pelas camadas.

Quanto à fusão de dados, Heydari e Mountrakis (2019) apontam diversas estratégias, entre as quais foram adotadas as de fusão multinível por meio de ramo paralelo e inicial no 1º nível do classificador de dados RGB. Essas adaptações na arquitetura U-Net para a criação de um ramo paralelo foi inspirado na FuseNet de Hazirbas et al. (2017) e a abordagem de adicionar outros canais na imagem RGB sem ramos paralelos no estudo de Arief et al. (2018). O propósito é fornecer informações de localização e geometria (primeiro ramo - entrada RGB) e de contexto (segundo ramo - imagens oriundas LiDAR) para complementação de informações na discriminação da classe rodovia.

A fusão inicial, frente a de ramo paralelo, adiciona menor quantidade de parâmetros na rede. Ou seja, somente no 1º nível serão adicionadas camadas de convolução e ativação

⁶ Floating Point Operations - quantidade de operações em ponto flutuante de multiplicações e adições entre vetores e matrizes, mais detalhes em: <https://mediatum.ub.tum.de/doc/625604/625604>

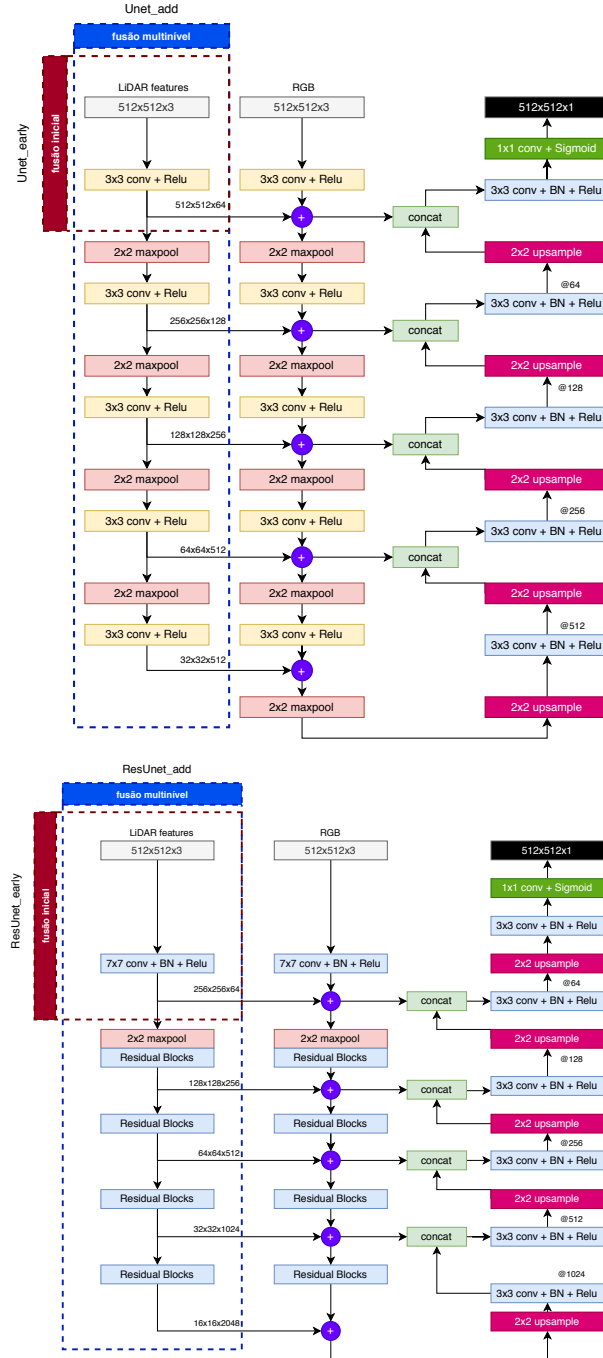
à segunda imagem de entrada, os demais níveis do classificador não serão alterados (quantidade de filtros permanece idêntica a extração sem fusão). Além disso, acredita-se que a separação em duas imagens de entrada possa melhorar a extração de características em vez de simplesmente adicionar bandas à imagem RGB. A 2ª abordagem utiliza a técnica de fusão da FuseNet para melhorar a ativação da função não-linear. A utilização das duas abordagens da fusão visam encontrar a melhor forma de aproveitamento da complementariedade dos dados LiDAR na predição de rodovias nas áreas de oclusão/obstrução por objetos ou sombras. A Figura 36 ilustra as duas estratégias em cada arquitetura (Unet e ResUnet).

A etapa de treinamento é realizada primeiramente com a inspeção dos hiperparâmetros (taxa de aprendizado, tamanho das subamostras, otimizador da função custo, recorte das imagens e distribuição para inicialização das camadas). No caso de uso de pesos pré-treinados, a quantidade de canais nas imagens de entrada deve obedecer o pré-treinamento de classificação (padrão de 3 canais R-G-B) ou então utilizar técnicas de adaptação nos pesos (não realizada). Outro detalhe importante refere-se a adequação da infraestrutura disponível. Por exemplo a memória disponível na placa de vídeo, leitura e escrita de arquivos, tempo de treinamento, entre outros.

O treinamento deve ser realizado de modo a evitar problemas de ajuste excessivo aos dados de treinamento e de falta de ajuste aos dados. As amostras de treino são efetivamente utilizadas para o ajuste da rede a partir dos hiperparâmetros. Enquanto os dados de validação são utilizados para avaliar a predição da rede durante o treinamento, de forma que os melhores hiperparâmetros sejam escolhidos. Por fim, as amostras de teste permitem verificar a acurácia do modelo em dados não utilizados no treinamento e a capacidade de generalização nestes novos dados.

A validação do método é composto por duas etapas. Na primeira etapa, a extração de vias é realizada a partir das ortoimagens sem o auxílio das informações do sensor LASER. Na etapa final, a fusão é efetuada com os dados imagem oriundos do LiDAR. A separação em duas etapas visa compreender a contribuição dos dados LiDAR na extração de rodovias na área urbana.

Figura 36 – Abordagens para fusão CNN



Fonte: De autoria própria. Nas camadas do classificador são exibidas a resolução dos mapas de características; @ - representa dimensão (quantidade de mapas de característica) nas camadas de expansão; concat - concatenação; conv - convolução; maxpool - *maxpooling*.

4.5 Avaliação dos resultados

A análise dos resultados obtidos será qualitativa (visual) e quantitativa com uso de métricas *Pixel Accuracy* (PA) (Eq. 4.3), *F-Measure* ($F1_{score}$) (Eq. 4.6), *Intersection over Union* (IoU) (Eq. 4.4). Arief et al. (2018) afirmam que PA é mais utilizada para

problemas de classificação de cenas, além disso não é uma boa métrica para problemas com classes desbalanceadas (o que ocorre na segmentação semântica), porém ainda é utilizada. As métricas $F1_{score}$ e IoU são extensamente utilizadas nos problemas de segmentação semântica, inclusive nos casos de classes desbalanceadas.

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.3)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (4.4)$$

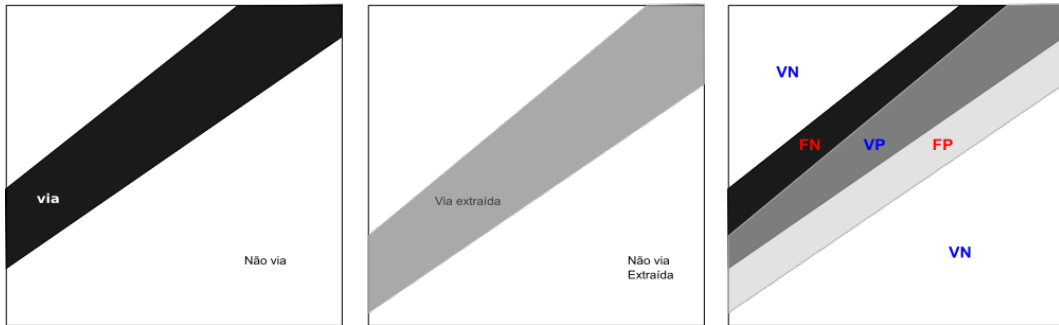
$$p = \frac{TP}{TP + FP}, \quad r = \frac{TP}{TP + FN} \quad (4.5)$$

$$F1_{score} = 2 \cdot \frac{p \times r}{p + r} = \frac{TP}{TP + \frac{FP + FN}{2}} \quad (4.6)$$

onde $F1_{score}$ é a média harmônica de $Precision(p)$ e $Recall(r)$.

onde FN (Falso Negativo) é o número de pixels da classe "via" classificados na classe "não-via", FP (Falso Positivo) os *pixels* de "não-via" classificados em "via", TP (Verdadeiro Positivo) e TN (Verdadeiro Negativo) os *pixels* classificados corretamente nas classes "via" e "não-via", respectivamente (Fig. 37).

Figura 37 – Comparação das classes extraídas (predição) com os dados de teste



TN - Verdadeiro Negativo; TP - Verdadeiro Positivo; FN - Falso Negativo e FP - Falso Positivo.

5 Resultados e Análises

Neste capítulo, os resultados dos testes nos conjuntos de dados são apresentados e discutidos. Com esse propósito, algumas combinações de hiperparâmetros foram preliminarmente testadas. Em cada conjunto de dados foram utilizadas variações na taxa de aprendizado (10^{-2} , 10^{-3} , 10^{-4}), quantidade total de iterações (Eq. 5.1) ou total de atualizações dos parâmetros na rede, classificadores (ResNet-50, VGG-16), otimizadores (SGD, ADAM), tipos de regularização (*Dropout*, *L2Norm* e GC), funções de perda (*dice*, *tversky*, *focal tversky* e *bce + dice*) para classes desbalanceadas, decaimento polinomial da taxa de aprendizado ao longo do treino (Eq. 5.2) e uso de transferência de aprendizado (parâmetros pré-treinados no conjunto ImageNet).

$$iteracoes = \frac{(n^\circ \text{ recortes treino})}{(n^\circ \text{ subamostras})} \times (n^\circ \text{ épocas}) \quad (5.1)$$

$$Lr = (Lr_i - Lr_f) \cdot \left(1 - \frac{Ep}{Ep_t}\right)^{0.9} + Lr_f, \quad Ep \in (0, 1, \dots, Ep_t - 1) \quad (5.2)$$

onde Lr representa a taxa de aprendizado calculada na época Ep ; Lr_i e Lr_f taxas de aprendizado inicial e final; Ep_t o total de épocas.

O uso de regularização na rede possibilitou maior iteração com os dados sem o risco de superajuste. Entretanto, o aumento de dimensionalidade, mais parâmetros ou classificador com mais camadas, não implicam necessariamente aumento da acurácia. Assim, todos os resultados apresentados nas seções de experimentos utilizam as melhores combinações dos hiperparâmetros mencionados acima.

Outro fator relevante é a reconstrução da imagem original. A literatura apresenta algumas estratégias (HUANG et al., 2018), tais como: média ou união de predições com recortes sobrepostos (geralmente sobreposição de 50%, também chamado de *grid shifting* (ZHAO; WANG; ZHANG, 2017)); média de predições de vários modelos (*ensemble models*); média de predições com aumento artificial do conjunto teste (*test augmentation*); entre outros. No caso da predição da segmentação de rodovias, a escolha da reconstrução interfere na acurácia resultante e pode causar descontinuidades. De forma a reduzir essa influência, todas as redes realizam a predição no tamanho da imagem original do conjunto teste. Ou seja, se a imagem original possui dimensões de 1500x1500, a predição será obtida nas mesmas dimensões sem "costura" (*stitching*) de recortes menores. Ressalta-se ainda que a mudança das dimensões da entrada dos dados na rede convolucional para treinamento altera a acurácia final, entretanto tal fato é discutido no experimento III.

5.1 Experimento I

Este experimento analisa a influência da banda do infravermelho (IR) na extração de via. Para isto, o conjunto de dados de ISPRS Potsdam foi utilizado devido à disponibilidade de IR e à altíssima resolução espacial.

Em virtude do conjunto ser de referência na literatura, a quantidade de imagens de treino (total de 24 amostras) e teste (total de 14 amostras) foram definidas anteriormente pelos autores. Então, as imagens de treino foram separadas na proporção de 90% para treinamento e 10% para validação e embaralhadas na entrada da rede. O otimizador utilizado foi SGD com *momentum* de 0,9, taxa de aprendizado inicial de 0,01 e final de 0,0001, regularização por GC, n° subamostras = 4 e aumento artificial dos dados com rotação de 270° e escolha aleatória das seguintes transformações na imagem : espelhamento vertical/horizontal, recortes de dimensões 512x512.

Após o pré-processamento, o conjunto de referência possui 1701 recortes para treinamento, 108 para validação e 504 para teste. As redes ResUnet e Unet foram treinadas com diferentes funções de perda, assim como em dados RGB/RGIR por 180 épocas (76.545 iterações), tempo médio de 4 horas de treinamento por modelo. O resultado é apresentado na Tabela 7 com limiar das predições acima de 0,5.

Tabela 7 – Resultado no conjunto teste de ISPRS Potsdam

Modelos	IoU		Precision		Recall		F1	
	RGB	RGIR	RGB	RGIR	RGB	RGIR	RGB	RGIR
ResUnet_ftv	79,65	79,95	86,79	87,19	90,60	90,54	88,61	88,78
ResUnet_dice	79,85	79,83	88,68	88,94	88,93	88,60	88,73	88,71
ResUnet_tv	79,04	77,95	86,10	83,98	90,58	91,55	88,23	87,53
ResUnet_bce+dice	79,78	81,12	89,40	90,77	88,08	88,39	88,67	89,52
Unet_ftv	77,10	78,47	82,62	84,35	92,10	91,81	87,00	87,88
Unet_dice	79,75	80,03	89,84	89,34	87,67	88,45	88,65	88,83
Unet_tv	78,96	79,24	86,23	85,26	90,34	91,75	88,18	88,33
Unet_bce+dice	78,85	80,01	89,88	88,97	86,61	88,75	88,11	88,81

Resultado da média das 14 amostras do conjunto teste. **Em negrito:** melhores resultados do modelo na métrica por tipo de imagem de entrada. **Em negrito e sublinhado:** melhor resultado na métrica. ftv: *focal tversky loss*; tv: *tversky loss*; bce+dice: *binary cross entropy* e *dice loss*

O conjunto de dados de Potsdam não apresenta especificamente a segmentação da classe 'via', e sim a superfícies impermeáveis, a qual engloba as vias, estacionamentos, caminhos e outros. No entanto, os testes realizados mostram que a escolha da função de perda e o uso do IR influenciam na predição das vias e das outras feições dentro da classe de superfícies na área urbana.

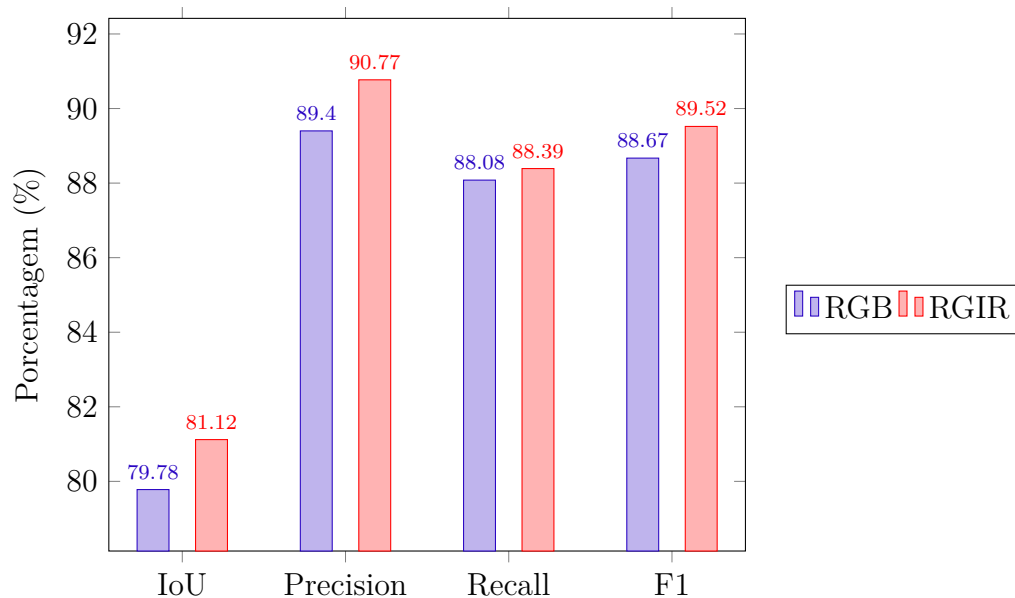
De acordo com a Tabela 7, modelos otimizados com funções de custo que possuem penalização para falsos negativos (*tversky* e *focal tversky*) proporcionam os valores mais

altos na métrica *recall*, por exemplo a Unet_ftv apresenta percentual de **92,10** no conjunto de imagens RGB. Entretanto, houve perda significativa de *precision* (**82,62%**). As redes otimizadas pelas funções de perda *dice* e *bce+dice* apresentaram os melhores resultados nas métricas, principalmente F1 e IoU. O modelo ResUnet_bce+dice obteve o melhor resultado no conjunto teste de Potsdam com **IoU = 81,12** e **F1 = 89,52**.

