



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Campus de Bauru

Daniel Felipe Silva Santos

**Detecção de Mudanças em Imagens Utilizando
Aprendizado Residual de Redes Neurais Convolucionais**

Bauru

2024

Daniel Felipe Silva Santos

**Detecção de Mudanças em Imagens Utilizando
Aprendizado Residual de Redes Neurais Convolucionais**

Tese apresentada como parte dos requisitos para obtenção do título de Doutor em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”.

Financiadoras: FUNDUNESP - 2662/2017 (de 2019 à 2020) e CAPES/DS - 88887.512828/2020-00 (de 2020 à 2023)

Orientador: Prof. Dr. João Paulo Papa

Coorientador: Prof. Dr. Aparecido Nilceu Marana

Bauru

2024

S237d	Santos, Daniel Felipe Silva Detecção de Mudanças em Imagens Utilizando Aprendizado Residual Profundo de Redes Neurais Convolucionais / Daniel Felipe Silva Santos. -- Bauru, 2024 125 p. Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru Orientador: João Paulo Papa Coorientador: Aparecido Nilceu Marana 1. Detecção de Mudanças. 2. Atenção Residual. 3. Rede Neural Convolucional. I. Título.
-------	--

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

Essa ficha não pode ser modificada.

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE DANIEL FELIPE SILVA SANTOS, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 05 dias do mês de janeiro do ano de 2024, às 10:00 horas, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de DANIEL FELIPE SILVA SANTOS, intitulada **Detecção de Mudanças em Imagens Utilizando Aprendizado Residual Profundo de Redes Neurais Convolucionais**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. JOAO PAULO PAPA (Orientador(a) - Participação Virtual) do(a) FC / UNESP Bauru SP, Prof. Dr. ALEXANDRE LUIS MAGALHÃES LEVADA (Participação Virtual) do(a) CCET / UFSCar/São Carlos (SP), Prof. Dr. JURANDY GOMES DE ALMEIDA JUNIOR (Participação Virtual) do(a) Instituto de Ciência e Tecnologia / Universidade Federal de São Paulo, Prof. Dr. ANDRE LUIS DEBIASO ROSSI (Participação Virtual) do(a) Engenharia de Produção / UNESP - Campus Experimental de Itapeva, Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Participação Virtual) do Departamento de Ciência da Computação / UNESP Bauru. Após a exposição pelo doutorando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final APROVADO. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo Presidente da Comissão Examinadora.

Prof. Dr. JOAO PAULO PAPA

Daniel Felipe Silva Santos

**Detecção de Mudanças em Imagens Utilizando
Aprendizado Residual de Redes Neurais Convolucionais**

Tese apresentada como parte dos requisitos para obtenção do título de Doutor em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”.

Financiadoras: FUNDUNESP - 2662/2017 (de 2019 à 2020) e CAPES/DS - 88887.512828/2020-00 (de 2020 à 2023)

Comissão Examinadora

Prof. Dr. João Paulo Papa (Orientador)
Universidade Estadual Paulista “Júlio de Mesquita Filho”

Prof. Dr. Jurandy Gomes de Almeida Junior
Universidade Federal de São Paulo

Prof. Dr. Alexandre Luis Magalhães Levada
Universidade Federal de São Carlos

Prof. Dr. André Luis Debiasso Rossi
Universidade Estadual Paulista “Júlio de Mesquita Filho”

Prof. Dr. Andre Luis Debiasso Rossi
Universidade Estadual Paulista “Júlio de Mesquita Filho”

Bauru
06 de fevereiro de 2024

AGRADECIMENTOS

Aos meus queridos pais Daniel e Creunice, por todo suporte e também pelos inúmeros conselhos nas horas mais difíceis;

Ao meu orientador, prof. Dr. João Paulo Papa, e também ao meu co-orientador, prof. Dr. Aparecido Nilceu Marana, por todo apoio, paciência e confiança que sempre depositaram em mim;

Agradeço aos meus amigos Lucas, Rafael e Marcos, com os quais tive a honra de partilhar inúmeros momentos de aprendizagem ao longo de mais essa etapa da minha jornada. Agradeço pela imensa generosidade em partilharem comigo suas visões únicas de mundo;

Finalmente, gostaria de agradecer pela oportunidade de trabalhar no projeto Sistema Supervisório de SMS para Sondas, financiado pela FUNDUNESP (2662/2017). Sem dúvidas uma chance ímpar de ampliar as fronteiras do meu conhecimento. Além disso, sou muito grato também ao amparo financeiro prestado pela CAPES/DS (88887.512828/2020-00). Sem esse apoio, não teria sido possível prosseguir com minhas atividades de pesquisa ao longo da segunda metade do meu doutorado.

RESUMO

Deteção de mudanças em vídeos e imagens diz respeito ao processo de separar os conteúdos de primeiro plano da cena observada, do inglês *foreground*, dos conteúdos de segundo plano, do inglês *background*. Bons métodos de detecção de mudanças necessitam trabalhar de forma rápida e precisa sob condições adversas intrínsecas, relacionadas a falhas dos dispositivos de captura utilizados, e extrínsecas, relacionadas a alterações ambientais de clima e luminosidade capazes de induzir processos de detecção falso positivos e falso negativos. Para superar tais desafios propõe-se, num primeiro momento, o mecanismo denominado de atenção residual implícita, a ser avaliado pelas técnicas propostas de detecção de mudanças em vídeos denominadas Rede Neural Convolutacional Residual em Cascata (CRCNN) e Rede Neural Convolutacional em Cascata Multiescala (MCRCNN). Em um segundo momento propõe-se o mecanismo de atenção residual explícita, avaliado pelo método proposto denominado de Rede Leve de Atenção Temporal (TITAN). Resultados quantitativos produzidos pelas técnicas CRCNN e MCRCNN, calculados sobre a base de dados de Detecção de Mudanças 2014 (CD2014), indicaram ser vantajosa a utilização do mecanismo de atenção residual implícita, sendo a técnica MCRCNN superior à técnica CRCNN e a demais técnicas em estado da arte, como RTSegNet e DeepBS. Resultados quantitativos produzidos pela técnica TITAN, calculados sobre as bases de dados SZATAKI, Onera, LEVIR-CD e SYSU-CD, demonstraram ser vantajosa a utilização de mecanismo de atenção residual explícita. Para as mesmas bases de dados, a técnica TITAN também superou técnicas em estado da arte como FLS e STANet e equiparou-se com técnicas como DMINet e RSCDNet. Além disso, obteve resultados superiores ao de técnicas como FgSegNet_S, FgSegNet_M, RtSegNet, MUNets, DeepBS, SEUNet3 e MCRCNN sobre a base de vídeos CD2014, enquanto equiparou-se com técnicas como DMINet e FgSegNet_v2, considerando a base de dados PetrobrasROUTES. Para quadros de dimensão 690×520 , nos piores casos analisados, a técnica TITAN realizou $25,8 \times 10^9$ menos operações de ponto flutuante (FLOPs) e necessitou armazenar em memória $3,04 \times 10^6$ menos parâmetros, em comparação com a técnica DMINet. Também conseguiu superar a taxa de processamento de quadros por segundo (FPS) da técnica MU-Net1 em 20 quadros.

Palavras-chave: Detecção de Mudanças; Atenção Residual; Rede Neural Convolutacional.

ABSTRACT

Video and image change detection relates to the process of decomposing an observed scene into foreground and background contents. Good change detection methods need to work in a fast and precise way under intrinsic adverse conditions, related to failures of the used capture devices, and extrinsic ones, regarding to climate and luminosity environment changes capable of inducing false positive and false negative detections. To overcome such challenges it is proposed, at first, a mechanism called implicit residual attention, to be evaluated using the proposed techniques named Cascade Residual Convolutional Neural Network (CRCNN) and Multiscale Cascade Residual Convolutional Neural Network (MCRCNN). In a second moment, it is proposed a mechanism of explicit residual attention, to be evaluated by the proposed method called LightWeight Temporal Attention Network (TITAN). Quantitative results produced by CRCNN and MCRCNN techniques, computed over Change Detection 2014 dataset (CD2014), indicate that the implicit residual attention mechanism usage was advantageous, in that MCRCNN technique surpassed CRCNN and other state-of-the-art techniques like RTSegNet and DeepBS. Quantitative results produced by TITAN technique, concerning SZATAKI, Onera, LEVIR-CD, and SYSU-CD datasets, shown that the explicit residual attention mechanism usage was advantageous. For the same datasets, the TITAN technique also surpassed other state-of-the-art techniques like FLS and STANet, and was on par with techniques such as DMINet and RSCDNet. Besides, it achieved on par results with DMInet and FgSegNet_v2 regarding PetrobrasROUTES dataset. Considering frames of sizes 690×520 , in the worst case scenarios, TITAN technique executes 25.8×10^9 less floating point operations (FLOPs) and needed to store 3.04×10^6 less parameters into memory in comparison against DMINet technique. It also overcame MU-Net1 frames per second processing rate by 20 frames.

Keywords: Change Detection; Residual Attention; Convolutional Neural Network.

Lista de ilustrações

Figura 1 – Exemplo de um trecho de vídeo.	24
Figura 2 – Geração de <i>background</i>	25
Figura 3 – Exemplo de máscara de <i>foreground</i>	26
Figura 4 – Diagrama do modelo <i>neocognitron</i>	28
Figura 5 – Aplicação de operação de convolução.	29
Figura 6 – Aplicação das operações de agregação.	30
Figura 7 – Convolução Espacial Dilatada	32
Figura 8 – Conexão Residual	33
Figura 9 – Arquitetura de rede DeepBS	35
Figura 10 – Arquitetura de rede CNN multiescala	36
Figura 11 – Arquitetura da rede CNN multiestágio	36
Figura 12 – Arquitetura de rede FgSegNet_M	37
Figura 13 – Arquitetura de módulo FPM	38
Figura 14 – Arquitetura de rede FgSegNet_v2	39
Figura 15 – Arquitetura de módulo M_FPM	40
Figura 16 – Arquitetura de rede SEUNet1	41
Figura 17 – Arquiteturas de redes SEUNet2 e SEUNet3	42
Figura 18 – Arquiteturas de redes MU-Net1 e ResNet18	43
Figura 19 – Arquitetura de rede MU-Net2	44
Figura 20 – Arquitetura de rede RTSegNet e Módulo Assimétrico	45
Figura 21 – Arquitetura de rede FLS	46
Figura 22 – Arquitetura de Rede STANet	47
Figura 23 – Arquitetura de rede RSCDNet	48
Figura 24 – Arquitetura de módulos MSA e GL-ASPP	49
Figura 25 – Arquitetura de rede DMINet e Módulo JointAtt	50
Figura 26 – Arquitetura de rede CRCNN	52
Figura 27 – Arquitetura de rede MCRCNN	54
Figura 28 – Arquitetura de módulo RPM	55
Figura 29 – Arquitetura de rede TITAN	56

Figura 30 – Arquitetura de módulo TCAM	58
Figura 31 – Base de dados CD2014	60
Figura 32 – Base de dados SZATAKI	62
Figura 33 – Base de dados Onera	63
Figura 34 – Base de dados LEVIR-CD	63
Figura 35 – Base de dados SYSU-CD	64
Figura 36 – Base de dados PetrobrasROUTES	66
Figura 37 – Subdivisão de base de dados SYSU-CD	69
Figura 38 – Subdivisão de bases de dados SZATAKI, Onera e LEVIR-CD	70
Figura 39 – Análises de Configurações Residuais Explícitas	75
Figura 40 – Ranqueamento PR e ROC etapa A_1	76
Figura 41 – Ranqueamento Medida-F e IOU etapa A_1	77
Figura 42 – Ranqueamento PWC etapa A_1	78
Figura 43 – Ranqueamento PR e ROC etapa A_2	79
Figura 44 – Ranqueamento Medida-F e IOU etapa A_2	80
Figura 45 – Ranqueamento PWC etapa A_2	81
Figura 46 – Ranqueamento PR e ROC etapa A_3	82
Figura 47 – Ranqueamento Medida-F, IOU e PWC etapa A_3	83
Figura 48 – Ranqueamento PR e ROC etapa A_4	84
Figura 49 – Ranqueamento Medida-F, IOU e PWC etapa A_4	85
Figura 50 – Ranqueamento PR e ROC etapa A_5	87
Figura 51 – Ranqueamento Medida-F, IOU e PWC etapa A_5	87
Figura 52 – Ranqueamento PR etapa A_6	89
Figura 53 – Ranqueamento Medida-F e IOU para etapa A_6	89
Figura 54 – Ranqueamento PWC para etapa A_6	90
Figura 55 – Ranqueamento Medida-F e PWC etapa A_8	91
Figura 56 – Ranqueamento IOU para etapa A_8	92
Figura 57 – Ranqueamento PR e ROC etapa A_9	92
Figura 58 – Ranqueamento Medida-F, IOU e PWC etapa A_9	93
Figura 59 – Atenção Residual Explícita	94
Figura 60 – Atenção Residual Implícita Nível 1	96
Figura 61 – Atenção Residual Implícita Nível 2	97

Figura 62 – Ranqueamento PR Vídeos Unitários Base CD2014	99
Figura 63 – Ranqueamento ROC Vídeos Unitários Base CD2014	100
Figura 64 – Ranqueamento IOU e Medida-F Vídeos Unitários Base CD2014	100
Figura 65 – Ranqueamento PWC Vídeos Unitários Base CD2014	101
Figura 66 – Resultados Qualitativos Vídeos Unitários Base CD2014 Exemplo 1 . .	102
Figura 67 – Resultados Qualitativos Vídeos Unitários Base CD2014 Exemplo 2 . .	103
Figura 68 – Ranqueamento PR e ROC Grupo de Vídeos Base CD2014	104
Figura 69 – Ranqueamento IOU, Medida-F e PWC Grupo de Vídeos Base CD2014	105
Figura 70 – Resultados Qualitativos Grupo de Vídeos Base CD2014 Exemplo 1 . .	106
Figura 71 – Resultados Qualitativos Grupo de Vídeos Base CD2014 Exemplo 2 . .	107
Figura 72 – Resultados Qualitativos Grupo de Vídeos Base CD2014 Exemplo 3 . .	108
Figura 73 – Ranqueamento Medida-F, IOU e PWC Imagens Aéreas	110
Figura 74 – Resultados Qualitativos Base de Imagens Aéreas Exemplo 1	110
Figura 75 – Resultados Qualitativos Base de Imagens Aéreas Exemplo 2	111
Figura 76 – Ranqueamento Medida-F e IOU Imagens PetrobrasROUTES	112
Figura 77 – Ranqueamento PWC Imagens PetrobrasROUTES	112
Figura 78 – Resultados Qualitativos Base de Imagens PetrobrasROUTES	113

Lista de tabelas

Tabela 1 – Resultados de Configuração Etapa A ₁	76
Tabela 2 – Resultados de Configuração Etapa A ₂	79
Tabela 3 – Resultados de Configuração Etapa A ₃	82
Tabela 4 – Resultados de Configuração Etapa A ₅	86
Tabela 5 – Resultados de Configuração Etapa A ₆	88
Tabela 6 – Resultados de Configuração Etapa A ₇	90
Tabela 7 – Resultados de Configuração Etapa A ₈	91
Tabela 8 – Resultados de Configuração TITAN e TITAN_b	95
Tabela 9 – Resultados de Comparação Unitária CD2014	99
Tabela 10 – Resultados de Comparação por Grupo CD2014	104
Tabela 11 – Resultados de Comparação Utilizando Imagens Aéreas	109
Tabela 12 – Resultados de Comparação Imagens PetrobrasROUTES	112
Tabela 13 – Comparação de Custo Computacional	114

Lista de abreviaturas e siglas

PCA	<i>Principal Component Analysis</i>
LBP	<i>Local Binary Pattern</i>
CNN	<i>Convolutional Neural Network</i>
CRCNN	<i>Cascade Residual Convolutional Neural Network</i>
MCRCNN	<i>Multiscale Cascade Residual Convolutional Neural Network</i>
TITAN	<i>highTwoIght Temporal Attention Network</i>
SubSENSE	Self-Balanced SENSitivity SEgmenter
FP	<i>False Positive</i>
FN	<i>False Negative</i>
TP	<i>True Positive</i>
TN	<i>True Negative</i>
PWC	<i>Percentage of Wrong Classification</i>
ReLU	<i>Rectified Linear Unity</i>
GPGPU	<i>General Purpose Graphical Processing Unit</i>
VGG	<i>Visual Geometry Group</i>
DeepBS	<i>Deep Background Subtraction</i>
FgSegNet_M	<i>Multiple Foreground Segmentation Networks</i>
FgSegNet_S	<i>Single Foreground Segmentation Network</i>
RGB	<i>Red Green Blue</i>
VGG-16	<i>Visual Geometry Group 16</i>
TCNN	<i>Transposed Convolutional Neural Network</i>

FPM	<i>Feature Pooling Module</i>
BN	<i>Batch Normalization</i>
SD	<i>Spatial Dropout</i>
FgSegNet_v2	<i>Single Foreground Segmentation Network version 2</i>
GAP	<i>Global Average Pooling</i>
IN	<i>Instance Normalization</i>
M_FPM	<i>Modified Feature Pooling Module</i>
SEUNet	<i>Semantic U-Net</i>
MU-Net	<i>Motion U-Net</i>
ResNet18	<i>Residual Network 18</i>
MOG2	<i>Mixtures of Gaussians</i>
RTSegNet	<i>Real-Time Segmentation Network</i>
FLS	<i>Focal Loss Segmenter</i>
FL	<i>Focal Loss</i>
STANet	<i>Spatial-Temporal Attention Network</i>
BAM	<i>Basic Spatial-Temporal Attention Module</i>
PAM	<i>Pyramid Spatial-Temporal Attention Module</i>
RSCDNet	<i>Remote Sensing Change Detection Network</i>
MSA	<i>Modified Self-Attention</i>
GL-ASPP	<i>Gated Linear Atrous Spatial Pyramid Pooling</i>
DMINet	<i>Dual-Branch Multilevel Intertemporal Network</i>
JointAtt	<i>Joint Attention</i>
BCNN	<i>Background Convolutional Neural Network</i>

SCNN	<i>Segmentation Convolutional Neural Network</i>
RPM	<i>Residual Processing Module</i>
TCAM	<i>Temporal Change Attention Module</i>
LFPM	<i>Lightweight Feature Processing Module</i>
TCE	<i>Temporal Change Expectation</i>
AG	<i>Attention Gate</i>
RAM	<i>Random Access Memory</i>
CD2014	<i>Change Detection Dataset 2014</i>
L. Base	Linha de Base
<i>Back. Din.</i>	<i>Background Dinâmico</i>
C. Trêm.	Câmera Trêmula
Mov. Int.	Movimento Intermitente
Mov. Int.	Movimento Intermitente de Objeto
Ter.	Térmica
Mau Tmp.	Mau Tempo
Exib. Lent.	Exibição Lenta
Vid. Notur.	Vídeos Noturnos
PTZ	<i>Pan Tilt Zoom</i>
Turb.	<i>Turbulência</i>
COCO	<i>Common Objects In Context</i>
ROC	<i>Reciever Operator Curve</i>
PR	<i>Precision Recall</i>
train.	<i>training</i>

val.	<i>validation</i>
Rand.	<i>Randomization</i>
IBN	<i>Interleaved Batch Normalization</i>
IIN	<i>Interleaved Instance Normalization</i>
FLOPs	<i>Floating-point Operations</i>
FPS	<i>Frames Per Second</i>

Sumário

1	INTRODUÇÃO	17
1.1	Organização do Trabalho	21
2	FUNDAMENTAÇÃO TEÓRICA	23
2.1	Deteccção de Mudanças em Vídeos e Imagens	23
2.1.1	Imagem Digital	23
2.1.2	Vídeo	23
2.1.3	Modelagem de Conteúdo de Segundo Plano	24
2.1.4	Destaque de Conteúdo de Primeiro Plano	25
2.2	Redes Neurais Convolucionais	27
2.2.1	Modelagem Computacional	27
2.2.2	Processo de Convolução	29
2.2.3	Ativação de Camadas	29
2.2.4	Mecanismos de Subamostragem de Características	30
2.2.5	Processo de Convolução Espacial Dilatada	31
2.2.6	Aprendizado Profundo Residual	32
2.3	Trabalhos Correlatos	34
2.3.1	Rede Neural de Segmentação Profunda de <i>Background</i>	34
2.3.2	Rede Neural Convolucional Multiestágio	35
2.3.3	Redes Neurais Codificadoras-Decodificadoras Multiescala	37
2.3.4	Rede Neural Autocodificadora Multiescala Versão 2	39
2.3.5	Redes Neurais Semânticas U	41
2.3.6	Redes Neurais de Deteccção de Movimento U	43
2.3.7	Rede de Deteccção de Mudanças de Tempo Real	44
2.3.8	Rede Convolucional Siamesa e Aprendizado Focalizado	45
2.3.9	Rede Convolucional de Atenção Espaço-Temporal	46
2.3.10	Rede de Deteccção de Mudanças para Sensoreamento Remoto	47
2.3.11	Rede de Deteccção de Mudanças Intertemporais Multiníveis	49
3	PROPOSTA	51

3.1	Atenção Residual Implícita	51
3.1.1	Rede Neural Convolucional Residual em Cascata	52
3.1.2	Rede Neural Convolucional Residual em Cascata Multiescala	54
3.2	Atenção Residual Explícita	55
3.2.1	Rede Leve de Atenção Temporal	56
4	RECURSOS E METODOLOGIA	60
4.1	Base de Detecção de Mudanças em Vídeos	60
4.2	Base de Imagens Aéreas de Sensoriamento Remoto	62
4.3	Base PetrobrasROUTES de Imagens de Obstruções	64
4.4	Protocolos de Treinamento e Avaliação	66
4.4.1	Protocolos de Estudo de Configuração	67
4.4.2	Protocolos de Estudo Comparativo	70
5	ANÁLISE DE RESULTADOS	74
5.1	Resultados dos Estudos de Configuração	74
5.1.1	Resultados de Configuração Residual Implícita	74
5.1.2	Resultados de Configuração Residual Explícita	95
5.2	Resultados dos Estudos Comparativos	98
5.2.1	Comparação em Base CD2014	98
5.2.2	Comparação em Base de Imagens Aéreas	109
5.2.3	Comparação em Base PetrobrasROUTES	111
5.2.4	Comparação de Custo Computacional	114
6	CONCLUSÃO	115
6.1	Propostas para Trabalhos Futuros	116
6.2	Trabalhos Desenvolvidos	117
	REFERÊNCIAS	119

1 Introdução

A detecção de mudanças em vídeos e imagens, também chamada de subtração de conteúdo de segundo plano, do inglês *background*, ou então de detecção de objetos que se movem, é uma tarefa que consiste na identificação e destaque de regiões ocupadas por *pixels* que sofreram alterações substanciais de intensidade, do inglês *foreground*, ao longo de um intervalo de tempo transcorrido entre sucessivas observações realizadas. Diante da definição apresentada, nota-se que o problema em se detectar mudanças em vídeos e imagens é algo extremamente abrangente, onde até mesmo pequenas interferências de natureza determinística ou não, quando introduzidas durante os intervalos de observação, são passíveis de induzir processos de detecção falso positivos e falso negativos.

Interferências de caráter determinístico relacionam-se com a baixa qualidade e falta de ajuste correto dos dispositivos utilizados para captura das imagens. Elas podem surgir na forma de ruídos, vide (BOVIK, 2000), gerados principalmente durante processos de conversão analógico-digital, na forma de oscilações eletrônicas causadas por variação de temperatura ambiente, ou mesmo geradas por falhas de ajuste focal das lentes dos dispositivos utilizados. Em contrapartida, condições adversas de caráter não determinístico relacionam-se com fatores extrínsecos aos dispositivos utilizados, tais como (a) condições climáticas de chuva, neblina, nevasca, (b) fenômenos óticos como reflexos, sombras, camuflagem e (c) dinâmicas imprevisíveis de movimento, tais como movimentação intermitente ou aparente¹ dos conteúdos de *foreground* e *background*. Além disso, há que se considerar problemas de ordem prática como a inviabilidade de separar manualmente *pixels* de *foreground* e *background* em imagens de dimensões muito grandes.

Levando-se em consideração as dificuldades apresentadas e a grande relevância que a detecção de mudanças possui nos processos automáticos de tomada de decisão, mediante interpretação de conteúdo visual, como monitoramento de tráfego (LIU et al., 2017), navegação autônoma de veículos (TARMIZI; AZIZ, 2018), rastreamento de objetos e pessoas (FANG et al., 2018), reconhecimento de ações (SUDHAKARAN; ESCALERA; LANZ, 2020), vigilância de ambientes (YANG; SU, 2018) e detecção de anomalias (SULTANI; CHEN; SHAH, 2018), diversos métodos têm sido propostos na tentativa de aumentar cada

¹ Sensação de movimentação da cena causada por deslocamento repentino de câmera.

vez mais a acurácia de detecção sem que o consumo de recursos computacionais, como memória e processamento, sejam elevados. Nesse sentido o desenvolvimento de técnicas para detecção de mudanças em vídeos e imagens, principalmente ao longo das últimas décadas, tem seguido por duas principais vertentes de pesquisa, uma mais especialista e outra mais generalista.

A vertente especialista concentra seus estudos no desenvolvimento de técnicas que buscam por padrões indicativos de constância (característica muito presente nos *pixels* de *background*) e inconstância (característica muito presente nos *pixels* de *foreground*) a serem utilizados, de maneira explícita, durante os processos de identificação das regiões de mudança. As técnicas especialistas são capazes de lidar com diferentes propriedades extraídas dos relacionamentos entre os *pixels* analisados, tais como (i) estatísticas paramétricas, à exemplo dos trabalhos de Stauffer e Grimson (1999) e de KaewTraKulPong e Bowden (2002), que utilizam misturas gaussianas para identificação de oscilações temporais de intensidades dos *pixels* monitorados, (ii) estatísticas não-paramétricas, onde Bevilacqua, Stefano e Lanza (2005) e também Lanza e Stefano (2011) tratam a detecção de mudanças em vídeos como processos estocásticos, cuja propriedade estacionária quando violada² ao longo do tempo permite inferir estados de mudanças, (iii) transformações lineares de espaço, como nos trabalhos de Dong e DeSouza (2011), de Javed et al. (2016) e de Lu et al. (2019), onde técnicas baseadas em Análise de Componentes Principais, do inglês *Principal Component Analysis* (PCA), são utilizadas para facilitar o processo de detecção de mudanças por meio da projeção das regiões de *foreground* em subespaços vetoriais gerados a partir de imagens de *background*, (iv) análise de texturas, onde Bilodeau, Jodoin e Saunier (2013), St-Charles e Bilodeau (2014) e também St-Charles, Bilodeau e Bergevin (2014) desenvolvem algoritmos capazes de utilizar Padrões Binários Locais, do inglês *Local Binary Patterns* (LBP), no processo de detecção de mudanças dos *pixels* e (v) agrupamento por diferenciação de características (cor, brilho e LBP dos *pixels*), onde Kim et al. (2005), Wu e Peng (2010) e também St-Charles, Bilodeau e Bergevin (2015) propõem técnicas auto-organizáveis capazes de criar agrupamentos identificáveis de *pixels* (grupos contendo *pixels* de *background* ou então *foreground*).

A vertente generalista se apoia em abordagens conexionistas de aprendizado, mais especificamente redes neurais convolucionais (FUKUSHIMA, 1980) e (LECUN, 1989), do

² Variações locais de intensidade de *pixels* muito intensas para serem julgadas apenas como ruído gerado durante o processo de captura.

inglês *Convolution Neural Networks* (CNNs), que por serem biologicamente inspiradas no comportamento do sistema visual dos primatas, vide (HUBEL; WIESEL, 1968), têm a habilidade de encontrar padrões de mudanças de maneira adaptativa, na forma de estruturas hierárquicas complexas, por meio de processo heurístico de otimização de parâmetros. A convergência do processo de treinamento de parâmetros das CNNs de detecção de mudanças dependerá de fatores como configuração de arquitetura, quantidade e diversidade de exemplos de *foreground* e *background* disponíveis.

Técnicas de natureza generalista começam a ser desenvolvidas a partir dos estudos de Braham e Droogenbroeck (2016), que utilizam uma arquitetura de CNN baseada na rede *LeNet-5*, de LeCun (1989), para detectar mudanças em vídeos. As mudanças são retornadas pela rede CNN treinada na forma de mapas locais de probabilidades, que após serem limiarizados, se tornam máscaras classificadoras de regiões de *foreground* (posições marcadas com valor um) e *background* (posições marcadas com valor zero). Apesar de demonstrar significativa superioridade em comparação com técnicas especialistas, principalmente quando anotações precisas de *foreground* e *background* são fornecidas para treinamento, a técnica de Braham e Droogenbroeck (2016) deixa espaço para melhorias em termos de aumento da velocidade de execução, capacidade adaptativa de aprendizado³, acurácia de detecção e diminuição da quantidade de parâmetros.

Dentre as diversas estratégias generalistas propostas para tentar solucionar os problemas apresentados pela técnica de Braham e Droogenbroeck (2016), como mostram os levantamentos mais recentes de Bouwmans et al. (2019), de Mandal e Vipparthi (2021) e de Zhao et al. (2022), levando-se em consideração os métodos de melhor desempenho, é possível destacar ao menos cinco grupos de técnicas, sendo esses grupos compostos pelas técnicas ditas (1) completamente convolucionais simples sem e com utilização de processamento multiestágio (redes em cascata), como as propostas de Babaei, Dinh e Rigoll (2018) e Wang, Luo e Jodoin (2017), respectivamente, pelas (2) técnicas de estrutura codificadora-decodificadora que utilizam camadas de subamostragem e dilatação de filtros de convolução, como as propostas de Lim e Keles (2018) e Lim e Keles (2019), por (3) CNNs codificadoras-decodificadores residuais, como as propostas por Rahmon et al. (2021) e Lin et al. (2021), pelas (4) redes siamesas que não utilizam mecanismos de atenção, como

³ Capacidade de aproveitamento de um mesmo modelo de CNN para vários contextos diferentes de mudanças (por exemplo vídeos distintos) e também de produzir resultado precisos mesmo que poucas imagens de treinamento sejam utilizadas.

as propostas por Santana et al. (2019) e Wang et al. (2021) e (5) pelas redes siamesas que utilizam mecanismos de atenção espacial e temporal, como as propostas por Chen e Shi (2020b), Barkur et al. (2022) e Feng et al. (2023).

Apesar das estratégias de construção e treinamento das técnicas generalistas solucionarem diversos problemas de ordem determinística, não-determinística e de ordem prática, nota-se que essas soluções muitas vezes não se coordenam entre si. Isso faz com que a solução proposta, para determinada instância do problema de treinamento ou de detecção de mudanças, acabe gerando outros problemas em contrapartida. Nesse sentido, as técnicas em (1) que não utilizam multiestágio representam soluções de baixo custo computacional, em termos de pouco gasto de memória e rápido processamento, mas também falham em capturar informação contextual, aumentando as chances de detecções falso negativas ocorrerem. As técnicas de (2) têm uma capacidade natural, proporcionada por sua estrutura composta por camadas de subamostragem, muito maior de extrair informação contextual, porém apresentam-se como soluções de maior custo computacional, principalmente no que diz respeito ao maior gasto de memória para armazenamento da estrutura de parâmetros. Além disso, as chances de detecções falso positivas e especialização exagerada de treinamento, do inglês *overfitting*, aumentam. Enquanto os métodos em (3) conseguem minimizar os efeitos nocivos da grande quantidade de parâmetros das técnicas em (2), eles não conseguem solucionar totalmente os problemas de detecções falso positivas. As redes em (4) e (5) apresentam-se como soluções mais adaptativas em comparação com (1) à (3), apresentando menos problemas de *overfitting*. Elas também utilizam uma estratégia de compartilhamento de camadas como alternativa ao problema de gasto de memória. Apesar disso, essas redes sofrem com um alto custo de processamento computacional, pois a redução de camadas implica na duplicação da quantidade de execuções das operações de convolução durante os processos de detecção de mudanças.

Reconhecendo as limitações ainda existentes, mesmo nas técnicas convolucionais de detecção de mudanças em vídeos e imagens, o principal objetivo desse trabalho é desenvolver uma nova técnica generalista, nos moldes de métodos pertencentes à (5), que satisfaça demandas ainda existentes nas técnicas de (1) à (5), tais como redução da quantidade de parâmetros e otimização de mecanismos de atenção e processamento de imagens.

As principais contribuições desse trabalho incluem:

- Modelagem computacional de mecanismo implícito de pré-atenção espacial, com base em (POSNER, 1980);
- Modelagem computacional de mecanismo explícito de pré-atenção espacial e temporal seletivas, com base em (LABERGE; BROWN, 1989);
- Criação de um novo método de detecção de mudanças de duplo estágio denominado Rede Neural Convolutacional Residual em Cascata, do inglês *Cascade Residual Convolutional Neural Network* (CRCNN);
- Rede Neural Convolutacional Residual em Cascata Multiescala, do inglês *Multiscale Cascade Residual Convolutional Neural Network* (MCRCNN);
- Criação de um novo método de detecção de mudanças conexionista de estágio único TITAN denominado Rede Leve de Atenção Temporal, do inglês *lightWeight Temporal Attention Network* (TITAN).

1.1 Organização do Trabalho

O trabalho subdivide-se em:

Capítulo 1: Apresenta uma breve introdução que define e contextualiza os principais desafios da atividade de detecção de mudanças em vídeos e imagens. Além disso, é traçado um panorama geral sobre as principais linhas de pesquisa, baseadas ou não em aprendizado de máquina, que oferecem soluções em estado da arte para tais desafios. Ao final apresentam-se as principais contribuições de nossa pesquisa;

Capítulo 2: Apresenta os fundamentos teóricos básicos necessários para se compreender o contexto de atuação das técnicas baseadas em aprendizado profundo de redes neurais convolucionais na resolução do problema de detecção de mudanças em vídeos e imagens;

Capítulo 3: Apresenta a proposta de desenvolvimento de um método eficiente de detecção de mudanças capaz de atuar de maneira ótima tanto em vídeos quanto em imagens.

Capítulo 4: Apresenta os recursos de *hardware*, *software* e bases de dados utilizados no desenvolvimento e avaliação de nossa pesquisa;

Capítulo 5: Apresenta, compara e discute os resultados quantitativos e qualitativos obtidos pelos métodos CRCNN, MCRCNN e TITAN propostos;

Capítulo 6: Apresenta as conclusões mediante análise dos resultados obtidos. Destacam-se as contribuições das técnicas propostas, além de serem sugeridas algumas direções futuras de continuidade dessa pesquisa.

2 Fundamentação Teórica

Nesse capítulo são apresentadas as bases teóricas necessárias para compreensão dos principais mecanismos computacionais utilizados no desenvolvimento de métodos baseados em aprendizado profundo de CNNs no contexto de detecção de mudanças em vídeos e imagens.

2.1 Detecção de Mudanças em Vídeos e Imagens

Para entender melhor o funcionamento do processo automático de detecção de mudanças em vídeos e imagens é necessário que sejam caracterizados alguns de seus elementos centrais, sendo eles a imagem digital, vídeo, modelagem de conteúdo de segundo plano, do inglês *background* e destaque de conteúdo de primeiro plano *foreground*.

2.1.1 Imagem Digital

Como apontado por Jain (1989) e por Gonzalez e Woods (2006), a imagem digital é um modelo matemático capaz de representar discretamente a intensidade luminosa capturada de determinada cena do ambiente real por dispositivo sensível ao espectro de luz. As intensidades luminosas são digitalmente representadas por elementos pictográficos, do inglês *pixels*, que matematicamente podem ser expressos na forma de $f(x, y, c)$, onde o par ordenado (x, y) indica respectivamente as posições coluna e linha no domínio espacial do plano cartesiano ao qual o *pixel* pertence e o parâmetro c indica a localização do *pixel* no domínio de cores, onde nas imagens monocromáticas¹ $c = 1$, nas imagens coloridas baseadas nas cores primárias² $c = 3$ e nas imagens geradas para representar informação além do espectro de luz visível, como profundidade e calor, $c > 3$.

2.1.2 Vídeo

Enquanto uma imagem digital compreende um conjunto espacialmente organizado de *pixels*, um vídeo compreende um conjunto de imagens ordenadas no eixo temporal. Em

¹ Imagens de única tonalidade, também conhecidas por imagens em tons de cinza.

² Imagens formadas a partir da combinação entre as cores primárias vermelho, verde e azul (RGB).

outras palavras, um vídeo pode ser definido como um conjunto $V = \{I_{t_1}, I_{t_2}, I_{t_3}, \dots, I_{t_n}\}$, composto por n imagens I coletadas sequencialmente durante um intervalo de tempo $T = \{t_1, t_2, t_3, \dots, t_n\}$. Em geral, especialmente nos dispositivos de captura mais modernos, t_n representa pequenas frações de tempo, da ordem de 1/30 segundos. A Figura 1 ilustra um conjunto de três imagens dispostas de maneira ordenada (da esquerda para direita) no eixo temporal, onde fica caracterizado o deslocamento de um objeto de *foreground*.