Em uma mesma arquitetura, a escolha da função de perda para otimização pode acarretar melhores resultados nas predições finais. Por exemplo, a ResUnet_bce+dice possui acréscimo de **4% em IoU e 2,3% em F1**, assim como o modelo Unet_dice tem **2% em IoU e 1,1% em F1** em relação aos otimizados pela função *tversky*. Ambos os casos no conjunto com dados IR.

Outro aspecto está relacionado ao uso da imagem IR no conjunto de dados. Praticamente todos os modelos apresentaram desempenho superior utilizando os dados com IR nas métricas F1 e IoU. Somente os modelos ResUnet_dice e ResUnet_tv apresentaram resultados superiores com os dados RGB. No caso do modelo ResUnet_bce+dice, o acréscimo foi de **1,8% em IoU e 1% em F1** em relação ao treinado com RGB. O gráfico da Figura 38 ilustra o resultado superior nas métricas utilizando o conjunto de dados com IR, principalmente na precisão final da predição da classe superfícies impermeáveis.

Figura 38 – Comparativo resultado modelo ResUnet_bce+dice com dados RGB x RGIR

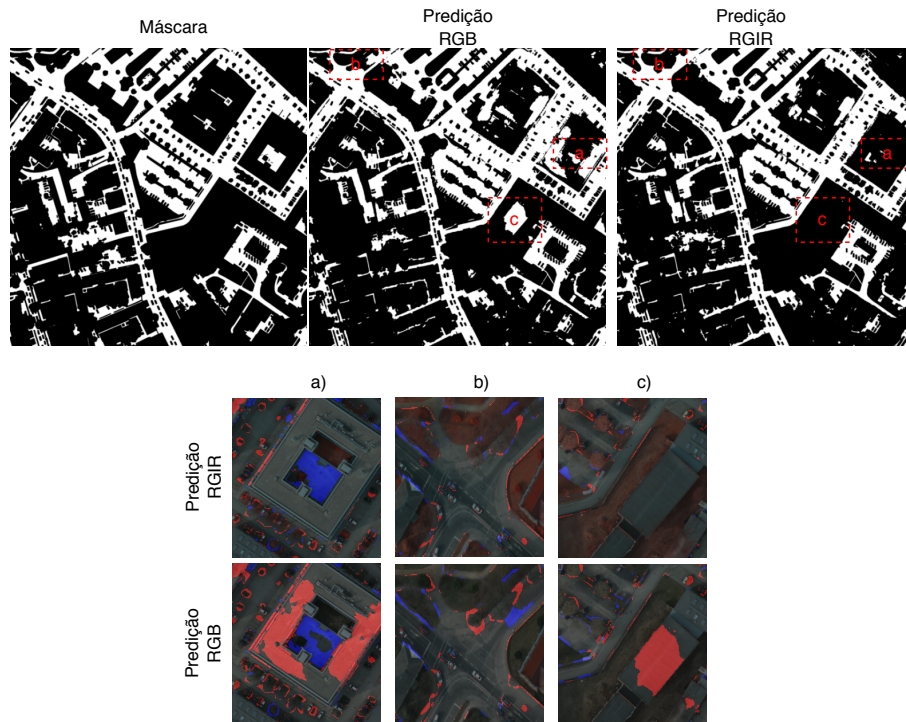


Fonte: De autoria própria.

Com isso, é possível observar que o canal IR agrega informação na separação de classes (especificamente na métrica *precision*), principalmente com a vegetação e na diferença de vias e telhados. Essa separação acarreta um melhor delineamento da via. A Figura 39 exibe essa tendência no aspecto visual das predições, onde a predição da classe superfícies impermeáveis com IR possui contornos mais definidos e também diminui a

confusão com texturas diferentes.

Figura 39 – Influência canal IR na predição



Fonte: De autoria própria. Predições realizadas com entrada RGB e RGIR do modelo de maior desempenho na métrica IoU e F1. Em azul: Falsos negativos; Em vermelho: Falsos positivos sobrepostos na imagem.

Embora os modelos treinados com funções de perda diferentes tenham influência no resultado, não foi encontrada uma relação entre a escolha e a predição final. Ou seja, não é possível afirmar que a função $bce+dice$ será a melhor opção de função na extração de vias em outros conjuntos de dados.

Por fim, de acordo com Mou, Hua e Zhu (2020) e Shang et al. (2020), os resultados na métrica F1 para a classe superfícies impermeáveis nas redes UNet e FCN variam de 88,1 a 91,33. Portanto, o resultado obtido está de acordo com a literatura.

5.2 Experimento II

Os objetivos neste experimento são realizar a segmentação semântica de vias em área urbana no conjunto de dados *benchmark* de Massachusetts e comparar o resultado com o estado da arte.

Após o pré-processamento descrito na Subseção 4.3.2, os dados são embaralhados antes do treinamento para evitar que a rede memorize os casos. O otimizador utilizado foi SGD com *momentum* de 0,9, taxa inicial de aprendizado de 0,01 e final de 0,0001, regularização por GC, n° de subamostras = 4 e aumento artificial de dados com rotação de 270° e escolha aleatória das seguintes transformações na imagem : espelhamento vertical/horizontal , recortes 512x512 e 224x224.

O conjunto *benchmark* possui 8037 recortes para treinamento, 126 para validação e 441 para teste. As redes ResUnet e Unet foram treinadas por 74 épocas (148.684 iterações), tempo médio de 7 horas de treinamento por modelo.

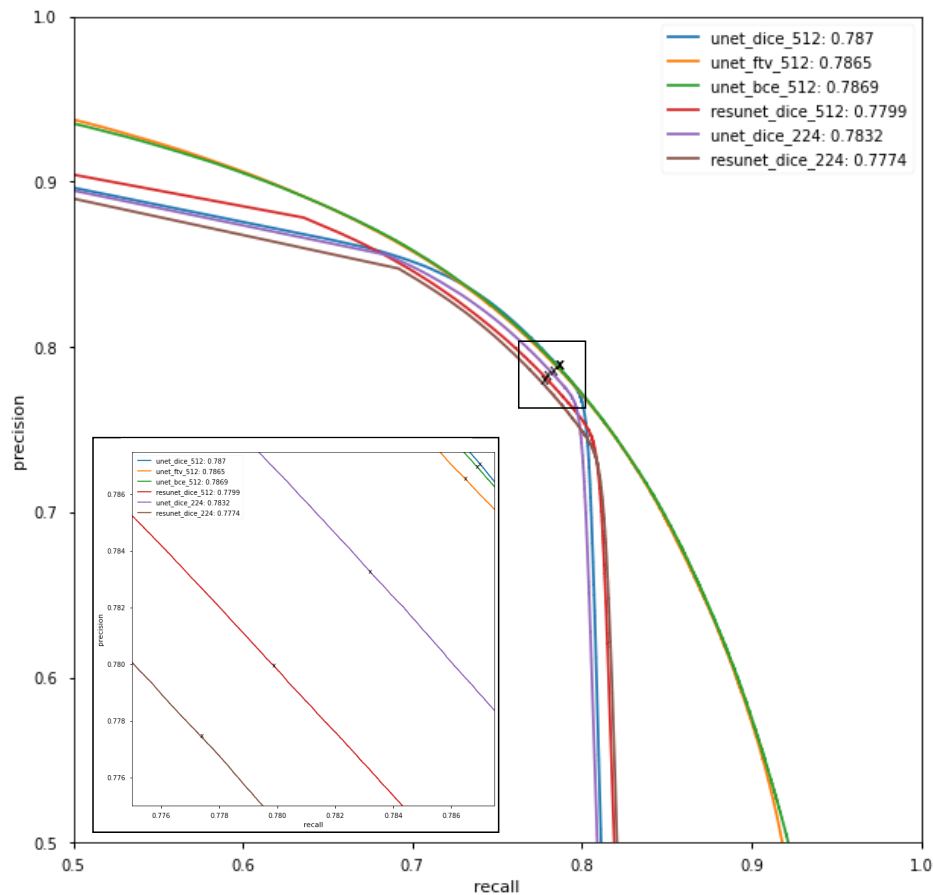
Mnih (2013) indica em sua tese a necessidade de avaliar os modelos no conjunto de dados pelo gráfico de comparação *precision* x *recall* em vários limiares. A predição final na segmentação de objetos possui valores no intervalo $[0,1]$ e no cálculo das métricas é necessário estabelecer previamente o limiar para geração da máscara binária. O melhor modelo é aquele onde o ponto de equilíbrio (*break-even*) tenha maior valor. Outro aspecto é o uso de uma tolerância para o cálculo das métricas (*relaxed precision/recall*). Com isso, é utilizado uma tolerância $\rho = 3 \text{ pixels}$ a partir da borda da máscara de 'via' original e qualquer *pixel* da predição definido como classe 'via' nessa região é igualmente considerado corretamente classificado.

Embora essa tolerância resulte em métricas de valores mais elevados, o *buffer* não foi utilizado. Todas as métricas foram calculadas em relação a máscara original com limiar acima de 0,5, além do cálculo do ponto de equilíbrio de cada modelo. A Tabela 8 apresenta os resultados no conjunto de dados de Massachusetts com limiares do *break-even* das métricas *precision* e *recall*. Além disso, o gráfico da Figura 40 mostra o ponto de equilíbrio dos modelos apresentados.

Tabela 8 – Comparativo dos principais modelos (%) no conjunto teste de Massachusetts

Modelos	IoU	Precision	Recall	F1
Unet_dice_512	64,94	80,27	77,27	78,74
Unet_bce+dice_512	64,87	80,46	77,0	78,69
Unet_ftv_512	64,76	80,8	76,54	78,61
Unet_dice_224	64,4	79,44	77,28	78,35
ResUnet_dice_512	63,94	78,95	77,09	78,01
ResUnet_dice_224	63,57	77,55	77,91	77,73

Resultado da média das 49 amostras do conjunto teste de Massachusetts; **Em negrito** - melhor resultado na métrica; 224 ou 512 - tamanho de recorte no treinamento.

Figura 40 – Gráfico *Precision x Recall* das predições dos modelos conjunto Massachusetts

Fonte: De autoria própria. No canto inferior esquerdo: ampliação região dos pontos de equilíbrio; No canto superior direito: modelos com o valor de equilíbrio.

De forma semelhante a Seção 5.1, a escolha da função de perda para otimização possui influência no resultado. No caso do conjunto de Massachusetts, o modelo Unet_dice_512 alcançou **IoU = 64,94** e **F1 = 78,74**, com resultado superior a ResUnet_dice_512 de **1,6%** e **0,9%** nas respectivas métricas. A função que penaliza falsos

negativos (*focal tversky*) obteve o pior desempenho na métrica *recall* sendo que foi utilizada a penalização de 0.7. Indicativo que a penalização foi excessiva.

De acordo com a Tabela 8, todos os modelos da arquitetura Unet apresentaram performance superior ao da ResUnet quanto as métricas IoU e F1. Como o classificador ResNet-50 possui mais camadas que a VGG-16, esse resultado pode indicar profundidade maior que a necessária, o que dificulta a otimização com poucos dados, consequentemente impacta a generalização final da rede.

Sob o aspecto visual, a Figura 41 apresenta a comparação entre os recortes e as predições do modelo de melhor desempenho nas métricas IoU e F1, isto é, Unet_dice_512.

Figura 41 – Resultados da predição imagens conjunto Massachusetts

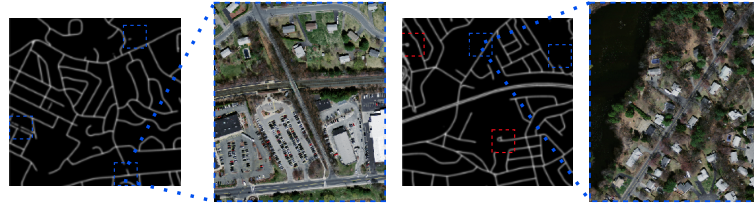


Fonte: De autoria própria. Imagens, máscaras de vias e as predições com melhor performance nas métricas. Em azul: falsos negativos; em vermelho: falsos positivos.

As predições apresentam poucas descontinuidades devido a oclusões e obstruções. Algumas omissões foram observadas em trechos de caminhos (casos das linhas (a), (b), (c)) e em regiões onde a copa de árvores causa oclusão na identificação da via (casos (d) e (e)). Falsos positivos foram detectados em regiões onde a máscara original possui omissões, por exemplo nos casos (c) e (d). A Figura 42 ilustra os casos de omissão por oclusão devido à

copa das árvores.

Figura 42 – Casos de oclusão copa de árvores

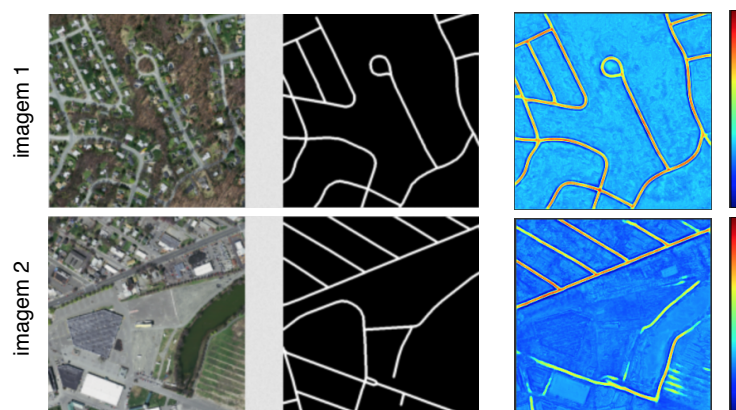


Fonte: De autoria própria.

Os modelos `Unet_dice_512` e `ResUnet_dice_512` apresentaram performance superior de **0,8% e 0,6% em IoU**, **0,5% e 0,4% em F1**, respectivamente, em relação aos modelos treinados com recortes de dimensão menor (224x224). Com isso, é possível afirmar que o contexto é fundamental para segmentação de vias. A melhor segmentação da rede viária está associada ao contexto que é apresentado ao classificador. Com mais contexto, melhor a discriminação das classes e redução das descontinuidades.

Além do resultado quantitativo, dois recortes (Fig. 43) foram selecionadas para interpretação do processo de segmentação no conjunto de dados. O primeiro pertence aos melhores resultados no conjunto teste ($\text{IoU} = 71,91$, $\text{F1} = 83,66$) e o segundo entre os piores ($\text{IoU} = 50,12$, $\text{F1} = 66,77$). Para isso, foi utilizado o método *Score-CAM* (WANG et al., 2020) que utiliza pontuação dos mapas de características ativados para interpretar a segmentação da classe na rede convolucional.

Figura 43 – Exemplos de inferência de vias no conjunto Massachusetts

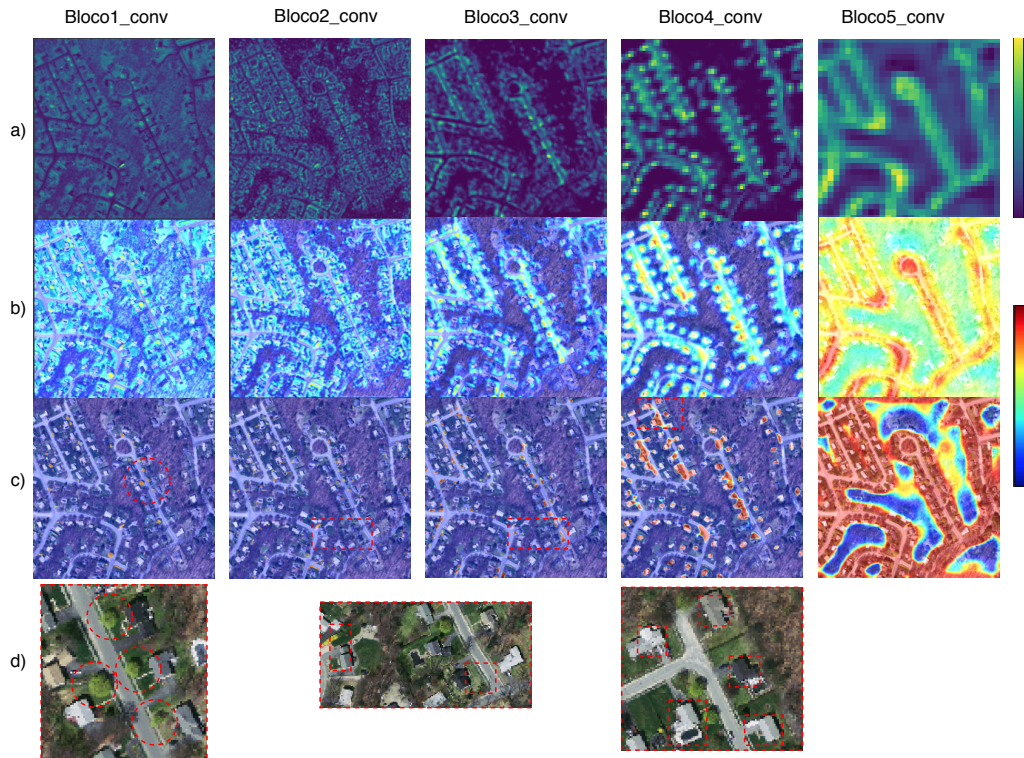


Fonte: De autoria própria. Imagem 1: a inferência de melhor performance; Imagem 2: a inferência de pior performance; Coluna à esquerda representa as imagens, ao centro as máscaras de referência e à direita as máscaras de probabilidades obtidas pela rede.