Figura 1 – Exemplo de trecho de vídeo.



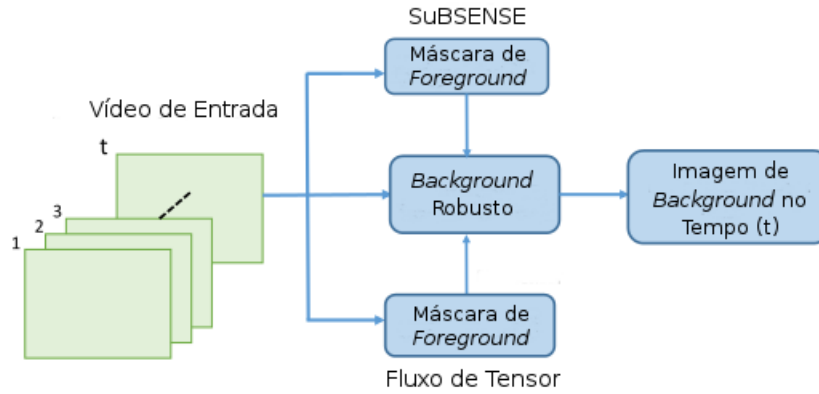
Fonte: Imagens retiradas da base de dados de Wang et al. (2014).

2.1.3 Modelagem de Conteúdo de Segundo Plano

Dado um vídeo V , definido na Subseção 2.1.2, seu conteúdo de segundo plano, do inglês *background*, é representado por uma imagem I_b cujas intensidades de seus *pixels*, idealmente, não variam ao longo de um intervalo de tempo T . Estimar corretamente tal modelo é uma tarefa bastante desafiadora, tendo em vista o caráter não determinístico que estas variações de intensidade possuem.

Como alternativa para geração de imagens de *background* para as técnicas que delas necessitarão, futuramente abordadas na Seção 2.3, tem-se o método proposto por Babaei, Dinh e Rigoll (2018), que por sua vez combina as técnicas Segmentador de Sensibilidade Autobalanceável, do inglês *Self-Balanced SENSitivity SEgmenter* (SuBSENSE), de (ST-CHARLES; BILODEAU; BERGEVIN, 2014), com Tensor de Fluxo, do inglês *Flux Tensor*, criada por Wang et al. (2014), para gerar I_b como mostra Figura 2.

A técnica SuBSENSE, que por sua vez é uma técnica de natureza especialista usada para detecção de mudanças em vídeos, nesse contexto serve para gerar máscaras de *foreground* auxiliares que serão utilizadas para geração de uma biblioteca de *pixels*

Figura 2 – Algoritmo de geração do modelo de *background* utilizado na rede DeepBS.

Fonte: Traduzida de Babaei, Dinh e Rigoll (2018, p. 639).

de *background*. A técnica Tensor de Fluxo, por ser capaz de segmentar os *pixels* com base na informação de deslocamento temporal, é utilizada para corrigir eventuais erros de armazenamento existentes nessa biblioteca.

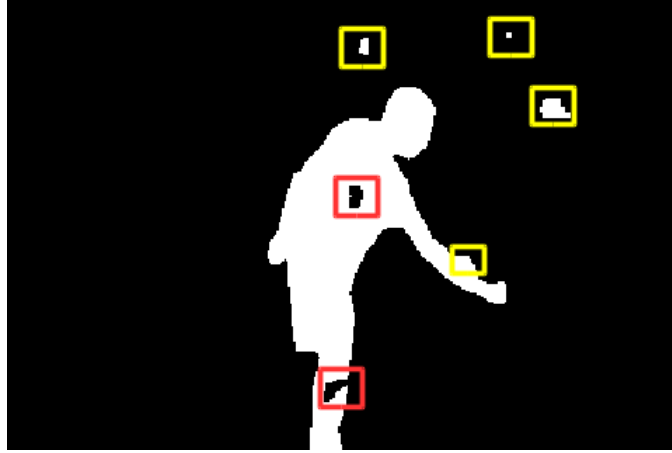
2.1.4 Destaque de Conteúdo de Primeiro Plano

Define-se como conteúdo de primeiro plano de uma imagem, do inglês *foreground*, toda a porção da imagem $I_{t_n} \in V$ que difere do conteúdo de *background* por uma margem relativamente grande, denominada limiar de diferença L . Em outras palavras, o conteúdo de *foreground* acaba sobrepondo-se ao conteúdo de *background* da imagem. Uma das formas mais usuais de se destacar as regiões de *foreground* é por meio da Equação (1),

$$I_m(x, y) = \begin{cases} 255, & \text{se } \| I_{t_n}(x, y) - I_b(x, y) \| \geq L \\ 0, & \text{caso contrário,} \end{cases} \quad (1)$$

onde I_m representa a imagem de máscara de *foreground*, onde os rótulos brancos indicam as posições (x, y) dos *pixels* pertencentes à I_{t_n} que de fato mudaram e os rótulos em preto correspondem às regiões que não mudaram, como mostra a Figura 3.

A Figura 3 também ilustra os principais problemas que as máscara de *foreground* podem apresentar, sendo que os retângulos em amarelo evidenciam detecções falso positivas (FPs), enquanto que os retângulos em vermelho indicam as detecções falso negativas (FNs). As regiões FPs são relativas aos *pixels* de *background* incorretamente detectados como

Figura 3 – Exemplo de máscara de *foreground* corrompida.

Fonte: Imagem rotulada retirada de Wang et al. (2014) e adaptada.

pertencentes à região de *foreground* da imagem I_m e as regiões FNs relacionam-se com os *pixels* de *foreground* incorretamente classificados como sendo de *background*.

As formas mais comuns utilizadas para se medir a qualidade de uma máscara de *foreground*, em termos de suas quantidades de FNs e FPs, são dadas pelas Equações (2) à (5),

$$\text{Precisão} = \frac{TP}{TP + FP}, \quad (2)$$

$$\text{Revocação} = \frac{TP}{TP + FN}, \quad (3)$$

$$\text{Medida-F} = 2 \times \frac{\text{Revocação} \times \text{Precisão}}{\text{Revocação} + \text{Precisão}}, \quad (4)$$

e

$$\text{PWC} = 100 \times \frac{FN + FP}{TP + FP + FN + TN}, \quad (5)$$

onde TP indica a quantidade de verdadeiros positivos (*pixels* de *foreground* corretamente classificados), TN a quantidade de verdadeiros negativos (*pixels* de *background* corretamente

classificados) e PWC, do inglês *Percentage of Wrong Classification*, indica a porcentagem de erros de classificação.

2.2 Redes Neurais Convolucionais

A construção das primeiras Redes Neurais Convolucionais foi, em grande parte, baseada nos trabalhos de Hubel e Wiesel (1962) e de Hubel e Wiesel (1968), que na época tentavam explicar melhor o comportamento do sistema visual dos vertebrados. Nos estudos de Hubel e Wiesel (1962), envolvendo análise do comportamento do sistema visual de gatos, em se tratando mais especificamente do campo receptivo³.

Posteriormente Hubel e Wiesel (1968) estendem seus estudos na direção de compreender melhor o comportamento do sistema visual cognitivo dos primatas, fazendo duas grandes descobertas. A primeira diz respeito à encontrar um terceiro tipo de campo receptivo, denominado campo receptivo hipercomplexo, formado por células corticais hipercomplexas. Estas células demonstraram maior capacidade de assimilação de características ainda mais complexas que suas células predecessoras simples e complexas. A segunda descoberta foi a organização hierárquica das células que compõem o sistema visual dos primatas na forma de camadas. Sendo que, do sentido de mais baixa ordem (regiões dos gânglios de retina) para mais alta ordem (região do córtex visual), encontram-se as células ganglionares de retina, células simples, complexas, hipercomplexas inferiores e hipercomplexas superiores.

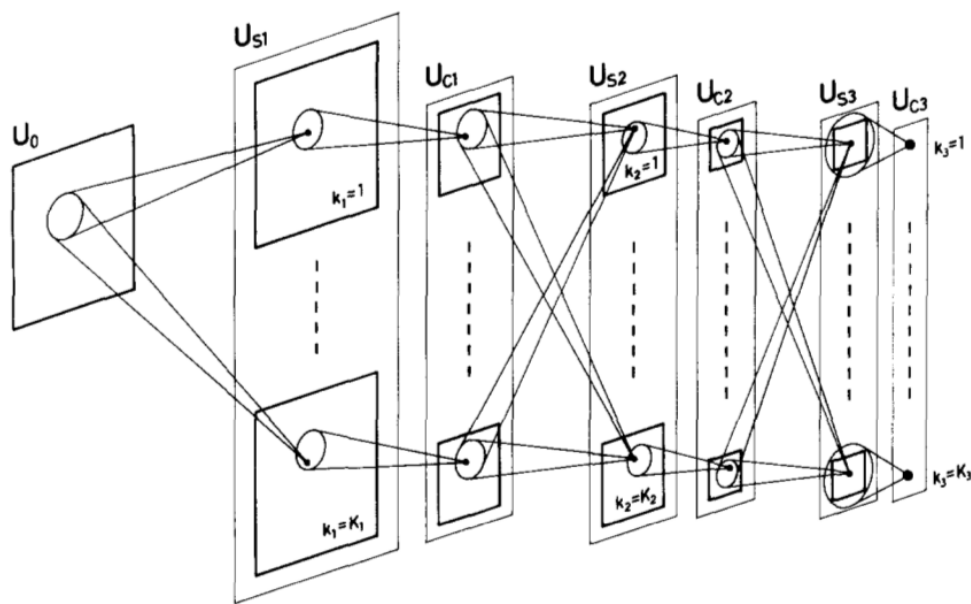
2.2.1 Modelagem Computacional

Na tentativa de melhor compreender o modelo biológico proposto por Hubel e Wiesel (1968), Fukushima (1980) propõe a construção do *neocognitron*, um dos primeiros modelos de rede neural capazes de traduzir um estímulo visual, capturado por um dispositivo fotorreceptor, em padrões de ativações artificiais. Como mostra a Figura 4 o *neocognitron* possui sete camadas, sendo as cinco primeiras inspiradas exatamente no modelo biológico proposto por Hubel e Wiesel (1968) onde, da mais baixa ordem hierárquica U_0 até a mais alta ordem U_{C2} , considerando também as camadas U_{S3} e U_{C3} , representam-

³ Campo receptivo, como definido por (HUBEL; WIESEL, 1962), é a região da retina (também chamada de campo visual) composta por células que ao serem estimuladas podem ativar.

se respectivamente as células hipercomplexas superiores especialistas e células avós⁴. No interior de cada camada U_S e U_C estão representadas, na forma de retângulos, agrupamentos de células simples e complexas (hipercomplexas a partir da camada U_{S2}). Cada plano bidimensional k recebe conexões aferentes da região elíptica demarcada na camada predecessora. Esse mecanismo faz com que o campo receptivo da rede tenda a aumentar progressivamente no sentido de U_0 à U_{C3} .

Figura 4 – Diagrama de interconexões de camadas no modelo *neocognitron*.



Fonte: Retirada de Fukushima (1980, p. 195).

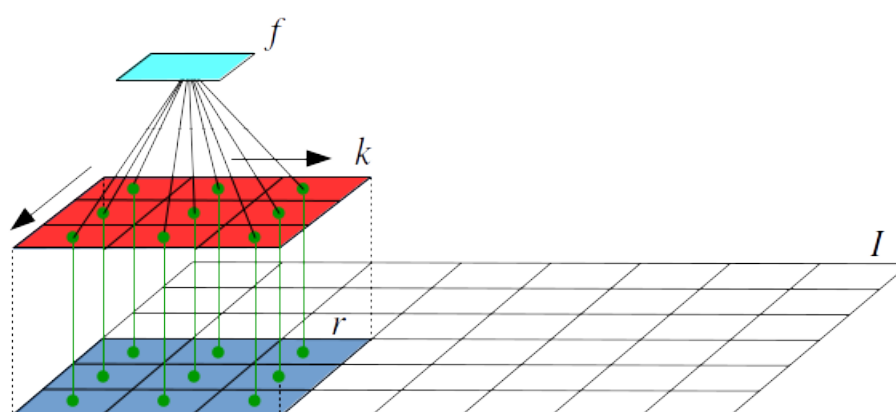
Apesar de Fukushima (1980) ter contribuído muito com a construção das redes neurais artificiais profundas, elas só ganharam notoriedade a partir dos trabalhos de LeCun (1989) e de LeCun et al. (1990). Dentre várias metodologias estudadas por LeCun (1989), a mais eficiente encontrada acabou sendo aquela que combina processamento espacial convolucional, compartilhamento de pesos e treinamento supervisionado utilizando otimização por retropropagação de erro, do inglês *backpropagation*. No trabalho de LeCun et al. (1990) fica demonstrada a grande efetividade de uma CNN de cinco camadas treinada para reconhecer caracteres numéricos em imagens de envelopes postais.

⁴ De acordo com Gross, Rocha-Miranda e Bender (1972), as células avós estariam localizadas na região do lóbulo temporal inferior do cérebro.

2.2.2 Processo de Convolução

As redes CNNs, nos moldes propostos por LeCun et al. (1990), tentam simular o comportamento das células do córtex visual por meio de estruturas denominadas filtros de convolução. Essas estruturas se apresentam na forma de matrizes bidimensionais e são responsáveis pela tarefa de extração local de características. A Figura 5 ilustra o processo de aplicação de um filtro k , de dimensões 3×3 , sobre um determinado plano I , sendo que esta aplicação segue uma orientação espacial, neste caso indicada pelas setas na cor preta. Na Figura 5 a região em azul r representa o campo receptivo do filtro, as hastes na cor verde representam operações de multiplicação ponto a ponto entre os elementos de k e r e a região f , na cor ciano, representa o resultado da soma destes produtos.

Figura 5 – Representação gráfica da aplicação de uma convolução.



Fonte: Elaborada pelo autor.

É importante observar que, na primeira camada da rede CNN, o plano I representará a imagem digital. No entanto, nas camadas subsequentes, I representará os mapas de características.

2.2.3 Ativação de Camadas

Para tentar simular da melhor forma possível o processo biológico de aprendizado a rede neural convolucional artificial utiliza também funções de ativação não lineares. Elas servem de artifício matemático para que a rede consiga aprender padrões mais complexos na medida em que mais camadas vão sendo adicionadas em sua estrutura.

Dentre as funções mais utilizadas estão a sigmoide, vide (HAN; MORAGA, 1995), tangente hiperbólica, vide (SHERVATOV et al., 2013) e retificação linear, do inglês *Rectified Linear Unit* (ReLU), vide (NAIR; HINTON, 2010), respectivamente representadas pelas Equações de (6) à (8),

$$\text{sig}(x) = \frac{1}{1 + e^{-x}}, \quad (6)$$

$$\text{tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad (7)$$

e

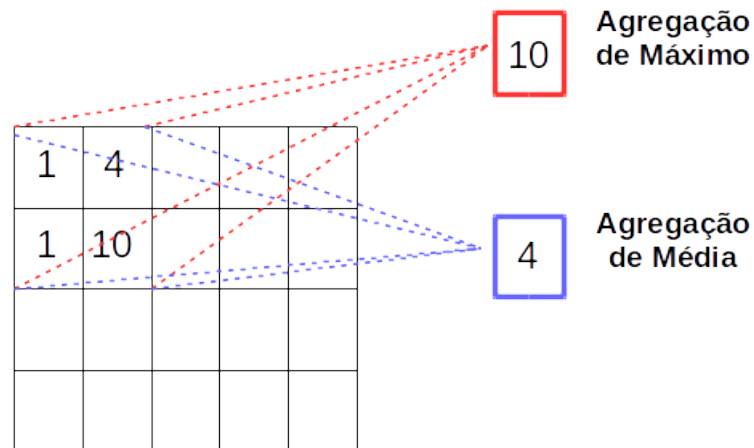
$$\text{ReLU}(x) = \max(0, x), \quad (8)$$

sendo que a função $\max(0, \cdot)$ retorna apenas valores positivos ou iguais à 0.

2.2.4 Mecanismos de Subamostragem de Características

Uma forma muito utilizada para reduzir o tamanho das dimensões dos mapas de características da rede CNN é a introdução de camadas de subamostragem, do inglês *subsampling*, entre as camadas de convolução da CNN.

Figura 6 – Representação gráfica das operações de agregação.



Fonte: Elaborada pelo autor.

Camadas de subamostragem realizam um tipo especial de operação⁵ não convolutiva de agregação de características locais, do inglês *pooling*, é aplicada sobre os mapas de características da rede. As técnicas de subamostragem mais utilizadas são a agregação por máximos locais, do inglês *maximum pooling* (*max pool*) e a agregação por média local, do inglês *average pooling* (*avg pool*). A Figura 6 ilustra o funcionamento de ambas operações, sendo mostrada em vermelho a operação *max pool* e em azul a *avg pool*.

Dentre as vantagens de se utilizar camadas de subamostragem de características estão a redução do esforço computacional nos processos de convolução, devido à redução das dimensões dos mapas de características, e o aumento do campo receptivo da rede convolucional, sendo esse aumento fundamental para que a rede consiga melhorar a capacidade de extração de padrões mediante contextualização.

2.2.5 Processo de Convolução Espacial Dilatada

Como observado por Chen et al. (2017), a introdução de muitas camadas de *pooling* nas redes de convolução acaba por reduzir a resolução espacial dos mapas de características e impactar negativamente nos processos de classificações densas⁶. Em virtude desse problema, Chen et al. (2017) propõem que sejam substituídas as últimas camadas de *pooling* das redes de classificações densas por filtros dilatados, provenientes de Holschneider et al. (1990).

Tomando-se como exemplo um caso unidimensional, para um dado elemento i de um sinal de entrada x de comprimento n , o processo de convolução de $x[i]$ por um filtro k é dado pela Equação (9) onde:

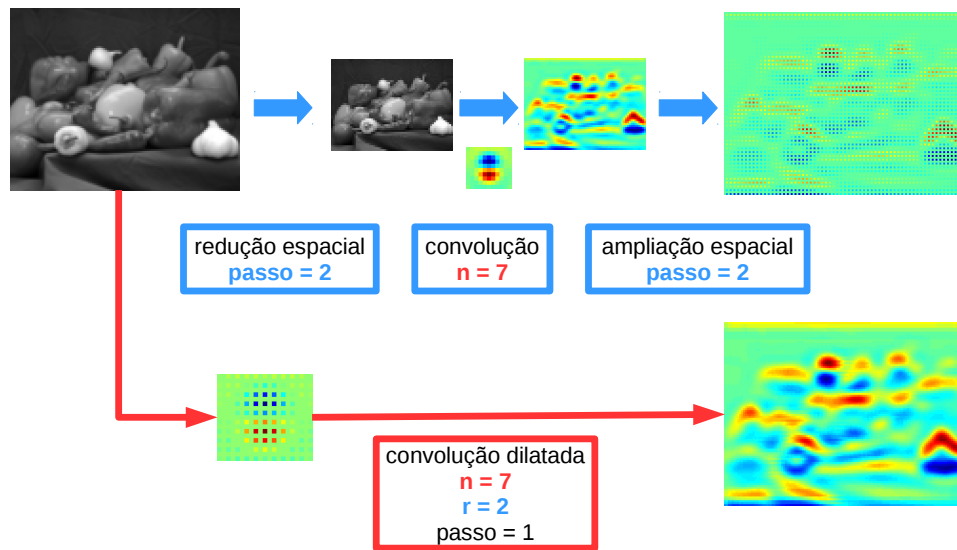
$$y[i] = \sum_{c=1}^n x[i + r \cdot c]k[c], \quad (9)$$

sendo que r corresponde à taxa de dilatação utilizada. Para $r = 1$, tem-se um processo de convolução unidimensional padrão.

Como mostra a Figura 7 a aplicação da convolução dilatada é feita na imagem em sua dimensão original, eliminando a necessidade de ter que redimensioná-la ou então de ter que redimensionar seu correspondente mapa de características. Além disso, a Figura 7

⁵ As operações de *pooling* são em grande parte inspiradas no comportamento das células complexas do córtex visual, vide Subseção 2.2.1.

⁶ Classificações ponto-à-ponto realizadas por intermédio da extração de características espacialmente afastadas porém correlacionadas.

Figura 7 – Comparação entre as aplicação espacial de convolução padrão e dilatada.

Fonte: Traduzida de Chen et al. (2017, p. 837).

Linha superior: extração de características utilizando convolução padrão. Linha inferior: extração de características utilizando convolução dilatada com taxa $r = 2$.

mostra que o resultado produzido pela aplicação da convolução dilatada sobre a imagem não redimensionada não teve falhas de preenchimento espacial.

2.2.6 Aprendizado Profundo Residual

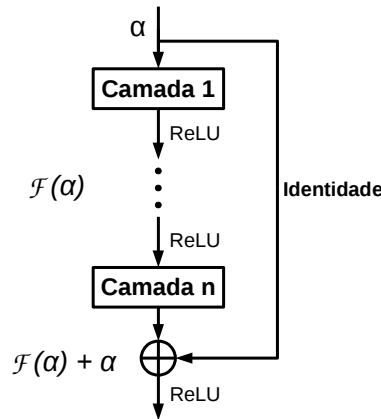
Juntamente com o aumento da compreensão do funcionamento das redes CNNs, como discutido nas Subseções 2.2.1 à 2.2.5, além do aumento da capacidade de processamento computacional, principalmente após o desenvolvimento das arquiteturas de processamento gráfico de propósito geral, do inglês *General Purpose Graphical Processing Units* (GPGPUs), surge uma nova vertente de estudos que explora o uso do aprendizado profundo em redes CNNs. Neste sentido trabalhos notáveis na área de visão computacional, como as redes AlexNet e VGGs de Krizhevsky, Sutskever e Hinton (2012) e Liu e Deng (2015), respectivamente, conseguem demonstrar o grande potencial do aprendizado profundo de máquinas em tarefas relacionadas à classificação de imagens, ao menos nesse primeiro momento.

No entanto, de maneira contraintuitiva, He et al. (2016) observam que nem sempre aumentar a profundidade (número de camadas) das CNNs gera um impacto positivo. Em

seus estudos He et al. (2016) notam que aumentar muito a quantidade de camadas de convolução pode gerar um problema de saturação de aprendizado. Desta maneira, após o treinamento atingir um estágio saturado, a acurácia da rede tenderia a degradar-se rapidamente.

Sendo tal fenômeno observado nos resultados de acurácia de treinamento⁷, He et al. (2016) notam que o problema não teria relação com *overfitting*, mas com a dificuldade da rede CNN em continuar assimilando novos padrões àqueles já aprendidos em etapas anteriores. Para tentar solucionar esse problema, He et al. (2016) sugerem o paradigma de aprendizado residual profundo.

Figura 8 – Representação Gráfica do Aprendizado Residual.



Fonte: Adaptada de He et al. (2016, p. 771).

Como mostra a Figura 8, para He et al. (2016), durante o aprendizado de uma rede CNN profunda, assim como as camadas convolucionais poderiam hipoteticamente aproximar-se (de maneira assintótica) de funções não lineares complexas (representadas por $\mathcal{H}(\alpha)$), elas também poderiam aproximar-se de funções residuais $\mathcal{F}(\alpha) = \mathcal{H}(\alpha) - \alpha$. Utilizando essa simples reformulação, He et al. (2016) demonstram que o problema de degradação do aprendizado das CNNs profundas pode ser corrigido por meio de mapeamentos de identidade. Dessa forma, sempre que o aprendizado das camadas pertencentes à um bloco residual estagnar haverá a possibilidade do fluxo de informação ser propagado pelas suas conexões identidades, como mostra a Figura 8. Isso faria com que o processo de correção de erros (desencadeado por mecanismos de *backpropagation*)

⁷ Acurácia medida sobre as mesmas amostras utilizadas no treinamento da rede CNN.

evitasse buscar por soluções subótimas para as camadas de mais baixo nível hierárquico ao tentar encontrar soluções ótimas para as camadas de mais alto nível hierárquico.

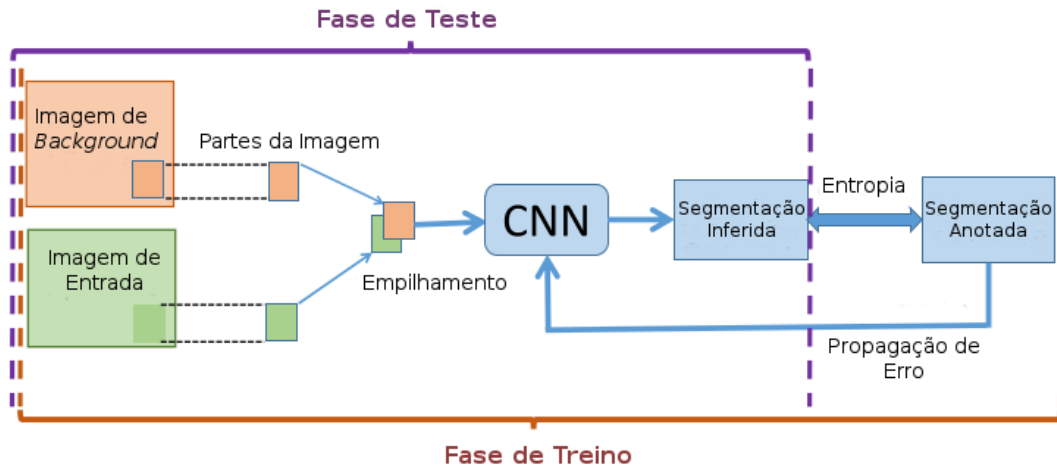
2.3 Trabalhos Correlatos

Ao longo dessa seção apresentam-se as principais técnicas baseadas em CNNs profundas de detecção de mudanças em vídeos e imagens utilizadas como referências para o desenvolvimento dessa tese. Nas Subseções 2.3.1 e 2.3.2 são abordadas as técnicas completamente convolucionais simples e de multiestágio. As Subseções 2.3.3 e 2.3.4 tratam das técnicas de estrutura codificadora-decodificadora multiescala. Posteriormente, na Subseção 2.3.5, são abordadas as redes siamesas de segmentação em formato de U. As Subseções 2.3.6 e 2.3.7 abordam redes convolucionais residuais. A seção 2.3.8 aborda técnica siamesa que utiliza aprendizagem de métrica. As seções que vão de 2.3.9 à 2.3.11 apresentam técnicas siamesas que utilizam mecanismos de atenção espacial e temporal para auxiliar na tarefa de detecção de mudanças em imagens.

2.3.1 Rede Neural de Segmentação Profunda de *Background*

A técnica DeepBS proposta por Babaee, Dinh e Rigoll (2018) é um método para detecção de mudanças em vídeos que utiliza como base dois elementos, uma única CNN e um modelo de *background*, sendo que o *background* é gerado a partir de uma sequência de imagens consecutivas coletadas de determinado trecho do vídeo analisado, como explicado na Subseção 2.1.3.

O treinamento da técnica DeepBS utiliza metodologia de extração de fragmentos, do inglês *patches*, sendo esses fragmentos pertencentes a ambas as imagens alvo de detecção e de *background* concatenados na dimensão de canais antes de serem processados pela técnica, como mostra Figura 9. Nesta fase a rede CNN aprende a reproduzir, em sua camada de saída, imagens contendo valores reais de probabilidade, valores pertencentes ao intervalo $[0, 1]$. A etapa de inferência das regiões de mudanças segue o mesmo fluxo de processamento mostrado pela Figura 9, sendo acrescido o passo de limiarização do mapa de probabilidades produzido pela técnica e removido o passo de propagação de erro.

Figura 9 – Arquitetura da rede DeepBS.

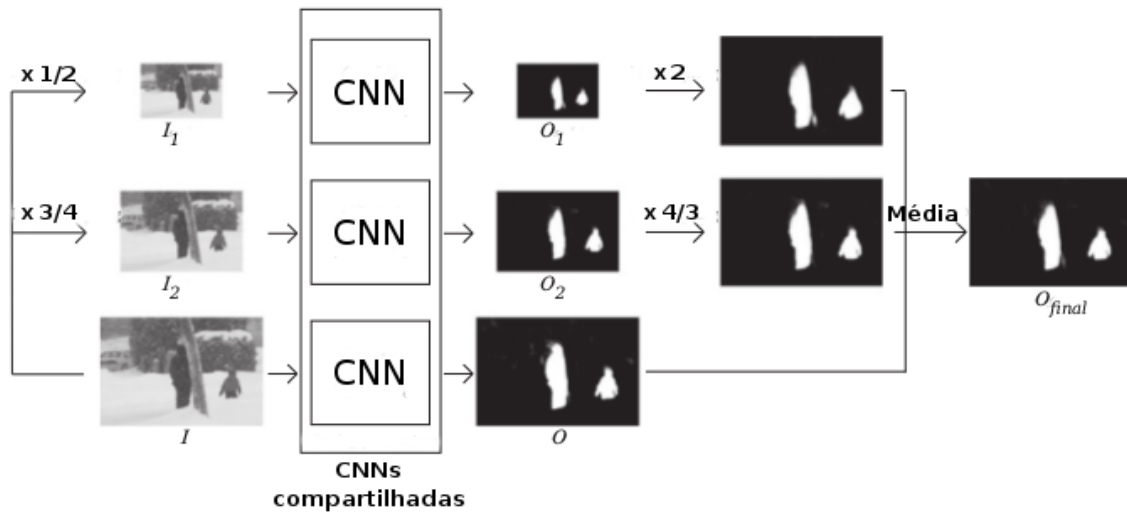
Fonte: Traduzida de Babaei, Dinh e Rigoll (2018, p. 639).

2.3.2 Rede Neural Convolutacional Multiestágio

No trabalho proposto por Wang, Luo e Jodoin (2017), é desenvolvido um sistema de rotulação automática para regiões de mudanças em vídeos, cujo núcleo central é uma CNN de detecção de mudanças, denominada Cascade, que engloba estratégias de processamento multiescala e multiestágio.

Processamento Multiescala

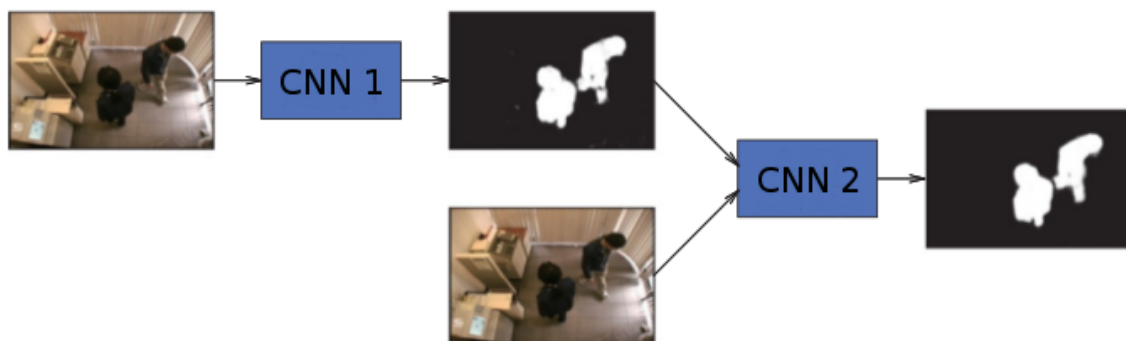
Observando a dificuldade das redes CNNs em lidar com a segmentação de regiões de mudança que excedem os limites espaciais de seus campos receptivos, vide Subseções 2.2 e 2.2.1, devido a falta de informação contextual, Wang, Luo e Jodoin (2017) propõe que as imagens alvo da detecção de mudanças sejam reduzidas de tamanho, dimensões espaciais, antes de passarem pelo método CNN. Nesse processo, dada uma imagem de entrada I , geram-se duas novas imagens I_1 e I_2 , cujas dimensões correspondem respectivamente à 50% e 75% do tamanho original de I . Como ilustra a Figura 10, após serem processadas pela mesma CNN e terem sido gerados seus respectivos mapas de probabilidades O_1 , O_2 e O , redimensionam-se espacialmente os mapas O_1 e O_2 , utilizando como referência a dimensão de O e gera-se um único mapa de saída O_{final} resultante da média ponderada calculada entre O_1 , O_2 e O .

Figura 10 – Diagrama da rede CNN simples estendida para sua versão multiescala.

Fonte: Adaptada de Wang, Luo e Jodoin (2017, p. 69).

Processamento Multiestágio

Tendo em vista que ambas as CNNs, simples e até mesmo multiescala, processam cada *pixel* de maneira independente, a qualidade da segmentação fica condicionada à extração de informação local, referente a fragmento de imagem processado.

Figura 11 – Diagrama da rede em cascata de dois estágios.

Fonte: Retirada de Wang, Luo e Jodoin (2017, p. 70).

Essa característica de aprendizado acaba gerando falsos positivos e até mesmo falsos negativos isolados na máscara de *foreground* gerada pela técnica. Para lidar com esse problema de incoerência espacial Wang, Luo e Jodoin (2017) propõem uma versão

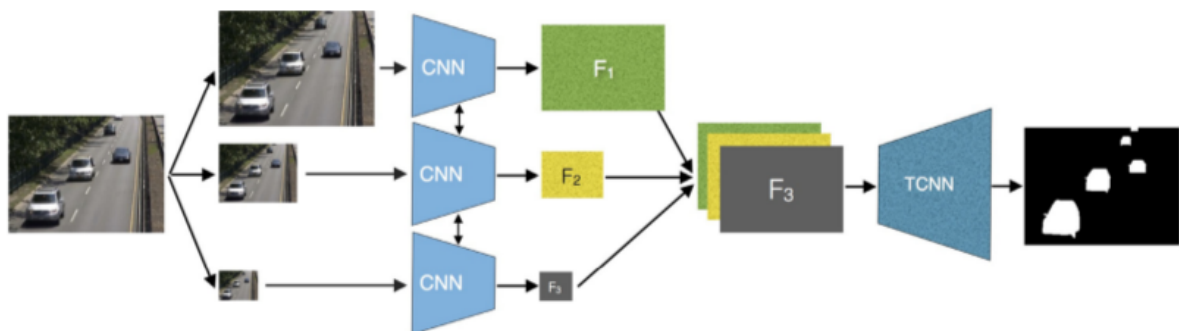
multiestágio do método abordado na Subseção 2.3.2. Como mostra a Figura 11, a abordagem multiestágio utiliza duas redes convolucionais **CNN 1** e **CNN 2** treinadas separadamente em dois estágios, sendo que no primeiro estágio a **CNN 1** é utilizada para gerar uma primeira versão do mapa de probabilidades do detector, que por sua vez, após ser concatenado à imagem alvo da detecção, deverá ser refinado pela **CNN 2**.

2.3.3 Redes Neurais Codificadoras-Decodificadoras Multiescala

Nesse trabalho Lim e Keles (2018) propõem alternativas mais eficiente para lidar com o problema de incoerência espacial, exibido nas técnicas DeepBS e Cascade, que envolve principalmente a utilização de redes codificadoras-decodificadoras para processamento multiescala e de estágio único. As alternativas de Lim e Keles (2018) para o problema de detecção de mudanças em vídeos são as técnicas denominadas FgSegNet_M e FgSegNet_S, cujas arquiteturas são apresentadas em maiores detalhes nas Subseções 2.3.3 e 2.3.3, respectivamente.

Rede de Múltiplas Entradas

Figura 12 – Arquitetura da rede FgSegNet_M.



Fonte: Retirada de Lim e Keles (2018, p. 258).

Como mostra a Figura 12, a arquitetura da rede FgSegNet_M é composta por um módulo CNN codificador⁸ compartilhado, que recebe como entrada três imagens no domínio de cores RGB. A primeira delas é a imagem alvo da detecção de mudanças em seu tamanho original, a segunda e a terceira são versões redimensionadas desta imagem, que

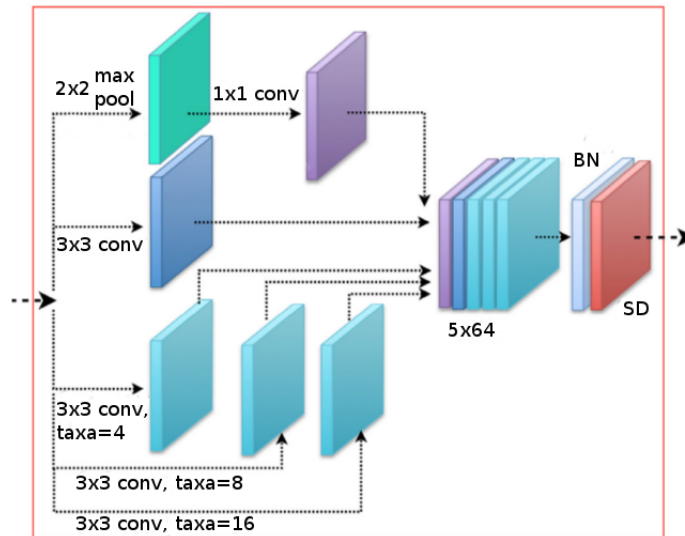
⁸ Lim e Keles (2018) utilizam a rede VGG-16 de Liu e Deng (2015) como módulo codificador.

correspondem respectivamente à imagens com 1/2 e 1/4 do tamanho da imagem original. Após serem gerados os mapas de projeção de características F_1 , F_2 e F_3 , para cada uma das imagens de entrada da rede CNN, F_2 e F_3 são redimensionados e concatenados com F_1 , formando o conjunto F de mapas. F é então submetido ao módulo⁹ de Rede Neural Convolutacional Transposto, do inglês *Transposed Convolutional Neural Network* (TCNN) de decodificação, que por sua vez, é responsável por gerar o mapa de probabilidades. Assim como nas redes anteriormente abordadas nas Seções 2.3.1 e 2.3.2, após limiarização do mapa de probabilidades gera-se a máscara de *foreground*.