A Unet possui 5 níveis de profundidade (ou abstração) para o classificador e outros 5 níveis para a reconstrução da máscara de predição da classe 'via' de mesmas dimensões (comprimento x largura) que a imagem original. Sendo que 4 níveis são conectados

por conexões diretas para auxiliar na recuperação da resolução espacial perdida nas camadas de *pooling*. De forma a interpretar quais características são extraídas e/ou objetos são identificados, foram seleccionados os mapas de características das camadas do classificador anteriores à de *pooling* e aplicado o método *Score-CAM*. A Figura 44 ilustra as características principais que a rede utiliza para classificação da via.

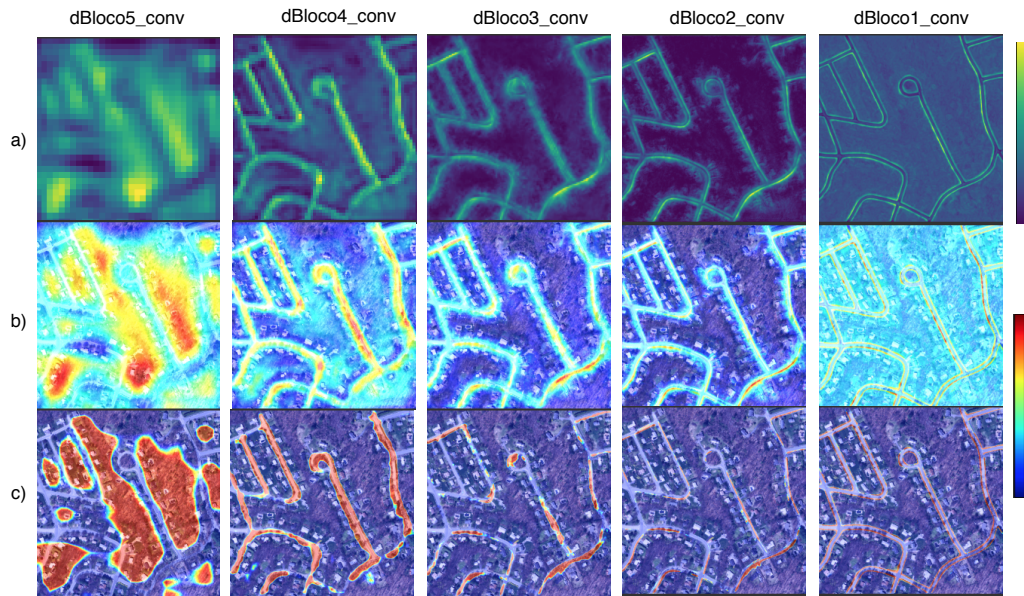
Figura 44 – Interpretação *encoder* da rede Unet da imagem 1



Fonte: De autoria própria. a) representam a ponderação dos mapas de ativação por *Score-CAM*; b) a ponderação obtida sobreposta a imagem; c) regiões de maior ponderação na ativação; d) as regiões e objetos delimitados em c).

Observa-se que somente no nível de maior abstração (maior contexto) a classe via foi detectada. Em níveis de menor contexto, o classificador busca identificar objetos de vegetação (árvores e cercas vivas) e telhados para então detectar a via. Ou seja, o classificador necessita de contexto de outros objetos na cena para identificar a via no ambiente urbano.

Nas camadas de reconstrução da máscara foram seleccionados os mapas de características anteriores às de concatenação das conexões diretas e aplicado o método *Score-CAM* (Fig. 45).

Figura 45 – Interpretação *decoder* da rede Unet da imagem 1

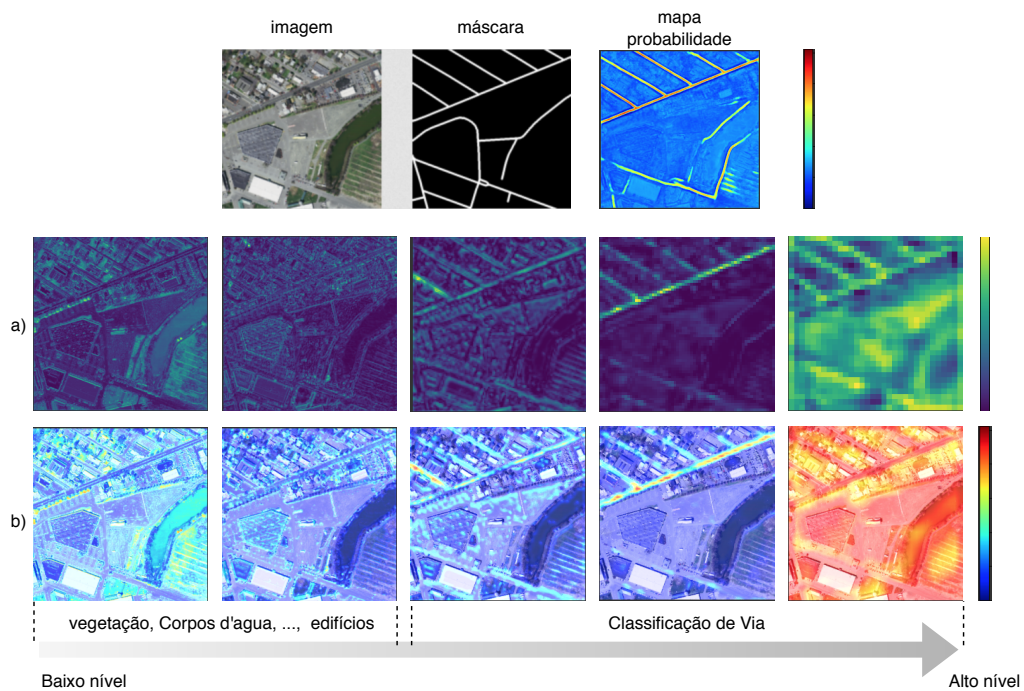
Fonte: De autoria própria. a) representam a ponderação dos mapas de ativação por *Score-CAM*; b) a ponderação obtida sobreposta a imagem; c) regiões de maior ponderação na ativação.

Ao nível mais abstrato (*dbloco5_conv*) a rede inverte a ativação para classe não-via. E as conexões diretas auxiliam na delimitação, principalmente porque identificaram os objetos da cena que não pertencem a rede viária. Ao nível de menor abstração e de maior resolução espacial (*dbloco1_conv*), a rede apresenta maior ativação e relevância nas regiões de borda da rede viária.

Da mesma forma que no caso da imagem 1 (melhor desempenho), para imagem 2 foram selecionados os mapas de características das camadas do classificador anteriores à de *pooling* e aplicado o método *Score-CAM*. A Figura 46 ilustra as características principais que a rede utiliza para classificação.

A característica de classificação residual é relevante para a compreensão do resultado de menor performance (imagem 2). Nos dois níveis de menor abstração (*bloco1_conv* e *bloco2_conv*) a ativação dos mapas está relacionada à identificação de classes de vegetação, corpos d'água e telhados, além das características geométricas de borda e textura. A partir do 3º nível, a rede identifica a classe via de forma parcial e não detecta outras características dentro da área do estacionamento. No nível mais abstrato (*bloco5_conv*) observa-se a ativação de área muito abrangente na cena, demonstrando a incerteza na identificação do objeto.

Mesmo com a incerteza final do classificador, a região onde a rede viária está delimitada pelas classes de vegetação, água e telhado apresentou a identificação e delimitação correta. Com isso, ressalta-se novamente a importância do contexto para segmentação da rede viária.

Figura 46 – Interpretação *encoder* da rede Unet da imagem 2

a) representam a ponderação dos mapas de ativação por *Score-CAM*; b) a ponderação obtida sobreposta a imagem.

De acordo com Zhang e Wang (2019), os resultados obtidos das redes Unet e ResUnet no conjunto de dados Massachusetts nas métricas F1/IoU foram de 78,07/61,85 e 77,93/61,40, respectivamente. Esses resultados estão abaixo dos modelos Unet_dice_512 (**78,74/64,94**) e ResUnet_dice_512 (**78,01/63,94**) apresentados anteriormente na Tabela 8. Além disso, o resultado na métrica IoU é superior ao obtido pela JointNet (79.35/64.00) dos autores.

Ressalta-se ainda que a diminuição de aproximadamente 20% dos dados de treino não resultou em perda de desempenho. A predição da rede apresentou resultados compatíveis com o estado da arte.

5.3 Experimento III

Os objetivos neste experimento são realizar a segmentação semântica de vias em área urbana no conjunto de dados do DF com/sem fusão de dados, analisar a influencia da fusão de dados no resultado final e comparar com outras arquiteturas do Estado da Arte na área de estudo.

A análise do conjunto do DF foi realizada em duas etapas. A primeira utilizou apenas ortoimagens, sem os dados oriundos do LiDAR, para definição de hiperparâmetros. A segunda etapa analisa o efeito da fusão com dados LiDAR utilizando as arquiteturas propostas na pesquisa.

5.3.1 Discussão sobre segmentação semântica sem fusão

Para a realização da segmentação, os dois conjuntos de dados (áreas 1 e 10) pré-processados (Subseção 4.3.3) são embaralhados antes do treinamento para evitar que a rede memorize os casos. O otimizador utilizado foi SGD com *momentum* de 0,9, taxa de aprendizado inicial de 0,01 e final de 0,0001, regularização por GC, n° subamostras = 4 e aumento artificial de dados com rotação de 270° e escolha aleatória das seguintes transformações na imagem : espelhamento vertical/horizontal , recortes 512x512. Todos os modelos foram otimizados com a função de perda *focal tversky*, a qual obteve os melhores resultados nas métricas. Ressalta-se que foram também realizados testes com recortes 224x224 para verificar a influência do contexto na acurácia da segmentação com a mesma quantidade de iterações dos modelos.

O conjunto área 1 possui 5675 recortes para treinamento, 1008 para validação e 1044 para teste. E o conjunto área 10 com 7700 recortes para treinamento, 1368 para validação e 1404 para teste. As redes ResUnet e Unet foram treinadas por 90 épocas (127.687 iterações na área 1 e 173.250 na área 10), tempo médio de 7 horas de treinamento por modelo na área 1 e de 10 horas na área 10. Os resultados obtidos nas áreas estão descritos na Tabela 9.

Tabela 9 – Resultado nos conjuntos testes das Área 10 e 1 - sem fusão

Modelos	Área 10				Área 1			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
ResUnet_224	71,40	81,77	83,21	81,97	66,32	78,62	80,25	78,21
ResUnet_512	73,32	82,21	86,24	83,71	<u>68,78</u>	81,01	<u>79,38</u>	<u>79,57</u>
Unet_224	<u>72,68</u>	83,02	84,79	<u>83,16</u>	67,33	<u>82,13</u>	77,88	79,05
Unet_512	72,31	82,17	<u>84,90</u>	82,89	69,41	84,10	78,95	80,73

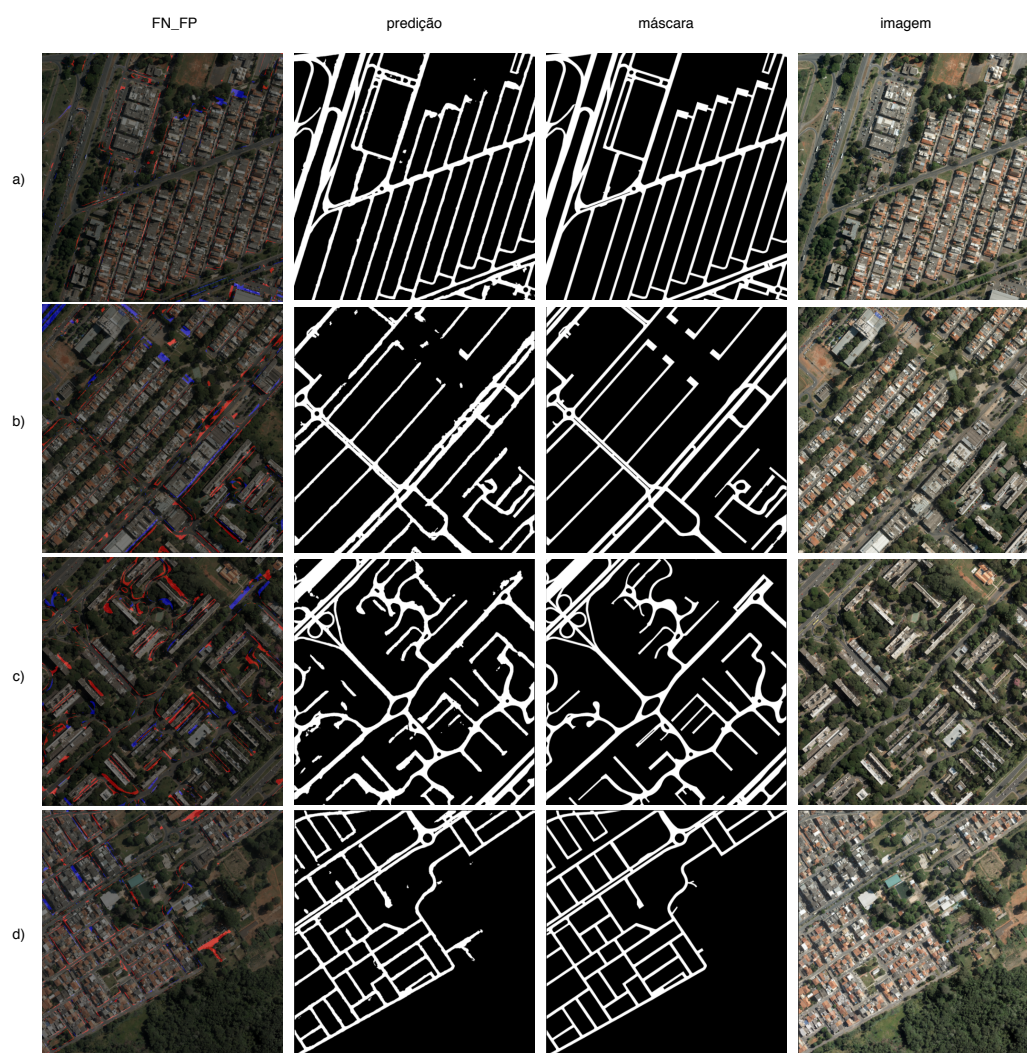
Média das métricas em 39 amostras da Área 10 e em 28 amostras da área 1; Os valores 224 e 512 representam os tamanhos de recortes 224x224 e 512x512 para treinamento; **Em Negrito**: melhor resultado na métrica; Sublinhado: segundo melhor resultado.

De acordo com a Tabela 9, os modelos ResUnet_512 e Unet_512 apresentaram melhor desempenho quanto a IoU e F1 nas áreas 10 e 1, respectivamente. Na área 10, a ResUnet_512 apresentou desempenho **0,9% IoU e 0,7% F1** superior a Unet_224. Entretanto, na área 1 a Unet_512 apresentou resultado **0,9% IoU e 1,4% F1** superior à ResUnet_512. No caso da área 10, a penalidade de 0,7 para falsos negativos foi efetiva para a métrica *recall* em detrimento da precisão em todos os casos. Entretanto, não foi determinante na área 1.

Outro aspecto refere-se ao contexto para a segmentação da via. O aumento da imagem no treinamento acarretou acréscimo de **2,7% IoU e 2,1% F1** para ResUnet na área 10, assim como **3,1% IoU e 2,1% F1** para Unet na área 1. Os dois melhores modelos foram treinados com recortes de 512x512, reafirmando o observado na Subseção 5.2.

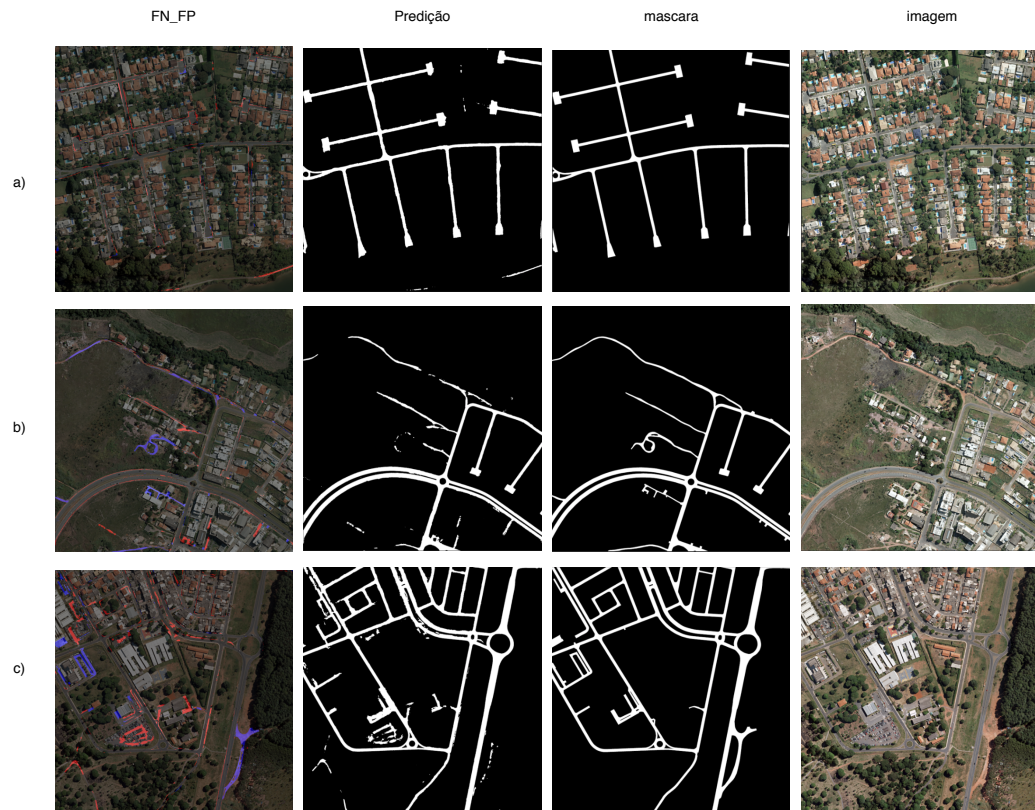
As Figuras 47 e 48 ilustram alguns exemplos de predições das áreas 10 e 1 pelos modelos ResUnet e Unet.

Figura 47 – Predições da rede ResUnet na área 10 - sem fusão



Fonte: De autoria própria. FN - Falsos Negativos representados pela cor azul; FP - Falsos Positivos representados pela cor vermelha

Figura 48 – Predições da rede Unet na área 1 - sem fusão

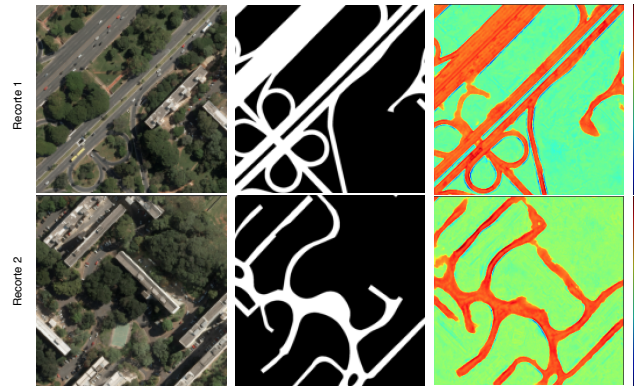


Fonte: De autoria própria. FN - Falsos Negativos representados pela cor azul; FP - Falsos Positivos representados pela cor vermelha

Interessante observar que os resultados obtidos nas métricas IoU e F1 pelos modelos sem a complementariedade dos dados LiDAR demonstram a excepcional capacidade de localização e extração de objetos por DL, inclusive em áreas com grande oclusão/obstrução. Porém, para analisar de que forma o contexto influencia na extração da via foi utilizado o método *Score-CAM* em dois recortes (1024x1024) do exemplo (c) da Figura 47.

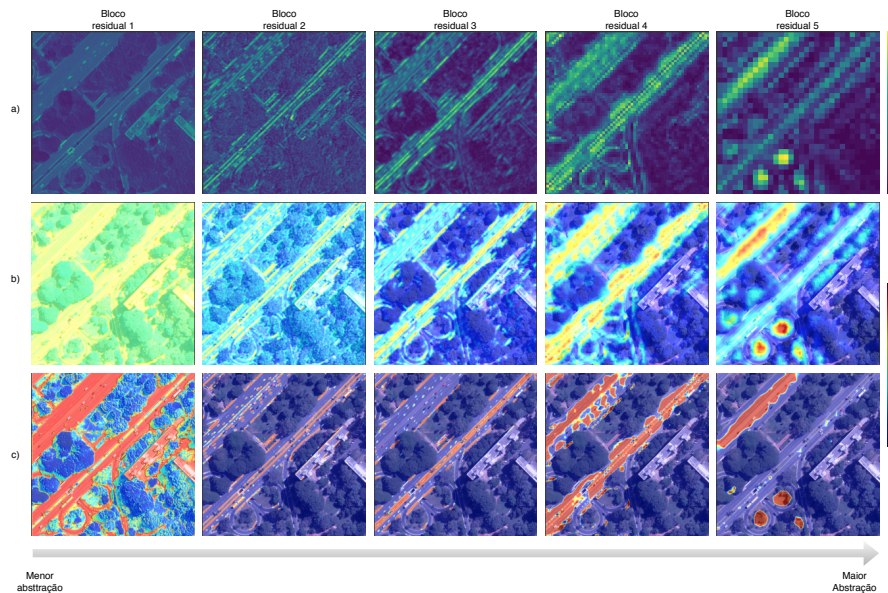
Os recortes selecionados, assim como as máscaras de referência e as de probabilidades das predições são apresentados na Figura 49.

Figura 49 – Exemplos de Inferência de vias no conjunto DF - área 10



Fonte: De autoria própria. Recorte 1: cena com maior predominância da classe via; Recorte 2: cena menor predominância da classe 'via'; Coluna à esquerda representa as imagens, ao centro as máscara de referência e à direita as máscaras de probabilidades obtidas pela rede

O primeiro recorte apresenta a 'via' com oclusões parciais pela copa das árvores e sombras, entretanto a classe predomina na área. Ou seja, maior quantidade de *pixels* pertencentes a rede viária. No segundo, as classes de vegetação e telhado predominam na cena. As Figuras 50 e 51 ilustram os mapas de características relevantes na classificação da 'via'.

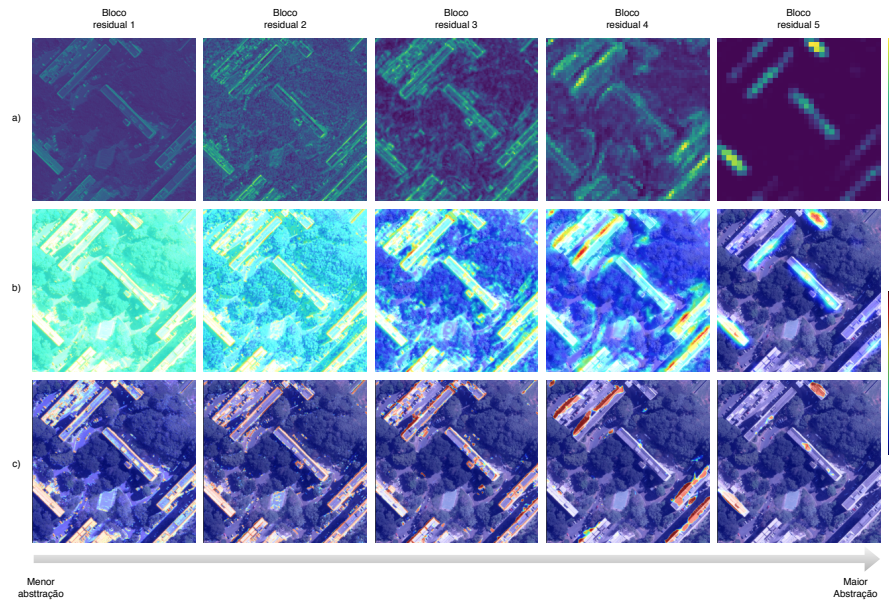
Figura 50 – Mapas de características extraídos do *encoder* do recorte 1

Fonte: De autoria própria. a) representam a ponderação dos mapas de ativação por Score-CAM; b) a ponderação obtida sobreposta a imagem; c) regiões de maior ponderação na ativação;

Em relação ao recorte 1, observa-se que o classificador desde o primeiro nível (menos abstrato) é ativado na região da classe via. Primeiramente, realiza a separação na cena da copa das árvores com auxílio de textura e cores. No segundo e terceiro níveis são extraídas informações de carros, sinalização (faixas na via), canteiro central e meio fio para buscar

informações internas à região inicialmente ativada. Com isso, ao quarto nível, a região da classe via é delimitada com maior confiança. Por fim, ao nível de maior abstração, o classificador detecta a via, bem como o padrão de trevo. Esse padrão auxilia a rede a realizar a predição de parte do trevo mesmo com pouca ou nenhuma informação aparente. Esse padrão detectado pela rede destaca a importância do contexto no treinamento para que o classificador possa inferir em áreas de muita oclusão e obstrução.

Figura 51 – Mapas de características extraídos do *encoder* do recorte 2

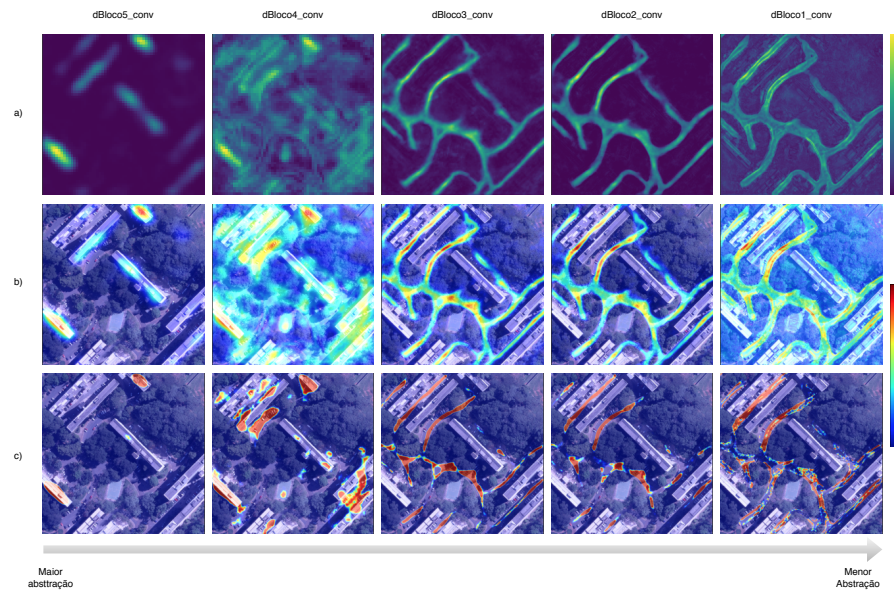


Fonte: De autoria própria. a) ponderação dos mapas de ativação por Score-CAM; b) ponderação obtida sobreposta a imagem; c) regiões de maior ponderação na ativação;

No caso do recorte 2, a Figura 51 ilustra um resultado singular quanto a relevância do contexto para segmentação de rede viária. Na região de maior oclusão, o classificador não detecta a via em nenhum nível do *encoder*. Ao nível de menor abstração, objetos que não sejam copa das árvores são ativados. Ao 2º e 3º níveis, classes carros, sombras, sinalização (p.ex: vagas de carros), meio-fio, muros e bordas de telhados são ativados para que nos níveis subsequentes os telhados sejam detectados. O classificador busca ativar a classe que possui maior semelhança com a rede viária. Com isso, a rede possui todos os elementos detectados da classe não-via.

Nos níveis de reconstrução, a rede delimita a classe via por exclusão. A Figura 52 mostra os mapas de características extraídos que auxiliam na delimitação da rede viária. Essa forma de reconstrução por exclusão acarreta muitos falsos positivos na predição final. Fato observado nos resultados na métrica *precision* da rede na área 10 (Tab. 9).

Figura 52 – Mapas de características extraídos do *decoder* do recorte 2



Fonte: De autoria própria. a) ponderação dos mapas de ativação por Score-CAM; b) ponderação obtida sobreposta a imagem; c) regiões de maior ponderação na ativação;

5.3.2 Discussão sobre segmentação semântica com dados fusionados

Para a realização da segmentação a partir de dados fusionados, os dois conjuntos de dados das áreas 1 e 10 (com a inclusão das imagens de intensidade e de modelos) pré-processados (Subseção 4.3.3) são embaralhados antes do treinamento e os hiperparâmetros são os mesmos da Subseção 5.3.1.

As redes ResUnet e Unet foram treinadas por 90 épocas (127.687 iterações na área 1 e 173.250 na área 10), 8 horas de treinamento por modelo na área 1 e de 11 horas na área 10. Os resultados obtidos na área 1 estão descritos nas Tabelas 10 e 11, enquanto os da área 10 estão nas Tabelas 12 e 13.

Tabela 10 – Comparativo resultados da rede ResUnet no conjunto teste da Área 01 - DF fusão com Imagem Intensidade e com Modelos

Modelos	Fusão Imagem Intensidade				Fusão Imagem Modelos			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
ResUnet_224	66,32	78,62	<u>80,25</u>	78,21	66,32	78,62	80,25	78,21
ResUnet_512	<u>68,78</u>	81,01	<u>79,38</u>	<u>79,57</u>	68,78	81,01	79,38	79,57
ResUnet_add_224	67,17	78,75	80,00	78,51	68,12	78,25	84,18	79,83
ResUnet_add_512	68,29	<u>81,33</u>	79,24	79,39	70,60	85,10	79,59	81,68
ResUnet_early_224	68,82	79,06	83,11	80,45	<u>69,53</u>	80,80	<u>83,39</u>	80,84
ResUnet_early_512	68,16	83,99	76,51	79,52	68,06	<u>84,14</u>	76,84	79,42

Em Negrito: melhor resultado na métrica; Sublinhado: segundo melhor resultado; add - fusão multinível (ramo paralelo); early - fusão inicial.

Tabela 11 – Comparativo resultados Unet no conjunto teste da Área 01 - DF fusão com Imagem Intensidade e com Modelos

Modelos	Fusão Imagem Intensidade				Fusão Imagem Modelos			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
Unet_224	67,33	82,13	77,88	79,05	67,33	82,13	77,88	79,05
Unet_512	69,41	84,10	78,95	80,73	69,41	84,10	78,95	80,73
Unet_add_224	<u>69,56</u>	84,05	<u>78,75</u>	<u>80,74</u>	70,09	83,51	<u>80,02</u>	81,03
Unet_add_512	69,49	<u>86,23</u>	77,04	80,71	<u>70,16</u>	86,12	77,87	81,06
Unet_early_224	70,15	86,20	78,22	81,33	71,01	83,86	81,10	81,95
Unet_early_512	67,39	87,32	73,81	79,01	69,97	<u>85,45</u>	78,64	81,09

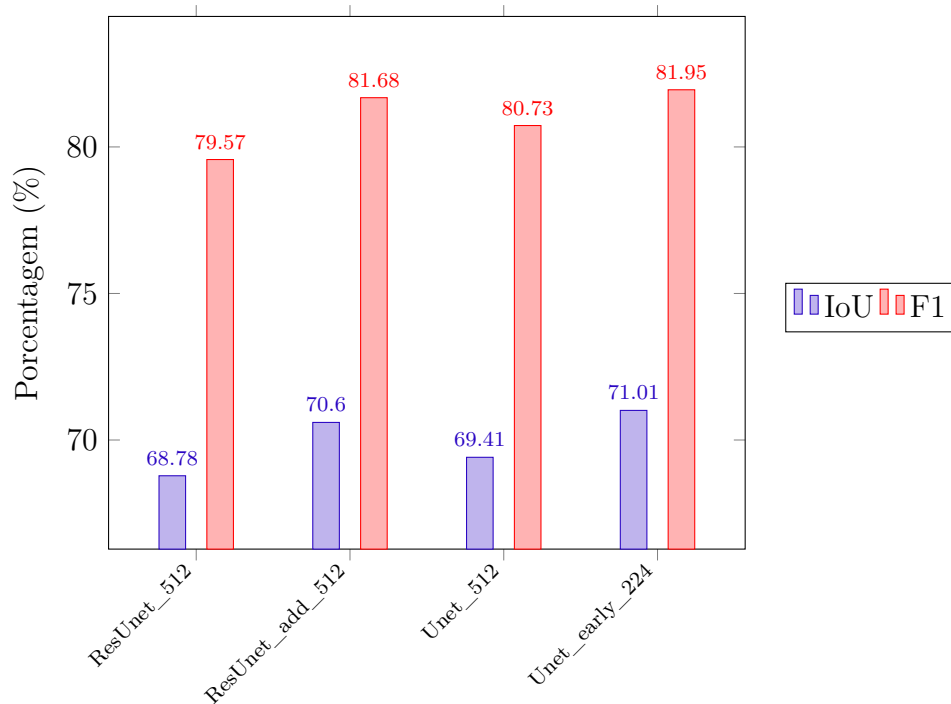
Em Negrito: melhor resultado na métrica; Sublinhado: segundo melhor resultado; add - fusão multinível (ramo paralelo); early - fusão inicial.