Rede de Única Entrada

A rede FgSegNet_S foi construída com o propósito de ser uma alternativa simplificada da rede FgSegNet_M. Sendo assim, a rede FgSegNet_S necessita apenas da imagem em escala original, sendo que as características multiescala são extraídas pelo Módulo de Agrupamento de Características, do inglês *Feature Pooling Module* (FPM).

Figura 13 – Diagrama do Módulo FPM da Rede FgSegNet_S.



Fonte: Adaptada de Lim e Keles (2018, p. 258).

A principal finalidade do módulo FPM, ilustrado pela Figura 13, é aumentar a capacidade da rede CNN codificadora de extrair padrões locais mais complexos por

⁹ Rede VGG-16 pré-treinada acoplada ao modelo em ordem reversa.

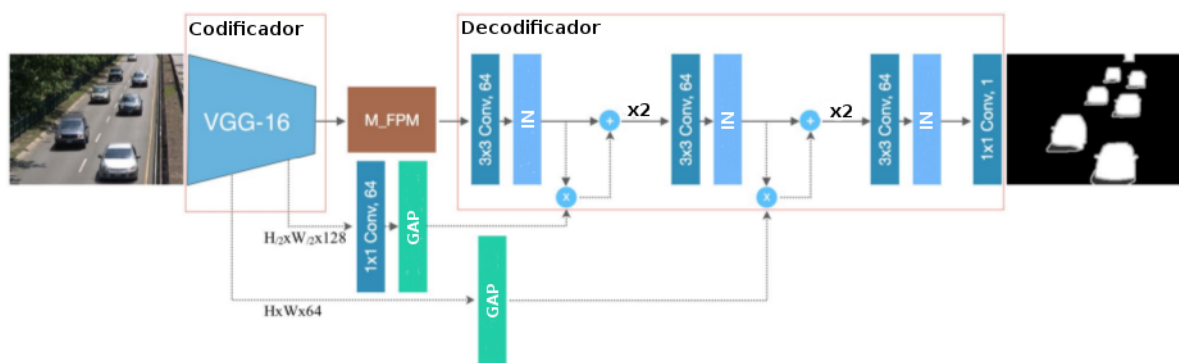
meio da simulação do aumento da resolução¹⁰ dos filtros de convolução, como explica a Subseção 2.2.5. Dessa forma a utilização de filtros convolucionais dilatados nas taxas 4, 8 e 16, como mostra a Figura 13, simula a aplicação de convoluções de tamanho 7×7 , 11×11 e 19×19 , respectivamente, sobre os mapas de características resultantes da rede codificadora. Nota-se que tal estratégia acaba sendo menos onerosa em comparação com as redes DeepBS, Cascade e FgSegNet_M.

Terminado o processo de dilatação de características, o módulo FPM entra em seu estágio final de execução, onde os mapas de características deverão ser concatenados e normalizados em lote, do inglês *Batch Normalization* (BN), segundo a aplicação da técnica desenvolvida por Ioffe e Szegedy (2015), e espacialmente regularizados por meio de Anulação Espacial de Ativação, do inglês *Spatial Dropout* (SD), segundo a técnica desenvolvida por Tompson et al. (2015).

2.3.4 Rede Neural Autocodificadora Multiescala Versão 2

A rede convolucional FgSegNet_v2, criada por Lim e Keles (2019), surge a partir de melhorias feitas na rede FgSegNet_S, vide Subseção 2.3.3. Como ilustrado pela Figura 14, a arquitetura da rede FgSegNet_v2 possui duas principais diferenças em relação à sua rede predecessora.

Figura 14 – Arquitetura da rede FgSegNet_v2.

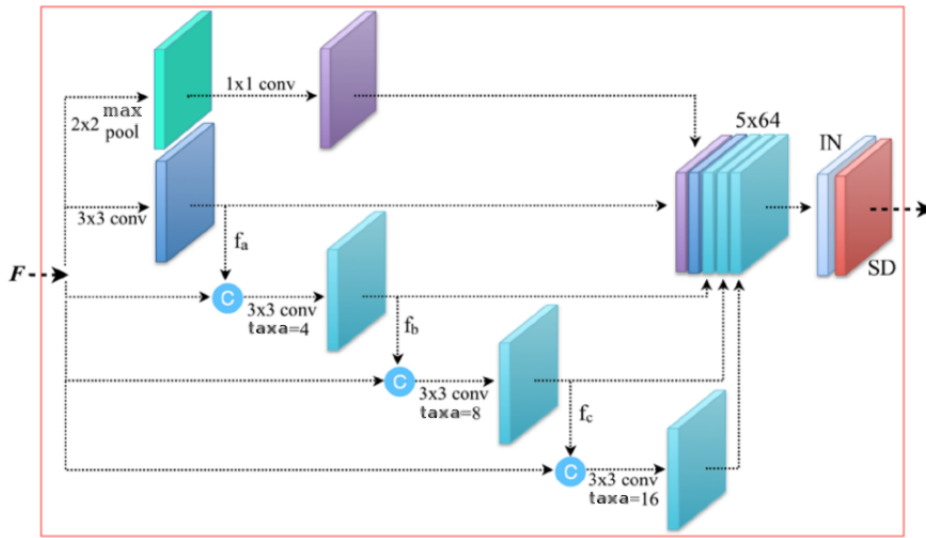


Fonte: Adaptada de Lim e Keles (2019, p. 3).

¹⁰ Sob os campos receptivos são aplicadas taxas de dilatação que simulam o efeito de aumentar a dimensão dos mapas de características.

A primeira diferença diz respeito à reestruturação do mecanismo de codificação e decodificação da rede FgSegNet_S, onde o módulo de decodificação TCNN é substituído por apenas três camadas de convolução e o mecanismo de autocodificação da rede passa a ser mediado por dois módulos de pré-processamento denominados Aglomerador de Médias Global, do inglês *Global Average Pooling* (GAP), desenvolvido por Lin, Chen e Yan (2013) e Normalizador de Instâncias, do inglês *Instance Normalization* (IN), desenvolvido por Ulyanov, Vedaldi e Lempitsky (2016).

Figura 15 – Diagrama do Módulo M_FPM da Rede FgSegNet_v2.



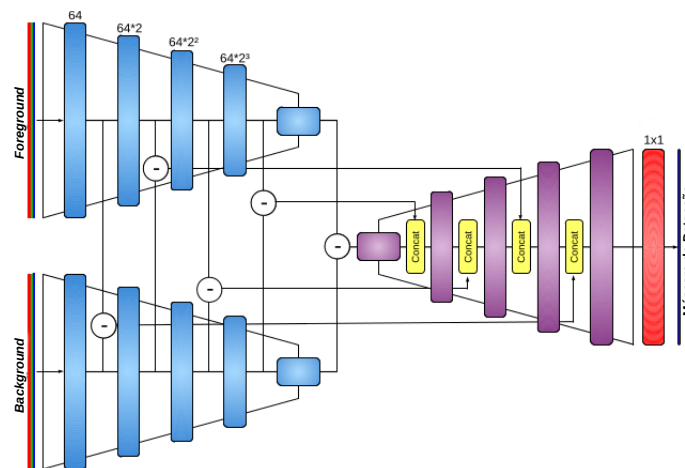
Fonte: Adaptada de Lim e Keles (2019, p. 3).

A segunda diferença observável na rede FgSegNet_v2 diz respeito à utilização do novo módulo de processamento multiescala M_FPM, como mostra a Figura 15. Nesse novo módulo os mapas F de características da metade codificadora da rede FgSegNet_v2 passam a ser processados de maneira estratificada. Dessa maneira cada nova convolução dilatada é responsável por acrescentar mais informação de contexto espacial sobre a pilha de mapas de características resultante da concatenação incremental feita entre mapas anteriormente processados.

2.3.5 Redes Neurais Semânticas U

Em (SANTANA et al., 2019) os autores desenvolvem as redes convolucionais codificadoras-decodificadoras de detecção de mudanças em vídeos denominadas redes de segmentação semântica em forma de U, do inglês *Semantic U-Nets* (SEUNets), onde o mecanismo de aprendizado relacional divide-se em duas etapas sendo elas (i) diferenciação inter-ramificações codificadas e (ii) agregação desses valores residuais produzidos às correspondentes camadas de decodificação das redes.

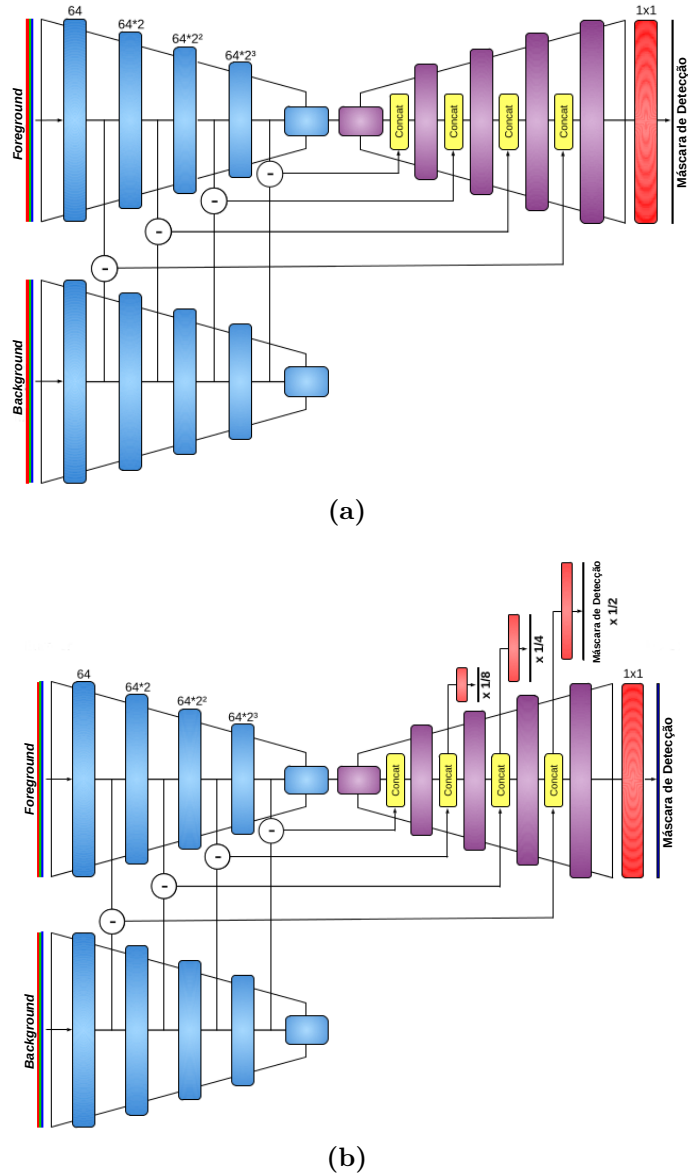
Figura 16 – Representação Gráfica da Rede SEUNet1.



Fonte: Traduzida de Santana et al. (2019, p. 47).

Como mostra a Figura 16, na arquitetura SEUNet1 os mapas de características das camadas de codificação, pertencentes a ramificações de *foreground* e *background*, são subtraídos entre si durante a execução da etapa (i). Durante a execução da etapa (ii), representada pelos retângulos amarelos de nome *Concat* na Figura 16, os mapas de características produzidos pelas subtrações são então concatenados, na dimensão de canais, aos mapas de características decodificadas de mesmo nível hierárquico. Ainda de acordo com a Figura 16, observa-se que o processo de decodificação da rede SEUNet1 começa a ser aplicado sobre o resultado da subtração calculada entre os mapas de características de mais alto nível hierárquico produzidas metade codificadora da rede.

Como mostra a Figura 17(a), na arquitetura SEUNet2 Santana et al. (2019) não realizam o cálculo relacional entre os mapas de características de mais alto nível

Figura 17 – Representação Gráfica das Redes SEUNet2 e SEUNet3.

Fonte: Elaborada pelo autor.

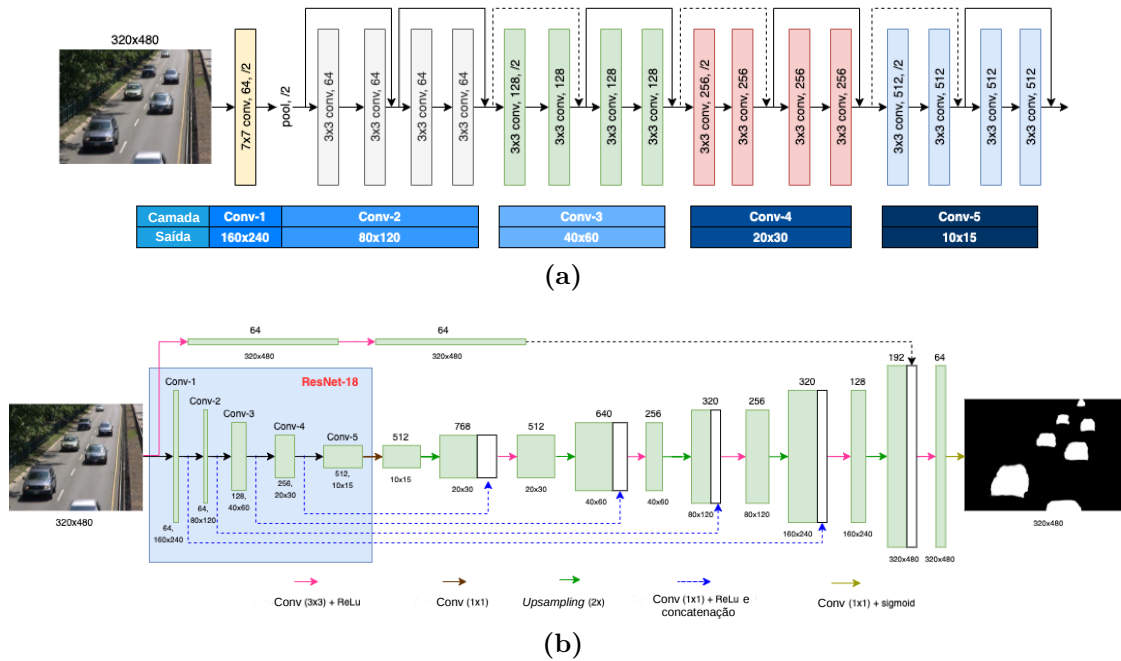
Em (a) tem-se representada a arquitetura da rede SEUNet2 e em (b) rede SEUNet3.

hierárquico codificados e dessa forma o processo de decodificação começa a ser aplicado nos mapas de características de mais alto nível hierárquico proveniente apenas da ramificação de *foreground* da metade codificadora da rede. No método SEUNet3, Como mostra a Figura 17(b), Santana et al. (2019) criam um mecanismo de reforço para o modelo relacional onde, paralelamente à otimização do resultado de saída do processo de decodificação, minimiza-se também as discrepâncias dos mapas de mudanças gerados nos processos intermediários de agregação, camadas *Concat* em amarelo na Figura 17(b).

2.3.6 Redes Neurais de Detecção de Movimento U

Em (RAHMON et al., 2021) os autores desenvolvem as redes U-Nets de detecção de mudanças por movimento, do inglês *Motion U-Nets* (MU-Nets). Constituídas por camadas codificadoras residuais ResNet18, de (HE et al., 2016), as redes MU-Nets são mais eficientes em termos de consumo de memória e custo de processamento quando comparadas com a rede U-Net convencional de Ronneberger, Fischer e Brox (2015).

Figura 18 – Representação Gráfica das Redes MU-Net1 e ResNet18.

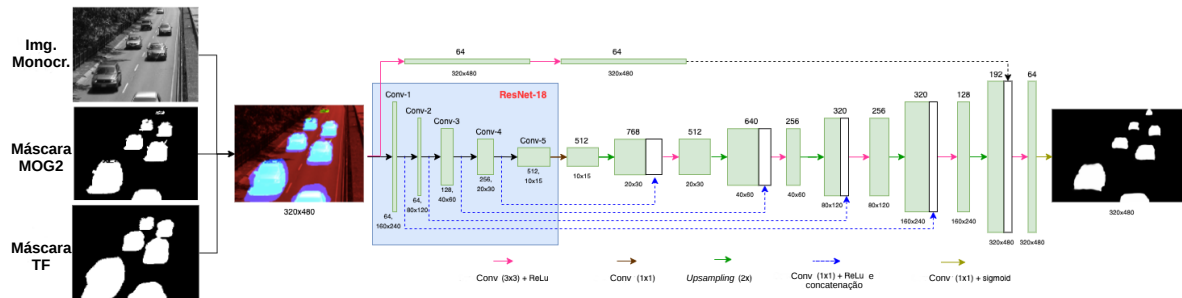


Fonte: Traduzida de Rahmon et al. (2021, p. 8127).

Em (a) tem-se a arquitetura da rede ResNet18 e em (b) da rede MU-Net1.

A Figura 18(b) apresenta a arquitetura da rede MU-Net1 que, dada uma imagem cromática de entrada, é capaz de primeiramente realizar a extração de características por meio do módulo codificador ResNet18, Figura 18(a), e após agregação dessas características com as camadas de decodificação, setas em azul na Figura 18(b), torna-se capaz de gerar a máscara de detecção de mudanças. Observando que, assim como no caso da rede U-Net de Ronneberger, Fischer e Brox (2015), o processo de agregação tem por finalidade a recuperação de informação espacial perdida durante o processo de redimensionamento dos mapas de características feito por camadas de *max pool*.

A rede MU-Net2, exibida pela Figura 19, fora desenvolvida para suprir as deficiências da rede MU-Net1 com relação à (a) impossibilidade de detectar objetos que não aparecem

Figura 19 – Representação Gráfica da Rede MU-Net2.

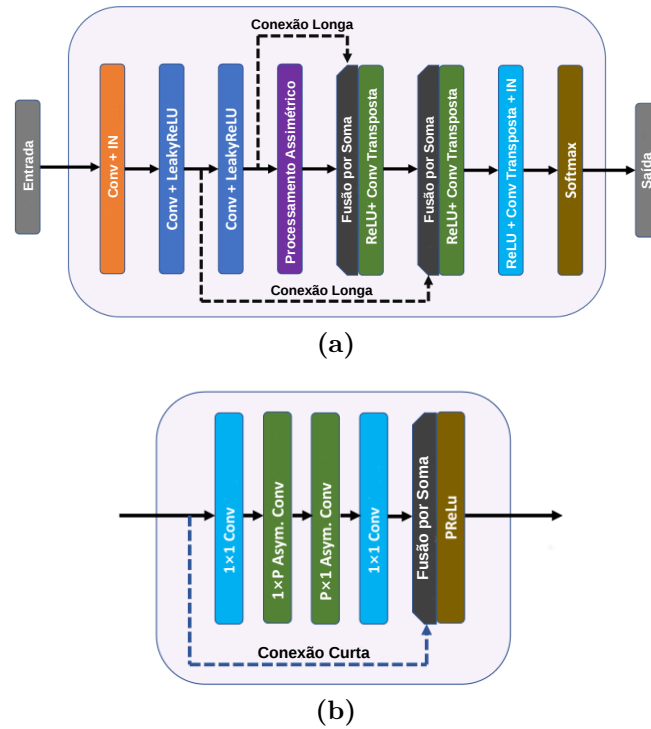
Fonte: Traduzida de Rahmon et al. (2021, p. 8128).

no treinamento, (b) falhas nos casos de detecção com pouco contexto de cena (objetos muito distantes da câmera) e (c) incapacidade de diferenciar objetos estacionários de objetos em movimento. A estratégia utilizada por Rahmon et al. (2021) para criar a MU-Net2 consiste em agregar à imagem monocromática de entrada (informação espacial) conteúdo de movimento (informação temporal), sendo que o conteúdo de movimento é gerado pelas técnicas MOG2, do inglês *Mixtures of Gaussians*, e Tensor de Fluxo desenvolvidas respectivamente por Zivkovic (2004a) e Wang et al. (2014).

2.3.7 Rede de Detecção de Mudanças de Tempo Real

Desenvolvida por Lin et al. (2021), a Rede de Detecção de Mudanças de Tempo Real, do inglês *Real-Time Segmentation Network* (RTSegNet), combina mecanismo de conexão residual curta e longa implementados entre os pontos terminais de uma estrutura denominada processamento assimétrico e as camadas de codificação e decodificação, respectivamente.

A Figura 20 mostra a configuração da arquitetura da rede RTSegNet, que por sua vez apresenta-se como uma solução muito mais eficiente em termos de custo computacional em comparação com outras técnicas de detecção de mudanças como FgSegNets, SEUNets e MU-Nets. Como mostra a Figura 20, Lin et al. (2021) constroem uma rede de detecção de mudanças arquiteturalmente mais simples, em comparação com as redes FgSegNets e MU-Nets por exemplo, para demonstrar a possibilidade de se obter bons resultados de segmentação de mudanças sem haver a necessidade de utilizar redes codificadoras pré-treinadas convencionais como redes VGGs.

Figura 20 – Representação Gráfica da Rede RTSegnet com seu módulo assimétrico.

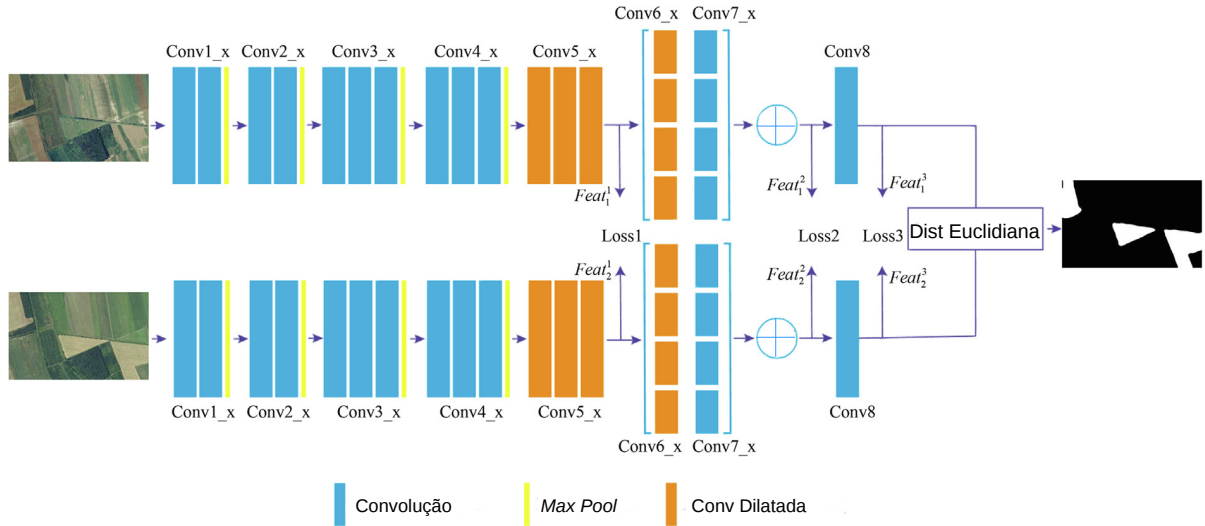
Fonte: Traduzida de Lin et al. (2021, p. 548).

Em (a) tem-se a arquitetura da rede RTSegNet e em (b) o seu módulo assimétrico.

2.3.8 Rede Convolutacional Siamesa e Aprendizado Focalizado

Wang et al. (2021) propõem a Rede de Segmentação de Custo Focalizado, do inglês *Focal Loss Segmenter* (FLS), uma nova variação da rede convolutacional siamesa de detecção de mudanças em imagens aéreas de sensoriamento remoto, muito similar ao método de Zhang et al. (2018), como mostra a Figura 21. Sua proposta tem como diferencial importante a utilização de três camadas no aprendizado semântico relacional de métricas, cujas distâncias euclidianas calculadas entre os mapas de características são minimizadas por meio de uma função de custo focalizado em mudanças, do inglês *Focal Loss* (FL). A FL tenta reforçar a capacidade de discriminação interclasses (capacidade de diferenciar *pixels* de *foreground* de *pixels* de *background*) ao mesmo tempo em que tenta aumentar a capacidade de aproximação intraclasses (capacidade de aproximar *pixels* de mesma classe entre si) da rede convolutacional principalmente nos casos de treinamento com rótulos desbalanceados¹¹.

¹¹ O desbalanceamento de rotulagem ocorre quando mais classes de mudanças são anotadas em detrimento de classes de *background* ou vice-versa.

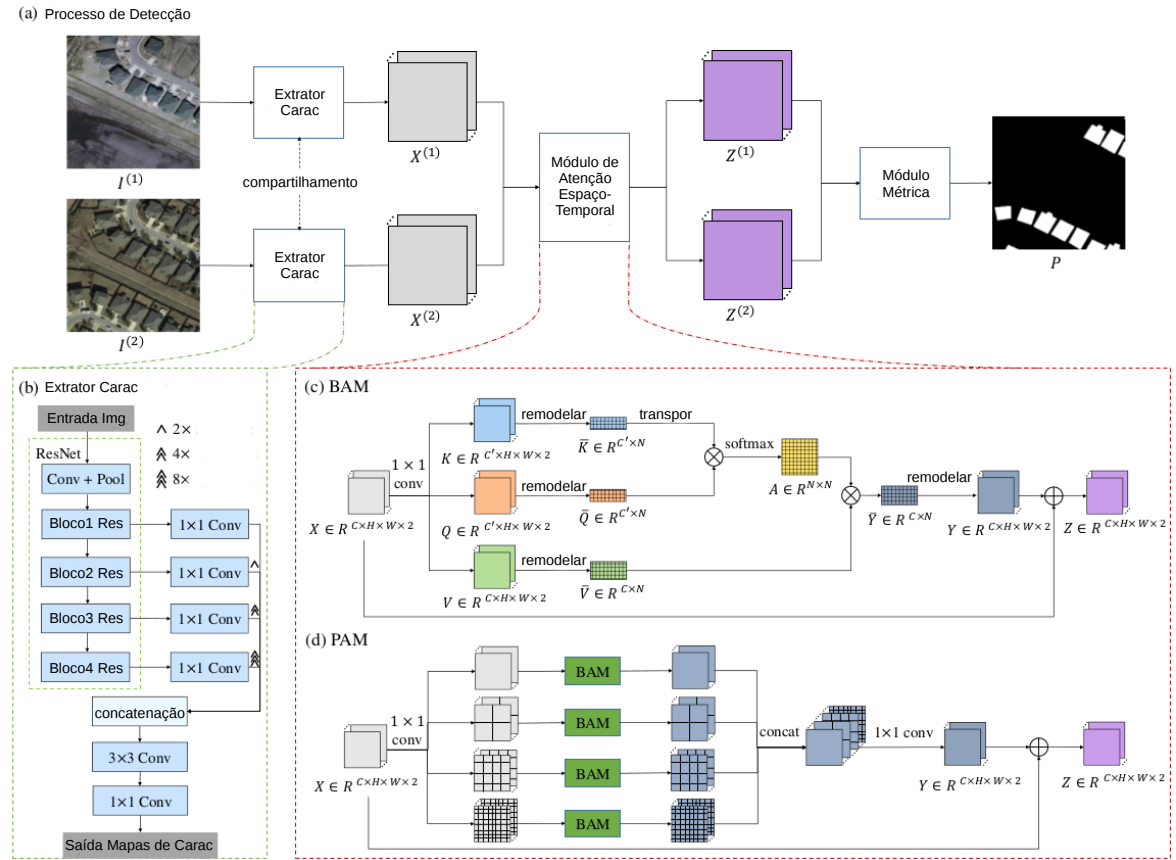
Figura 21 – Representação Gráfica da Rede FLS.

Fonte: Traduzida de Wang et al. (2021, p. 158).

2.3.9 Rede Convolutacional de Atenção Espaço-Temporal

Chen e Shi (2020b) propõem a Rede de Atenção Espaço-Temporal, do inglês *Spatial-Temporal Attention Network* (STANet), que corresponde à uma rede convolutacional siamesa utilizada para detecção de mudanças em imagens aéreas de sensoriamento remoto.

A arquitetura da parte codificadora da rede STANet assemelha-se muito com as redes residuais convolutacionais siamesas de Zhang et al. (2018) e de Wang et al. (2021), como mostra a Figura 22. Num primeiro momento a rede STANet utiliza um módulo residual ResNet18 para extrair características de alto nível hierárquico das imagens alvo da detecção de mudanças. Em um segundo momento, a rede STANet busca correlacionar as características previamente extraídas na tentativa de evidenciar regiões de possíveis mudanças ocorridas no domínio espaço-temporal. Para realizar essa tarefa Chen e Shi (2020b) propõem a utilização do Módulo de Atenção Espaço-Temporal Básica, do inglês *Basic Spatial-Temporal Attention Module* (BAM), e do Módulo de Atenção Espaço-Temporal Piramidal, do inglês *Pyramid Spatial-Temporal Attention Module* (PAM). Enquanto o módulo BAM viabiliza a comparação não-local entre os mapas de características extraídos das imagens de entrada, o módulo PAM tenta melhorar ainda mais o processo de análise

Figura 22 – Representação Gráfica da Rede STANet.

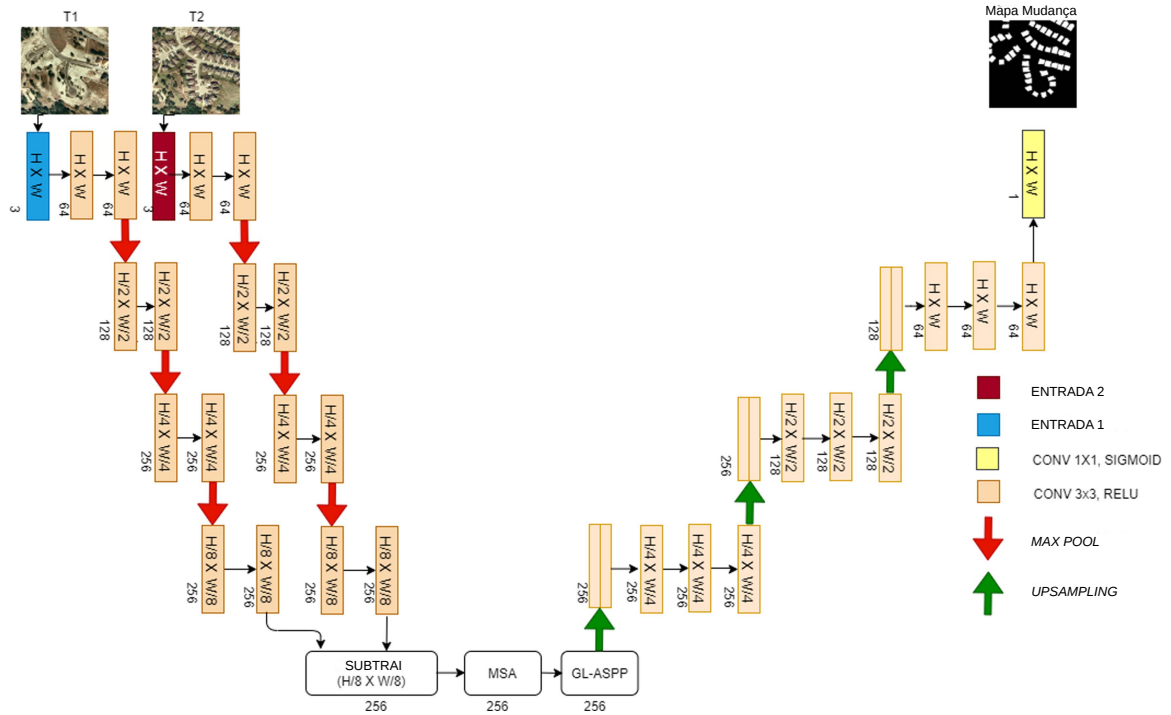
Fonte: Traduzida de Chen e Shi (2020b, online).

semântica de relacionamentos por meio da aplicação em multiescala dos módulos de atenção básica (BAMs).

2.3.10 Rede de Detecção de Mudanças para Sensoriamento Remoto

Barkur et al. (2022) propõem a Rede de Detecção de Mudanças para Sensoriamento Remoto, do inglês *Remote Sensing Change Detection Network* (RSCDNet), que consiste em uma rede codificadora-decodificadora em formato de U, como mostra a Figura 23.

Diferentemente das redes SEUNet, no método RSCDNet o cálculo das diferenças semânticas é feito apenas entre os mapas de características extraídos das camadas siamesas de mais alto nível hierárquico da rede neural. Além disso, o processo de agregação das diferenças codificadas às decodificadoras também é substituído pelos módulos de Auto-atenção Modificada, do inglês *Modified Self-Attention* (MSA), e de Aglutinação Piramidal Espacial Dilatada Linear Limitada do inglês *Gated Linear Atrous Spatial Pyramid Pooling*

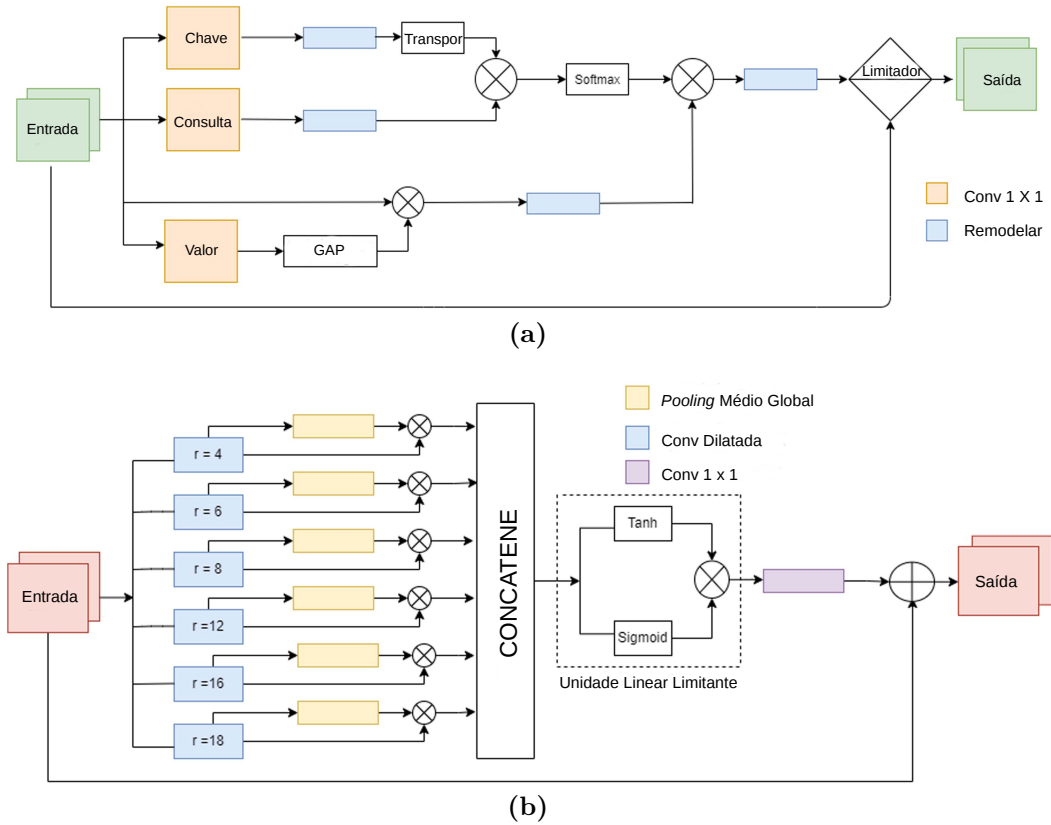
Figura 23 – Representação Gráfica da Rede RSCDNet.

Fonte: Traduzida de Barkur et al. (2022, p. 542).

(GL-ASPP).

Como mostra a Figura 24(a), o módulo MSA é uma versão modificada do mecanismo de auto-atenção não local de Vaswani et al. (2017). As modificações propostas por Barkur et al. (2022) incluem a adição de um módulo de GAP para, num primeiro momento, destacar as características que em média estão mais ativas nos mapas produzidos pela etapa Valor. Em um segundo momento, tais características evidenciadas são filtradas e ponderadas pelos mapas de correlações não locais calculadas entre os elementos das etapas de Consulta e Chave, vide Figura 24(a). Já o módulo GL-ASPP, apresentado na Figura 24(b), busca recuperar a informação perdida no processo de codificação da rede RSCDNet (perda de resolução) por meio da extração de informação contextual, sendo que para isso utilizam-se convoluções dilatadas, processamentos de GAP e agregação local por combinação linear.¹²

¹² Convolução com filtros de dimensão 1×1 .

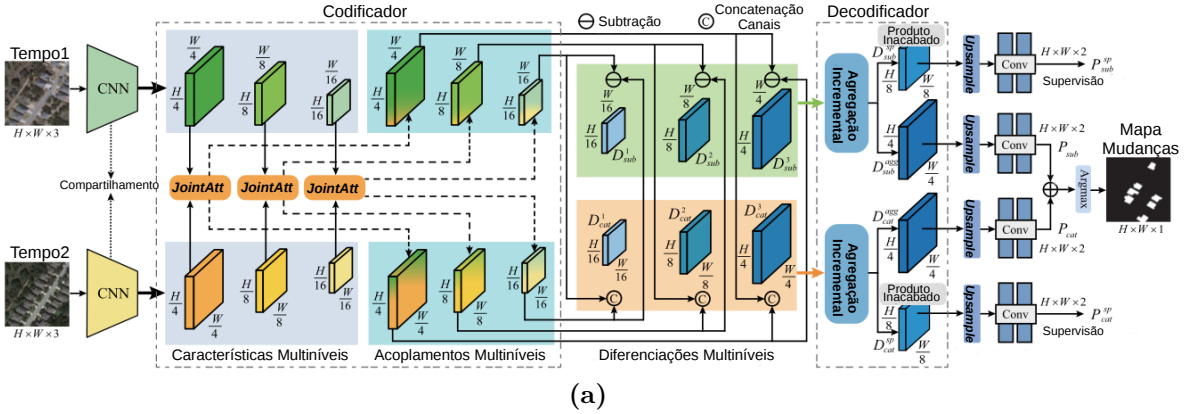
Figura 24 – Representação Gráfica dos módulos MSA e GL-ASPP.

Fonte: Traduzida de Lin et al. (2021, p. 540).

Em (a) tem-se a arquitetura do módulo MSA e em (b) do módulo GL-ASPP.

2.3.11 Rede de Detecção de Mudanças Intertemporais Multiníveis

A Rede Intertemporal Multinível de Dupla Ramificação, do inglês *Dual-Branch Multilevel Intertemporal Network* (DMINet), foi construída por Feng et al. (2023) para solucionar, de maneira eficiente, problemas de detecção de mudanças em imagens bitemporais de sensoriamento remoto. Como mostra a Figura 25(a), para uma dada entrada bitemporal a rede DMINet segue um fluxo de processamento bem definido que combina (i) extração de mapas de características de alto nível utilizando rede residual pré-treinada ResNet18, (ii) modelagem de correlação intertemporal, estágio de nome Acoplamentos Multiníveis na Figura 25(a), entre os mapas bitemporais por meio de módulo nomeado de Atenção Conjunta, Figura 25(b), do inglês *Joint Attention* (JointAtt), (iii) subtração espacial local entre os mapas de características bitemporais, representado pela etapa de nome Diferenciações Multiníveis na Figura 25(a) e (iv) agregação multinível das subtrações com as camadas de decodificação, etapa de nome Agregação Incremental na Figura 25(a).

Figura 25 – Representação Gráfica da rede DMINet e JointAtt.

Fonte: Traduzida de Feng et al. (2023, online).

Em (a) tem-se a arquitetura da rede DMINet e em (b) do módulo JointAtt.

O módulo de atenção JointAtt presente, como mostra a Figura 25(b), diferentemente dos módulos BAM e MSA das redes STANet e RSCDNet, respectivamente, oferece um mecanismo de correlação intertemporal combinada, operação *Concat* na Figura 25(b), que por sua vez aumenta a capacidade da rede DMINet em detectar apenas mudanças genuínas, sendo menos provável que alterações nas imagens bitemporais provenientes de condições adversas, tais como variações de luminosidade e movimentação intermitente de *background*, sejam incorretamente classificadas como regiões de *foreground*.

3 Proposta

Essa tese propõe-se verificar a hipótese de que adequadas estratégias de modelagem computacional de mecanismos de pré-atenção visual, biologicamente inspirados, seriam capazes de impactar positivamente a construção de redes CNNs profundas de detecção de mudanças em termos de eficiência e precisão. Na Seção 3.1 apresentam-se as técnicas Rede Neural Convolucional Residual em Cascata, do inglês *Cascade Residual Convolutional Neural Network* (CRCNN) e Rede Neural Convolucional Residual em Cascata Multiescala, do inglês *Multiscale Cascade Residual Convolutional Neural Network* (MCRCNN), sendo ambos os métodos desenvolvidos utilizando, de maneira implícita, estratégia de pré-atenção visual espacialmente orientada, vide (POSNER, 1980). Na Seção 3.2 apresenta-se o método de detecção de mudanças em vídeos e imagens denominado Rede Leve de Atenção Temporal, do inglês *lightWeight Temporal Attention Network* (TITAN), que além de reunir em um só lugar muitas das melhores práticas observadas nos métodos analisados na Seção 2.3 utiliza, de maneira explícita, mecanismo de pré-atenção visual espacial de filtragem seletiva, vide (LABERGE; BROWN, 1989).

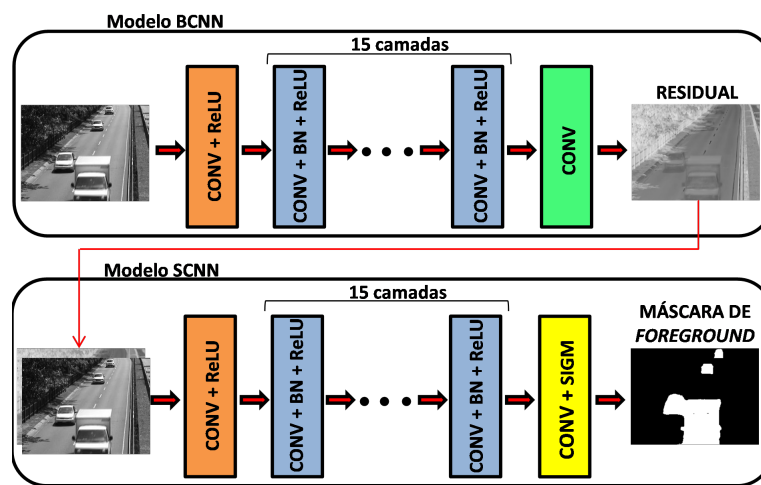
3.1 Atenção Residual Implícita

A terminologia atenção residual implícita é utilizada para designar as técnicas de detecção de mudanças em vídeos CRCNN e MCRCNN, aqui propostas, que utilizam processamento multiestágio para conseguir separar conteúdos de *background* e *foreground* das imagens processadas, sendo apresentadas em maiores detalhes nas Subseções 3.1.1 e 3.1.2, respectivamente. Ambas as técnicas são compostas por CNNs de primeiro e segundo estágios, sendo que no primeiro estágio de processamento a Rede Neural Convolucional de Segundo Plano, do inglês *Background Convolutional Neural Network* (BCNN), fica responsável por orientar espacialmente o mecanismo de atenção, via mapa residual, de sua correspondente rede CNN de segundo estágio, denominada Rede Neural Convolucional de Segmentação, do inglês *Segmentation Convolutional Neural Network* (SCNN).

3.1.1 Rede Neural Convolucional Residual em Cascata

Baseando-se no método de detecção de mudança Cascade, vide Subseção 2.3.2, propõem-se um novo método denominado Rede Neural Convolucional Residual em Cascata, do inglês *Cascade Residual Convolutional Neural Network* (CRCNN), ilustrado pela Figura 26.

Figura 26 – Arquitetura da rede CRCNN.



Fonte: Traduzida de Santos et al. (2019, p. 2).

Como mostra a Figura 26, o primeiro estágio de processamento da rede CRCNN, Módulo BCNN, visa extrair um mapa residual¹ da imagem de entrada, que por sua vez será utilizado pelo segundo estágio de processamento, Módulo SCNN, como uma espécie de guia para a rede SCNN.

Geração de Mapa Residual

No primeiro estágio de processamento do modelo CRCNN, denominado Rede Neural Convolucional de Imagem de Segundo Plano, do inglês *Background Convolutional Neural Network* (BCNN), utiliza-se de treinamento residual, com base no trabalho de Zhang et al. (2017), para gerar um mapa de valores reais denominado mapa residual. Sua principal função é servir como um mecanismo de pré-atenção espacial atuando como um guia para

¹ Matriz de valores reais denotando segmentações parciais

a manutenção da coerência espacial dos *pixels* da máscara de segmentação a ser gerada ao final da execução do segundo estágio SCNN de processamento.

Para que a rede BCNN aprenda como extrair a informação residual presente na imagem de entrada² f , informação não inclusa na imagem de *background* de referência b , seu treinamento consiste na minimização da função de custo:

$$L_B(b, f; \Theta_1) = \frac{1}{2m} \sum_{i=1}^m \|b_i - a_i\|_F^2, \quad (10)$$

sendo,

$$a = sig(f - BCNN(f; \Theta_1)), \quad (11)$$

onde Θ_1 representa o conjunto de parâmetros a serem otimizados, $BCNN(\cdot)$ o mapa residual de saída da rede BCNN, m a quantidade de amostras de treinamento utilizadas e $\|\cdot\|_F^2$ a aplicação da norma *Frobenius*.

Segmentação Guiada por Mapa Residual

Neste segundo estágio de processamento, como ilustra a Figura 26, aplica-se sobre a imagem monocromática alvo da detecção, concatenada com seu respectivo mapa residual, a Rede Neural Convolutiva de Segmentação, do inglês *Segmentation Convolutional Neural Network* (SCNN). A arquitetura da rede SCNN é muito parecida com a da rede BCNN, diferindo apenas com relação à última camada de convolução, retângulo em amarelo na Figura 26. O principal objetivo da última camada é gerar um mapa de probabilidades a partir da combinação linear dos mapas de característica da décima quinta camada de convolução da rede SCNN, seguida da aplicação de ativação por função linear logística, Equação (6).

Para fazer com que a rede aprenda a gerar o mapa de probabilidades utiliza-se um processo de otimização baseado na minimização da função de custo:

$$L_S(g, c; \Theta_2) = -\frac{1}{m} \sum_{i=1}^m [g_i \log(\hat{g}_i) + (1 - g_i) \log(1 - \hat{g}_i)], \quad (12)$$

sendo,

² Os valores da imagem de entrada são normalizados e pertencem ao intervalo $[0, 1]$.

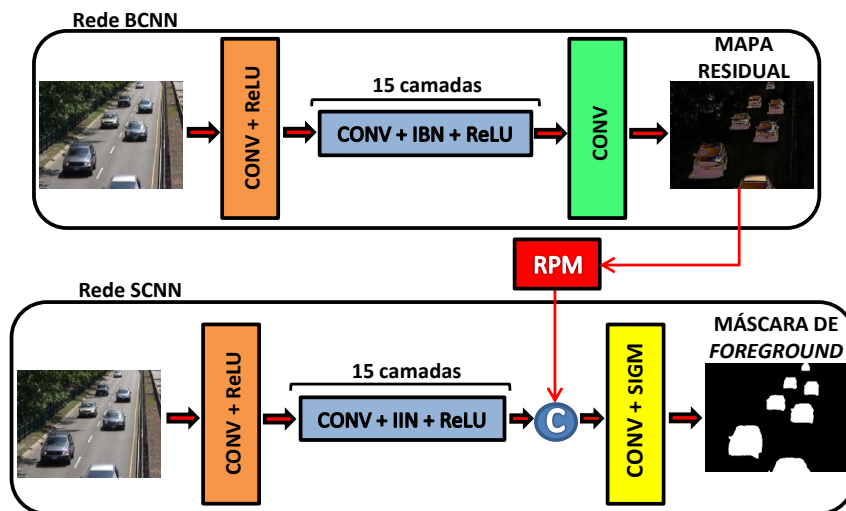
$$\hat{g} = SCNN(c; \Theta_2), \quad (13)$$

onde g indica a imagem rotulada, Θ_2 representa o conjunto de parâmetros da rede SCNN a serem otimizados e c indica a entrada concatenada da rede SCNN, sendo esta concatenação feita entre a imagem monocromática f e seu respectivo mapa residual $BCNN(f; \Theta_1)$.

3.1.2 Rede Neural Convolutiva Residual em Cascata Multiescala

Como mostra a Figura 27, a rede MCRCNN é uma extensão da técnica CRCNN. Também composta por dois estágios, tem acrescentada em sua arquitetura um Módulo de Processamento Residual, do inglês *Residual Processing Module* (RPM).

Figura 27 – Arquitetura da rede MCRCNN.



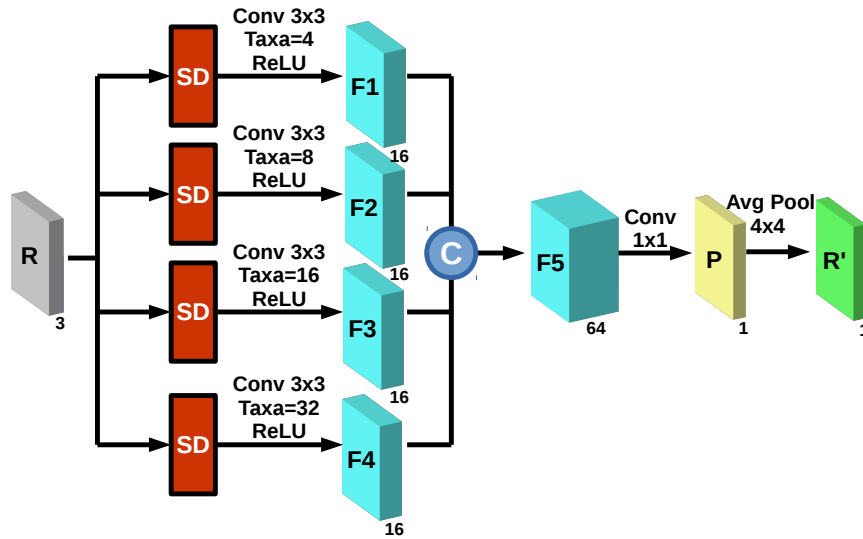
Fonte: Adaptada de Santos et al. (2019, p. 2).

Módulo de Processamento Residual

Visando solucionar o problema da coerência espacial do mapa de probabilidades gerado pela rede SCNN, construiu-se o módulo RPM utilizando como inspiração os módulos FPM e M_FPM presentes nas redes de detecção de mudança FgSegNets, vide Subseções 2.3.3 e 2.3.4. No entanto, diferentemente dos módulos FPMs, o módulo RPM

utiliza menos filtros convolucionais e processamento não estratificado e logo sua arquitetura acaba sendo mais compacta em comparação com os módulos FPMs.

Figura 28 – Arquitetura do Módulo de Processamento Residual RPM.



Fonte: Elaborada pelo autor.

Na Figura 28 observar-se que o módulo RPM é constituído por quatro diferentes camadas de convolução com capacidade de dilatação espacial de filtros nas taxas de valores 4, 8, 16 e 32 vezes, sendo que transformações espaciais SD são aplicadas antes de cada convolução e ativações ReLU, vide Equação (8), logo após. Após aplicação das dilatações, gera-se o conjunto de mapas de característica $\mathbf{F5} = \{\mathbf{F1} \cdot \mathbf{F2} \cdot \mathbf{F3} \cdot \mathbf{F4}\}$, onde $\cdot$ indica concatenação entre os mapas. O conjunto $\mathbf{F5}$ passa então por um processo de combinação local de características por meio da aplicação de convolução pontual, convolução com filtro de dimensão 1×1 . Sobre o resultado da combinação $\mathbf{P}$ aplica-se operação de *avg pool*, vide Subseção 2.2.4, buscando produzir o mapa residual R , representado na Figura 28 por $\mathbf{R'}$.

3.2 Atenção Residual Explícita

Utilizamos a terminologia atenção residual explícita para nos referir à estratégia adotada pelo método de detecção de mudanças proposto, técnica TITAN, para tentar replicar o comportamento de pré-atenção visual por filtragem seletiva proposto por LaBerge

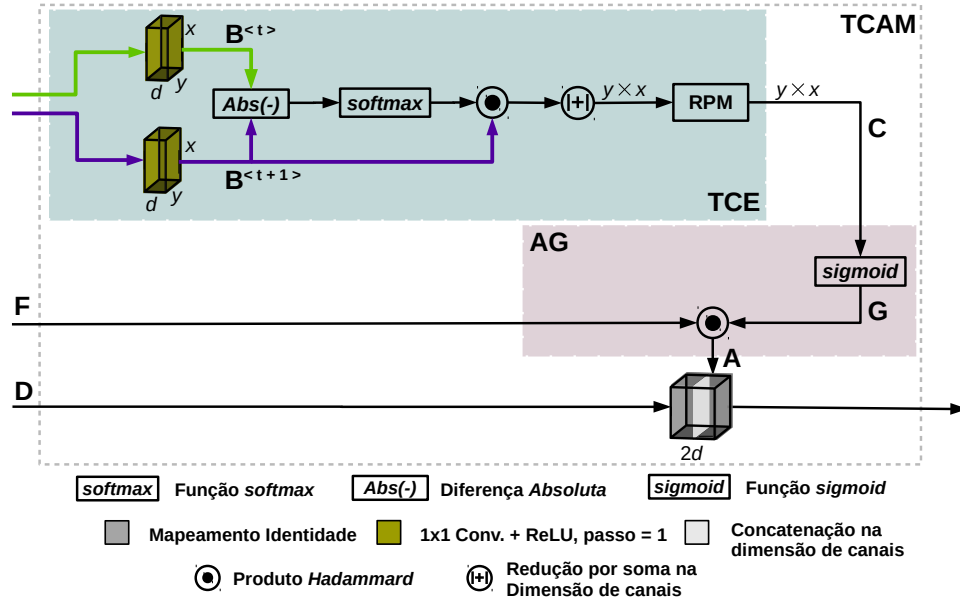
fora criada a Rede Leve de Atenção Temporal, do inglês *lighTweIght Temporal Attention Network* (TITAN), cuja arquitetura é mostrada pela Figura 29.

De acordo com a Figura 29, para um dado par de imagens $p = (I^t, I^{t+1})$, tal que $p \in \mathcal{R}^{h \times w \times 3}$, onde h e w denotam as respectivas dimensões em termos de altura e largura e as variáveis t e $t + 1$ denotam os respectivos instantes de captura das imagens, o método TITAN tem como objetivo detectar em I^{t+1} as regiões de mudança com relação à I^t . Isso é feito utilizando uma arquitetura de rede codificadora e decodificadora, parcialmente siamesa, com o auxílio de um módulo de processamento de características, posicionado no mais baixo nível do codificador da rede (módulo LFPM), e de um mecanismo de auto-observação chamado de Módulo de Atenção a Mudanças Temporais, do inglês *Temporal Change Attention Module* (TCAM). Assim como no caso das redes de codificação e decodificação das Subseções 2.3.3 e 2.3.4, a utilização do módulo LFPM na rede TITAN tem como objetivo aumentar o alcance do campo receptivo da rede, para que ela seja capaz de incorporar mais informação contextual presente ao redor das regiões de mudança e assim minimizar os problemas de inconsistências (falso positivos e falso negativos) no processo de detecção de uma maneira computacionalmente mais eficiente, justificada pela utilização de convoluções dilatadas, vide Subseção 2.2.5.

Módulo de Atenção a Mudanças Temporais

O módulo TCAM, apresentado na Figura 30, fora desenvolvido como intuito de simplificar o mecanismo de aprendizado residual adotado nos métodos CRCNN e MCRCNN. Contando com dois estágios de processamento, denominados de Expectativa de Mudança Temporal, do inglês *Temporal Change Expectation* (TCE), e Limitador de Atenção, do inglês *Attention Gate* (AG), o módulo TCAM busca preservar a mesma efetividade dos resultados de detecção obtidos por técnicas cuja arquitetura pode ser decomposta em camadas codificadoras e decodificadoras, como apresentado nas Subseções 2.3.3 à 2.3.5, porém a um custo computacional menor.

Como mostra a Figura 30, dados os tensores de entrada do módulo TCAM $\mathbf{B}^t \in \mathcal{R}^{x \times y \times d}$ e $\mathbf{B}^{t+1} \in \mathcal{R}^{x \times y \times d}$, onde $d = 16$ e $d = 64$ para TCAM1 e TCAM2, respectivamente, sendo que x e y correspondem às respectivas dimensões de altura e largura dos tensores, o estágio TCE calcula as expectativas de mudanças temporais $\mathbf{C} \in \mathcal{R}^{x \times y \times 1}$. O processo utilizado para gerar $\mathbf{C}$ é iniciado pelo cálculo das distribuições temporais de probabilidades

Figura 30 – Módulo de Atenção a Mudanças Temporais (TCAM).

Fonte: Traduzida de Santos e Papa (2023, p. 3).

de mudanças dadas por:

$$p_i = \frac{e^{|b_i^{<t+1>} - b_i^{<t>}|}}{\sum_{j=1}^d e^{|b_j^{<t+1>} - b_j^{<t>}|}}, \quad (14)$$

onde $i \in [1, x \times y]$, sendo que $b_i^{<t>}$, $b_i^{<t+1>}$, e p_i representam os elementos de $\mathbf{B}^{<t>}$, $\mathbf{B}^{<t+1>}$, e $\mathbf{P}$ tomados da i -ésima posição espacial, respectivamente. Tendo sido calculadas as probabilidades de mudanças temporais $\mathbf{P}$, o mapa das expectativas de mudanças passa a ser definido por:

$$\mathbf{C} = \sum_{i=1}^d b_i^{<t+1>} \odot p_i, \quad (15)$$

sendo que $\odot$ simboliza o produto *Hadamard*.

Após completar a etapa de processamento TCE, o módulo TCAM entra em seu segundo estágio de execução, responsável por gerar o limitador de atenção $\mathbf{G} \in \mathcal{R}^{x \times y \times 1}$ definido por:

$$g_i = \frac{1}{1 + e^{-c_i}}, \quad (16)$$

onde g_i representa cada elemento de $\mathbf{G}$ tomado da i -ésima posição espacial.

O mecanismo de atenção às mudanças temporais é então criado a partir da aplicação do limitador $\mathbf{G}$ sobre o tensor $\mathbf{F} \in \mathbb{R}^{x \times y \times d}$, sendo $\mathbf{F}$ o subproduto da combinação linear dos tensores $\mathbf{B}^{<t>}$ e $\mathbf{B}^{<t+1>}$, vide Figura 29. Desta forma, a equação que define o mecanismo de atenção é dada por:

$$\mathbf{A} = \mathbf{G} \odot \mathbf{F}. \quad (17)$$

Em linhas gerais, a geração implícita do mapa de expectativas de mudanças $\mathbf{C}$ pelo módulo TCAM substitui a necessidade de usar uma rede auxiliar, como no caso dos métodos CRCNN e MCRCNN (vide Subseções 3.1.1 e 3.1.2, respectivamente), apenas para estimar mapas de ruídos a serem posteriormente utilizados no ajuste de inconsistências no processo de detecção de mudanças.

Otimização de Função de Custo

O processo de otimização da rede TITAN consiste na minimização de uma função de custo de entropia cruzada, do inglês *Cross-Entropy*, de modo a fazer com que os valores contidos nos tensores $\mathbf{Z} \in \mathbb{R}^{x \times y \times 2}$, produzidos pela rede, se aproximem dos seus respectivos valores binários de referência⁴ contidos em $\mathbf{M} \in \mathbb{R}^{x \times y \times 2}$. Dessa maneira a função de custo da rede TITAN passa a ser representada por:

$$L(\mathbf{M}, \mathbf{Z}; \Theta) = -\frac{1}{x \times y} \sum_{i=1}^{i=x \times y} \sum_{k=1}^2 m_{i,k} \log(z_{i,k}), \quad (18)$$

onde $m_{i,k}$ e $z_{i,k}$ indicam as localizações dos valores de $\mathbf{M}$ e $\mathbf{Z}$, que variam espacialmente e em profundidade de acordo com i e k , respectivamente. O termo Θ representa o conjunto de hiperparâmetros a serem otimizados.

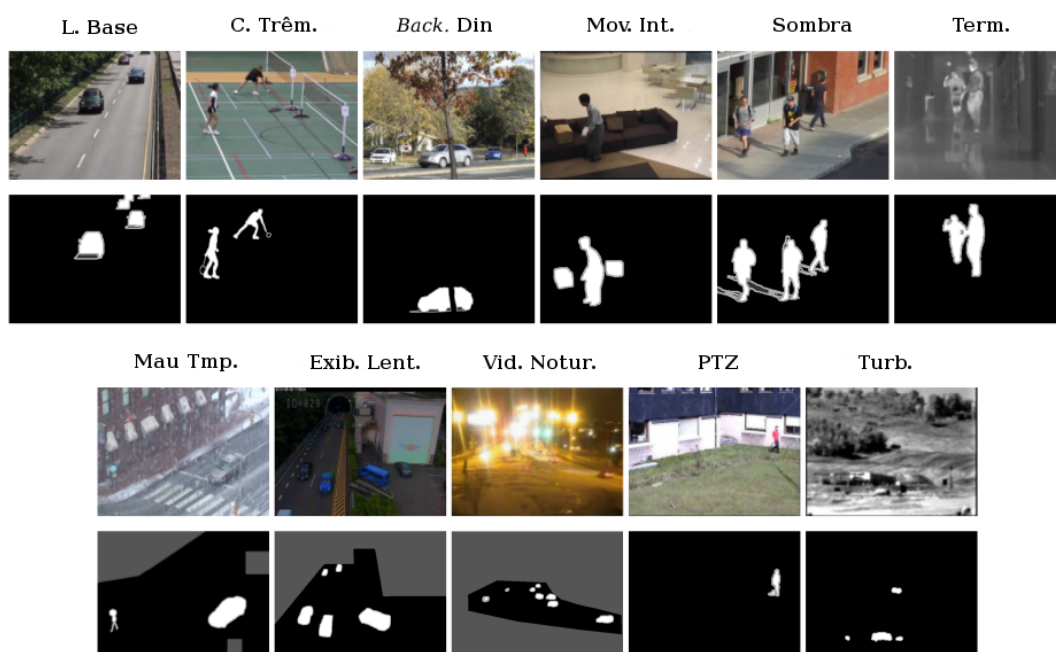
⁴ Anotações manuais dos *pixels* como sendo de mudança (valor unitário) ou não (valor nulo).

4 Recursos e Metodologia

Para avaliar os mecanismos e técnicas propostos no Capítulo 3 foram utilizados (a) um computador com processador intel i7 de oito núcleos de processamento, (b) 32GB de Memória de Acesso Randômico, do inglês *Random Access Memory* (RAM), e (c) uma placa de processamento gráfico de propósito geral Nvidia GeForce 1080Ti. Os métodos foram codificados em linguagem de programação *python*, versão 3.6, utilizando o ambiente *pytorch* de Paszke et al. (2017), versão 1.10.0. As bases de dados utilizadas no estudo subdividem-se nos grupos voltados à (i) detecção de mudanças em vídeos e (ii) detecção de mudanças em imagens aéreas de sensoriamento remoto, abordadas respectivamente nas Seções 4.1 e 4.2.

4.1 Base de Detecção de Mudanças em Vídeos

Figura 31 – Amostras de imagens das categorias de vídeos da base CD2014.



Fonte: Adaptada de Wang et al. (2014, p 390).

Como mostra a Figura 31, a base de Detecção de Mudanças 2014, do inglês *Change Detection Dataset 2014* (CD2014), elaborada por Wang et al. (2014), é composta por

vídeos coletados de ambientes realísticos, não gerados por meio de computação gráfica, sob influência de condições adversas. Os vídeos se distribuem em 11 categorias, sendo elas:

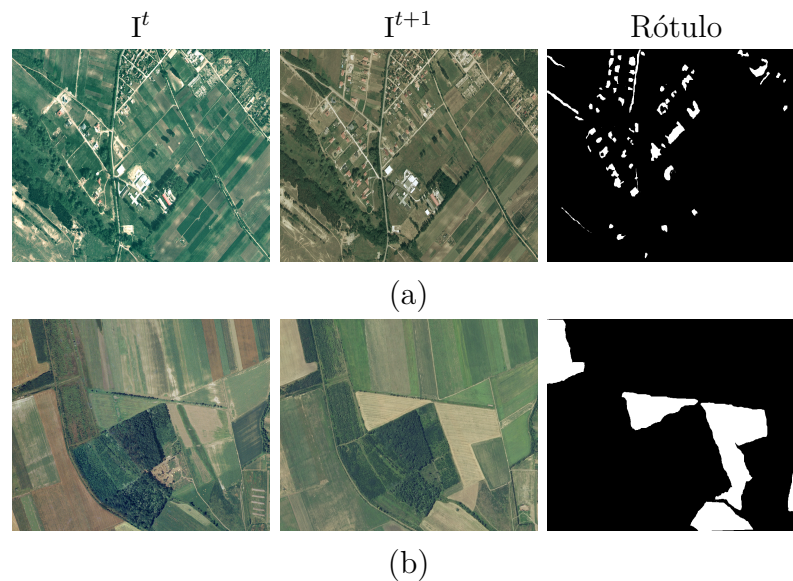
- **Linha de Base** (L. Base): Combina imagens com características similares a das categorias *Background* Dinâmico (*Back. Din.*), Câmera Trêmula (*C. Trêm.*), Movimento Intermitente de Objeto (*Mov. Int.*) e Sombra. Possui um total de quatro vídeos, sendo eles autoestradaH, pedestres, escritório e PETS2006;
- **Background Dinâmico**: Inclui cenas com movimentação de elementos de *background*, por exemplo movimentação de galhos de árvores. Os seis vídeos pertencentes à categoria *Back. Din.* são botes, canoa, queda, fonte01, fonte02 e ultrapassagem;
- **Câmera Trêmula**: Contém cenas de ambientes fechados e abertos capturadas por dispositivos desestabilizados, por exemplo pelo impacto de vibrações externas;
- **Movimento Intermitente de Objeto**: Possui cenas onde objetos de *foreground* se movem, param por um longo período de tempo e depois reiniciam seu movimento;
- **Sombra**: Contém vídeos de ambientes fechados e abertos onde objetos de *foreground* encontram-se envoltos por regiões de sombra;
- **Térmica** (Ter.): Conjunto de vídeos capturados por câmeras infravermelho de longo alcance;
- **Mau Tempo** (Mau Tmp.): Contém filmagens de ambientes abertos sob influência de forte nevasca;
- **Exibição Lenta** (Exib. Lent.): Possui um conjunto de vídeos de baixa taxa de exibição, que variam de 0,17 à 1 quadro por segundo;
- **Vídeos Noturnos** (Vid. Notur.): Vídeos coletados de ambientes noturnos sob condições precárias de luminosidade;
- **PTZ**: Categoria composta por cenas coletadas de câmeras de rotação horizontal, vertical e zoom, do inglês *Pan Tilt Zoom*;
- **Turbulência** (Turb.): Vídeos coletados de ambientes abertos em que algumas cenas foram corrompidas por ruído de turbulência atmosférica, gerado pela movimentação de ar quente.

4.2 Base de Imagens Aéreas de Sensoriamento Remoto

Base de dados de imagens capturadas de um mesmo local em diferentes instantes de tempo, onde modificações ocorridas durante o intervalo de tempo de captura, tais como prédios recém construídos, alterações de vegetação, construção de estradas, são considerados como mudanças. Subdivide-se em quatro grupos de imagens sendo eles:

- **SZTAKI**: Criada por Benedek e Szirányi (2008) e por Benedek e Szirányi (2009), contém 13 pares de imagens de dimensões 952×640 com resolução de $1,5m/pixel$ cada, sendo que o intervalo de tempo entre as capturas das imagens de cada par varia de 5 e 7 anos, respectivamente. Exemplos referentes aos subgrupos **Szada7** e **Tiszadob5**, contidos na base **SZTAKI**, são mostrados na Figura 32.

Figura 32 – Imagens aéreas de sensoriamento remoto **SZATAKI**.



Fonte: Imagens retiradas de Benedek e Szirányi (2013).

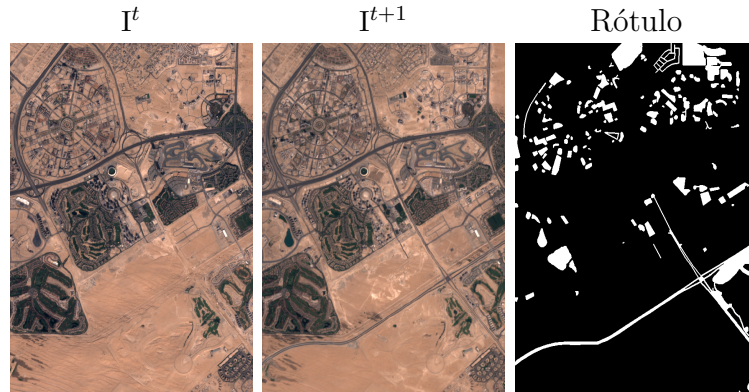
Exemplos de capturas bitemporais (a) **Szada** e (b) **Tiszadob**.

Em nossos experimentos utilizamos as nomenclaturas **Szada7** e **Tiszadob5**, que também servem para indicar as quantidades de pares de imagens pertencentes às categorias **SZADA** e **TISZADOB**, respectivamente.

- **Satélite Onera**: Gerada por Daudt et al. (2018b), possui 24 pares de imagens coletadas pelo satélite *Sentinel-2* entre os anos de 2015 e 2018, sendo que a Figura 33

apresenta uma par de imagens capturadas da cidade de Dubai.

Figura 33 – Imagens aéreas de sensoriamento remoto **Onera**.



Fonte: Imagens retiradas de Daudt et al. (2018a).

Satélite Onera também contém cidades localizadas no Brasil, Estados Unidos, Europa, Oriente Médio e Ásia. Em nossos experimentos a base **Satélite Onera** fora subdividida em *Onera10* e *Onear14* contendo 10 e 14 imagens, respectivamente;

- **LEVIR-CD:** Criada por Chen e Shi (2020b), é composta por 637 pares de imagens coletadas do *Google Earth*, sendo que cada imagem possui uma alta resolução ($0,5m/pixel$) e dimensões 1024×1024 . Intervalos de captura das imagens variam de 5 à 14 anos.

Figura 34 – Imagens aéreas de sensoriamento remoto **LEVIR-CD**.



Fonte: Imagens retiradas de Chen e Shi (2020a).

A base **LEVIR-CD** subdivide-se em três grupos denominados **LEVIR64**, **LEVIR128** e **LEVIR445**, sendo compostos por 64, 128 e 445 pares de imagens, respectivamente. Na Figura 34 é apresentado um exemplo pertencente ao grupo **LEVIR445**;

- **SYSU-CD**: Desenvolvida por Shi et al. (2021), contém 20.000 pares de imagens com resolução espacial de $0,5m/pixel$ e dimensões 256×256 , tendo sido coletadas entre os anos de 2004 e 2007 dos arredores de Hong Kong. As mudanças anotadas incluem: (a) construções recentes de prédios, (b) crescimentos de áreas suburbanas, (c) zoneamentos urbanos, (d) mudanças de vegetação, (e) construções de estradas e (f) construções marítimas. A base **SYSU-CD** é disponibilizada por Shi et al. (2021) subdividida em três partes, denominadas **SYSUtrain**, **SYSUval** e **SYSUtest**, sendo que cada parte contém 12.000, 4.000 e 4.000 pares de imagens, respectivamente. A Figura 35 apresenta um exemplo pertencente ao grupo **SYSUtest**.

Figura 35 – Imagens aéreas de sensoriamento remoto **SYSU-CD**.