De acordo com os resultados das Tabelas 10 e 11 na área 1, a fusão com a **imagem de modelos** agregou mais informação à segmentação da 'via'. Os melhores desempenhos foram obtidos pela ResUnet_add_512 (**70,60 IoU/81,68 F1**) e Unet_early_224 (**71,01 IoU/81,95 F1**). Sendo o último o de melhor resultado nas métricas.

Os melhores modelos fusionados adicionaram **2,6% IoU e F1**, no caso da ResUnet, assim como **2,3% IoU e 1,5% F1** na caso da Unet em relação aos melhores modelos

não fusionados (ResUnet_512 e Unet_512). Esse resultado demonstra que a fusão foi efetiva mesmo com a influência do contexto, sendo o melhor modelo treinado com recortes menores (224x224). O gráfico da Figura 53 ilustra os resultados dos modelos fusionados e os sem fusão.

Figura 53 – Gráfico resultados das métricas dos modelos sem fusão e fusionados - Área 1

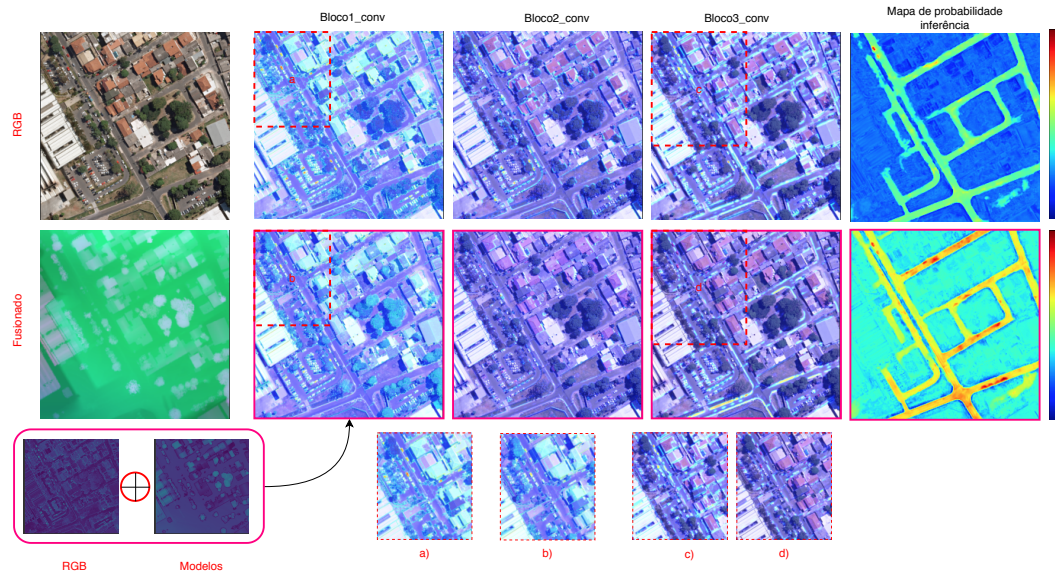


Fonte: De autoria própria.

A maioria dos modelos fusionados agrega informação na precisão, reduzindo os falsos positivos e consequentemente, diminui a incerteza na delimitação da rede viária. Entretanto, existe uma deterioração no resultado da métrica *recall*, mesmo com a penalidade em falsos negativos da função de perda *focal tversky*, indicativo de penalidade excessiva. Eventualmente, alguns modelos apresentaram desempenho pior do que os sem fusão. Esse efeito é verificado principalmente nos modelos treinados com recortes de dimensões maiores. Além disso, o melhor resultado da fusão nas métricas foi treinado com pouco contexto e fusão somente no nível menos abstrato. Ou seja, na área 1, é possível afirmar que a adição dos dados LiDAR (fusão multinível) esteja inserindo ruído na extração de características dos níveis mais abstratos, prejudicando a análise de contexto com outras classes na cena, impactando a otimização da rede. Com exceção das classes de vegetação e telhado, onde a fusão inicial impacta diretamente no nível menos abstrato. A característica de área suburbana também é relevante, pois as classes de vegetação (copa) e telhado pouco auxiliam a delimitação da via pelo contexto. Além disso, a imagem intensidade introduz diferença na classe via pela informação radiométrica. Ou seja, materiais como areia, cimento, água sobre a superfície da via causam falsos negativos. Característica que

não ocorre com a imagem de modelos pois adicionam somente informação geométrica. Quanto ao aspecto visual, a Figura 54 ilustra os mapas de características do classificador com a fusão inicial da rede Unet.

Figura 54 – Mapas de características extraídos do *encoder*



Fonte: De autoria própria. Primeira linha: a ponderação mapas de características sobreposta a imagem (sem fusão); Segunda linha: a ponderação mapas de características sobreposta a imagem (fusionado).

Observa-se nos extratos (a) e (b) que a imagem de modelos adiciona informações geométricas (contorno) para copa da árvores, telhados, carros e meio-fio no primeiro bloco convolucional. Sendo que no terceiro bloco, os extratos (c) e (d) permitem visualizar a definição de borda por carros e meio-fio.

Em relação à área 10, as Tabelas 12 e 13 apresentam os resultados obtidos com modelos fusionados utilizando imagem intensidade e de modelos.

Tabela 12 – Comparativo resultados da rede ResUnet no conjunto teste da Área 10 - DF fusão com Imagem Intensidade e com Modelos

Modelos	Fusão Imagem Intensidade				Fusão Imagem Modelos			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
ResUnet_224	71,40	81,77	83,21	81,97	71,40	81,77	83,21	81,97
ResUnet_512	73,34	82,21	86,24	83,71	73,32	82,21	86,24	83,71
ResUnet_add_224	72,18	82,86	83,39	82,59	72,51	82,13	<u>84,73</u>	82,89
ResUnet_add_512	<u>74,63</u>	<u>84,75</u>	<u>84,73</u>	<u>84,24</u>	<u>74,19</u>	<u>86,00</u>	<u>83,18</u>	<u>84,01</u>
ResUnet_early_224	73,20	84,16	82,81	82,94	72,81	84,30	82,92	82,88
ResUnet_early_512	74,92	86,84	83,31	84,68	74,55	86,34	84,21	84,56

Em Negrito: melhor resultado na métrica; Sublinhado: segundo melhor resultado; add - fusão multinível (ramo paralelo); early - fusão inicial.

Tabela 13 – Comparativo resultados da rede Unet no conjunto teste da Área 10 - DF fusão com Imagem Intensidade e com Modelos

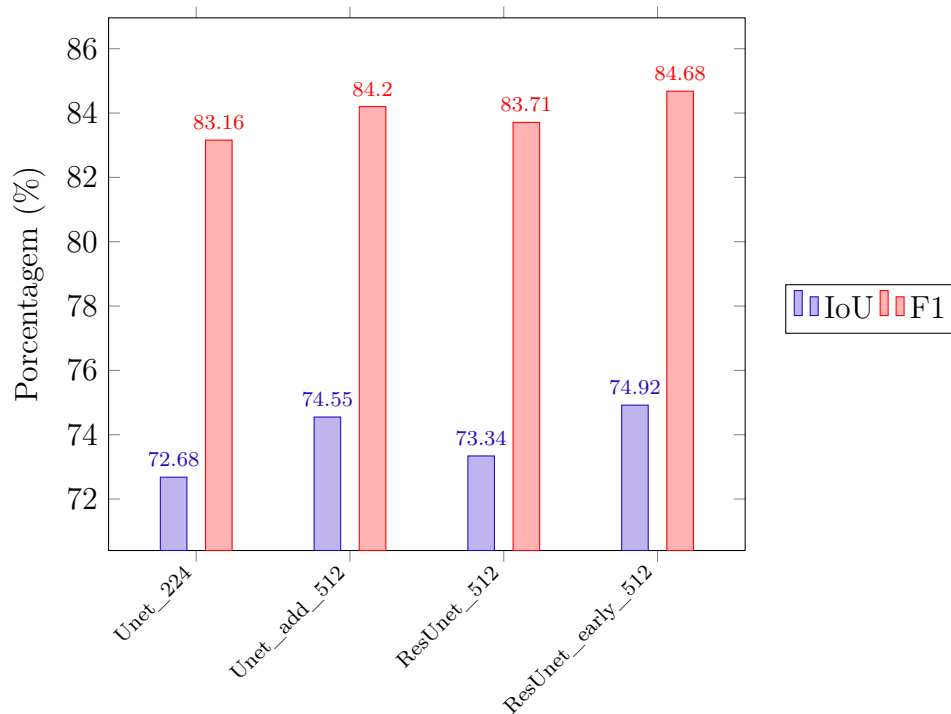
Modelos	Fusão Imagem Intensidade				Fusão Imagem Modelos			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
Unet_224	72,68	83,02	<u>84,79</u>	83,16	72,68	83,02	<u>84,79</u>	83,16
Unet_512	72,31	82,17	84,90	82,89	72,31	82,17	84,90	82,89
Unet_add_224	<u>73,76</u>	84,54	84,10	<u>83,88</u>	73,51	85,00	83,36	83,64
Unet_add_512	74,29	87,10	82,83	84,22	74,55	87,61	82,48	<u>84,20</u>
Unet_early_224	73,64	85,21	83,22	83,72	74,42	86,23	83,71	84,35
Unet_early_512	73,02	<u>86,36</u>	81,68	83,31	73,03	<u>86,98</u>	81,43	83,16

Em Negrito: melhor resultado na métrica; Sublinhado: segundo melhor resultado; add - fusão multinível (ramo paralelo); early - fusão inicial.

De acordo com os resultados das tabelas 12 e 13, a fusão com a **imagem de intensidade** agregou mais informação à segmentação da 'via'. Os melhores desempenhos foram da ResUnet_early_512 (**74,92 IoU/84,68 F1**) e Unet_add_512 (**74,55 IoU/84,20 F1**). Sendo o primeiro de melhor resultado nas métricas.

Os melhores modelos fusionados adicionaram **2,1% IoU e 1,2% F1**, no caso da ResUnet, assim como **2,6% IoU e 1,2% F1** no caso da Unet em relação aos melhores modelos não fusionados (ResUnet_512 e Unet_224). Esse resultado demonstra que a fusão também foi efetiva mesmo com a influência do contexto, principalmente nas oclusões. O gráfico da Figura 55 ilustra os resultados dos modelos fusionados e os sem fusão.

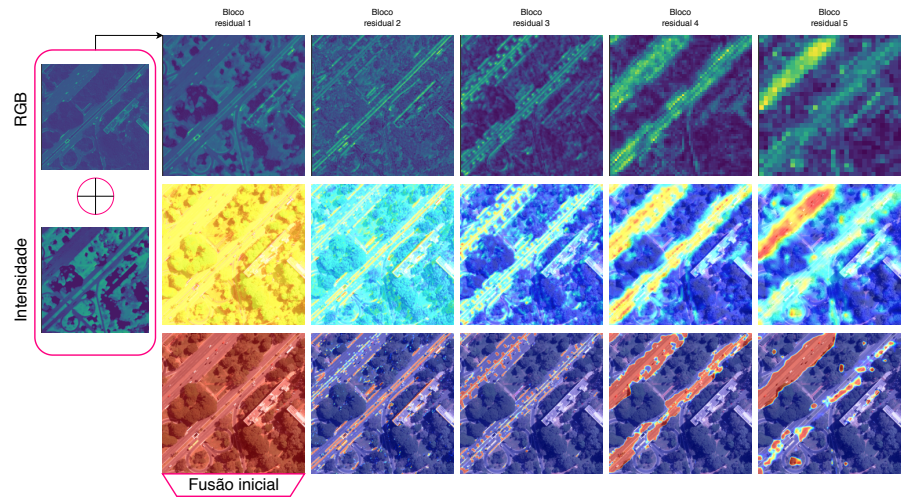
Figura 55 – Gráfico resultados métricas dos modelos sem fusão e fusionados - Área 10



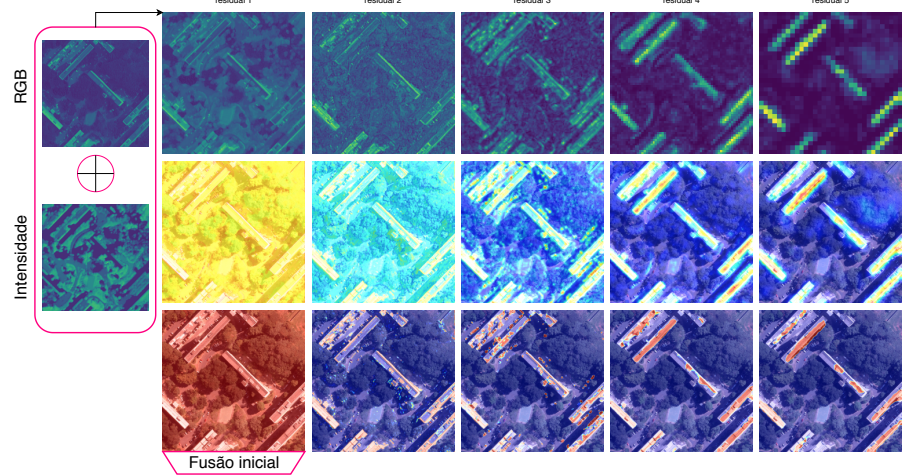
Fonte: De autoria própria.

Todos os modelos fusionados agregam informação na precisão, reduzindo os falsos positivos e consequentemente, diminui a incerteza na delimitação da rede viária. Entretanto, da mesma forma que na área 1, existe uma deterioração no resultado da métrica *recall*, mesmo com a penalidade em falsos negativos da função perda *focal tversky*. Sendo todos os modelos fusionados com desempenho pior, nessa métrica, do que os sem fusão treinados com maior contexto. Além disso, semelhante ao resultado da área 1, o melhor desempenho ocorre na fusão ao nível menos abstrato. Com isso, a adição multinível dos dados LiDAR (ramo paralelo) insere ruído na extração de características dos níveis mais abstratos e afeta a análise de contexto com outras classes.

Entretanto, no caso da área 10 (região urbana), as classes vegetação e telhado são mais relevantes na delimitação da via pelo contexto. Sendo a fusão inicial com a imagem intensidade mais efetiva por aumentar a ativação do classificador diretamente nestas classes. Quanto ao aspecto visual, as Figuras 56 e 57 ilustram os mapas de características do classificador com a fusão inicial da rede ResUnet. A fusão incide diretamente na classe de vegetação e nas sombras, além de melhorar o delineamento da classe com informações de borda e informações oclusas.