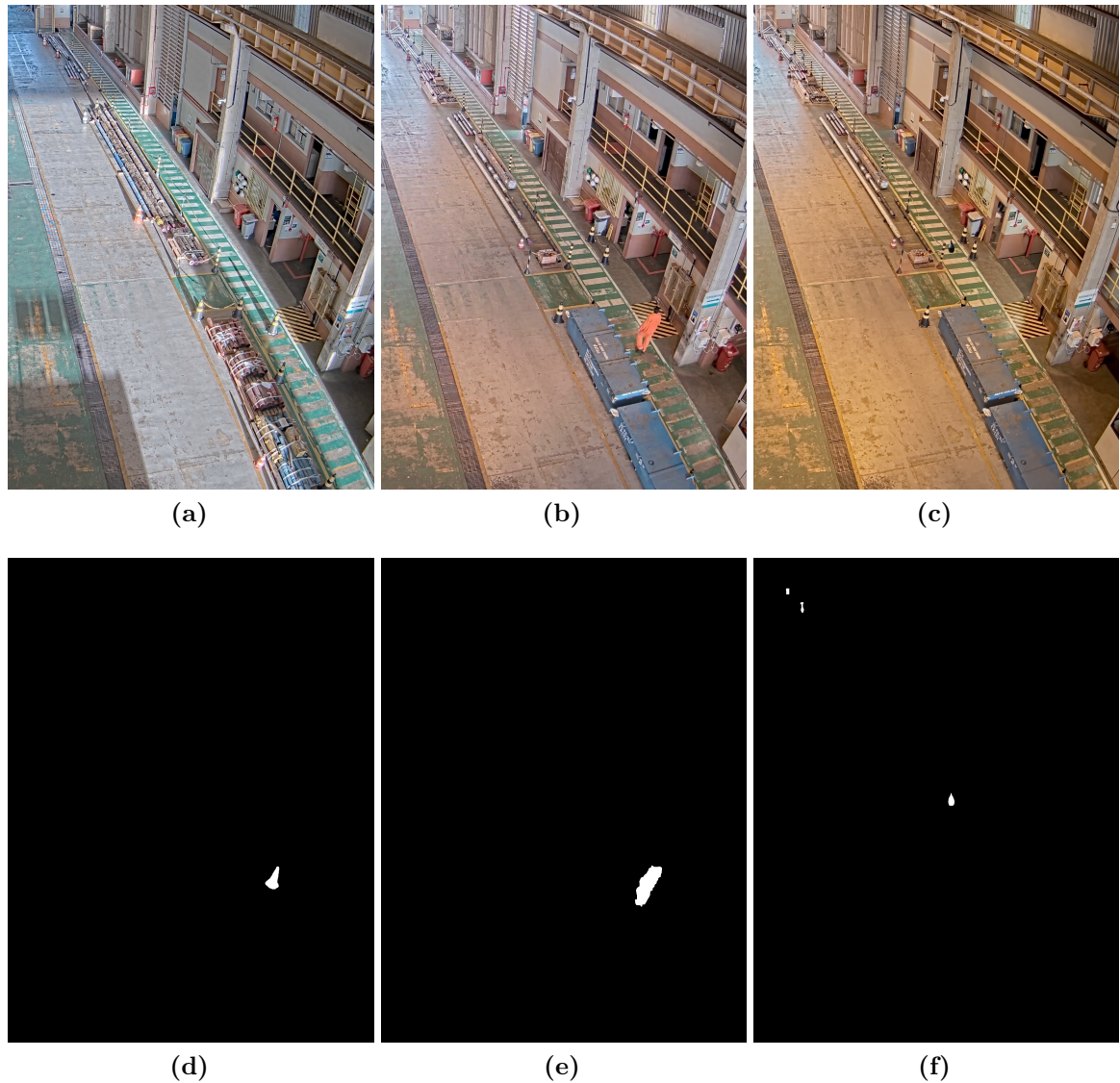
Fonte: Imagens retiradas de Shi et al. (2020).

4.3 Base PetrobrasROUTES de Imagens de Obstruções

Criada por Santos et al. (2021), durante o desenvolvimento de um sistema de detecção de obstruções de rota de escape emergencial (localizada em uma das plataformas de exploração de petróleo da Petrobras em Macaé). Composta por imagens coletadas de uma câmera fixa de ultra resolução, imagens de dimensão 3.840×2.160 *pixels*, a base PetrobrasROUTES é composta por quatro conjuntos de imagens, sendo eles denominados:

- Macae: Conjunto composto por 55 imagens manualmente selecionadas e rotuladas de gravações intermitentes realizadas, em intervalos de 24 horas, durante um período de três dias. Dentre as classes de objetos geradoras de obstruções encontram-se (i) itens de limpeza (vassouras, mangueiras, lixeiras e sacola plástica), (ii) ferramentas de trabalho (chave de fenda, caixa de ferramentas e carrinho de suporte), (iii) itens de segurança (capacetes, luvas e cones de sinalização), (iv) pessoas e (v) cadeira de escritório;
- MacaeA1: Conjunto formado por 65 imagens, sendo 55 pertencentes ao conjunto Macae e 10 tendo sido manualmente coletadas a partir de gravações intermitentes com duração média de dez minutos. Onde a diferença de tempo entre as coletas das imagens da base Macae e das outras dez imagens é de no máximo um ano e seis meses. A rotulagem do conjunto originário das 10 imagens acrescidas à base Macae foi feita de maneira automática, utilizando o método estatístico paramétrico de detecção de mudanças de Zivkovic (2004b) Misturas Gaussianas Melhoradas, do inglês *Improved Mixture of Gaussians* (MOG2);
- MacaeA2: Conjunto formado por 65 imagens, sendo 55 pertencentes ao conjunto Macae e 10 tendo sido artificialmente geradas. Para a geração artificial fora utilizada uma ferramenta, desenvolvida por Santos et al. (2021), capaz de inserir objetos já rotulados da base de imagens Objetos Comuns em Contexto, do inglês *Common Objects In Context* (COCO) de Lin et al. (2014) ao longo da rota de escape contida em imagens de *background* semiautomaticamente coletadas.
- MacaeA3: Conjunto formado por 75 imagens, sendo composto pela combinação entre os conjuntos Macae, MacaeA1 e MacaeA2.

A Figura 36 exemplifica as três situações possíveis de obstruções encontrados na base de dados PetrobrasROUTES, onde nas Figuras 36(a) à 36(c) têm-se exemplos pertencentes aos grupos Macae, MacaeA1 e MacaeA2, respectivamente. Sendo que nas Figuras 36(d) à 36(f) apresentam-se as correspondentes regiões de mudanças rotuladas. Nas Figuras 36(a) à 36(c) a rota de escape aparece como uma região listrada nas cores branco e verde.

Figura 36 – Obstruções de rota de escape para base PetrobrasROUTES.

Fonte: Elaborado por (SANTOS et al., 2021).

4.4 Protocolos de Treinamento e Avaliação

Para uma maior organização metodológica, os protocolos utilizados durante os experimentos foram subdivididos em dois grandes grupos, sendo eles o grupo de Protocolos de Estudo de Configuração e de Estudo Comparativo, como mostram as correspondentes Subseções 4.4.1 e 4.4.2. No grupo dos Estudos de Configuração são definidas as estratégias de treinamento e avaliação utilizadas para medir os impactos gerados pela utilização dos mecanismos de Atenção Residual Implícita e Explícita nos métodos CRCNN, MCRCNN e TITAN, propostos no Capítulo 3. No grupo dos Protocolos de Estudo Comparativo definem-se as metodologias de treinamento e avaliação utilizadas para viabilizar o processo

de comparação entre os métodos correlatos, abordados na Seção 2.3, e as melhores versões das técnicas CRCNN, MCRCNN e TITAN, encontradas após execução dos Estudos de Configuração.

Os resultados dos Estudos de Configuração e dos Estudos Comparativos foram avaliados utilizando-se as medidas quantitativas abordadas na Subseção 2.1.4, sendo elas Medida-F, PWC, IOU, ROC e PR. Observando que as medidas de área sob curva, vide (DAVIS; GOADRICH, 2006), para Característica do Operador Receptor, do inglês *Receiver Operator Characteristic* (ROC) e para precisão-revocação, do inglês *Precision Recall* (PR), são aplicadas apenas nos casos em que os métodos produziram como saídas mapas de probabilidades¹. Para dar suporte estatístico às análises quantitativas, seguindo as recomendações de Demšar (2006), foram utilizados os testes não paramétricos de Friedman (1937), de Nemenyi (1963) e de Dunnett (1955), para análises de significância estatística envolvendo três ou mais técnicas e de Wilcoxon (1992), para os casos envolvendo a comparação entre duas técnicas. O limiar de significância estatística adotado foi 0,05, exceto em nossos Estudos de Configuração aplicados sobre as técnicas CRCNN e MCRCNN onde fora adotado um limiar de 0,1. Esse ajuste de limiar para as variações CRCNN e MCRCNN foi necessário para que fosse possível detectar diferenças muito pequenas entre elas, sem com tudo haver a possibilidade de aumentar ainda mais a quantidade de amostras de vídeos fornecidos pela base CD2014.

Resultados qualitativos são apresentados por meio de um código de cores que, após ser aplicado sobre as máscaras de detecção geradas pelos métodos analisados, é capaz de indicar nas cores azul, preto, vermelho e amarelo, as regiões correspondentes a TPs, TNs, FNs e FPs, respectivamente.

4.4.1 Protocolos de Estudo de Configuração

Considerando a base de detecção de mudanças CD2014, vide Seção 4.1, para avaliar as melhores configurações de treinamento, com e sem aplicação do mecanismo de atenção residual implícita sobre os métodos CRCNN e MCRCNN, utilizou-se o seguinte protocolo:

1. Selecionar 200 imagens de cada vídeo pertencente às categorias L. Base, Ter., Mau Tmp., Exib. Lent., Vid. Notur., PTZ e Turb., sendo seus índices iguais ao das imagens também utilizadas por Lim e Keles (2018) para treinamento;

¹ Estruturas de dados planas contendo valores reais no intervalo [0; 1].

2. Randomizar três vezes cada conjunto de 200 imagens e ao final de cada randomização subdividi-los em quatro partes, a serem utilizadas nas etapas de treinamento, validação, ajuste de limiar de binarização e avaliação (quantitativa e qualitativa) de cada nova configuração de rede. Os subconjuntos deverão ter um total de 20, 5, 35 e 140 imagens cada, respectivamente;
3. Iniciar o processo sequencial e iterativo de configuração, treinamento e avaliação. Cada modelo recém configurado deverá ser treinado sobre os subconjuntos de treino e validação definidos na etapa 2, onde o conjunto de validação é utilizado para determinar o melhor momento para armazenar em disco os parâmetros otimizados para cada vídeo escolhido na etapa 1;
4. Encerrado o treinamento para determinada randomização e vídeo específico, iniciar o processo de busca pelo melhor valor de limiar de binarização para o mapa de probabilidades de saída do modelo. Isso se dá por meio de uma busca sequencial no intervalo de limiares $[0,05; 0,95]$ em passos de 0,05, realizada sobre o subconjunto de ajuste da etapa 2, sendo escolhido como limiar o valor correspondente à maior Medida-F encontrada;
5. Testar o modelo recém treinado sobre o subconjunto de avaliação definido na etapa 2, utilizando para binarização os limiares encontrados após execução da etapa 4;
6. Comparar as médias² dos resultados quantitativos (Medida-F, PWC, IOU, ROC e PR) com os resultados obtidos por configurações anteriores. Voltar à etapa 3 enquanto existirem novas possibilidades de configurações. Para os casos de indeterminação de escolha das melhores configurações, utilizar o protocolo mais abrangente da Subseção 4.4.2.

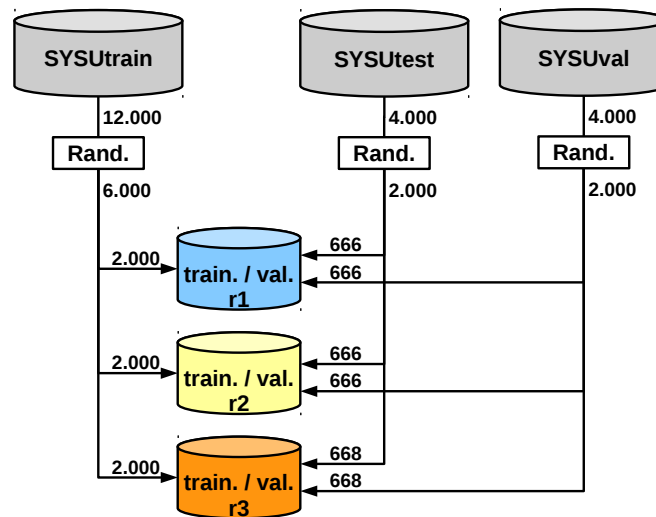
Considerando as bases de imagens aéreas de sensoriamento remoto, vide Seção 4.2, para avaliar as melhores configurações de treinamento, com e sem aplicação do mecanismo de atenção residual explícita sobre o método TITAN, utilizou-se o seguinte protocolo:

1. Randomizar três vezes o conteúdo das bases **SYSUtrain**, **SYSUtest** e **SYSUval**, sendo que a cada nova randomização 50% das imagens serão utilizadas para execução

² Assim como em (DEMŠAR, 2006), cada vídeo da base CD2014 fora tratado como uma base de dados diferente.

das etapas de treinamento, validação e teste de configuração do método TITAN. Observando que as etapas de validação e teste serão executadas sobre o mesmo³ conjunto de imagens não utilizadas na etapa de treinamento. A Figura 37 ilustra o processo de construção dos conjuntos de treinamento e validação/teste, sendo que dos 3.332, 3.332 e 3.336 pares de imagens presentes, respectivamente, nos conjuntos **train. / val. r1**, **train. / val. r2** e **train. / val. r3**, 80% foram utilizados para treinamento e 20% para validação/teste;

Figura 37 – Método de subdivisão da base de dados **SYSU-CD** utilizado como parte dos protocolos de estudo de configuração e comparativo.

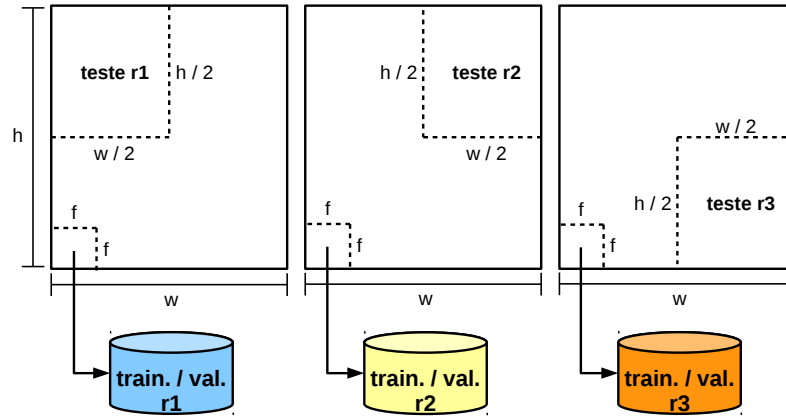


Fonte: Elaborada pelo autor.

2. Subdividir as bases **SZATAKI**, **Onera** e **LEVIR-CD** em três partes, como mostra a Figura 38.

Além disso, após a extração de fragmentos, os conjuntos formados por **train. / val. r1**, **train. / val. r2** e **train. / val. r3** deverão ser randomizados, por três vezes, sendo cada randomização seguida por subdivisão em partes de 80% e 20% dos totais de imagens. Assim como em 1, as frações de 80% serão utilizadas para treinamento e as de 20% para validação/teste;

³ Assim como em (SANTOS; PAPA, 2023), compartilham-se as imagens utilizadas nas etapas de validação e teste apenas para ajuste de configuração do modelo TITAN.

Figura 38 – Método de subdivisão das bases de dados **SZATAKI**, **Onera** e **LEVIR-CD**.

Fonte: Elaborado pelo autor.

As dimensões de altura e largura das imagens são indicadas por h e w , respectivamente, e f indica a dimensão dos fragmentos de treinamento e validação, sendo $f = 128$ para as bases **SZATAKI** e **Onera** e $f = 256$ para a base **LEVIR-CD**.

3. Iniciar o processo sequencial e iterativo de configuração, treinamento e avaliação. Cada modelo recém configurado é treinado sobre os subconjuntos de treino e validação definidos nas etapas 1 e 2, onde o conjunto de validação é utilizado para determinar o melhor momento para armazenar em disco os parâmetros otimizados e também para escolha das melhores configurações;
4. Ao final da etapa de treinamento, aplica-se o modelo treinado sobre os conjuntos de validação/teste gerados nas etapas 1 e 2. Observando que nessa mesma etapa, para fins de análises estatísticas, os conjuntos **SZATAKI**, **Onera** e **LEVIR-CD**, são subdivididos nos conjuntos **Szada7**, **Tiszadob5**, **Onera10**, **Onera14**, **LEVIR64**, **LEVIR128** e **LEVIR445**;
5. Comparar as médias dos resultados quantitativos⁴ (Medida-F, PWC, IOU) com os resultados obtidos por configurações anteriores.

4.4.2 Protocolos de Estudo Comparativo

Para comparar as melhores configurações dos métodos encontrados após realização dos Estudos de Configuração da Subseção 4.4.1, com as técnicas correlatas Cascade,

⁴ Por não gerarem um mapa de probabilidades de saída, os resultados obtidos por diferentes configurações da rede TITAN não foram avaliados pelas métricas PR e ROC.

FgSegNets, RTSegNet, DeepBS, SEUNets e MUNets, utilizou-se a base de detecção de mudanças CD2014, vide Seção 4.1, e o seguinte protocolo:

1. Selecionar 200 imagens de cada vídeo da base CD2014, considerando as 11 categorias da Seção 4.1, sendo seus índices iguais ao das imagens também utilizadas por Lim e Keles (2018) para treinamento;
2. Para cada vídeo utilizado, randomizar três vezes os respectivos conjuntos de treinamento e validação contendo 200 imagens e também, pelo mesmo número de vezes, o restante das imagens não utilizadas para treino ou validação. Ao final de cada randomização terão sido gerados quatro subconjuntos, a serem utilizados nas etapas de treinamento, validação, ajuste de limiar para binarização e avaliação (quantitativa e qualitativa) de cada modelo. Os subconjuntos de treinamento e validação serão formados por 80% e 20%, respectivamente, do total de imagens do conjunto de 200. Os subconjuntos de ajuste de limiar e avaliação serão formados por 20% e 80%, respectivamente, do total de imagens não pertencentes ao conjunto de 200;
3. Iniciar o processo sequencial e iterativo de treinamento e avaliação, onde nas comparações por vídeo unitário, envolvendo métodos Cascade, MCRCNN, TITAN, FgSegNets e RTSegNet, serão treinados e avaliados modelos individuais, sendo um para cada vídeo da base CD2014. Nas comparações por grupos de vídeos, envolvendo técnicas DeepBS, SEUNets, MUNets, MCRCNN e TITAN, serão treinados e avaliados sobre os vídeos da base CD2014 apenas um único modelo por técnica;
4. Ao concluir do treinamento de cada randomização, iniciar o processo de busca pelo melhor valor de limiar de binarização para o mapa de probabilidades de saída dos modelos. Esse processo segue o mesmo protocolo da etapa 4 dos estudos de configuração aplicados sobre as técnicas CRCNN e MCRCNN.

Para comparar a melhor configuração da técnica TITAN, encontrada após realização dos Estudos de Configuração da Seção 4.4.1, com as técnicas correlatas de detecção de mudanças DMINet, RSCDNet, STANet e FLS, com base nas imagens aéreas de sensoriamento remoto, vide Seção 4.2, utilizou-se o seguinte protocolo:

1. Aplicar o mesmo critério da etapa 1 do protocolo de subdivisão da base **SYSU-CD** utilizado nos Estudos de Configuração do método TITAN. Agora as imagens

utilizadas na etapa de teste não serão mais compartilhadas pela etapa de validação, sendo então utilizadas apenas os outros 50% do total de imagens da base **SYSU-CD** que não compõem os conjuntos **train. / val. r1**, **train. / val. r2** e **train. / val. r3**, como mostra a Figura 37;

2. Aplicar o mesmo critério da etapa 2 do protocolo de subdivisão da base **SZATAKI** utilizado nos Estudos de Configuração do método TITAN, porém utilizar para teste os conjuntos **teste r1**, **teste r2** e **teste r3** exibidos pela Figura 38;
3. Iniciar o processo sequencial e iterativo de treinamento e avaliação, onde cada método comparado será treinado sobre os subconjuntos de treino e validação definidos nas etapas 1 e 2, onde o conjunto de validação é utilizado apenas para determinar o melhor momento para armazenar em disco os parâmetros otimizados;
4. Concluído o treinamento, aplicar o método recém treinado sobre os conjuntos de teste gerados nas etapas etapas 1 e 2. Observando que nessa mesma etapa, para fins de análises estatísticas, os conjuntos **SZATAKI**, **Onera** e **LEVIR-CD**, são subdivididos nos conjuntos **Szada7**, **Tiszadob5**, **Onera10**, **Onera14**, **LEVIR64**, **LEVIR128** e **LEVIR445**;
5. Comparar as médias dos resultados quantitativos (Medida-F, PWC, IOU) obtidas pelos diferentes métodos treinados utilizando as bases de imagens aéreas de sensoriamento remoto.

Sobre as técnicas de detecção de mudanças que desempenharam melhor nas bases CD2014 e de imagens aéreas de sensoriamento remoto, realizou-se um Estudo de Comparação sobre a base PetrobrasROUTES de imagens de obstruções de rotas de escape, vide Seção 4.3, onde o seguinte protocolo fora adotado:

1. Subdividir os grupos Macae, MacaeA1, MacaeA2 e MacaeA3 em conjuntos de treinamento, validação e teste, sendo 35 imagens separadas para teste (as mesmas 35 imagens serão utilizadas para avaliar cada grupo) e 20, 30, 30 e 40 separadas para treino e validação de cada grupo, respectivamente. As imagens de teste foram manualmente escolhidas de modo a representarem as seis maiores variações de luminosidade ambiente encontradas num período de vinte e quatro horas de filmagem da região de interesse (rota de escape emergencial);

2. Randomizar cada conjunto de imagens de treinamento e validação três vezes e, a cada randomização, selecionar 80% dessas imagens para compor o conjunto de treinamento e 20% o de validação;
3. Redimensionar as imagens (de treino e validação) para um terço⁵ do tamanho original e extrair fragmentos de dimensões 256×256 , em passos de tamanho 256;
4. Iniciar o processo sequencial e iterativo de treinamento e validação, onde cada modelo será treinado sobre os subconjuntos de fragmentos de imagens de treino e validação definidos nas etapas 1 à 3, onde o conjunto de validação será utilizado apenas para determinar o melhor momento para armazenar em disco os parâmetros otimizados;
5. Após o término do treinamento, aplicar o modelo recém treinado sobre o conjunto de teste gerado na etapa 1. Observando que nessa etapa, as imagens de teste deverão ser apenas redimensionadas, sem a necessidade de extração de fragmentos, para um terço do seu tamanho original antes de serem submetidas ao processo de detecção de mudanças;
6. Comparar as médias dos resultados quantitativos (Medida-F, PWC, IOU) obtidas pelos diferentes métodos treinados.

⁵ Como as imagens possuem ultra definição, ao serem redimensionada para um terço do tamanho original se tornam imagens de alta definição (1280×720).

5 Análise de Resultados

Nesse capítulo faz-se uma análise do desempenho obtido por diferentes configurações dos métodos de detecção de mudanças apresentados nas Subseções 3.1.1 à 3.2.1. Além disso, as melhores configurações obtidas são comparadas com as técnicas correlatas descritas no Capítulo 3. Para fins de organização, as análises foram subdivididas nos grupos que tratam dos (a) resultados dos estudos de configuração, Seção 5.1, e dos (b) resultados dos estudos de comparação, Seção 5.2.

5.1 Resultados dos Estudos de Configuração

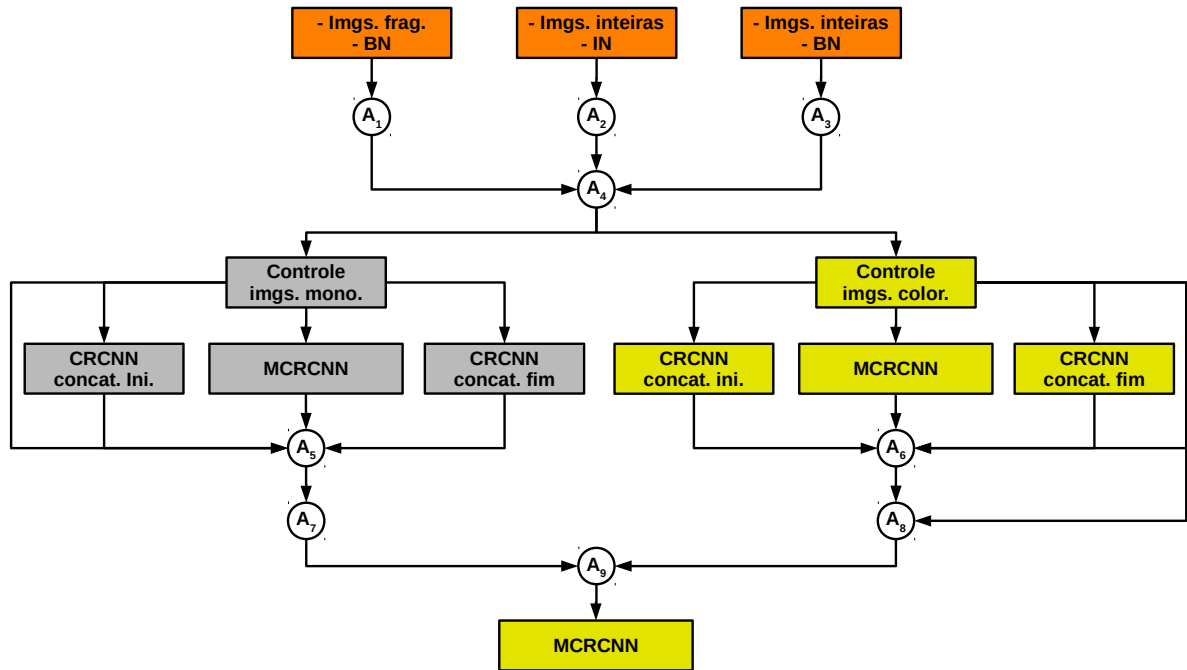
Nessa seção, orientando-se pelos protocolos pré-estabelecidos na Subseção 4.4.1, são apresentados os resultados quantitativos obtidos por meio de variações de estratégias de treinamento e de configuração dos métodos CRCNN, MCRCNN e TITAN analisados. Tais resultados, a serem apresentados respectivamente pelas Subseções 5.1.1 e 5.1.2, têm caráter eliminatório, sendo que além de contribuir para verificação da nossa hipótese, vide Capítulo 3, objetivam filtrar apenas as melhores configurações e métodos analisados.

5.1.1 Resultados de Configuração Residual Implícita

Diante das inúmeras possibilidades de configuração dos métodos CRCNN e MCRCNN, optamos por dar um enfoque maior a algumas das características mais proeminentes também encontradas nos métodos de detecção de mudança correlatos, vide Capítulo 3, sendo elas (i) cromaticidade da imagem alvo da detecção, (ii) treinamento utilizando imagens completas ou fragmentos, (iii) métodos de normalização de camadas (BN, IN, IBN e IIN) e (iv) local de concatenação do mapa residual nas redes CRCNN e MCRCNN.

O processo de análise de configurações do mecanismo de atenção residual implícita proposto segue estritamente o protocolo de treinamento definido na Subseção 4.4.1, para base de dados CD2014. A Figura 39 ilustra o processo de escolha de configurações, que se subdivide nas etapas:

Figura 39 – Etapas de análise de configurações, sendo que cada A_n , para $n \in [1, 9]$, representa uma etapa de comparação executada.



Fonte: Elaborada pelo autor.

- **Etapa A_1 :** Etapa de análise que considera o treinamento de rede SCNN, vide Subseção 3.1.1, onde são analisadas as seguintes configurações:
 - $S_g_nbtm128$: Imagem de entrada monocromática, treinamento utilizando fragmentos de dimensão 35×35 e blocos de tamanho 128, não utiliza camadas de normalização;
 - $S_g_b_nbtm128$: Mantém as mesmas configurações de $S_g_nbtm128$ exceto pela utilização de camadas BN na rede;
 - $S_g_ib_nbtm128$: Mantém as mesmas configurações de $S_g_nbtm128$ exceto pela utilização de camadas IBN na rede;
 - $S_c_nbtm128$: Mantém as mesmas configurações de $S_g_nbtm128$ exceto pela utilização de imagens cromáticas durante o treinamento e avaliação;
 - $S_c_b_nbtm128$: Mantém as mesmas configurações de $S_c_nbtm128$ exceto pela utilização de camadas BN na rede;
 - $S_c_ib_nbtm128$: Mantém as mesmas configurações de $S_c_nbtm128$ exceto pela utilização de camadas de IBN na rede.

Tabela 1 – Resultados quantitativos médios para Etapa A₁ de análise.

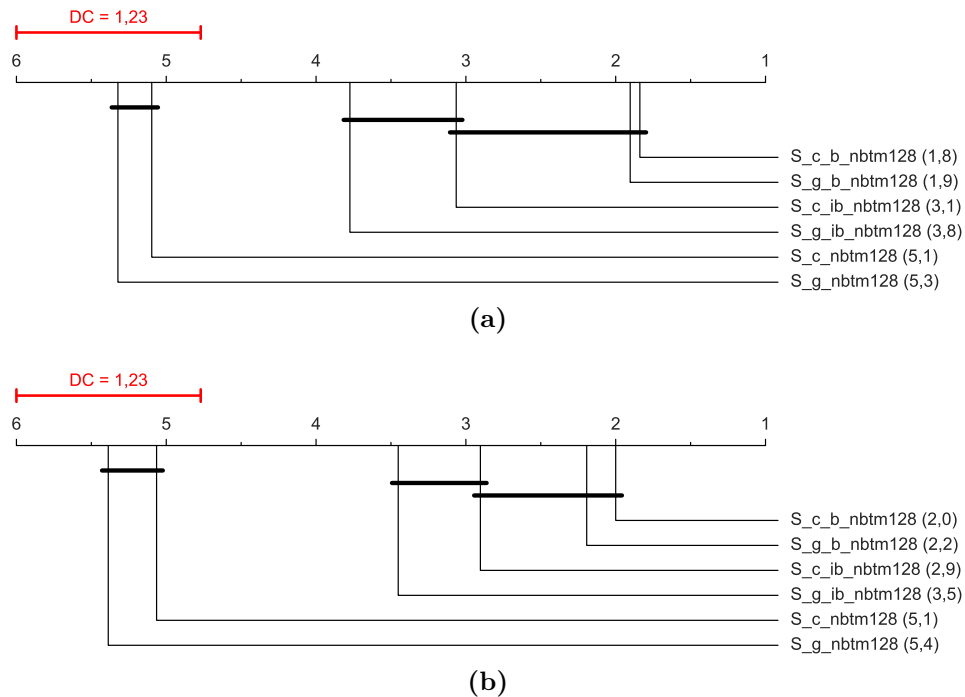
Métodos	Medida-F	IOU	PWC	PR	ROC
S_g_nbtm128	0,5022	0,4142	17,6533	0,4794	0,7243
S_g_b_nbtm128	0,8357	0,7433	0,8839	0,8618	0,9471
S_g_ib_nbtm128	0,7300	0,6220	2,7199	0,7239	0,9261
S_c_nbtm128	0,5077	0,4230	18,3014	0,4959	0,7272
S_c_b_nbtm128	0,8345	0,7432	0,8554	0,8529	0,9541
S_c_ib_nbtm128	0,7772	0,6711	1,5046	0,7723	0,9433

Fonte: Elaborada pelo autor.

Com base nos resultados da Tabela 1, para um valor de significância $\rho = 0,1$, é possível dizer, num primeiro momento, que existem diferenças significantes entre as configurações analisadas em A₁, tendo em vista que após aplicação do teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F, PWC, PR e ROC foram obtidos os valores de significância $\rho_F = 1,32e^{-32}$, $\rho_F = 1,32e^{-32}$, $\rho_F = 5,56e^{-28}$, $\rho_F = 8,13e^{-33}$ e $\rho_F = 3,01e^{-27}$, respectivamente.

Figura 40 – Ranqueamento de configurações para etapa A₁.

Consideram-se as medidas (a) PR e (b) ROC.

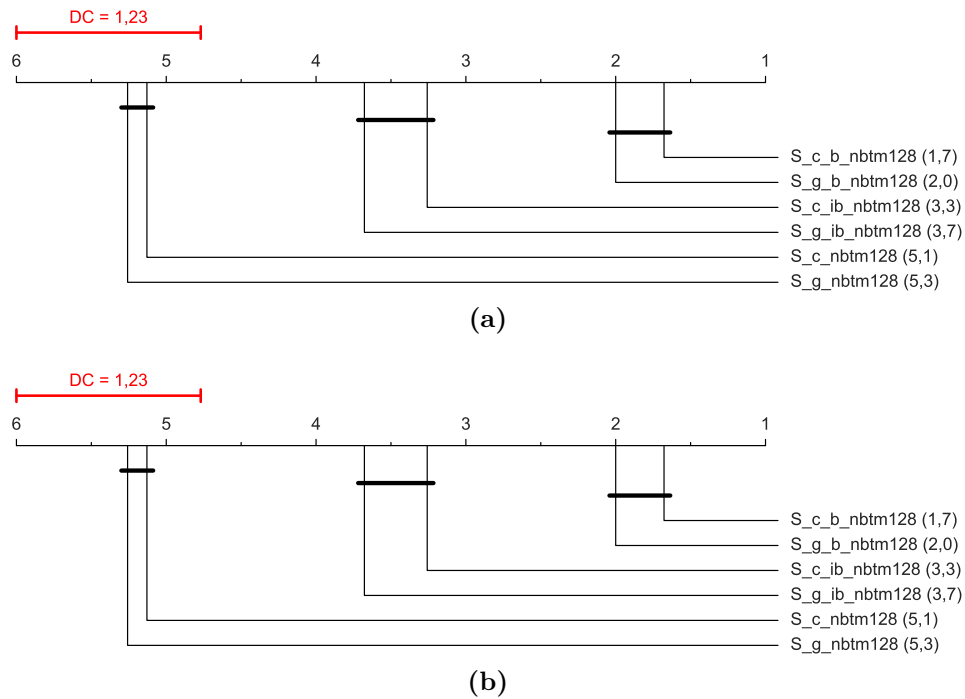


Fonte: Elaborada pelo autor.

Em um segundo momento, para um valor de significância de $\rho = 0,1$, o teste estatístico post-hoc de Nemenyi (1963) mostra, de acordo com a Figura 40, que as

configurações melhor ranqueadas para as medidas PR e ROC foram S_c_b_nbtm128, S_g_b_nbtm128 e S_c_ib_nbtm128. A Figura 40 ainda mostra que, apesar de não ser possível concluir quantitativamente qual configuração dentre S_c_b_nbtm128, S_g_b_nbtm128 e S_c_ib_nbtm128 desempenhou melhor, é possível dizer que elas foram quantitativamente melhores que S_c_nbtm128 e S_g_nbtm128. Além disso, a Figura 40 também mostra que S_c_b_nbtm128 e S_g_b_nbtm128 tiveram um melhor desempenho em comparação com S_g_ib_nbtm128.

Figura 41 – Ranqueamento de configurações para etapa A₁.
Consideram-se as medidas (a) IOU e (b) Medida-F.

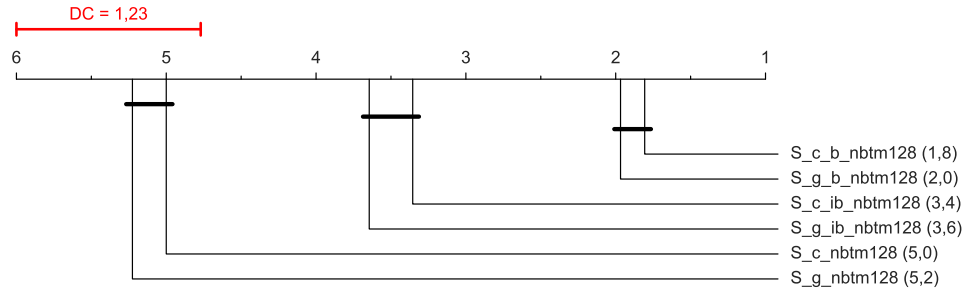


Fonte: Elaborada pelo autor.

De acordo com a Figura 41 as configurações que aparecem no grupo de técnicas melhor ranqueadas para as medidas IOU e Medida-F são S_c_b_nbtm128 e S_g_b_nbtm128. A Figura 41 ainda mostra que apesar de não ser possível concluir quantitativamente se S_c_b_nbtm128 desempenhou melhor que S_g_b_nbtm128, é possível dizer que ambas superaram S_c_nbtm128, S_g_nbtm128, S_g_ib_nbtm128 e S_c_ib_nbtm128. A Figura 42, assim como a Figura 41, também indica que S_c_b_nbtm128 e S_g_b_nbtm128 foram quantitativamente melhores que as demais configurações analisadas. Diante dos resultados obtidos selecionam-se as

configurações $S_c_b_nbtm128$ e $S_g_b_nbtm128$ como sendo as de melhor desempenho após execução da etapa A_1 de análises.

Figura 42 – Ranqueamento de configurações para etapa A_1 .
Considera-se a medida PWC.



Fonte: Elaborada pelo autor.

- **Etapa A_2 :** Etapa de análise que considera o treinamento da rede SCNN, onde comparam-se as seguintes configurações:

- S_g_nbt1 : Imagem de entrada monocromática, não fragmentada durante o treinamento, e redimensionada para 224×224 . Treinamento sequencial (uma imagem a cada iteração de otimização) e arquitetura de rede sem camadas de normalização;
- $S_g_b_nbt1$: Mantém as mesmas configurações de S_g_nbt1 exceto pela utilização de camadas IN;
- $S_g_ib_nbt1$: Mantém as mesmas configurações de S_g_nbt1 exceto pela utilização de camadas IIN;
- S_c_nbt1 : Mantém as mesmas configurações de S_g_nbt1 exceto pela utilização de imagens cromáticas durante o treinamento e avaliação;
- $S_c_b_nbt1$: Mantém as mesmas configurações de S_c_nbt1 exceto pela utilização de camadas IN;
- $S_c_ib_nbt1$: Mantém as mesmas configurações de S_c_nbt1 exceto pela utilização de camadas IIN.