Figura 56 – Mapas de características extraídos do *encoder* fusionado do recorte 1

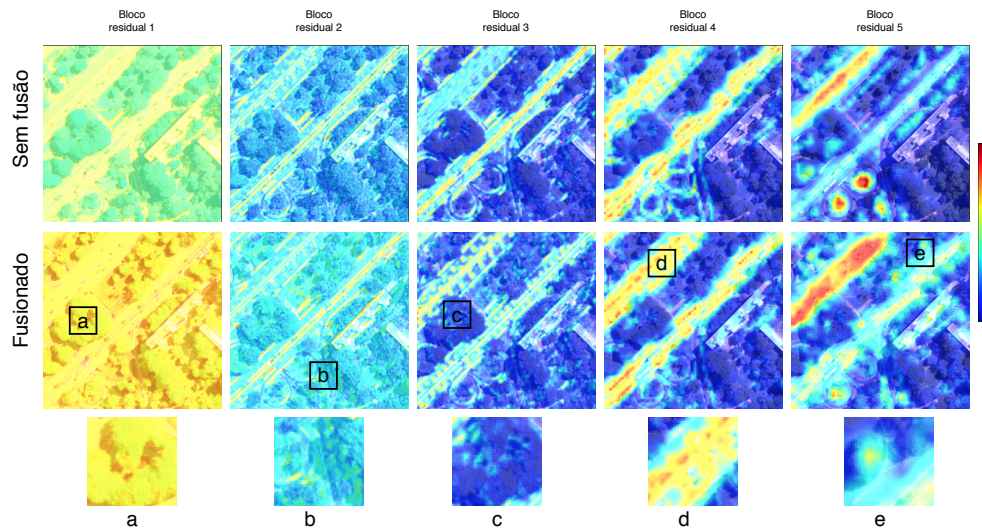
Fonte: De autoria própria. Primeira linha: representam a ponderação dos mapas de ativação por Score-CAM; Segunda linha: a ponderação obtida sobreposta a imagem; Terceira linha: regiões de maior ponderação na ativação;

Figura 57 – Mapas de características extraídos do *encoder* fusionado do recorte 2

Fonte: De autoria própria. Primeira linha: representam a ponderação dos mapas de ativação por Score-CAM; Segunda linha: a ponderação obtida sobreposta a imagem; Terceira linha: regiões de maior ponderação na ativação;

Para melhor entendimento da fusão, a Figura 58 mostra o comparativo entre mapas de características extraídos do recorte 1 no qual a fusão acrescenta ao classificador ativação na classe via, diminui area da vegetação, melhora detecção de vegetação, sinalização (efeito decorrente da radiometria da imagem intensidade) e acarreta diminuição de busca de contexto pela rede. A adição de informação oculta pode ser observada no extrato (a),(c) e (e), melhor delineamento de sombra em (b), bem como mais ativação da região interna da via (d).

Figura 58 – Comparativo Mapas de características extraídos do *encoder* fusionado x não fusionado do recorte 1



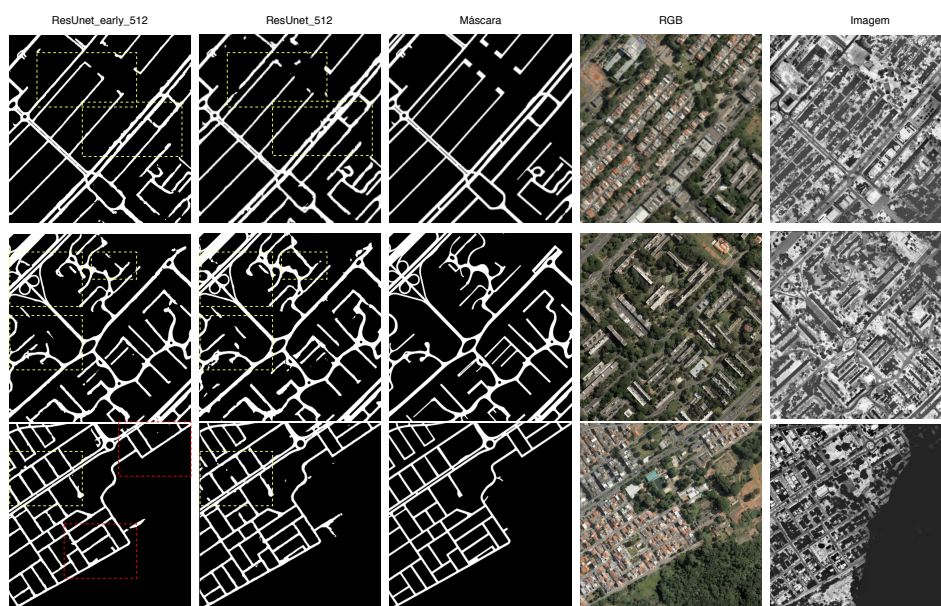
Fonte: De autoria própria. Primeira linha: representam a ponderação dos mapas de ativação por Score-CAM sobreposta a imagem da rede sem fusão; Segunda linha: ponderação dos mapas de ativação da rede fusionada;

No caso do recorte 2 ocorre de maneira semelhante, entretanto o classificador permanece com a ativação nas demais classes e de forma residual obtém a classe rede viária. E a fusão auxilia na ativação das demais classes (principalmente telhados).

Outro aspecto observado refere-se ao classificador fusionado apresentar extração de características semelhantes ao sem fusão. Apenas a confiança na detecção aumentou. Com isso, a inserção dos dados LiDAR ocorreu pontualmente e não impactou a detecção de outras características nas classes mais abstratas. Entretanto, problemas de falta de dados (algumas regiões sem informação do LiDAR ao nível do solo) e diferença radiométrica dos materiais na superfície (semelhante ao ocorrido na área 1) aumentam os falsos negativos.

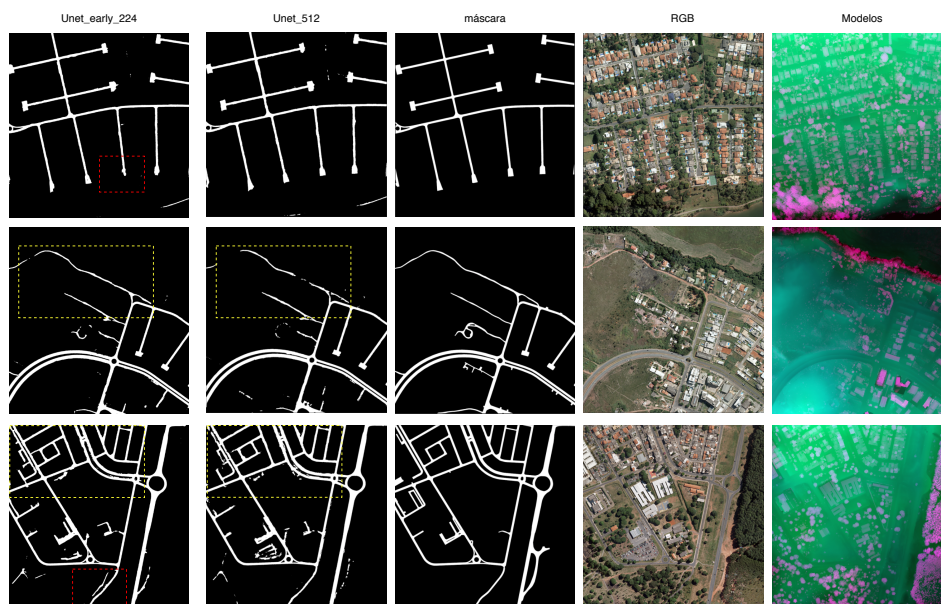
Por fim, as Figuras 59 e 60 apresentam a comparação entre as predição das redes fusionadas e sem fusão. A primeira referente a área 10, mais densa de rodovias (tendência de região urbana) com mais oclusões e obstruções, e a área 1, região menos densa (tendência de região suburbana).

Figura 59 – Comparativo Predições das redes fusionadas x não fusionadas - área 10



Fonte: De autoria própria. Em amarelo: casos que a fusão foi superior. Em vermelho: problema de falta de dados imagem Intensidade.

Figura 60 – Comparativo Predições das redes fusionadas x não fusionadas - área 1



Fonte: De autoria própria. Em amarelo: casos que fusão foi superior. Em vermelho: problema fusão com imagem modelos.

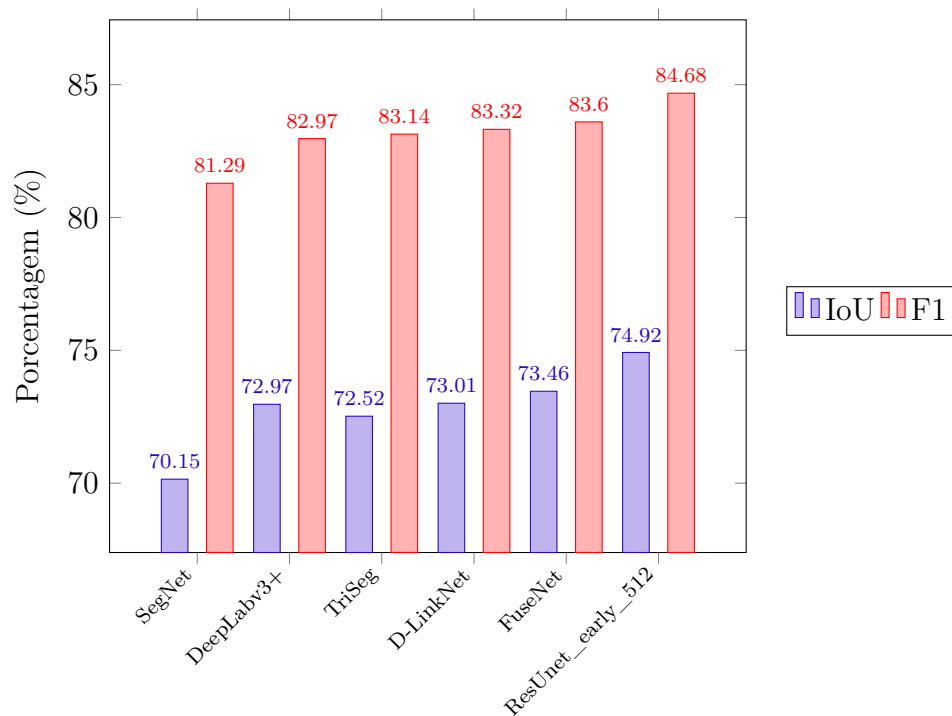
5.3.3 Comparação com métodos do Estado da Arte

Esta subseção apresenta a comparação do resultado obtido da ResUnet_early_512 na área 10 (área com mais problemas de oclusões) com arquiteturas do Estado da Arte, com dados fusionados ou não. As redes sem dados fusionados escolhidas foram: **D-LinkNet** (ZHOU; ZHANG; WU, 2018) que utiliza como base a arquitetura da LinkNet (CHAURASIA; CULURCIO, 2018) com convolução dilatada (*Dilated* ou *Atrous convolution*) para aumentar o campo receptivo sem alterar a resolução espacial dos mapas de características; **DeepLabv3+** (CHEN et al., 2018) que utiliza o módulo de várias convoluções dilatadas em paralelo com diferentes taxas de dilatação (*Atrous Spatial Pyramid Pooling - ASPP*) e ramo de reconstrução com poucos parâmetros; e **SegNet** (BADRINARAYANAN; KENDALL; CIPOLLA, 2017) que possui arquitetura semelhante à U-Net, porém não utiliza conexões diretas.

Quanto às redes com dados fusionados, foram escolhidas as arquiteturas da **FuseNet** (HAZIRBAS et al., 2017) que utiliza fusão multinível com dois classificadores e um ramo de reconstrução e a **TriSeg** (PARAJULI; CHOUDHURY; KUMAR, 2020) que consiste em 3 redes SegNet para extração de rede viária em área urbana com oclusões e obstruções.

Para a comparação foram utilizados os mesmos hiperparâmetros utilizados na ResUnet_early_512, bem como as imagens intensidade para as redes com fusão. As arquiteturas D-LinkNet, DeepLabv3+ foram implementadas com classificador ResNet-50 e as demais com VGG-16. Os resultados das redes foram resumidos no Gráfico 61.

Figura 61 – Gráfico comparativo dos diferentes modelos na Área 10



Fonte: De autoria própria.

Os resultados obtidos no comparativo demonstram que a arquitetura proposta na pesquisa obteve a melhor performance nas métricas. Inclusive superior a TriSeg, rede pioneira na extração de rede viária com fusão de dados em DL. O modelo ResUnet_early_512 alcançou resultado **2% IoU e 1,3% F1** superior frente a FuseNet (melhor modelo fusionado), bem como **2,6% IoU 1,6% F1** em relação a D-LinkNet (melhor modelo não fusionado).

6 Conclusões e Recomendações Futuras

A pesquisa está baseada na hipótese que a fusão de imagens (ortoimagens + LiDAR) por DL pode contribuir com informações complementares nas áreas de oclusões e obstruções na extração automática de rede viária em ambiente urbano. Para este propósito, uma metodologia de extração automática de rede viária a partir de dados fusionados foi desenvolvida de modo a aproveitar a complementariedade entre os dois modais.

A arquitetura escolhida UNet mostrou-se adequada ao propósito pretendido devido aos resultados alcançados. Além da facilidade nas adaptações (mudança de classificador, alteração de camadas no ramo de reconstrução e adição de módulo de fusão), possui implementação simples e pode ser treinada em diferentes conjuntos de dados para segmentação de objetos.

Todas as fases preparatórias executadas nos conjuntos de dados coletados foram relevantes para redução de ruídos e inconsistências. A redução de 20% dos dados no conjunto de Massachusetts e processamento das imagens oriundas do LiDAR no conjunto do DF são exemplos da necessidade de análise prévia dos dados. Embora a pesquisa tenha a área de estudo do DF, a utilização de *benchmarks* (seções 5.1 e 5.2) foi útil na validação de hiperparâmetros e na implementação da rede convolucional. Além disso, permitiu ampliar o conhecimento sobre a segmentação semântica por DL em diferentes conjuntos de dados.

Os resultados obtidos pela fusão por DL na extração da rede viária indicam 3 aspectos fundamentais: contexto de cena, separação de classes e escolha da função perda.

O contexto é fundamental para segmentação da rede viária em ambiente urbano. Os resultados obtidos na seção 5.2 e subseção 5.3.1 indicaram que a classe via é obtida de forma residual, ou seja, outras classes dentro do contexto apresentado são identificadas preliminarmente para se obter a delimitação da rede viária. Em cenas onde a classe é predominante, características geométricas (p.ex: meio-fio, canteiro) e radiométricas (sinalização de via, pinturas de vagas no estacionamento) são úteis para a delimitação em ambiente urbano.

Estratégias para separação de classes são importantes para aperfeiçoar a segmentação da rede viária. Na subseção 5.3.2, foi apresentado que a utilização da imagem intensidade fusionada com ortoimagens RGB no nível menos abstrato (fusão inicial) da DCNN permitiu melhor identificação da vegetação, de telhados e da superfície da via. Com isso, a classe via é melhor delimitada e a necessidade de obtenção da classe por contexto é reduzida para regiões de oclusões e obstruções no ambiente urbano. De forma semelhante, a informação geométrica advinda de modelos de superfície agrega informação na discriminação de classes, porém não possui o aspecto radiométrico para auxiliar na

diferença de rodovia e estacionamento por exemplo. Entretanto, ao reduzir a influência do contexto na predição, a rede tem maior acurácia nas características extraídas do insumo LiDAR, verificada pelo acréscimo de **2,6%** e **1,2%** nas métricas IoU e F1, respectivamente. Consequentemente, maior precisão na separação da classe via, maior discriminação intra-classe (materiais na superfície são discriminados) e, por outro lado, inserção de eventuais descontinuidades por falta de informação em grandes áreas.

Sob o aspecto da otimização das DCNN, como visto nos experimentos das seções 5.1, 5.2 e 5.3, a escolha da função de perda impacta o desempenho final da segmentação da rede viária. A pesquisa utilizou funções que ponderam o desbalanceamento de classes. Porém, não foi encontrado um indicativo ou uma tendência para melhor escolha.

O uso de algoritmos de aprendizado profundo de máquina nas tarefas de Visão Computacional aplicadas nas áreas de Fotogrametria e Sensoriamento remoto é relevante para se obter informações dos objetos de forma ágil, acurada e automática.

Como recomendações futuras, alguns temas relevantes:

- A tese utilizou dados oriundos de sensor ativo para a realização da fusão. Para isto, algumas transformações foram necessárias para adaptar a nuvem de pontos (espaço objeto) para entrada na rede convolucional. Entretanto, a extração de características obtida diretamente da nuvem de pontos pode aperfeiçoar a fusão dos sensores;
- A inspiração da técnica de fusão na pesquisa foi a utilizada na arquitetura FuseNet. Entretanto, sugere-se que outros métodos da literatura sejam analisados, como por exemplo uso de *Attention*;
- A extração completa da rede viária depende da transformação do espaço-imagem para o espaço-objeto. Sugere-se a pesquisa de métodos para obtenção do eixo central e contornos para obtenção da rede viária no espaço de conjuntos de dados vetoriais;
- A arquitetura utilizada foi a Unet. Porém, existem outros tipos de arquiteturas na literatura que podem ser objeto de pesquisa para melhor aproveitamento de contexto, outros tipos de camadas convolucionais (p.ex: *pyramidal convolutional*) e outros classificadores.

Referências

- ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <<http://tensorflow.org/>>. Citado na página 62.
- ABDOLLAHI, A. et al. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sensing*, v. 12, n. 9, p. 1444, may 2020. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/12/9/1444>>. Citado 3 vezes nas páginas 22, 23 e 24.
- ABRAHAM, N.; KHAN, N. M. A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation. oct 2018. Disponível em: <<http://arxiv.org/abs/1810.07842>>. Citado na página 46.
- ALSHEHHI, R. et al. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), v. 130, p. 139–149, 2017. ISSN 09242716. Disponível em: <<http://dx.doi.org/10.1016/j.isprsjprs.2017.05.002>>. Citado 2 vezes nas páginas 24 e 33.
- AMINI, J. et al. Automatic road-side extraction from large scale imagemaps. *International Journal of Applied Earth Observation and Geoinformation*, v. 4, n. 2, p. 95–107, 2002. ISSN 03032434. Citado na página 22.
- ARIEF, H. A. et al. Land cover segmentation of airborne LiDAR data using stochastic atrous network. *Remote Sensing*, v. 10, n. 6, 2018. ISSN 20724292. Citado 2 vezes nas páginas 72 e 74.
- ASHRAF, I.; HUR, S.; PARK, Y. An Investigation of Interpolation Techniques to Generate 2D Intensity Image from LIDAR Data. *IEEE Access*, v. 5, n. c, p. 8250–8260, 2017. ISSN 21693536. Citado na página 49.
- ASPRS. *LAS Specification 1 . 4 - R14*. Maryland 20814-2160, 2019. 47 p. Citado na página 54.
- AUDEBERT, N.; SAUX, B. L.; LEFÈVRE, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 140, p. 20–32, 2018. ISSN 09242716. Citado na página 25.
- BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 39, n. 12, p. 2481–2495, 2017. ISSN 01628828. Citado 2 vezes nas páginas 57 e 103.
- BAJCSY, R.; TAVAKOLI, M. Computer Recognition of Roads from Satellite Pictures. *IEEE Transactions on Systems, Man and Cybernetics*, v. 6, n. 9, p. 623–637, 1976. ISSN 21682909. Citado na página 20.

- BALTSAVIAS, E. P. *Airborne laser scanning: Existing systems and firms and other resources*. [S.l.: s.n.], 1999. v. 54. 164–198 p. ISSN 09242716. ISBN 4116333042. Citado 2 vezes nas páginas 50 e 54.
- BANDARA, W. G. C.; VALANARASU, J. M. J.; PATEL, V. M. SPIN Road Mapper: Extracting Roads from Aerial Images via Spatial and Interaction Space Graph Reasoning for Autonomous Driving. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022. p. 343–350. ISBN 978-1-7281-9681-7. Disponível em: <<http://arxiv.org/abs/2109.07701><https://ieeexplore.ieee.org/document/9812134/>>. Citado na página 24.
- BAUMGARTNER, A.; HINZ, S. Multi-Scale Road Extraction Using Local and Global Grouping Criteria. *International Archives of Photogrammetry and Remote Sensing*, v. 33, n. B3/1; PART 3, p. 58–65, 2000. ISSN 1096-987X. Citado 3 vezes nas páginas 19, 20 e 22.
- BENGIO, Y. Deep Learning of Representations for Unsupervised and Transfer Learning. *JMLR: Workshop and Conference Proceedings*, v. 7, p. 1–20, 2011. Citado na página 42.
- BIANCO, S. et al. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, v. 6, p. 64270–64277, 2018. ISSN 21693536. Citado 2 vezes nas páginas 39 e 41.
- BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. ACM Press, 1992. Disponível em: <<http://dx.doi.org/10.1145/130385.130401>>. Citado na página 29.
- BOSSLER, J. D.; HABIB, A. *Manual of Geospatial Science and Technology*. 2. ed. [S.l.]: CRC Press Taylor & Francis Group, 2010. ISBN 9781420087345. Citado 4 vezes nas páginas 50, 51, 53 e 55.
- BOTTOU, L. Online Learning and Stochastic Approximations (revised 5/2018). *Online Learning in Neural Networks*, p. 1–35, 1998. Citado 2 vezes nas páginas 34 e 35.
- BREIMAN, L. Random Forests. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <<http://dx.doi.org/10.1023/A:1010933404324>>. Citado na página 29.
- BRENNAN, R.; WEBSTER, T. L. Object-oriented land cover classification of lidar-derived surfaces. *Canadian Journal of Remote Sensing*, v. 32, n. 2, p. 162–172, 2006. ISSN 17127971. Citado na página 56.
- BRIDLE, J. S. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in Neural Information Processing Systems*, v. 2, n. M1, p. 211–217, 1990. Citado na página 39.
- CARUANA, R. Learning Many Related Tasks at the Same Time With Backpropagation. *Advances in Neural Information Processing Systems 7*, p. 657–664, 1995. Disponível em: <<https://papers.nips.cc/paper/959-learning-many-related-tasks-at-the-same-time-with-backpropagation.pdf>>. Citado na página 42.

CHAURASIA, A.; CULURCIELLO, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. *2017 IEEE Visual Communications and Image Processing, VCIP 2017*, v. 2018-Janua, p. 1–4, 2018. Citado na página 103.

CHEN, C. et al. Performance comparison of filtering algorithms for high-density airborne lidar point clouds over complex landscapes. *Remote Sensing*, v. 13, n. 14, 2021. ISSN 20724292. Citado na página 69.

CHEN, L.-C. et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: FERRARI, V. et al. (Ed.). Cham: Springer International Publishing, 2018, (Lecture Notes in Computer Science, v. 11211). p. 833–851. Disponível em: <http://doi.org/10.1007/978-3-030-01234-2_46<http://link.springer.com/10.1007/978-3-030-01234-2>http://link.springer.com/10.1007/978-3-030-01234-2_49>. Citado na página 103.

CHENG, G. et al. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, v. 55, n. 6, p. 3322–3337, 2017. ISSN 01962892. Citado na página 24.

CHOLLET, F. *Deep Learning with Python*. [S.l.]: manning publications co., 2018. 386 p. ISBN 9781617294433. Citado na página 23.

CHOLLET, F. et al. *Keras*. 2015. <<https://keras.io>>. Citado na página 62.

CLOUDCOMPARE. *CloudCompare (version 2.X)*. 2020. [GPL software]. Disponível em: <[Retrievedfromhttp://www.cloudcompare.org/](http://www.cloudcompare.org/)>. Citado na página 62.

CNT. *Pesquisa CNT de Rodovias 2019*. [S.l.], 2019. Citado na página 18.

CONTRIBUTORS, P. *PDAL Point Data Abstraction Library*. 2018. Disponível em: <<https://doi.org/10.5281/zenodo.2556738>>. Citado na página 62.

COSTEA, D. et al. Creating Roadmaps in Aerial Images with Generative Adversarial Networks and Smoothing-Based Optimization. In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2017. p. 2100–2109. ISBN 978-1-5386-1034-3. Disponível em: <<http://ieeexplore.ieee.org/document/8265456/>>. Citado na página 24.

CRAMER, M. The dgpf-test on digital airborne camera evaluation overview and test design. *Photogrammetrie - Fernerkundung - Geoinformation*, Schweizerbart, v. 2010, n. 2, p. 73–82, May 2010. Disponível em: <<http://dx.doi.org/10.1127/1432-8364/2010/0041>>. Citado na página 61.

Dal Poz, A. P. *Metodologias semi-automáticas para extração de rodovias em imagens digitais usando programação dinâmica, análise de bordas de rodovia e teste ativo*. 154 p. Tese (Doutorado) — Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, 2003. Citado 2 vezes nas páginas 20 e 22.

Dal Poz, A. P. Synergy Between LiDAR and Image Data in Context of Building Extraction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-1, n. November, p. 89–93, nov 2014. ISSN 2194-9034. Disponível em: <<https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-1/89/2014/><http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-1/89/2014/>>. Citado na página 20.

Dal Poz, A. P.; ZANIN, R. B.; Do Vale, G. M. Automated extraction of road network from medium-and high-resolution images. *Pattern Recognition and Image Analysis*, v. 16, n. 2, p. 239–248, 2006. ISSN 10546618. Citado na página 22.

DAS, S.; MIRNALINEE, T. T.; VARGHESE, K. Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 49, n. 10 PART 2, p. 3906–3931, 2011. ISSN 01962892. Citado 2 vezes nas páginas 20 e 22.

DING, L.; BRUZZONE, L. DiResNet: Direction-Aware Residual Network for Road Extraction in VHR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, v. 59, n. 12, p. 10243–10254, dec 2021. ISSN 0196-2892. Disponível em: <<https://ieeexplore.ieee.org/document/9257405/>>. Citado na página 24.

DUCHI, J. C.; BARTLETT, P. L.; WAINWRIGHT, M. J. Randomized smoothing for (parallel) stochastic optimization. *Proceedings of the IEEE Conference on Decision and Control*, v. 12, p. 5442–5444, 2012. ISSN 01912216. Citado na página 36.

EL-ASHMAWY, N.; SHAKER, A.; YAN, W. Pixel Vs Object-Based Image Classification Techniques for Lidar Intensity Data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5/, n. August, p. 43–48, 2012. ISSN 1682-1777. Citado na página 56.

EL-SHEIMY, N.; VALEO, C.; HABIB, A. *Digital terrain modeling: acquisition, manipulation, and applications*. 2005. Disponível em: <<http://www.lavoisier.fr/livre/notice.asp?id=RASWOAA263ROWI>>. Citado 2 vezes nas páginas 52 e 53.

FILHO, A. G. G. et al. Quality control relevance on acquisition of large scale geospatial data to urban territorial management. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, XLII-3/W12-2020, p. 25–29, Nov 2020. Disponível em: <<http://dx.doi.org/10.5194/isprs-archives-XLII-3-W12-2020-25-2020>>. Citado na página 66.

FUKUSHIMA, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, v. 36, n. 4, p. 193–202, 1980. ISSN 03401200. Citado na página 29.

GALO, M. *Notas de Aula de Fotogrametria III Princípio dos sistemas de varredura a LASER aerotransportado*. Presidente Prudente - SP: [s.n.], 2018. Citado na página 52.

GAMBA, P.; DELL'ACQUA, F.; LISINI, G. Improving urban road extraction in high-resolution images exploiting directional filtering, perceptual grouping, and simple topological concepts. *IEEE Geoscience and Remote Sensing Letters*, v. 3, n. 3, p. 387–391, 2006. ISSN 1545598X. Citado na página 22.

GARCIA-GARCIA, A. et al. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing Journal*, Elsevier B.V., v. 70, p. 41–65, 2018. ISSN 15684946. Disponível em: <<https://doi.org/10.1016/j.asoc.2018.05.018>>. Citado 5 vezes nas páginas 37, 41, 42, 43 e 56.

GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. [S.l.], 2021. Disponível em: <<https://gdal.org/>>. Citado na página 62.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2. ed. Canadá: O'REILLY, 2019. 611 p. ISBN 9781492032649. Citado na página 42.

GHANMA, M. *Integration of Photogrammetry and LIDAR*. 139 p. Tese (Doutorado) — University of calgary, 2006. Citado 2 vezes nas páginas 25 e 55.

GHARIBI, H.; HABIB, A. True orthophoto generation from aerial frame images and LiDAR Data: An update. *Remote Sensing*, v. 10, n. 4, p. 1–28, 2018. ISSN 20724292. Citado na página 55.

GHOSH, S. et al. Understanding deep learning techniques for image segmentation. *ACM Computing Surveys*, v. 52, n. 4, p. 1–58, 2019. ISSN 15577341. Citado na página 43.

GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, v. 9, p. 249–256, 2010. ISSN 15324435. Citado na página 40.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. *Journal of Machine Learning Research*, v. 15, p. 315–323, 2011. ISSN 15324435. Citado na página 33.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [s.n.], 2016. 1–777 p. Disponível em: <www.deeplearningbook.org>. Citado 6 vezes nas páginas 23, 34, 35, 36, 37 e 42.

GOODFELLOW, I. J. et al. Generative Adversarial Networks. *Communications of the ACM*, v. 63, n. 11, p. 139–144, jun 2014. ISSN 15577317. Disponível em: <<http://arxiv.org/abs/1406.2661>>. Citado na página 29.

HAZIRBAS, C. et al. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture. *Computer Vision – ACCV 2016*, Springer International Publishing, p. 1–15, 2017. ISSN 1460373X. Disponível em: <<http://link.springer.com/10.1007/978-3-319-54181-5>>. Citado 6 vezes nas páginas 38, 57, 58, 59, 72 e 103.

HE, K. et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the IEEE International Conference on Computer Vision*, p. 1026–1034, feb 2015. ISSN 15505499. Disponível em: <<http://arxiv.org/abs/1502.01852>>. Citado na página 40.

HE, K. et al. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 2016-Decem, p. 770–778, 2016. ISSN 10636919. Citado 4 vezes nas páginas 23, 29, 40 e 41.

HEYDARI, S. S.; MOUNTRAKIS, G. Meta-analysis of deep neural networks in remote sensing: A comparative study of mono-temporal classification to support vector machines. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 152, n. May, p. 192–210, 2019. ISSN 09242716. Disponível em: <<https://doi.org/10.1016/j.isprsjprs.2019.04.016>>. Citado 2 vezes nas páginas 25 e 72.

- HOESER, T.; KUENZER, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sensing*, v. 12, n. 10, p. 1667, may 2020. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/12/10/1667>>. Citado 8 vezes nas páginas 23, 24, 28, 33, 34, 40, 43 e 44.
- HORNIK, K. Approximation Capabilities of Multilayer Neural Network. *Neural Networks*, v. 4, n. 1991, p. 251–257, 1991. Citado na página 30.
- HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, v. 2, n. 5, p. 359–366, jan 1989. ISSN 08936080. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/0893608089900208>>. Citado na página 30.
- HU, Y.; CHEN, Z.; LIN, W. RGB-D Semantic segmentation: A review. *2018 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2018*, IEEE, p. 1–6, 2018. Citado na página 56.
- HUANG, B. et al. Tiling and Stitching Segmentation Output for Remote Sensing: Basic Challenges and Recommendations. 2018. Disponível em: <<http://arxiv.org/abs/1805.12219>>. Citado na página 76.
- IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, v. 1, p. 448–456, 2015. Citado 3 vezes nas páginas 29, 38 e 40.
- KAISER, P. et al. Learning Aerial Image Segmentation From Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*, v. 55, n. 11, p. 6054–6068, nov 2017. ISSN 0196-2892. Disponível em: <<http://ieeexplore.ieee.org/document/7987710/>>. Citado 2 vezes nas páginas 10 e 65.
- KASHANI, A. G. et al. A review of LIDAR radiometric processing: From ad hoc intensity correction to rigorous radiometric calibration. *Sensors (Switzerland)*, v. 15, n. 11, p. 28099–28128, 2015. ISSN 14248220. Citado 2 vezes nas páginas 55 e 56.
- KINGMA, D. P.; BA, J. L. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, p. 1–15, 2015. Citado na página 36.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. ImageNet Classification with Deep Convolutional Neural Networks Alex. *Advances in Neural Information Processing Systems*, p. 45–1–45–16, 2012. Citado 4 vezes nas páginas 23, 29, 34 e 39.
- Kumar Mishra, R. A Review of Optical Imagery and Airborne LiDAR Data Registration Methods. *The Open Remote Sensing Journal*, v. 5, n. 1, p. 54–63, 2012. ISSN 18754139. Citado na página 55.
- LAPTEV, I. et al. Automatic extraction of roads from aerial images based on scale-space and snakes Chair for Photogrammetry and Remote Sensing. p. 23–31, 2000. Disponível em: <<https://www.diva-portal.org/smash/get/diva2:338656/FULLTEXT01.pdf>>. Citado na página 22.

- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, 2015. ISSN 14764687. Citado 4 vezes nas páginas 23, 30, 32 e 33.
- LECUN, Y. et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, p. 541–551, 1989. Citado na página 29.
- LEE, S.; PARK, S. J.; HONG, K. S. RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, v. 2017-October, p. 4990–4999, 2017. ISSN 15505499. Citado na página 57.
- LIN, T. Y. et al. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 42, n. 2, p. 318–327, 2020. ISSN 19393539. Citado na página 46.
- LIN, T. Y. et al. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 8693 LNCS, n. PART 5, p. 740–755, 2014. ISSN 16113349. Citado na página 23.
- LIU, L. et al. Aerial Images Meet Crowdsourced Trajectories: A New Approach to Robust Road Extraction. *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–15, 2022. ISSN 2162-237X. Disponível em: <<http://arxiv.org/abs/2111.15119https://ieeexplore.ieee.org/document/9696168/>>. Citado na página 25.
- LIU, Y. et al. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Information Fusion*, Elsevier, v. 42, n. September 2017, p. 158–173, 2018. ISSN 15662535. Disponível em: <<https://doi.org/10.1016/j.inffus.2017.10.007>>. Citado 2 vezes nas páginas 21 e 25.
- LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2015. v. 07-12-June, p. 3431–3440. ISBN 9781467369640. Citado 3 vezes nas páginas 23, 43 e 44.
- MA, J. Segmentation Loss Odyssey. may 2020. Disponível em: <<http://arxiv.org/abs/2005.13449>>. Citado na página 46.
- MA, L. et al. ISPRS Journal of Photogrammetry and Remote Sensing Deep learning in remote sensing applications : A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 152, n. March, p. 166–177, 2019. ISSN 0924-2716. Disponível em: <<https://doi.org/10.1016/j.isprsjprs.2019.04.015>>. Citado 4 vezes nas páginas 21, 23, 24 e 25.
- MACHADO, M. V. *Sincronização por Pós-Processamento de um Sistema de Varredura a LASER embarcado em VANT Sincronização por Pós-Processamento de um Sistema de Varredura a LASER embarcado em VANT*. Tese (Doutorado), 2018. Citado na página 51.
- MARTINS, E. fernando de oliveira. *ATUALIZAÇÃO DE MALHA RODOVIÁRIA EM ÁREA RURAL VIA INTEGRAÇÃO DE DADOS GEOESPACIAIS E SNAKES DE REDE*. 153 p. Tese (Doutorado) — UNESP, 2017. Citado na página 22.