Com base nos resultados da Tabela 2, para um valor de significância $\rho = 0,1$, é possível dizer, num primeiro momento, que existem diferenças significativas entre as

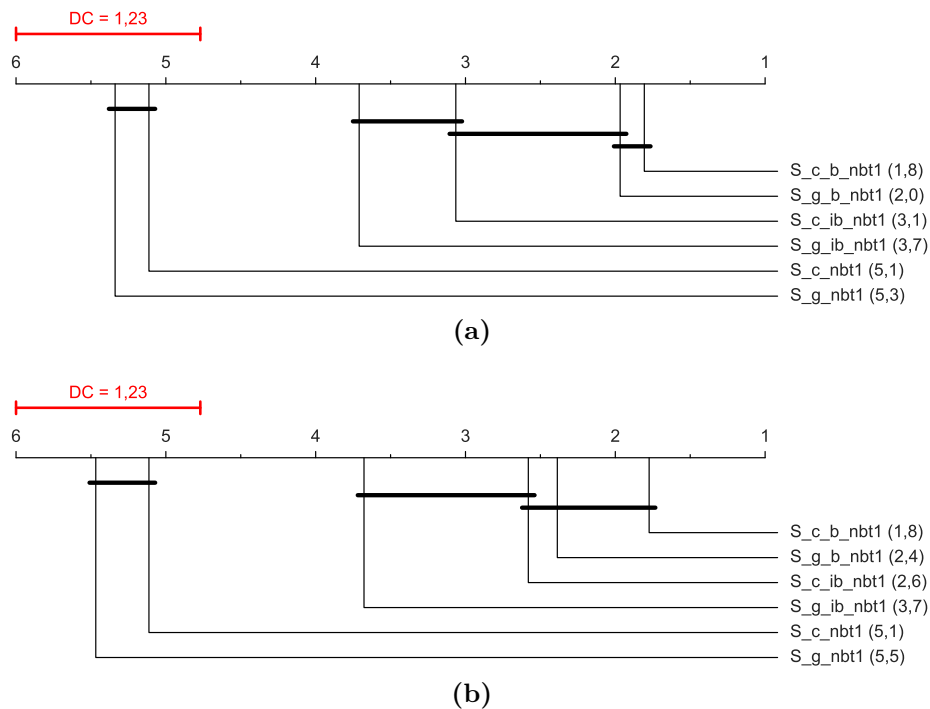
Tabela 2 – Resultados quantitativos médios para Etapa A₂ de análise.

Métodos	Medida-F	IOU	PWC	PR	ROC
S_g_nbt1	0,4950	0,4414	26,8875	0,4202	0,6446
S_g_b_nbt1	0,8899	0,8194	0,5853	0,8772	0,9529
S_g_ib_nbt1	0,8638	0,7862	0,7166	0,7559	0,9451
S_c_nbt1	0,5070	0,4382	24,8642	0,4333	0,6696
S_c_b_nbt1	0,8979	0,8292	0,5572	0,8673	0,9575
S_c_ib_nbt1	0,8613	0,7882	0,7116	0,7821	0,9464

Fonte: Elaborada pelo autor.

configurações analisadas em A₂, tendo em vista que após aplicação do teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F, PWC, PR e ROC foram obtidos os valores de significância $\rho_F = 1,12e^{-44}$, $\rho_F = 2,31e^{-46}$, $\rho_F = 6,37e^{-46}$, $\rho_F = 1,10e^{-32}$ e $\rho_F = 2,06e^{-36}$, respectivamente.

Figura 43 – Ranqueamento de configurações para etapa A₂.
Consideram-se as medidas (a) PR e (b) ROC.

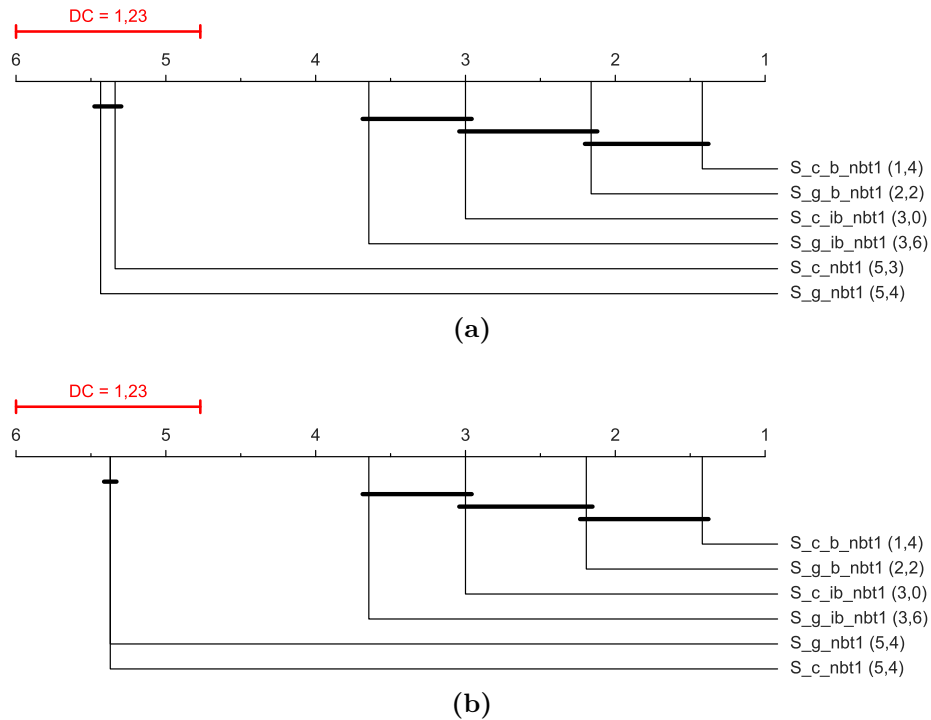


Fonte: Elaborada pelo autor.

Em um segundo momento, utilizando um valor de significância $\rho = 0,1$, o teste estatístico post-hoc de Nemenyi (1963) mostra que, de acordo com a Figura 43, as configurações que aparecem no grupo de técnicas melhor ranqueadas, considerando-se

ambas as medidas PR e ROC, são apenas $S_c_b_nbt1$ e $S_g_b_nbt1$, sendo que $S_c_ib_nbt1$ aparece entre as configurações melhor ranqueadas apenas considerando-se a métrica ROC. A Figura 43 ainda mostra que, apesar de não ser possível concluir quantitativamente qual configuração dentre $S_c_b_nbt1$, $S_g_b_nbt1$ e $S_c_ib_nbt1$ desempenhou melhor, é possível dizer que tais configurações foram quantitativamente melhores que as configurações S_c_nbt1 e S_g_nbt1 . Além disso, a Figura 43 também mostra que $S_c_b_nbt1$ e $S_g_b_nbt1$ tiveram um melhor desempenho em comparação com $S_g_ib_nbt1$.

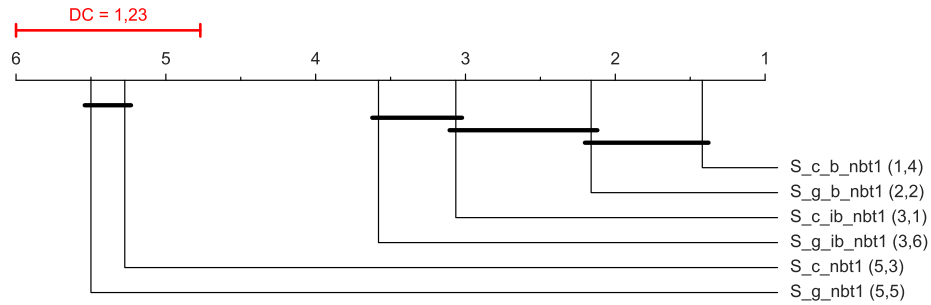
Figura 44 – Ranqueamento de configurações para etapa A_2 .
Consideram-se as medidas (a) IOU e (b) Medida-F.



Fonte: Elaborada pelo autor.

De acordo com a Figura 44 as configurações melhor ranqueadas para as medidas IOU e Medida-F foram $S_c_b_nbt1$ e $S_g_b_nbt1$. A Figura 44 ainda mostra que apesar de não ser possível concluir quantitativamente se $S_c_b_nbt1$ desempenhou melhor que $S_g_b_nbt1$, é possível dizer que ambas as configurações foram quantitativamente melhores que S_c_nbt1 , S_g_nbt1 e $S_g_ib_nbt1$.

Figura 45 – Ranqueamento de configurações para etapa A_2 .
Considera-se a medida PWC.



Fonte: Elaborada pelo autor.

A Figura 45 confirma que $S_c_b_nbt1$ e $S_g_b_nbt1$ foram quantitativamente melhores que as configurações S_c_nbt1 , S_g_nbt1 e $S_g_ib_nbt1$, levando-se em consideração a medida PWC.

- **Etapa A_3 :** Etapa de análise que considera o treinamento da rede SCNN comparando-se as seguintes configurações:
 - S_g_nbtm8 : Imagem de entrada monocromática, se ser fragmentada, redimensionada para 224×224 , treinamento com blocos de tamanho 8 e não utiliza camadas de normalização;
 - $S_g_b_nbtm8$: Mantém as mesmas configurações de S_g_nbtm8 exceto pela utilização de camadas BN;
 - $S_g_ib_nbtm8$: Mantém as mesmas configurações de S_g_nbtm8 exceto pela utilização de camadas IBN;
 - S_c_nbtm8 : Mantém as mesmas configurações de S_g_nbtm8 exceto pela utilização de imagens cromáticas durante o treinamento e detecção de mudanças;
 - $S_c_b_nbtm8$: Mantém as mesmas configurações de S_c_nbtm8 exceto pela utilização de camadas BN;
 - $S_c_ib_nbtm8$: Mantém as mesmas configurações de S_c_nbtm8 exceto pela utilização de camadas IBN.

Com base nos resultados da Tabela 3, para um valor de significância $\rho = 0,1$, é possível dizer, num primeiro momento, que existem diferenças significantes entre as

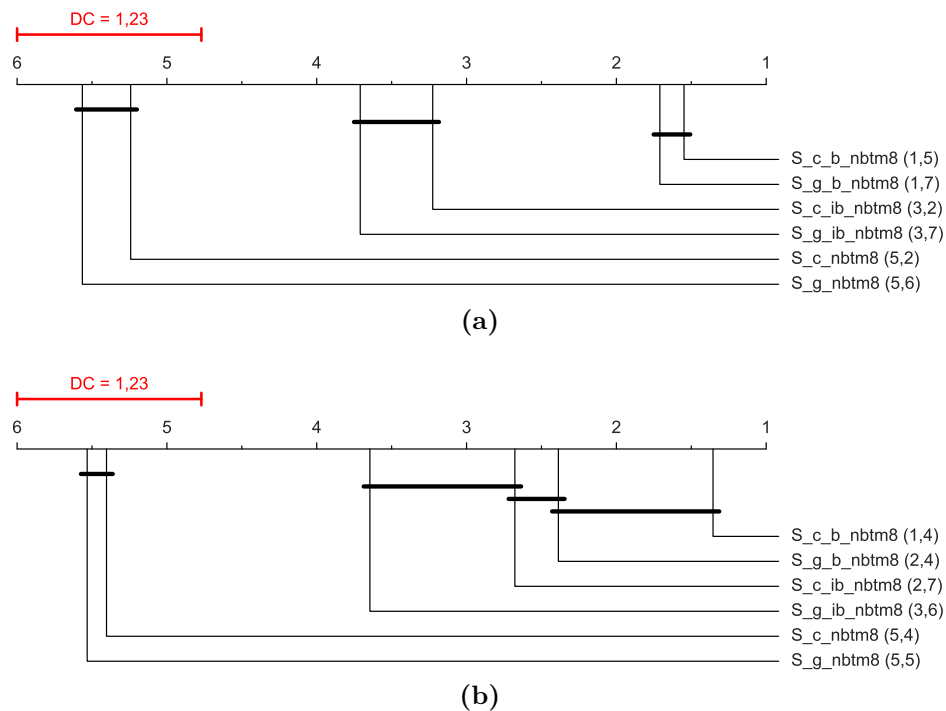
Tabela 3 – Resultados quantitativos médios para Etapa A₃ de análise.

Métodos	Medida-F	IOU	PWC	PR	ROC
S_g_nbtm8	0,4188	0,3267	7,1906	0,4098	0,7594
S_g_b_nbtm8	0,8592	0,7739	0,8267	0,8881	0,9568
S_g_ib_nbtm8	0,7649	0,6583	1,9419	0,7825	0,9339
S_c_nbtm8	0,4454	0,3630	8,2527	0,4363	0,7660
S_c_b_nbtm8	0,8734	0,7961	0,7350	0,8823	0,9640
S_c_ib_nbtm8	0,7824	0,6848	1,9189	0,7967	0,9385

Fonte: Elaborada pelo autor.

configurações analisadas em A₃, tendo em vista que após aplicação do teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F, PWC, PR e ROC foram obtidos os valores de significância $\rho_F = 4,91e^{-65}$, $\rho_F = 9,91e^{-65}$, $\rho_F = 1,01e^{-43}$, $\rho_F = 7,18e^{-55}$ e $\rho_F = 7,10e^{-53}$, respectivamente.

Figura 46 – Ranqueamento de configurações para etapa A₃.
Consideram-se as medidas (a) PR e (b) ROC.



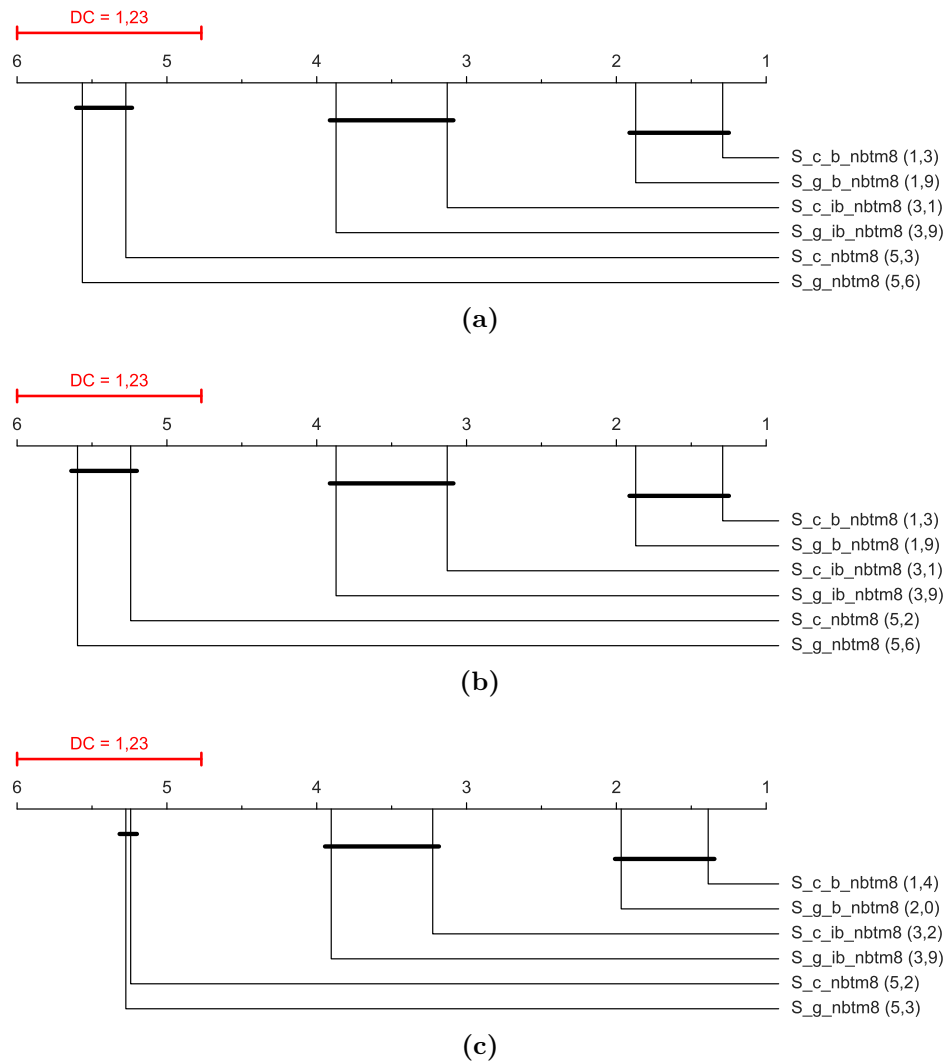
Fonte: Elaborada pelo autor.

Em um segundo momento, de maneira complementar ao teste de Friedman (1937) e utilizando um valor de significância $\rho = 0,1$, aplica-se o teste estatístico post-hoc de Nemenyi (1963). De acordo com a Figura 46, as configurações melhor

ranqueadas para as medidas PR e ROC foram S_c_b_nbtm8 e S_g_b_nbtm8. A Figura 46 ainda mostra que, apesar de não ser possível concluir quantitativamente se S_c_b_nbtm8 desempenhou melhor que S_g_b_nbtm8, ambas as configurações foram quantitativamente superiores à S_c_nbtm8, S_g_nbtm8 e S_g_ib_nbtm8.

Figura 47 – Ranqueamento de configurações para etapa A₃.

Consideram-se as medidas (a) IOU, (b) Medida-F e (c) PWC.



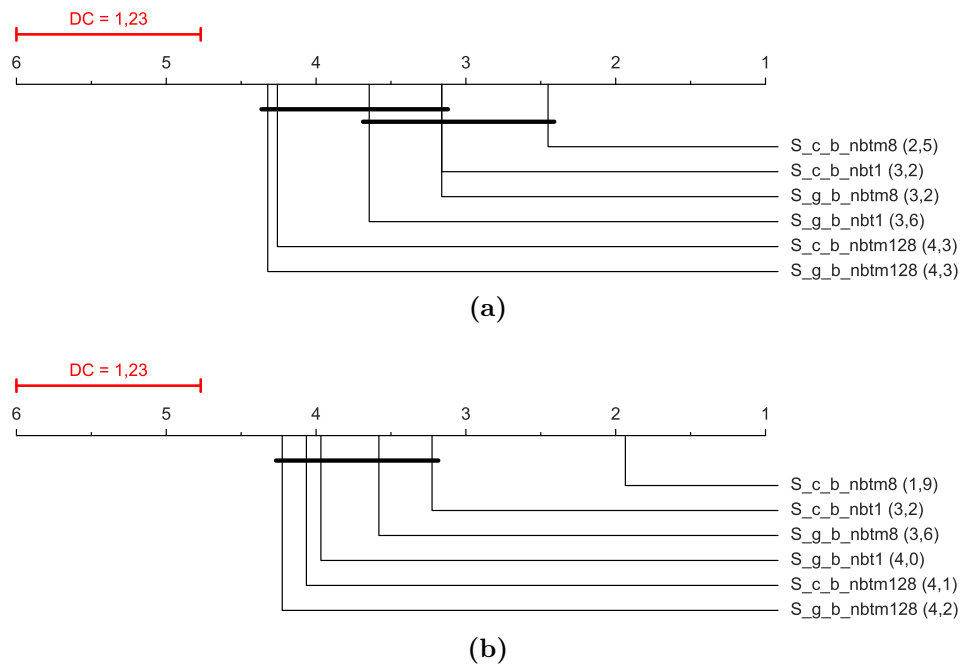
Fonte: Elaborada pelo autor.

De acordo com a Figura 47, as configurações que aparecem no grupo de técnicas melhor ranqueadas para as medidas IOU, Medida-F e PWC são S_c_b_nbtm8 e S_g_b_nbtm8. Além disso é possível notar na Figura 47 que apesar de não ser possível concluir quantitativamente se S_c_b_nbtm8 desempenhou melhor que

$S_g_b_nbtm8$, é possível dizer que as duas configurações superaram quantitativamente S_c_nbtm8 , S_g_nbtm8 , $S_g_ib_nbtm8$ e $S_c_ib_nbtm8$.

- **Etapa A₄**: Etapa de análise onde são comparadas entre si as melhores configurações obtidas após execução das etapas A₁, A₂ e A₃, sendo elas $S_c_b_nbtm128$, $S_g_b_nbtm128$, $S_c_b_nbt1$, $S_g_b_nbt1$, $S_c_b_nbtm8$ e $S_g_b_nbtm8$. Utiliza-se para tal o teste estatístico post-hoc de Nemenyi (1963), para um valor de significância $\rho = 0,1$.

Figura 48 – Ranqueamento de configurações para etapa A₄. Consideram-se as medidas (a) PR e (b) ROC.

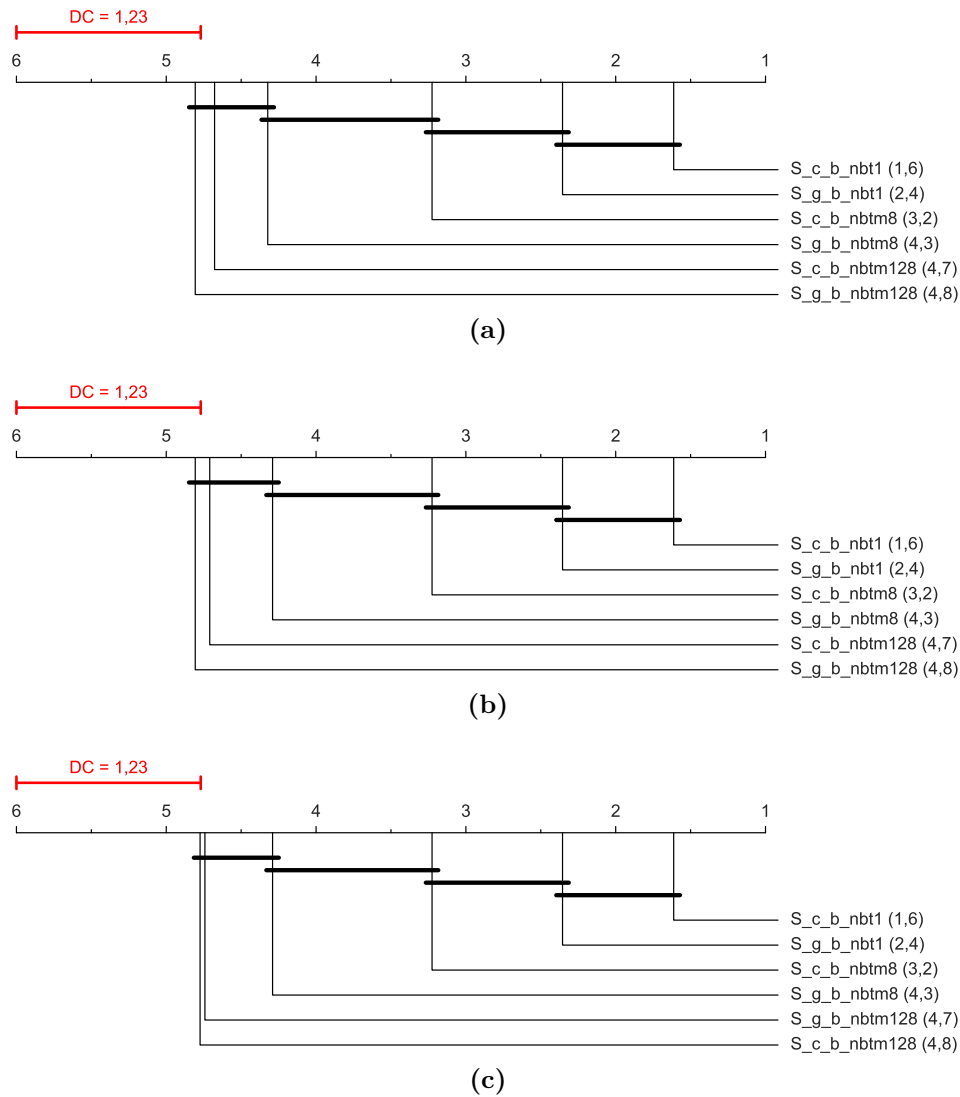


Fonte: Elaborada pelo autor.

De acordo com a Figura 48 (a) as configurações que aparecem no grupo de técnicas melhor ranqueadas, considerando-se a medida PR, são $S_c_b_nbt1$, $S_g_b_nbt1$, $S_c_b_nbtm8$ e $S_g_b_nbtm8$. Considerando-se a medida ROC, a Figura 48 (b) mostra que a técnica $S_c_b_nbtm8$ obteve o melhor desempenho em comparação com todas as demais configurações. Além disso, a Figura 48 (a) também mostra que, considerando a medida PR, a configuração $S_c_b_nbtm8$ obteve um melhor desempenho em comparação com as configurações $S_c_b_nbtm128$ e $S_g_b_nbtm128$.

De acordo com a Figura 49 as configurações que aparecem no grupo de técnicas melhor ranqueadas para as medidas IOU e Medida-F são S_c_b_nbt1 e S_g_b_nbt1. A Figura 49 ainda mostra que apesar de não ser possível concluir quantitativamente se S_c_b_nbt1 desempenhou melhor que S_g_b_nbt1, é possível dizer que ambas as configurações foram quantitativamente melhores que S_g_b_nbtm8, S_c_b_nbtm128 e S_g_b_nbtm128.

Figura 49 – Ranqueamento de configurações para etapa A₄.
Consideram-se as medidas (a) IOU, (b) Medida-F e (c) PWC.



Fonte: Elaborada pelo autor.

- **Etapa A₅:** Etapa de análise inicial de variações das redes multiestágio propostas, CRCNN e MCRCNN, onde comparam-se as seguintes configurações:

- C_sc_g_nbt1: Utilização de imagem de entrada monocromática, não fragmentada, e redimensionada para 224×224 . Treinamento sequencial (uma imagem a cada iteração de otimização), concatenação de mapa residual com a imagem de entrada da rede SCNN (mecanismo de atenção residual implícita) e utilização de camadas IN;
- C_ec_g_nbt1: Mantém as mesmas configurações de C_sc_g_nbt1 exceto pela concatenação do mapa residual ser feita com os mapas de característica da décima quinta camada da rede SCNN, vide Figura 27;
- C_ec_g_dl4_nbt1: Mantém as mesmas configurações de C_ec_g_nbt1 exceto pela aplicação do módulo RPM sobre o mapa residual antes de sua concatenação com os mapas de característica da décima quinta camada da rede SCNN, vide Figura 27.

Além disso, comparam-se as configurações multiestágio com a configuração S_g_b_nbt1, utilizada como técnica controle na verificação da superioridade das técnicas multiestágio.

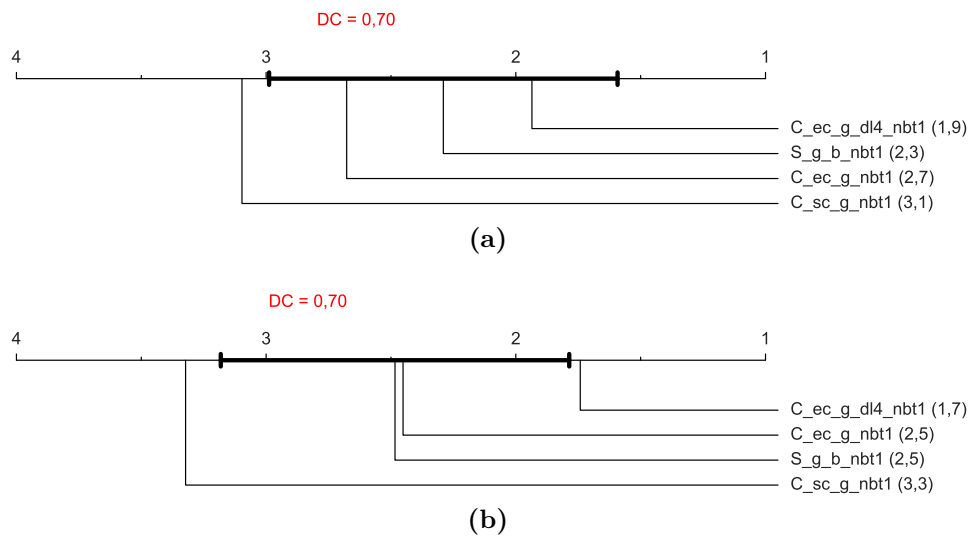
Tabela 4 – Resultados quantitativos médios para Etapa A₅ de análise.

Métodos	Medida-F	IOU	PWC	PR	ROC
C_sc_g_nbt1	0,8751	0,7975	0,6862	0,8619	0,9477
C_ec_g_nbt1	0,8838	0,8121	0,6434	0,8669	0,9527
C_ec_g_dl4_nbt1	0,8642	0,7902	0,7301	0,8621	0,9560

Fonte: Elaborada pelo autor.

Com base nos resultados da Tabela 4 e incluindo também os resultados da configuração S_g_b_nbt1 da Tabela 2, para um valor de significância $\rho = 0,1$, é possível dizer, num primeiro momento, que existem diferenças significantes entre as configurações analisadas em A₅, tendo em vista que após aplicação do teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F, PWC, PR e ROC foram obtidos os valores de significância $\rho_F = 5,16e^{-6}$, $\rho_F = 3,09e^{-6}$, $\rho_F = 9,19e^{-6}$, $\rho_F = 2,23e^{-3}$ e $\rho_F = 1,08e^{-5}$, respectivamente.

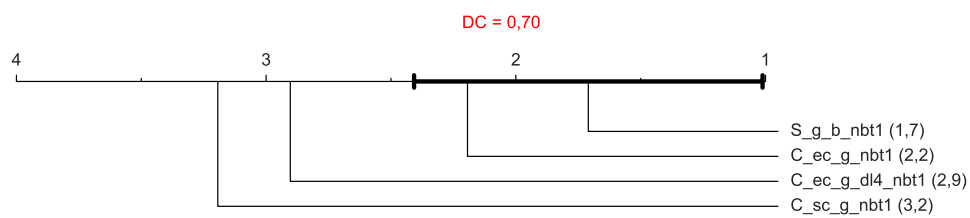
Figura 50 – Ranqueamento de configurações para etapa A₅.
Consideram-se as medidas (a) PR e (b) ROC.



Fonte: Elaborada pelo autor.

Em um segundo momento, utilizando um valor de significância $\rho = 0,1$, o teste estatístico post-hoc de Dunnett (1955) mostra que, de acordo com a Figura 50, a configuração C_sc_g_nbt1 exibiu resultados inferiores aos da técnica de controle S_g_b_nbt1 para ambas as medidas analisadas.

Figura 51 – Ranqueamento de configurações para etapa A₅.
Consideram-se as medidas IOU, Medida-F e PWC representadas pelo mesmo gráfico, tendo elas exibido resultados estatísticos equivalentes.



Fonte: Elaborada pelo autor.

A Figura 50 (a) também indica que a configuração C_ec_g_dl4_nbt1 não apresentou diferenças significativas em termos de PR com relação à configuração S_g_b_nbt1. Ainda sim, é possível observar na Figura 50 (b) que a configuração C_ec_g_dl4_nbt1 desempenhou melhor que S_g_b_nbt1 em termos de ROC. Já para a configuração

C_ec_g_nbt1, não se observam diferenças estatisticamente significativas com relação à S_g_b_nbt1 para ambas as medidas analisadas.

De acordo com a Figura 51, é possível dizer que as configurações multiestágio C_sc_g_nbt1 e C_ec_g_dl4_nbt1 tiveram um pior desempenho em comparação com a configuração de controle S_g_b_nbt1 para as medidas IOU, Medida-F e PWC analisadas. A Figura 51 ainda mostra que a comparação quantitativa entre as configurações C_ec_g_dl4_nbt1 e S_g_b_nbt1 foi inconclusiva.

Após execução da etapa A₅ de análises, selecionam-se as configurações S_g_b_nbt1 e C_ec_g_nbt1 como sendo as de melhor desempenho, tendo em vista que elas foram as únicas configurações melhor ranqueadas em quatro das cinco medidas analisadas.

- **Etapa A₆:** Etapa de análises de variações das redes multiestágio CRCNN e MCRCNN, semelhantes às da etapa A₅, exceto pela utilização de imagens de entrada cromáticas. Além disso, nessa etapa comparam-se as configurações multiestágio com a configuração S_c_b_nbt1, utilizada como técnica de controle.

Tabela 5 – Resultados quantitativos médios para Etapa A₆ de análise.

Métodos	Medida-F	IOU	PWC	PR	ROC
C_sc_c_nbt1	0,8830	0,8097	0,6261	0,8555	0,9544
C_ec_c_nbt1	0,8944	0,8248	0,5660	0,9039	0,9559
C_ec_c_dl4_nbt1	0,8921	0,8230	0,5827	0,8905	0,9567

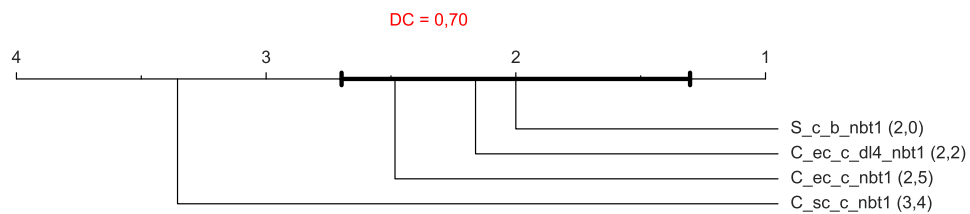
Fonte: Elaborada pelo autor.

Com base nos resultados da Tabela 5 e também nos resultados obtidos pela configuração S_c_b_nbt1 da Tabela 2, para um valor de significância $\rho = 0,1$, é possível dizer, num primeiro momento, que existem diferenças significantes entre as configurações analisadas em A₆, tendo em vista que após aplicação do teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F, PWC, PR e ROC foram obtidos os valores de significância $\rho_F = 1,13e^{-4}$, $\rho_F = 1,13e^{-4}$, $\rho_F = 6,33e^{-5}$, $\rho_F = 7,66e^{-8}$ e $\rho_F = 0,133$, respectivamente.

Em um segundo momento, utilizando um valor de significância $\rho = 0,1$, o teste estatístico post-hoc de Dunnett (1955) para todas as medidas exceto a ROC, cujo valor crítico retornado pelo teste de Friedman (1937) excede o limiar de $\rho = 0,1$, mostra

que, de acordo com a Figura 52, a configuração C_sc_c_nbt1 exibiu resultados inferiores aos da técnica de controle S_c_b_nbt1 para a medida PR.

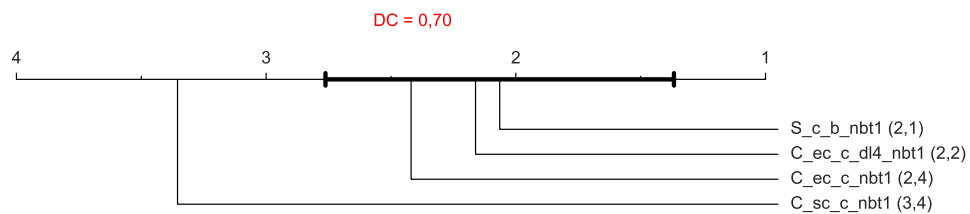
Figura 52 – Ranqueamento de configurações para etapa A₆.
Considera-se a medida PR.



Fonte: Elaborada pelo autor.

Paralelamente, nota-se também que as configuração C_ec_c_nbt1 e C_ec_c_dl4_nbt1 não exibiram diferenças significativas com relação à S_c_b_nbt1.

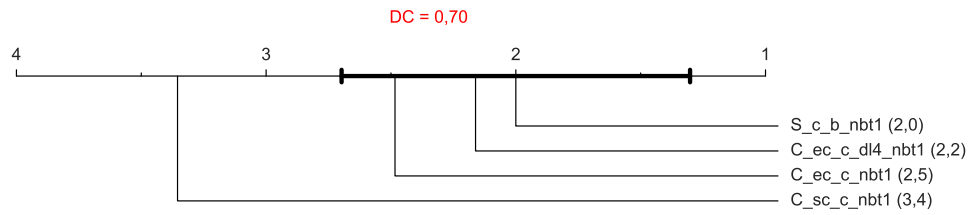
Figura 53 – Ranqueamento de configurações para etapa A₆.
Consideram-se as medidas IOU e Medida-F representadas pelo mesmo gráfico, tendo elas exibido resultados estatísticos equivalentes.



Fonte: Elaborada pelo autor.

A Figura 53 reforça a indicação de que a configuração em cascata C_sc_c_nbt1 obteve um pior desempenho em relação à S_c_b_nbt1 e também que as configurações C_ec_c_nbt1 e C_ec_c_dl4_nbt1 não exibiram diferenças significativas com relação à S_c_b_nbt1 levando-se em considerações as medidas IOU e Medida-F analisadas. A Figura 54 corrobora os resultados exibidos pelas Figuras 52 e 53. Por meio dos resultados obtidos segundo a medida PWC é possível dizer que as melhores configurações encontradas após análise da etapa A₆ foram S_c_b_nbt1, C_ec_c_nbt1 e C_ec_c_dl4_nbt1.

Figura 54 – Ranqueamento de configurações para etapa A₆.
Considera-se a medida PWC.



Fonte: Elaborada pelo autor.

- **Etapa A₇:** Etapa de análise onde são comparadas as melhores configurações obtidas após execução da etapa A₅, sendo elas S_g_b_nbt1 e C_ec_g_nbt1. Na tentativa de se obter resultados mais conclusivos e que permitissem definir qual das duas configurações obteve o melhor desempenho na tarefa de detecção de mudanças na base de dados CD2014, vide Seção 4.1. Na etapa A₇ de análises adotou-se o protocolo de treinamento de estudo comparativo descrito na Subseção 4.4.2.