- MCCULLOCH, W. S.; PITTS, W. A logical calculus nervous activity. *Bulletin of Mathematical Biology*, v. 52, n. 1, p. 99–115, 1943. ISSN 00074985. Citado na página 28.
- MENA, J. B. State of the art on automatic road extraction for GIS update: A novel classification. *Pattern Recognition Letters*, v. 24, n. 16, p. 3037–3058, 2003. ISSN 01678655. Citado 2 vezes nas páginas 19 e 25.
- MENDES, T. S. G. *DETECÇÃO DE REGIÕES DE VIA INTEGRANDO DADOS DE VARREDURA A LASER AEROTRANSPORTADO E IMAGENS DE ALTA RESOLUÇÃO*. 161 p. Tese (Doutorado) — UNESP, 2012. Citado na página 26.
- MENDES, T. S. G.; Dal Poz, A. P. Urban road classification in geometrically integrated high-resolution RGB aerial and laser-derived images using the artificial neural network classification method. *International Journal of Image and Data Fusion*, Taylor & Francis, v. 10, n. 1, p. 58–78, jan 2019. ISSN 1947-9832. Disponível em: <<https://doi.org/10.1080/19479832.2018.1469547><https://www.tandfonline.com/doi/full/10.1080/19479832.2018.1469547>>. Citado 2 vezes nas páginas 22 e 26.
- MERKEL, D. Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, Belltown Media, Houston, TX, v. 2014, n. 239, mar 2014. ISSN 1075-3583. Citado na página 62.
- MILLETARI, F.; NAVAB, N.; AHMADI, S. A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, p. 565–571, 2016. Citado na página 46.
- MINSKY, M.; PAPERT, S. Perceptrons: expanded edition. *MIT Press Cambridge MA*, v. 522, p. 20, 1969. ISSN 00199958. Disponível em: <<http://mitpress.mit.edu/book-home.tcl?isbn=0262631113>>. Citado na página 29.
- MNIH, V. *Machine Learning for Aerial Image Labeling*. Tese (Doutorado) — University of Toronto, 2013. Citado 3 vezes nas páginas 61, 65 e 80.
- MNIH, V.; HINTON, G. E. Learning to detect roads in high-resolution aerial images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 6316 LNCS, n. PART 6, p. 210–223, 2010. ISSN 16113349. Citado 3 vezes nas páginas 21, 22 e 24.
- MOU, L.; HUA, Y.; ZHU, X. X. Relation Matters: Relational Context-Aware Fully Convolutional Network for Semantic Segmentation of High-Resolution Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, p. 1–13, 2020. ISSN 0196-2892. Disponível em: <<https://ieeexplore.ieee.org/document/9076866/>>. Citado na página 79.
- NUNES, D. M. *EXTRAÇÃO SEMIAUTOMÁTICA DE MALHA VIÁRIA EM IMAGENS DIGITAIS UTILIZANDO CLASSIFICAÇÃO BASEADA EM OBJETOS E OPERADORES MORFOLÓGICOS* Dissertação. 76 p. Tese (Doutorado) — Universidade Federal de Viçosa, 2016. Citado na página 22.
- OpenStreetMap contributors. *Planet dump retrieved from <https://planet.osm.org>*. 2017. <<https://www.openstreetmap.org>>. Citado na página 20.

- PARAJULI, B.; CHOUDHURY, A. R.; KUMAR, P. Reconstructing road network graphs from both aerial lidar and images. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, p. 1726–1735, 2020. Citado 2 vezes nas páginas 25 e 103.
- PENG, T. et al. Incorporating generic and specific prior knowledge in a multiscale phase field model for road extraction from VHR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 1, n. 2, p. 139–146, 2008. ISSN 21511535. Citado na página 22.
- PETRIE, G.; TOTH, C. K. *Topographic Laser Ranging and Scanning: principles and processing*. 2nd. ed. [S.l.]: Taylor & Francis, CRC Press, 2018. 638 p. ISBN 9781498772273. Citado na página 50.
- PINGEL, T. J.; CLARKE, K. C.; MCBRIDE, W. A. An improved simple morphological filter for the terrain classification of airborne LIDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 77, n. October 2017, p. 21–30, 2013. ISSN 09242716. Disponível em: <<http://dx.doi.org/10.1016/j.isprsjprs.2012.12.002>>. Citado na página 68.
- POHL, C.; Van Genderen, J. L. *Review article Multisensor image fusion in remote sensing: Concepts, methods and applications*. [S.l.: s.n.], 1998. v. 19. 823–854 p. ISSN 13665901. ISBN 0143116982157. Citado na página 48.
- PRATT, L.; MOSTOW, J.; KAMM, C. Direct Transfer of Learned Information Among Neural Networks. *AAAI-91 Proceedings*, p. 584–589, 1991. Disponível em: <www.aaai.org>. Citado na página 42.
- QGIS Development Team. *QGIS Geographic Information System*. [S.l.], 2009. Disponível em: <<http://qgis.osgeo.org>>. Citado na página 62.
- QIAN, N. On the momentum term in gradient descent learning algorithms. *Neural Networks*, v. 12, n. 1, p. 145–151, 1999. ISSN 08936080. Citado na página 36.
- QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986. ISSN 0885-6125. Citado na página 29.
- RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 9351, p. 234–241, 2015. ISSN 16113349. Citado 3 vezes nas páginas 37, 44 e 45.
- ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958. ISSN 0033295X. Citado na página 29.
- ROTTENSTEINER, F. et al. ISPRS Test Project on Urban Classification and 3D Building Reconstruction. *ISPRS - Commission III - Photogrammetric Computer Vision and Image Analysis Working Group III / 4 - 3D Scene Analysis*, p. 1–16, 2013. ISSN 18770428. Citado 2 vezes nas páginas 62 e 63.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, oct 1986. ISSN 0028-0836. Disponível em: <<http://www.nature.com/articles/323533a0>>. Citado 2 vezes nas páginas 29 e 30.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, v. 115, n. 3, p. 211–252, 2015. ISSN 15731405. Citado na página 23.

SALEHI, S. S. M.; ERDOGMUS, D.; GHOLIPOUR, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: *Lecture Notes in Computer Science book series*. Springer International Publishing, 2017. p. 379–387. Disponível em: <<http://link.springer.com/10.1007/978-3-319-67389-944>>. Citado na página 46.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, jul 1959. ISSN 0018-8646. Disponível em: <<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=5392560http://ieeexplore.ieee.org/document/53925>>. Citado na página 29.

SANG, D. V.; MINH, N. D. Fully residual convolutional neural networks for aerial image segmentation. *ACM International Conference Proceeding Series*, p. 289–296, 2018. Citado na página 25.

SANTOS, R. cézar. *TELHADOS DE EDIFICAÇÕES A PARTIR DE DADOS LiDAR USANDO O ALGORITMO ALPHA-SHAPE E CD-SPLINE TELHADOS DE EDIFICAÇÕES A PARTIR DE DADOS LiDAR USANDO O ALGORITMO ALPHA-SHAPE E CD-SPLINE*. 88 p. Tese (Doutorado) — UNESP, 2019. Citado 3 vezes nas páginas 51, 52 e 55.

SCHENK, T.; CSATHÓ, B. Fusion of Lidar Data and Aerial Imagery for a More Complete Surface Description. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, v. 34, n. 3/A, p. 310–317, 2002. Disponível em: <http://rsl.geology.buffalo.edu/documents/schenk{_}csatho{_}gr>. Citado na página 25.

SHAN, J.; TOTH, C. K. *Topographic Laser Ranging and Scanning: principles and processing*. 2nd. ed. [S.l.]: Taylor & Francis, CRC Press, 2018. 638 p. ISBN 9781498772273. Citado 3 vezes nas páginas 51, 54 e 55.

SHANG, R. et al. Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images. *Remote Sensing*, v. 12, n. 5, p. 1–20, 2020. ISSN 20724292. Citado na página 79.

SHI, W.; MIAO, Z.; DEBAYLE, J. An integrated method for urban main-road centerline extraction from optical remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 52, n. 6, p. 3359–3372, 2014. ISSN 01962892. Citado na página 19.

SHI, W.; ZHU, C. The line segment match method for extracting road network from high-resolution satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, v. 40, n. 2, p. 511–514, 2002. ISSN 01962892. Disponível em: <<http://ieeexplore.ieee.org/document/992826/>>. Citado na página 22.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, Springer International Publishing, v. 6, n. 1, 2019. ISSN 21961115. Disponível em: <<https://doi.org/10.1186/s40537-019-0197-0>>. Citado na página 37.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, p. 1–14, 2015. Citado 5 vezes nas páginas 23, 29, 34, 39 e 45.

SONG, J.-H. et al. Assessing the possibility of land-cover classification using lidar intensity data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, v. 34, 2002. ISSN 16821750. Citado 2 vezes nas páginas 55 e 56.

SONG, S.; LICHTENBERG, S. P.; XIAO, J. SUN RGB-D: A RGB-D scene understanding benchmark suite. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 07-12-June-2015, p. 567–576, 2015. ISSN 10636919. Citado na página 59.

SRIVASTAVA, N. et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, v. 15, n. 56, p. 1929–1958, 2014. Disponível em: <<http://jmlr.org/papers/v15/srivastava14a.html>>. Citado 2 vezes nas páginas 29 e 37.

SUN, W.; WANG, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images. *IEEE Geoscience and Remote Sensing Letters*, v. 15, n. 3, p. 474–478, mar 2018. ISSN 1545-598X. Disponível em: <<http://ieeexplore.ieee.org/document/8281008/>>. Citado na página 25.

SUN, Y. et al. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 143, n. September 2017, p. 3–14, sep 2018. ISSN 09242716. Disponível em: <<https://doi.org/10.1016/j.isprsjprs.2018.06.005>><<https://linkinghub.elsevier.com/retrieve/pii/S0924271618301680>>. Citado 3 vezes nas páginas 20, 24 e 25.

SUNDELIUS, C. *Deep Fusion of Imaging Modalities for Semantic Segmentation of Satellite Imagery*. 98 p. Tese (Doutorado) — Linköping University, 2017. Citado na página 29.

SZEGEDY, C. et al. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. v. 91, n. 8, p. 1–9. ISBN 978-1-4673-6964-0. Disponível em: <<http://ieeexplore.ieee.org/document/7298594/>>. Citado 2 vezes nas páginas 23 e 29.

TRASK, A. W. *Grokking Deep Learning*. Shelter Island, NY: manning publications co., 2019. 309 p. ISBN 9781617293702. Citado na página 28.

TURING, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX, n. 236, p. 433–460, oct 1950. ISSN 1460-2113. Disponível em: <<https://academic.oup.com/mind/article/LIX/236/433/986238>>. Citado na página 28.

- VALE, g. M. do. *EXTRAÇÃO SEMI-AUTOMÁTICA DO EIXO DE RODOVIA EM IMAGENS DE MÉDIA E ALTA RESOLUÇÃO USANDO PROGRAMAÇÃO DINÂMICA*. 115 p. Tese (Doutorado), 2003. Citado na página 22.
- VOLPI, M.; TUIA, D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, v. 55, n. 2, p. 881–893, 2017. ISSN 01962892. Citado na página 24.
- VOSSELMAN, G.; HANS-GERD, M. *Airborne and terrestrial laser scanning*. [S.l.]: Whittles publishing, 2010. ISBN 9781904445876. Citado na página 49.
- WANG, H. et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, v. 2020-June, p. 111–119, 2020. ISSN 21607516. Citado na página 84.
- WANG, J. et al. Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing*, v. 36, n. 12, p. 3144–3169, 2015. ISSN 13665901. Citado na página 24.
- WANG, W. et al. A review of road extraction from remote sensing images. *Journal of Traffic and Transportation Engineering (English Edition)*, Elsevier Ltd, v. 3, n. 3, p. 271–282, 2016. ISSN 20957564. Disponível em: <<http://dx.doi.org/10.1016/j.jtte.2016.05.005>>. Citado 3 vezes nas páginas 19, 20 e 22.
- WARNEST, M. *A Collaboration Model for National Spatial Data Infrastructure in Federated Countries*. [S.l.], 2005. Citado na página 18.
- WEHR, A. *Topographic Laser Ranging and Scanning: principles and processing*. 2nd. ed. [S.l.]: Taylor & Francis, CRC Press, 2018. 638 p. ISBN 9781498772273. Citado 2 vezes nas páginas 50 e 51.
- WEHR, A.; LOHR, U. Airborne laser scanning - An introduction and overview. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 54, n. 2-3, p. 68–82, 1999. ISSN 09242716. Citado 3 vezes nas páginas 49, 50 e 51.
- WEI, Y.; WANG, Z.; XU, M. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 14, n. 5, p. 709–713, 2017. ISSN 1545598X. Citado 2 vezes nas páginas 19 e 24.
- WU, H. et al. DeepDualMapper: A Gated Fusion Network for Automatic Map Extraction Using Aerial Images and Trajectories. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, n. 01, p. 1037–1045, apr 2020. ISSN 2374-3468. Disponível em: <<https://aaai.org/ojs/index.php/AAAI/article/view/5453>>. Citado na página 25.
- XU, Y. et al. Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sensing*, v. 10, n. 9, 2018. ISSN 20724292. Citado na página 24.
- YANG, X. et al. Road Detection and Centerline Extraction Via Deep Recurrent Convolutional Neural Network U-Net. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 57, n. 9, p. 7209–7220, 2019. ISSN 0196-2892. Citado na página 24.

- YONG, H. et al. Gradient Centralization: A New Optimization Technique for Deep Neural Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 12346 LNCS, p. 635–652, 2020. ISSN 16113349. Citado na página 38.
- YOSINSKI, J. et al. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, v. 4, n. January, p. 3320–3328, 2014. ISSN 10495258. Citado na página 42.
- ZHANG, J. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, v. 1, n. 1, p. 5–24, mar 2010. ISSN 1947-9832. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/19479830903561035>>. Citado 3 vezes nas páginas 25, 48 e 49.
- ZHANG, Y. et al. Density and diversity of OpenStreetMap road networks in China. *Journal of Urban Management*, Elsevier, v. 4, n. 2, p. 135–146, 2015. ISSN 25890360. Disponível em: <<http://dx.doi.org/10.1016/j.jum.2015.10.001>>. Citado na página 69.
- ZHANG, Z.; WANG, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sensing*, v. 11, n. 6, p. 696, 2019. ISSN 2072-4292. Citado na página 87.
- ZHANG, Z. et al. Road centerline extraction from very-high-resolution aerial image and LiDAR data based on road connectivity. *Remote Sensing*, v. 10, n. 8, 2018. ISSN 20724292. Citado na página 22.
- ZHAO, G.; WANG, J.; ZHANG, Z. Random shifting for CNN: A solution to reduce information loss in Down-sampling layers. *IJCAI International Joint Conference on Artificial Intelligence*, v. 0, p. 3476–3482, 2017. ISSN 10450823. Citado na página 76.
- ZHOU, L.; ZHANG, C.; WU, M. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, v. 2018-June, p. 192–196, 2018. ISSN 21607516. Citado na página 103.
- ZHU, Q. et al. A Global Context-aware and Batch-independent Network for road extraction from VHR satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier B.V., v. 175, n. April, p. 353–365, 2021. ISSN 09242716. Disponível em: <<https://doi.org/10.1016/j.isprsjprs.2021.03.016>>. Citado na página 24.
- ZHU, X. X. et al. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, v. 5, n. 4, p. 8–36, dec 2017. ISSN 2168-6831. Disponível em: <<http://dx.doi.org/10.1109/MGRS.2017.2762307><http://ieeexplore.ieee.org/document/8113128/>>. Citado 2 vezes nas páginas 21 e 26.