Tabela 6 – Resultados quantitativos médios para Etapa A₇ de análise.

Métodos	Medida-F	IOU	PWC	PR	ROC
C_ec_g_nbt1	0,8983	0,8439	0,4961	0,8188	0,9887
S_g_b_nbt1	0,8959	0,8395	0,4386	0,5401	0,6344

Fonte: Elaborada pelo autor.

Com base nos resultados da Tabela 6, para um valor de significância $\rho = 0,1$, não é possível dizer qual configuração obteve o melhor desempenho. Isso porque o teste estatístico de Wilcoxon (1992) aplicado sobre as medidas IOU, Medida-F, PWC, PR e ROC retornou os seguintes valores de significância $\rho_W = 5,09e^{-1}$, $\rho_W = 6,36e^{-1}$, $\rho_W = 8,28e^{-1}$, $\rho_W = 6,87e^{-9}$ e $\rho_W = 1,44e^{-9}$, respectivamente. Dessa forma apenas as medidas PR e ROC não foram suficientes para se chegar a uma conclusão.

- **Etapa A₈:** Etapa de análise onde são comparadas as melhores configurações obtidas após execução da etapa A₆, sendo elas S_c_b_nbt1, C_ec_c_nbt1 e C_ec_c_dl4_nbt1, utilizando o protocolo de estudo comparativo como feito na

etapa A₇.

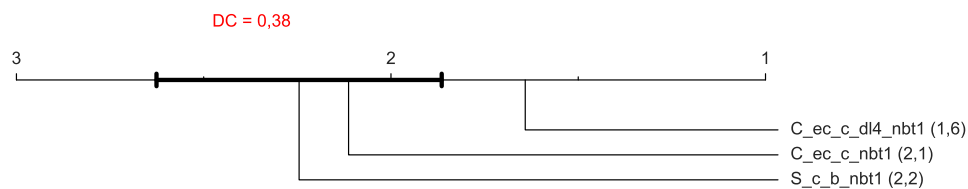
Tabela 7 – Resultados quantitativos médios para Etapa A₈ de análise.

Métodos	Medida-F	IOU	PWC	PR	ROC
C_ec_c_nbt1	0,8988	0,8467	0,3893	0,8119	0,9894
C_ec_c_dl4_nbt1	0,9080	0,8614	0,3596	0,8066	0,9892
S_c_b_nbt1	0,8909	0,8377	0,4408	0,8011	0,9893

Fonte: Elaborada pelo autor.

Com base nos resultados da Tabela 7, para um valor de significância $\rho = 0,1$, é possível dizer, num primeiro momento, que existem diferenças significativas entre as configurações analisadas em A₈, tendo em vista que após aplicação do teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F, PWC, PR e ROC foram obtidos os valores de significância $\rho_F = 3,60e^{-3}$, $\rho_F = 4,24e^{-3}$, $\rho_F = 4,79e^{-3}$, $\rho_F = 3,01e^{-1}$ e $\rho_F = 3,07e^{-1}$, respectivamente.

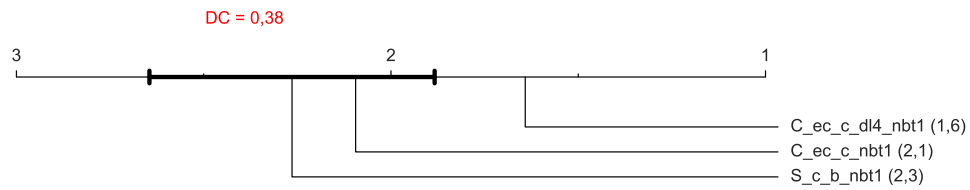
Figura 55 – Ranqueamento de configurações para etapa A₈. Consideram-se as medidas Medida-F e PWC representadas pelo mesmo gráfico, tendo elas exibido resultados estatísticos equivalentes.



Fonte: Elaborada pelo autor.

Em um segundo momento, utilizando um valor de significância $\rho = 0,1$, o teste estatístico post-hoc de Dunnett (1955) aplicado em todas as medidas, exceto em PR e ROC, cujos valores críticos retornados pelo teste de Friedman (1937) excedem o limiar de $\rho = 0,1$, indicou que, de acordo com a Figura 55, a configuração C_ec_c_dl4_nbt1 teve um melhor desempenho em comparação com a configuração de controle S_c_b_nbt1 nas medidas PWC e Medida-F analisadas, ainda que o resultado da comparação entre as configurações C_ec_c_nbt1 e S_c_b_nbt1 tenha sido inconclusivo.

Figura 56 – Ranqueamento de configurações para etapa A₈.
Considera-se a medida IOU.

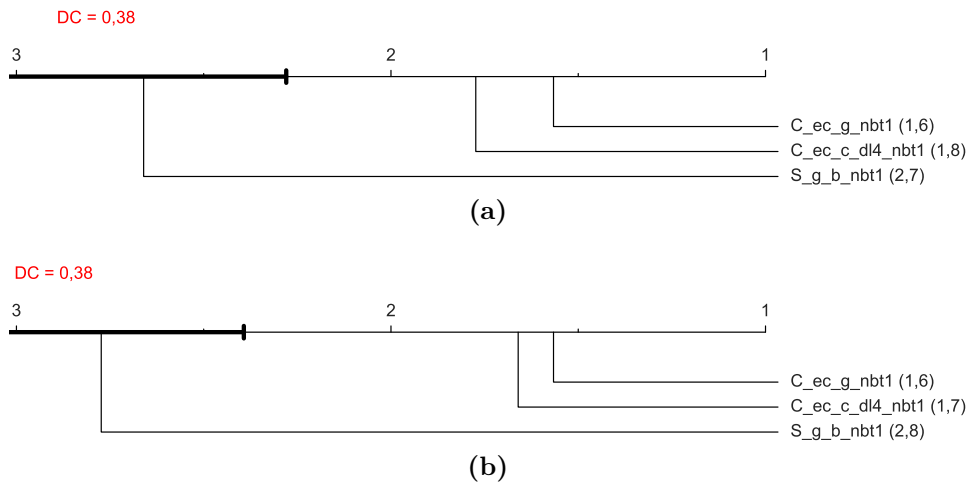


Fonte: Elaborada pelo autor.

A Figura 56 corrobora os resultados exibidos pela Figuras 55, ao se analisar a medida PWC, o que por sua vez torna possível dizer que a melhor configuração encontrada após análise da etapa A₈ foi C_ec_c_dl4_nbt1.

- **Etapa A₉:** Etapa de análise onde são comparadas as melhores configurações obtidas após execução das etapas A₇ e A₈. Utiliza-se para tal o teste estatístico post-hoc de Dunnett (1955), para um valor de significância $\rho = 0,1$

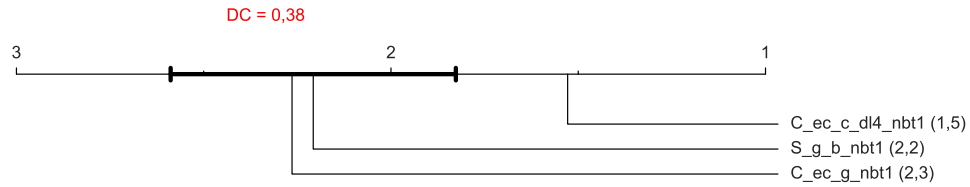
Figura 57 – Ranqueamento de configurações para etapa A₉.
Consideram-se as medidas (a) PR e (b) ROC.



Fonte: Elaborada pelo autor.

De acordo com a Figura 57, ambas as configurações multiestágio C_ec_g_nbt1 e C_ec_c_dl4_nbt1 exibiram resultados superiores aos da técnica de controle S_g_b_nbt1 para as medidas PR e ROC analisadas.

Figura 58 – Ranqueamento de configurações para etapa A_9 .
Consideram-se as medidas IOU, Medida-F e PWC representadas pelo mesmo gráfico, tendo elas exibido resultados estatísticos equivalentes.

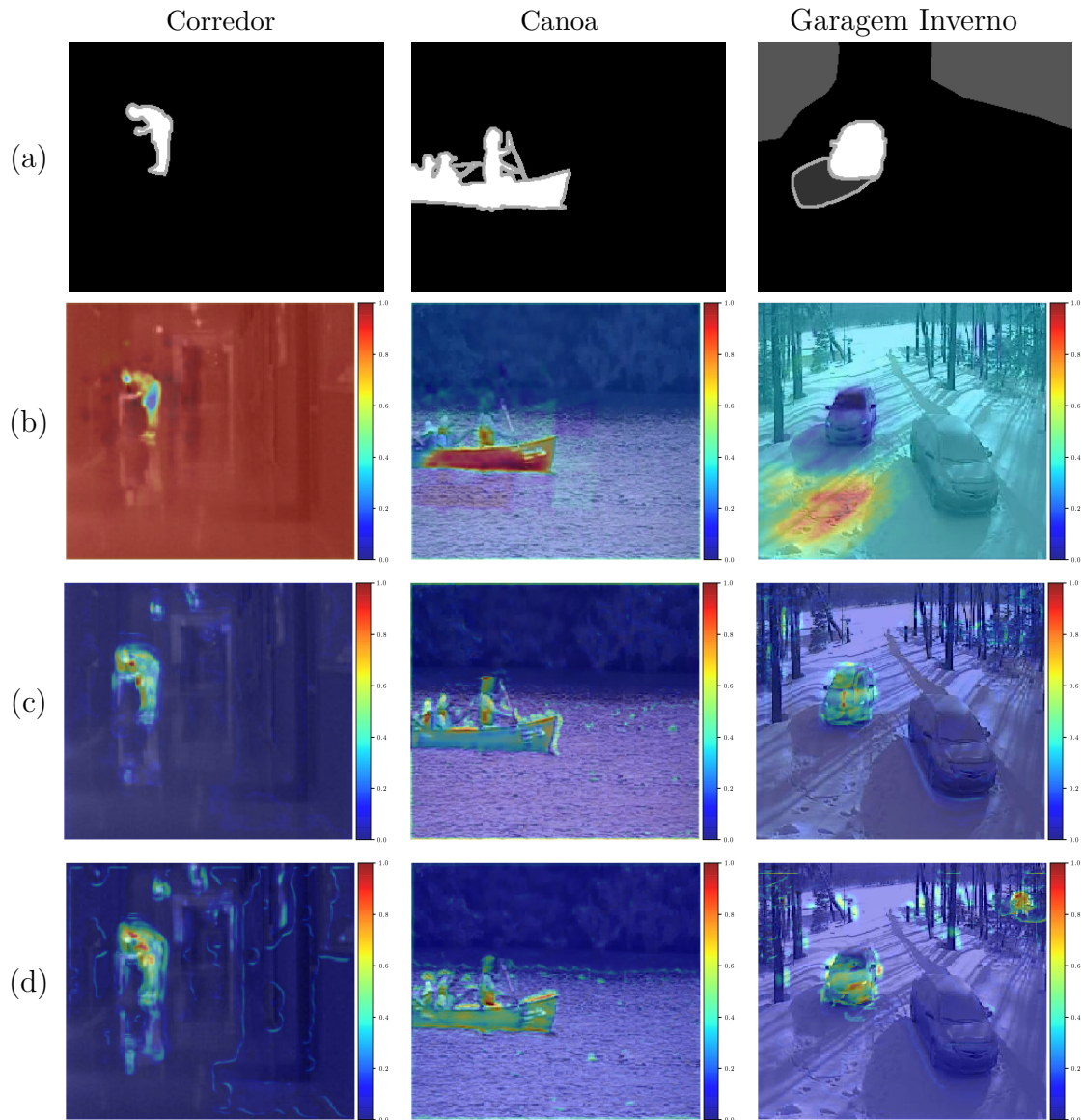


Fonte: Elaborada pelo autor.

Como mostra a Figura 58, a configuração multiestágio $C_ec_c_dl4_nbt1$ desempenhou melhor do que a configuração de controle $S_g_b_nbt1$ considerando as medidas IOU, Medida-F e PWC analisadas. A Figura 58 ainda mostra que a comparação quantitativa entre as configurações $C_ec_g_nbt1$ e $S_g_b_nbt1$ foi inconclusiva. Após execução da etapa A_9 de análises e de acordo também com todas as outras etapas de análise realizadas, é possível dizer que a configuração $C_ec_c_dl4_nbt1$, também chamada de MCRCNN, obteve o melhor desempenho em termos quantitativos na tarefa de detecção de mudanças.

Em termos qualitativos, a Figura 59(b) exhibe as representações visuais, normalizadas no intervalo $[0, 1]$, dos mapas residuais de atenção da rede MCRCNN, levando-se em consideração os vídeos Corredor, Canoa e Garagem Inverno. Subdivididos em ao menos duas regiões de cores equidistantes (tons de vermelho e roxo), esses mapas parecem orientar a rede MCRCNN na busca por separar os conteúdos de *foreground* e *background* das imagens de entrada. De maneira contraintuitiva, apesar da evidência de separação observada na Figura 59(b), as regiões destacadas nos mapas residuais não parecem condizer necessariamente com as regiões anotadas nas imagens da Figura 59(a). Isso pode ser verificado nos mapas residuais correspondentes aos exemplos dos vídeos Corredor e Garagem Inverno, onde as regiões em tons de vermelho e ciano correspondem ao conteúdo de *background* das imagens.

Diante de tal constatação, a funcionalidade plausível dos mapas residuais seria a de limitadores de ativações locais. Isso quer dizer que os mapas residuais teriam a capacidade de restringir as ativações dos mapas de características da penúltima camada da rede de segmentação (décima quinta camada da rede SCNN, vide Figura 27) do modelo

Figura 59 – Efeitos da atenção residual explícita.

Fonte: Imagens de (a) retiradas de (WANG et al., 2012). O restante elaborado pelo autor.

Onde têm-se em (a) Anotações manuais, (b) Atenção Residual Implícita e ativações máximas em profundidade das configurações (c) `C_ec_c_dl4_nbt1` e (d) `S_g_b_nbt1`.

multiestágio MCRCNN em regiões de *background* e liberar as ativações para regiões de *foreground*, induzindo de maneira implícita um processo de atenção espacial. Isso se verifica por meio das Figuras 59(c) e 59(d), que representam as ativações máximas locais calculadas na dimensão canal dos mapas de características das penúltimas camadas das configurações `C_ec_c_dl4_nbt1` e `S_g_b_nbt1`, respectivamente. Nos exemplos da Figura 59(c) é possível observar que as ativações da rede `C_ec_c_dl4_nbt1` são mais coerentes com as

respectivas anotações nas Figuras 59(a) em comparação com os exemplos da Figura 59(d).

5.1.2 Resultados de Configuração Residual Explícita

Tendo em vista que o método TITAN compartilha muitos dos padrões de configurações já analisados na construção das técnicas correlatas DMINet, RSCDNet, STANet e FLS, tais como quantidades e estratégias envolvendo aplicação de convoluções dilatadas, treinamento por blocos, utilização de imagens cromáticas e de camadas de normalização, optamos por dar enfoque à análise dos resultados obtidos pela rede TITAN quando configurada para utilizar ou não o mecanismo proposto de atenção residual explícita, sintetizado no módulo TCAM, vide Subseção 3.2.1. Desta forma as configurações analisadas foram TITAN_b, onde os mapas de características $\mathbf{F}_1$ e $\mathbf{F}_2$ da Figura 29 são concatenados, na dimensão de canais, com os mapas de características $\mathbf{D}_1$ e $\mathbf{D}_2$, respectivamente, e TITAN. Sendo TITAN_b a versão da rede TITAN que não utiliza mecanismo de atenção residual explícita.

Tabela 8 – Resultados quantitativos médios para análise de configurações TITAN e TITAN_b.

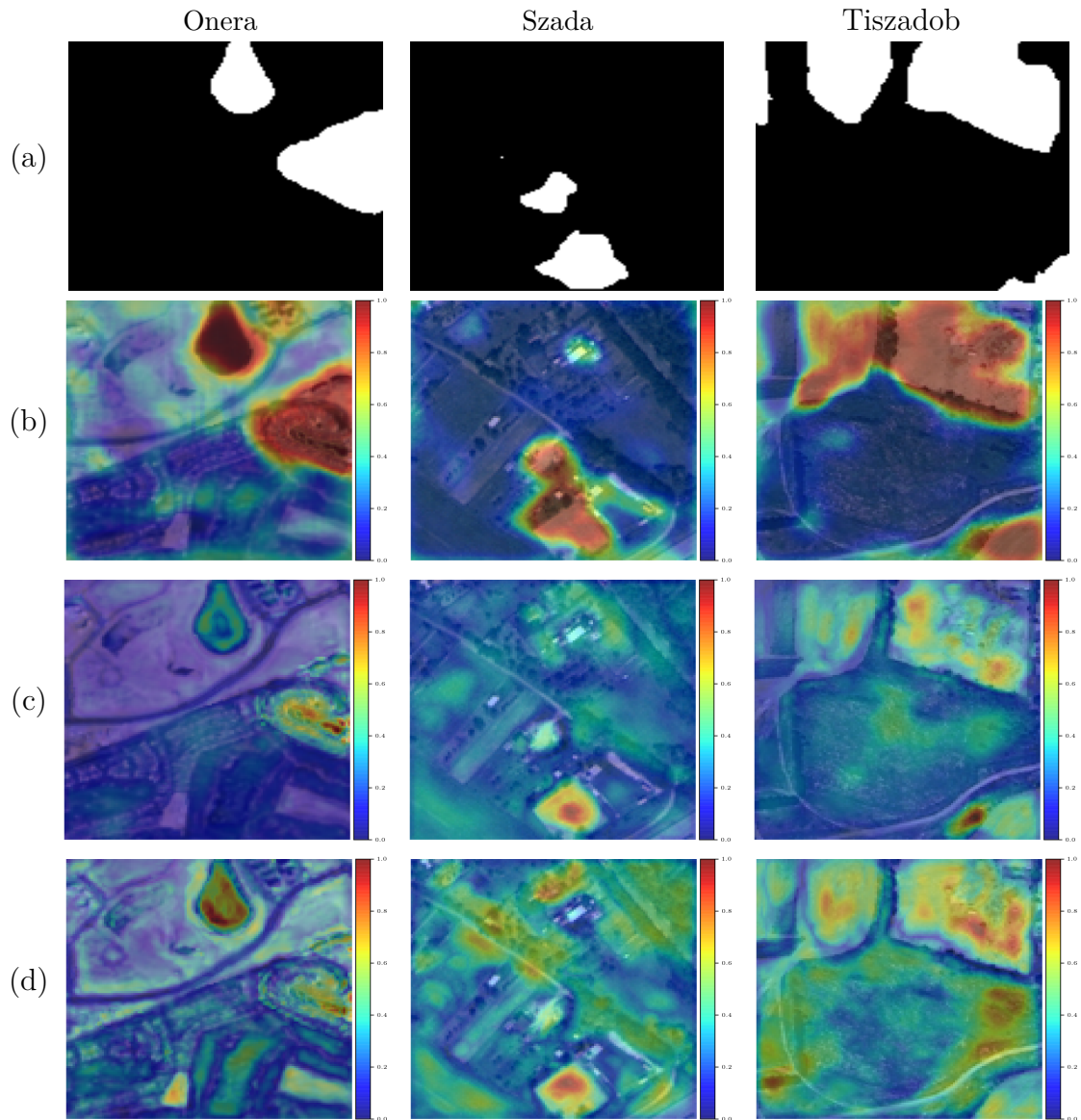
Métodos	Medida-F	IOU	PWC
TITAN	0,8933	0,8123	4,7957
TITAN_b	0,8846	0,7997	5,2341

Fonte: Elaborada pelo autor.

Com base nos resultados da Tabela 8, para um valor de significância $\rho = 0,05$, é possível dizer que a configuração TITAN obteve um melhor desempenho que TITAN_b. Isso porque o teste estatístico de Wilcoxon (1992) aplicado sobre as medidas IOU, Medida-F e PWC retorna os valores de significância $\rho_W = 9,8e^{-3}$, $\rho_W = 5,8e^{-3}$ e $\rho_W = 9,8e^{-3}$, respectivamente.

Na Figura 60(b) é possível visualizar o funcionamento do limitador de ativações $\mathbf{G}$ do módulo TCAM1, vide Subseção 3.2.1. Para diferentes exemplos de imagens aéreas de sensoriamento remoto, conjuntos **Onera**, **Szada** e **Tisza Dob**, submetidas para a rede TITAN já treinada, nota-se que o limitador torna-se mais ativo (cor avermelhada) em regiões correspondentes a mudanças, como mostra a Figura 60(a).

De uma maneira mais explícita, ao contrário do comportamento observado nos mapas residuais da rede multiestágio MCRCNN, Figura 59(b), os limitadores de TCAM1

Figura 60 – Efeitos da atenção residual implícita de primeiro nível.

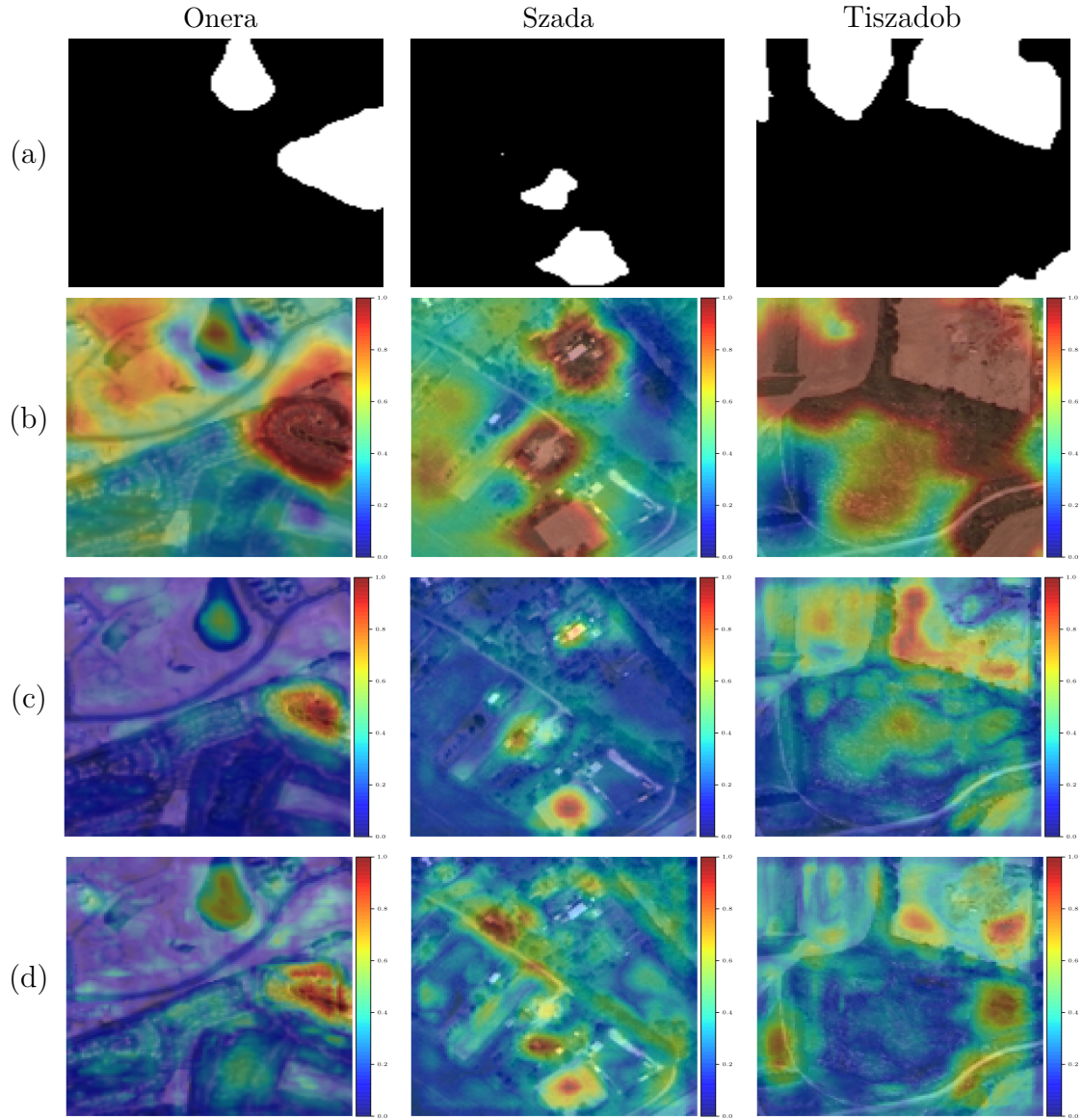
Fonte: Imagens de (a) retiradas de (DAUDT et al., 2018a) e (BENEDEK; SZIRÁNYI, 2013). O restante elaborado pelo autor.

Onde têm-se em (a) Anotações manuais, (b) Mapa limitador $\mathbf{G}$ de módulo TCAM1 e ativações máximas de decodificação de nível 1 para configurações (c) TITAN e (d) TITAN_b.

parecem estimular a ativação das camadas decodificadoras nas regiões de mudança. Isso se verifica por meio das Figuras 60(c) e 60(d), que representam as ativações máximas locais dos mapas de características gerados pelas camadas decodificadoras produzidas pela convolução das fusões $\mathbf{F}_1$ e $\mathbf{D}_1$, vide Figura 29, sendo os valores máximos extraídos da dimensão canal. Nos exemplos da Figura 60(c) é possível ainda observar que as

ativações da rede TITAN foram mais intensificadas nas respectivas regiões correspondentes a mudanças anotadas nas Figuras 60(a) em comparação com as correspondentes ativações na Figura 60(d) para a rede TITAN_b.

Figura 61 – Efeitos da atenção residual implícita de segundo nível.



Fonte: Imagens de (a) retiradas de (DAUDT et al., 2018a) e (BENEDEK; SZIRÁNYI, 2013).
O restante elaborado pelo autor.

Onde têm-se em (a) Anotações manuais, (b) Mapa limitador $\mathbf{G}$ de módulo TCAM2 e ativações máximas de decodificação de nível 2 para configurações (c) TITAN e (d) TITAN_b.

Na Figura 61(b) é possível visualizar o funcionamento do limitador de ativações

G do módulo TCAM2, vide Subseção 3.2.1. Para os mesmos exemplos de imagens da Figura 60, nota-se que a região ativa do limitador foi expandida, devido ao aumento do campo visual da rede TITAN em um nível hierárquico mais profundo. Corroborando as observações feitas sobre a Figura 60, os limitadores de TCAM2 também parecem estimular a ativação das camadas decodificadoras nas regiões de mudanças. Isso se verifica por meio das Figuras 61(c) e 61(d), que representam as ativações máximas locais dos mapas de características gerados pelas camadas decodificadoras produzidas pela convolução das fusões $\mathbf{F}_2$ e $\mathbf{D}_2$, vide Figura 29, sendo os valores máximos extraídos da dimensão canal. Nos exemplos da Figura 61(c) é possível observar que as ativações máximas da rede TITAN também foram mais intensas nas respectivas regiões anotadas nas Figuras 61(a) em comparação com os exemplos da Figura 60(d), que representam as respectivas ativações máximas observadas na rede TITAN_b.

5.2 Resultados dos Estudos Comparativos

Nessa seção são apresentados os resultados quantitativos e qualitativos obtidos mediante treinamento e comparação das melhores configurações encontradas, técnicas MCRCNN e TITAN, com as técnicas correlatas abordadas na Seção 2.3. Para tornar mais clara a exposição dos resultados os estudos comparativos foram subdivididos em três grupos, sendo eles de Comparação em Base CD2014, Comparação em Bases de Imagens Aéreas e Comparação em Base PetrobrasROUTES, apresentados respectivamente pelas Subseções 5.2.1 à 5.2.3.

5.2.1 Comparação em Base CD2014

Seguindo os protocolos de treinamento, validação e teste pré-estabelecidos na Subseção 4.4.2, comparam-se as técnicas sobre a base CD2014 de duas formas distintas. A primeira abordagem, apresentada pela Subseção 5.2.1, consiste na realização de treinamentos unitários, ou seja, para cada vídeo da base de dados CD2014 uma instância diferente do método de detecção analisado deverá ser treinada. Na segunda abordagem, Subseção 5.2.1, todos os vídeos são utilizados, de uma só vez, para treinamento dos métodos de detecção a serem comparados entre si.

Comparação por Vídeos Unitários

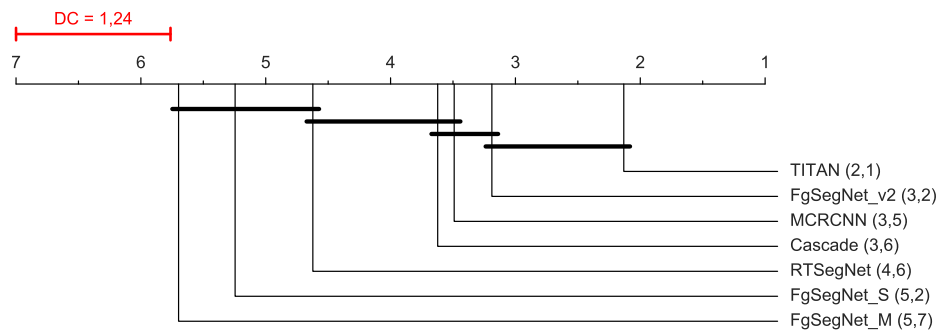
Tabela 9 – Resultados quantitativos médios de comparação.
Utilização de base CD2014 com treinamentos individuais por vídeo.

Métodos	Medida-F	IOU	PWC	PR	ROC
MCRCNN	0,9080	0,8614	0,3596	0,8066	0,9892
TITAN	0,9576	0,9269	0,0959	0,8798	0,9909
Cascade	0,7942	0,7364	1,2231	0,7557	0,9270
FgSegNet_v2	0,9576	0,9267	0,1032	0,8065	0,9919
FgSegNet_S	0,9552	0,9242	0,1092	0,5571	0,9951
FgSegNet_M	0,9546	0,9216	0,1154	0,5258	0,9914
RTSegNet	0,7828	0,6979	2,0202	0,6591	0,9870

Fonte: Elaborada pelo autor.

A Tabela 9 mostra os resultados de comparação entre técnicas considerando vídeos unitários. Para um valor de significância $\rho = 0,05$, o método estatístico de Friedman (1937) nos diz que as medidas IOU, Medida-F, PWC, PR e ROC possuem valores de significância $\rho_F = 7,13e^{-75}$, $\rho_F = 1,68e^{-74}$, $\rho_F = 7,23e^{-77}$, $\rho_F = 4,88e^{-25}$ e $\rho_F = 1,53e^{-14}$, respectivamente, indicando diferenças significativas entre as técnicas comparadas.

Figura 62 – Ranqueamento de técnicas sobre base CD2014 treinadas por vídeos unitários.
Considera-se medida PR.

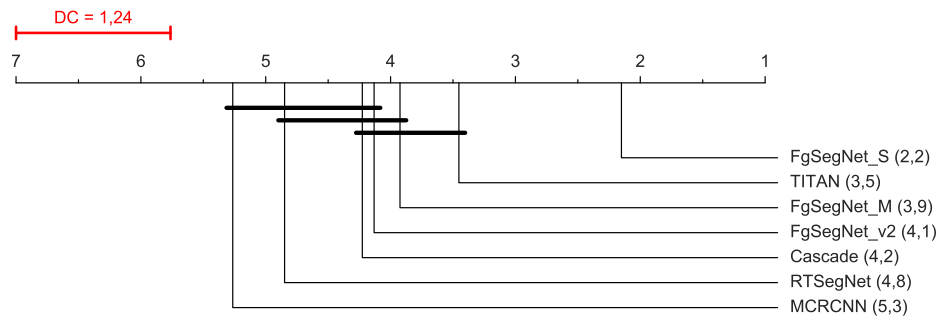


Fonte: Elaborada pelo autor.

No mesmo sentido, utilizando um valor de significância $\rho = 0,05$, o teste estatístico post-hoc de Nemenyi (1963), como mostra a Figura 62, indica que a técnica proposta TITAN coloca-se entre as melhores técnicas ranqueadas e supera todas as demais, exceto pela FgSegNet_v2, sendo os resultados de PR estatisticamente inconclusivos. A Figura 62 ainda mostra que a técnica MCRCNN obteve um melhor desempenho em comparação

com FgSegNet_S e FgSegNet_M.

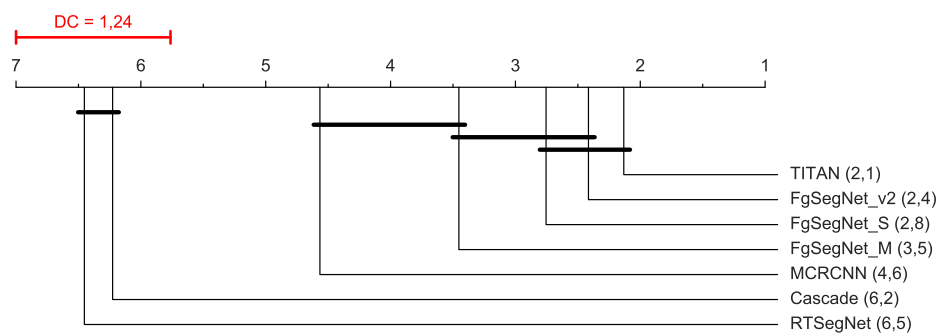
Figura 63 – Ranqueamento de técnicas sobre base CD2014 treinadas por vídeos unitários. Considera-se medida ROC.



Fonte: Elaborada pelo autor.

A Figura 63 apenas nos permite afirmar que a técnica proposta TITAN, em termos de medida ROC, supera as técnicas RTSegNet e MCRCNN, sendo pior que o método FgSegNet_S.

Figura 64 – Ranqueamento de técnicas sobre base CD2014 treinadas por vídeos unitários. Consideram-se as medidas IOU e Medida-F representadas pelo mesmo gráfico, tendo elas exibido resultados estatísticos equivalentes.

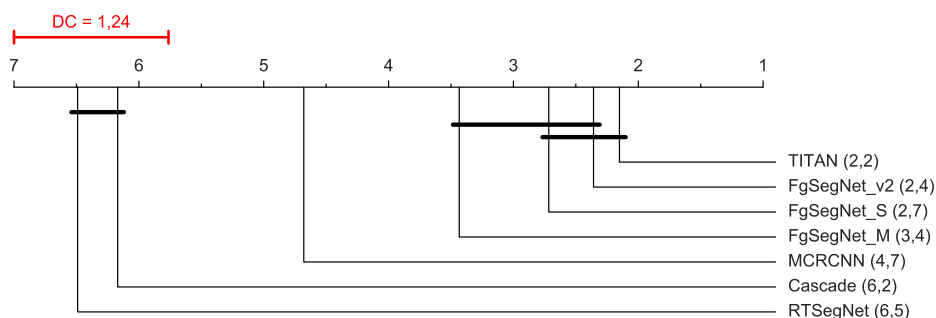


Fonte: Elaborada pelo autor.

De acordo com a Figura 64, levando-se em consideração ambas as medidas IOU e Medida-F, é possível dizer que a técnica proposta MCRCNN apresentou um pior desempenho em comparação com as técnicas TITAN, FgSegNet_v2 e FgSegNet_S, porém produziu resultados quantitativos melhores que as técnicas Cascade e RTSegNet. A

Figura 64 ainda mostra que a técnica proposta TITAN, além de superar as técnicas MCRCNN, FgSegNet_v2 e FgSegNet_S, também obteve um desempenho melhor do que a técnica FgSegNet_M.

Figura 65 – Ranqueamento de técnicas sobre base CD2014 treinadas por vídeos unitários. Considera-se a medida PWC.

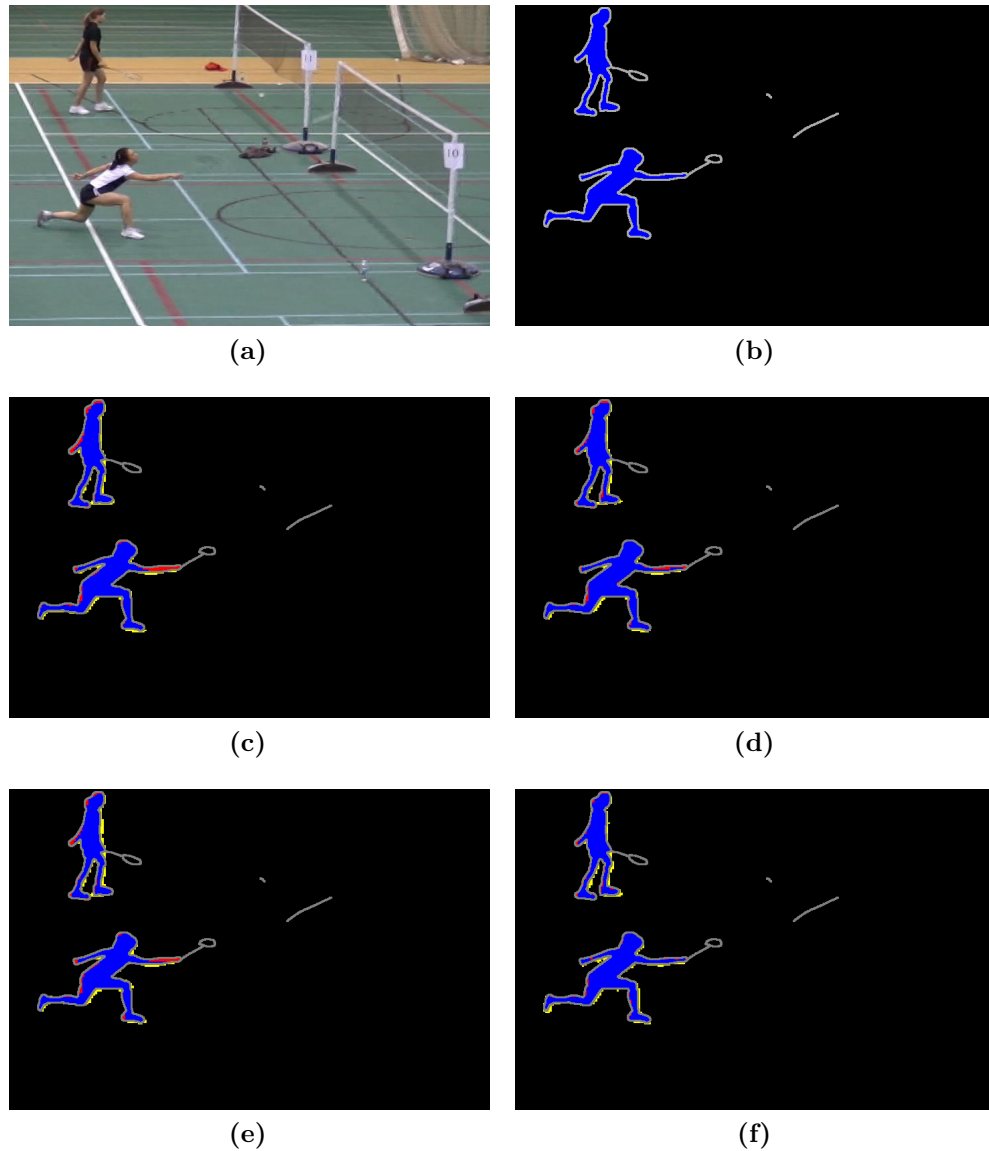


Fonte: Elaborada pelo autor.

A Figura 65, além de confirmar os resultados obtidos em termos de IOU e Medida-F, também indica que as diferenças entre as medidas PWC das técnicas MCRCNN e FgSegNet_M comparadas são suficientemente grandes para que se possa afirmar que a técnica MCRCNN teve um pior desempenho em comparação com a técnica FgSegNet_M. De forma geral, nota-se que a técnica TITAN, em termos de resultados quantitativos, exibiu melhores resultados em comparação com as técnicas Cascade, MCRCNN, FgSegNet_M e RTSegNet para a maioria das medidas comparadas.

Ao analisarmos a qualidade das detecções apresentadas pela Figura 66, é possível notar que os métodos FgSegNet_M, FgSegNet_S e FgSegNet_v2 produziram mais detecções falso negativas em regiões de contorno, em comparação com o método TITAN, com relação aos jogadores presentes no vídeo Jogo de Peteca. Como apontado por Ronneberger, Fischer e Brox (2015) esses problemas podem surgir após execução de sucessivos processos de subamostragem espacial durante o estágio de codificação das redes codificadoras-decodificadoras.

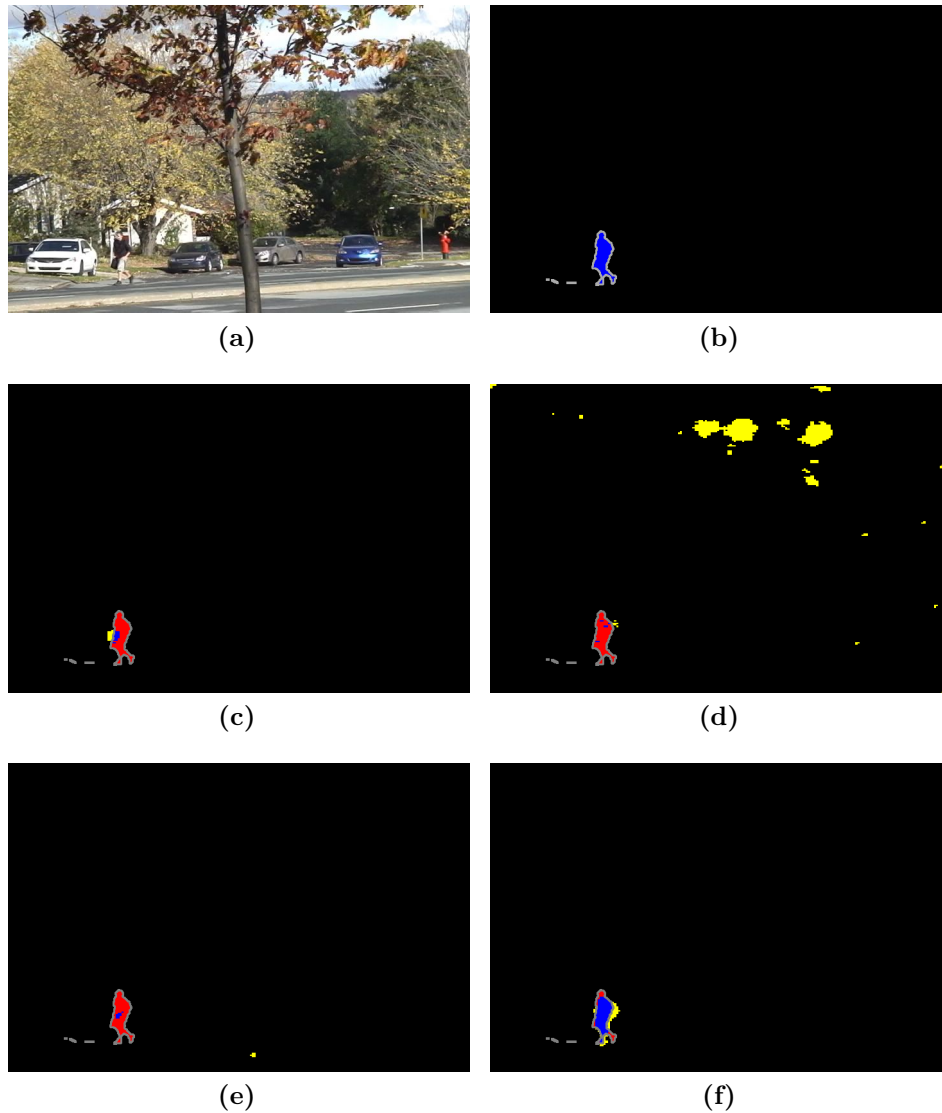
Por meio da Figura 67 é possível observar que os métodos Cascade, MCRCNN e RTSegNet, em alguns casos, tiveram uma maior dificuldade para diferenciar corretamente os conteúdos de *foreground* dos de *background* em situações adversas envolvendo movimentação de conteúdo de *background*. Como mostra a Figura 66, os resultados de detecção das redes Cascade, MCRCNN e RTSegNet exibem muitos FNs, em comparação com o método

Figura 66 – Resultados qualitativos considerando amostras da base de dados CD2014.

Fonte: Figuras (c) à (f) elaboradas pelo autor e as demais retiradas de Wang et al. (2012).

Detecções para vídeo Jogo de Peteca (categoria C. Trêm.) apresentados na seguinte ordem: (a) Imagem alvo de detecções, (b) Máscara rotulada, resultados gerados por (c) FgSegNet_M, (d) FgSegNet_S, (e) FgSegNet_v2 e (f) TITAN.

TITAN, sendo que a detecção de *foreground* praticamente não acontece. Além disso, nota-se que a rede RTSegNet também acabou confundindo o movimento das folhas das árvores, elementos de *background*, com regiões de mudança.

Figura 67 – Resultados qualitativos considerando amostras da base de dados CD2014.

Fonte: Figuras (c) à (f) elaboradas pelo autor e as demais retiradas de Wang et al. (2012).

Detecções para o vídeo Outono (categoria *Back. Din.*) apresentados na seguinte ordem: (a) Imagem alvo de detecções, (b) Máscara rotulada, resultados gerados por (c) Cascade, (d) RTSegNet, (e) MCRCNN e (f) TITAN.

Comparação por Grupo de Vídeos

Na Tabela 10 são apresentados os resultados quantitativos médios das medidas IOU, Medida-F, PWC, PR e ROC analisadas, obtidos de acordo com o treinamento das técnicas MCRCNN, TITAN, SEUNet1, SEUNet2, SEUNet3, MUNet1, MUNet2 e DeepBS.

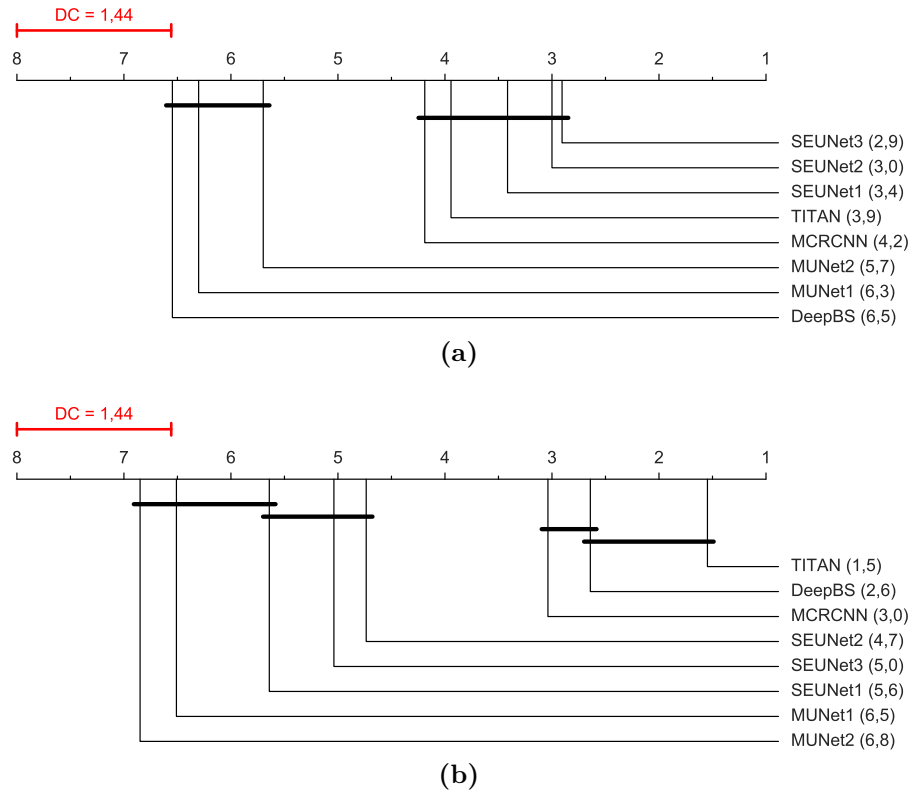
Com base nos resultados da Tabela 10, para um valor de significância $\rho = 0,05$, é possível dizer, num primeiro momento, que existem diferenças significativas entre as técnicas comparadas, tendo em vista que aplicando-se o teste estatístico de Friedman (1937) sobre

Tabela 10 – Resultados quantitativos médios de comparação por grupo de vídeos para base CD2014.

Métodos	Medida-F	IOU	PWC	PR	ROC
MCRCNN	0,9249	0,8767	0,1899	0,8568	0,9853
TITAN	0,9524	0,9176	0,1172	0,8678	0,9928
DeepBS	0,8320	0,7474	0,9675	0,7012	0,9865
SEUNet1	0,9472	0,9085	0,1515	0,8956	0,9811
SEUNet2	0,9468	0,9081	0,1436	0,8950	0,9831
SEUNet3	0,9365	0,8924	0,1771	0,9002	0,9797
MUNet1	0,9419	0,9059	0,1870	0,7514	0,9710
MUNet2	0,9413	0,8998	0,1818	0,7646	0,9711

Fonte: Elaborada pelo autor.

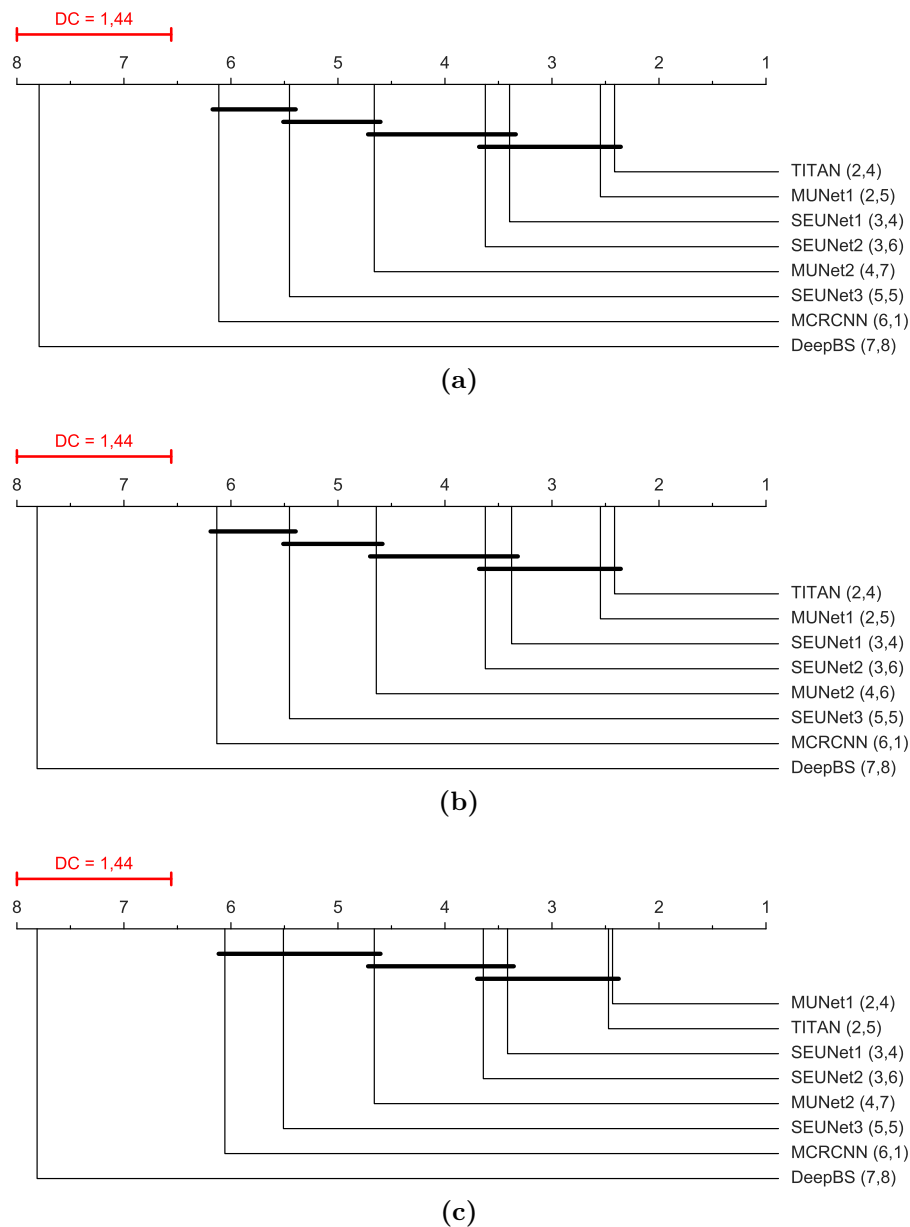
as medidas IOU, Medida-F, PWC, PR e ROC obtêm-se os valores de significância $\rho_F = 3,90e^{-66}$, $\rho_F = 3,87e^{-65}$, $\rho_F = 4,69e^{-66}$, $\rho_F = 3,59e^{-32}$ e $\rho_F = 1,10e^{-69}$, respectivamente.

Figura 68 – Ranqueamento de técnicas sobre base CD2014 comparadas por grupos de vídeos. Consideram-se as medidas (a) PR e (b) ROC.**Fonte:** Elaborada pelo autor.

Em um segundo momento, utilizando um valor de significância $\rho = 0,05$, o teste estatístico post-hoc de Nemenyi (1963) indica que, de acordo com a Figura 68(a), as

técnicas propostas MCRCNN e TITAN aparecem no grupo das técnicas melhor ranqueadas, superando as técnicas DeepBS, MUNet1 e MUNet2. De acordo com a Figura 68(b), é possível dizer que ambas as técnicas TITAN e MCRCNN superaram quantitativamente as técnicas SEUNets e MUNets.

Figura 69 – Ranqueamento de técnicas treinadas, por grupos de vídeos, sobre base CD2014. Consideram-se as medidas (a) Medida-F, (b) IOU e (c) PWC.

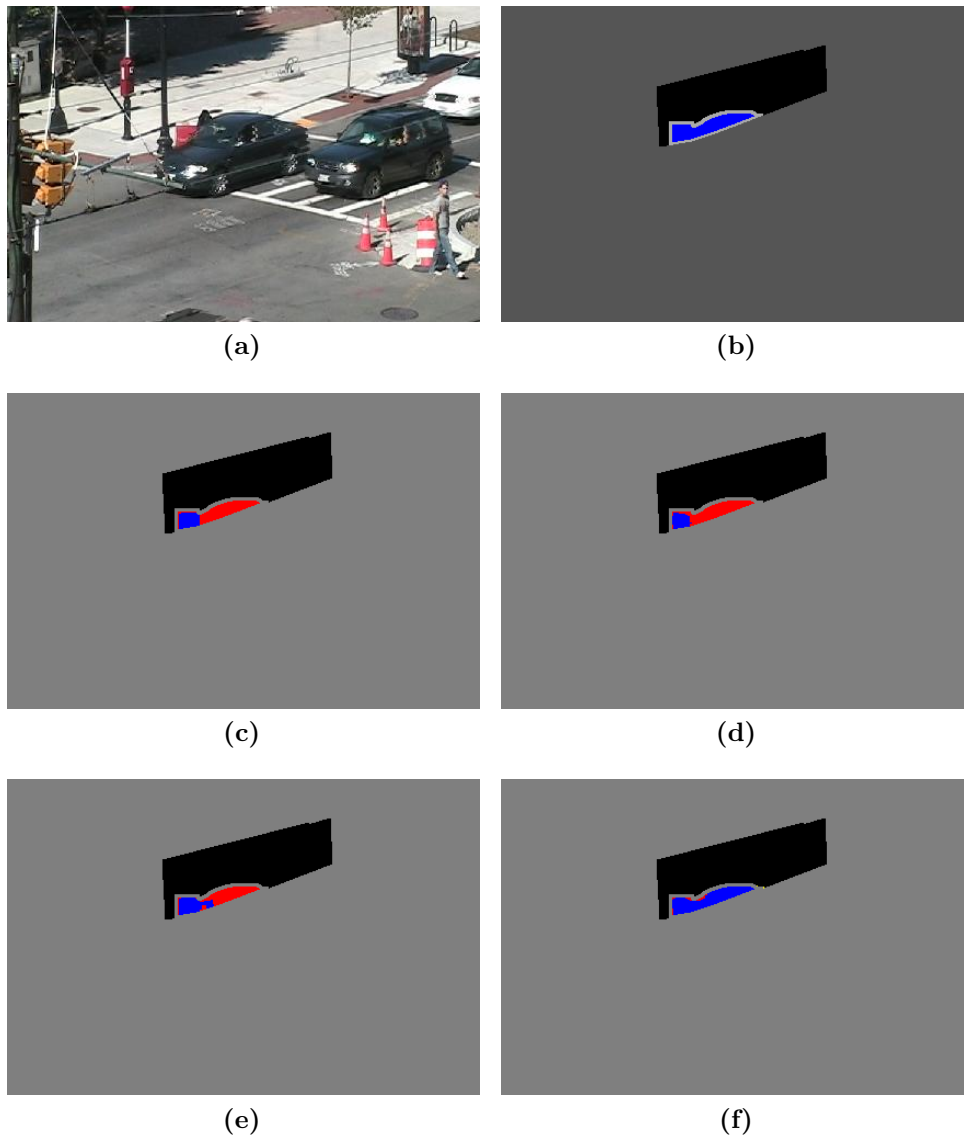


Fonte: Elaborada pelo autor.

De acordo com as Figura 69, é possível dizer que a técnica proposta MCRCNN apresentou um pior desempenho em comparação com as demais técnicas, exceto em

comparação com a técnica DeepBS. O mesmo não se verifica no caso da técnica TITAN, tendo ela sido ranqueada no grupo das técnicas com melhor desempenho superando os métodos MUNet2, SEUNet3, MCRCNN e DeepBS.

Figura 70 – Resultados qualitativos considerando vídeo Caixa Abandonada.



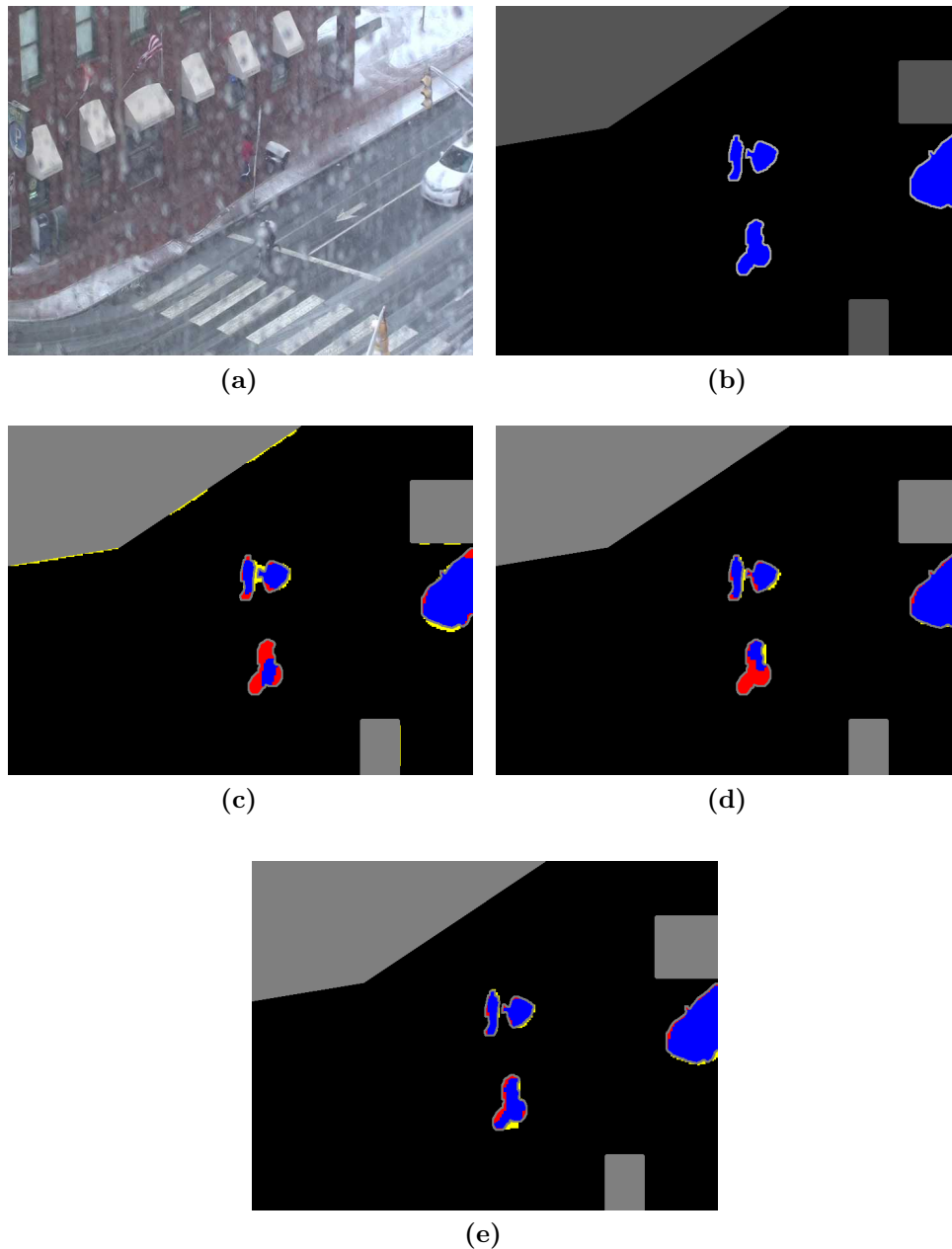
Fonte: Figuras (c) à (f) elaboradas pelo autor e as demais retiradas de Wang et al. (2012).

Detecções para o vídeo Caixa Abandonada (categoria Mov. Int.) apresentados na seguinte ordem: (a) Imagem alvo de detecções, (b) Máscara rotulada, resultados gerados por (c) SEUNet1, (d) SEUNet2, (e) SEUNet3 e (f) TITAN.

Ao analisarmos a qualidade das detecções apresentadas pela Figura 70, é possível observar que os métodos SEUNet1, SEUNet2 e SEUNet3 tiveram maior dificuldade para detectar regiões de mudanças. Tais detecções falso negativas seriam um indicativo de

que as SEUNets possuem uma grande dependência de informação espacial para fins de contextualização, porém os exemplos de treinamento que contemplam situações de intermitência de movimento, levando-se em consideração o vídeo Caixa Abandonada, são restritos à regiões espaciais muito reduzidas (regiões de interesse).

Figura 71 – Resultados qualitativos considerando vídeo Neve Molhada.

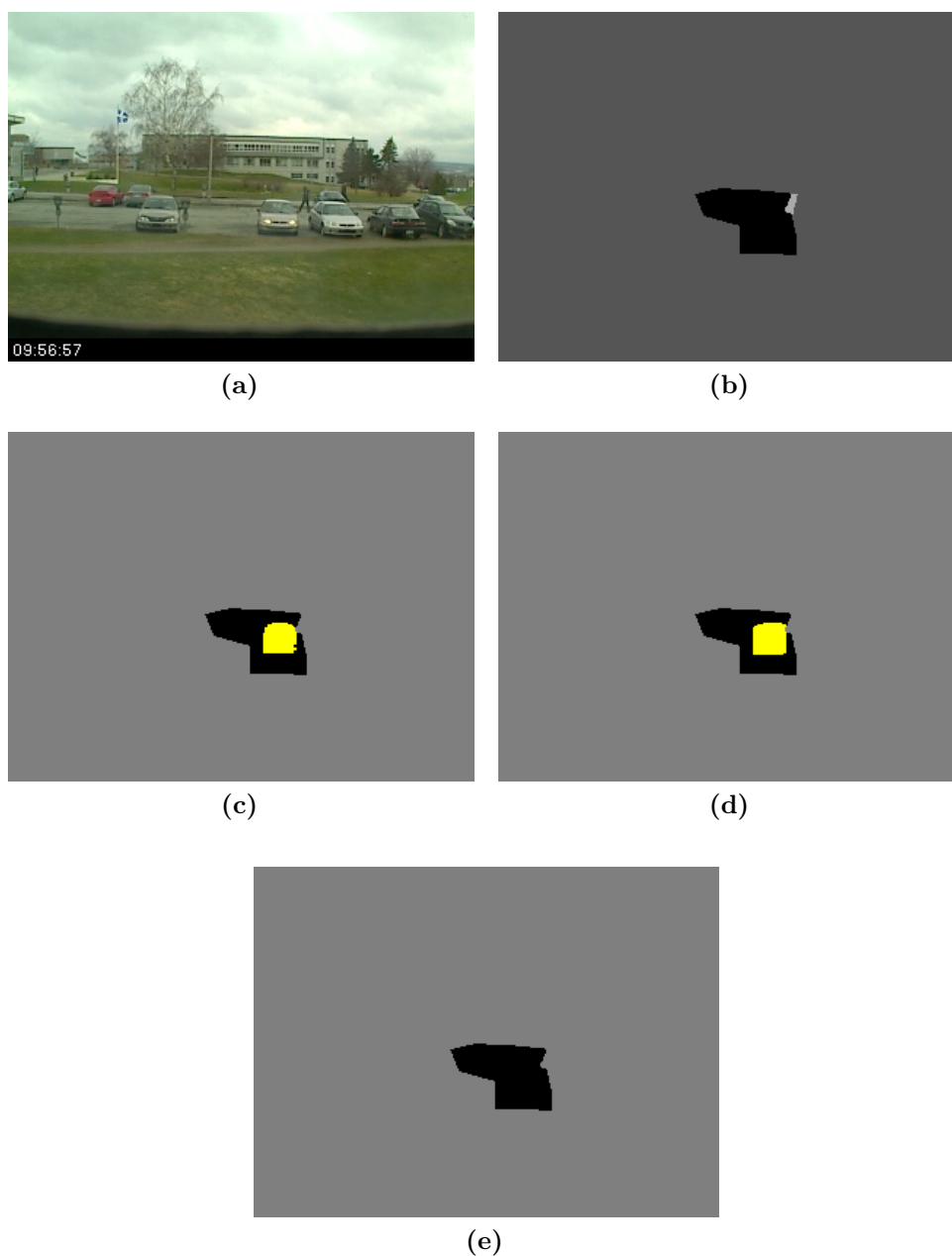


Fonte: Figuras (c) à (e) elaboradas pelo autor e as demais retiradas de Wang et al. (2012).

Detecções para o vídeo Neve Molhada (categoria Mau Tmp.) apresentados na seguinte ordem: (a) Imagem alvo de detecções, (b) Máscara rotulada, resultados gerados por (c) DeepBS, (d) MCRCNN e (e) TITAN.

Por meio da Figura 71 é possível observar que os métodos DeepBS e MCRCNN, em alguns casos, tiveram mais dificuldade para detectar corretamente regiões de mudanças sob condições adversas de camuflagem.

Figura 72 – Resultados qualitativos considerando vídeo Estacionamento.



Fonte: Figuras (c) à (e) elaboradas pelo autor e as demais retiradas de Wang et al. (2012).

Detecções para o vídeo Estacionamento (categoria Mov. Int.) apresentados na seguinte ordem:
(a) Imagem alvo de detecções, (b) Máscara rotulada, resultados gerados por (c) MUNet1, (d) MUNet2 e (e) TITAN.

Além disso, observando a Figura 71, é possível notar também que o processo de detecção das redes DeepBS e MCRCNN gerou muitos FNs, em comparação com o método TITAN, principalmente nas regiões da imagem onde encontram-se um ciclista e uma pessoa levando um carrinho de bebê. No caso do ciclista, nota-se que as cores de sua roupa e bicicleta se assemelham muito com as cores do asfalto, faixa de pedestres e com os flocos de neve que caem ao fundo. No caso da pessoa com o carrinho de bebê, é possível confundir as cores de suas roupas com as cores da calçada e do estabelecimento ao lado.

A Figura 72 mostra que os métodos MUNet1 e MUNet2, sob condições adversas envolvendo movimentação intermitente, para o vídeo Estacionamento, produziram mais detecções falso positivas em comparação com o método TITAN. Nessas circunstâncias as redes MUNets detectaram a existência de um carro em uma vaga de estacionamento totalmente desocupada.

5.2.2 Comparação em Base de Imagens Aéreas

Na Tabela 11 são apresentados os resultados quantitativos médios obtidos pelas técnicas MCRCNN, TITAN, DMINet, RSCDNet e FLS treinadas de acordo com os protocolos da Seção 5.2.2.

Tabela 11 – Resultados quantitativos médios de comparação de técnicas sobre imagens aéreas.

Métodos	Medida-F	IOU	PWC
TITAN	0,6895	0,5491	5,2030
DMINet	0,6910	0,5586	4,4192
RSCDNet	0,6277	0,4850	6,0240
STANet	0,5792	0,4446	6,1562
FLS	0,5295	0,3813	10,0305

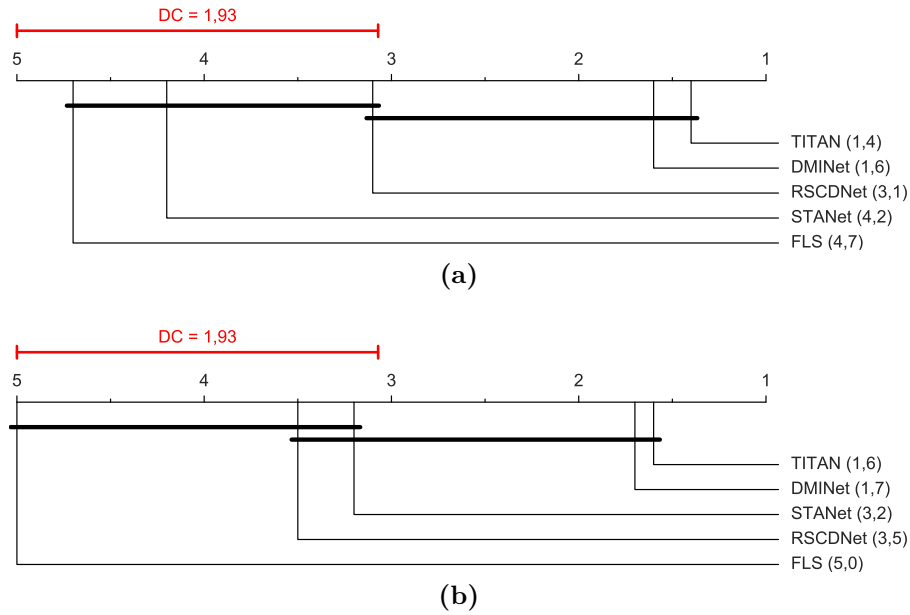
Fonte: Elaborada pelo autor.

Por meio de teste estatístico de Friedman (1937), considerando limiar de significância $\rho = 0,05$, obtém-se os valores de significância $\rho_F = 9,47e^{-16}$, $\rho_F = 9,47e^{-16}$ e $\rho_F = 2,14e^{-11}$, para as medidas IOU, Medida-F e PWC, respectivamente, indicando diferenças significativas nos resultados da Tabela 11.

Por meio de teste estatístico post-hoc de Nemenyi (1963) é possível afirmar que essas diferenças indicam que, de acordo com a Figura 73, a técnica proposta TITAN foi

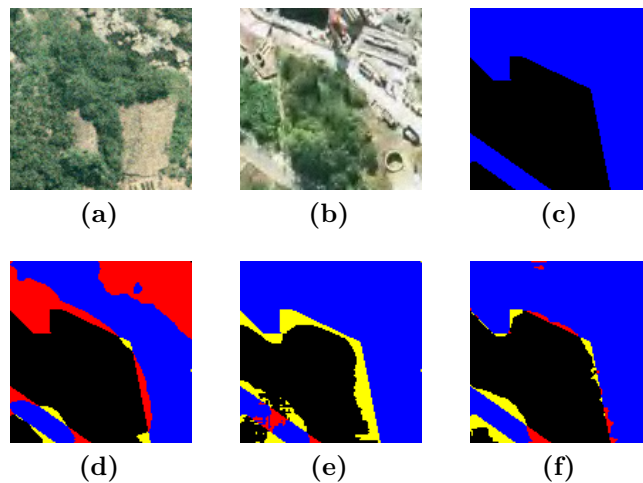
Figura 73 – Ranqueamento de técnicas treinadas sobre base de imagens aéreas.

Consideram-se as medidas (a) IOU e Medida-F e (b) PWC. Medidas IOU e Medida-F equivalentes e representadas pelo mesmo gráfico.



Fonte: Elaborada pelo autor.

capaz de superar os métodos STANet e FLS.

Figura 74 – Resultados qualitativos sobre amostra da base **SYSU-CD**.

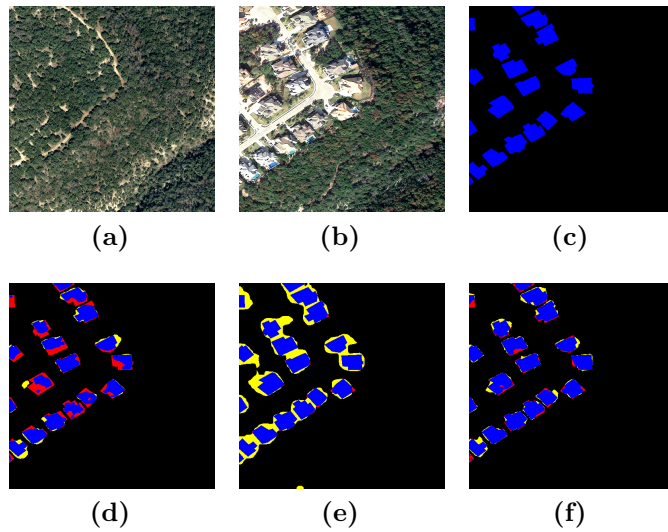
Fonte: Elaborado pelo autor.

Em (a) tem-se as imagens alvo de detecção (a) I^t , (b) I^{t+1} , a respectiva (c) máscara rotulada, além dos resultados produzidos por (d) DMINet, (e) RSCDNet e (f) TITAN.

A Figura 74 mostra que a técnica TITAN apresentou menos detecções falso negativas

(regiões em vermelho na Figura 74) em comparação com a técnica DMINet. Ao mesmo tempo, em comparação com a técnica RSCDNet, ela apresentou menos regiões falso positivas. Já a Figura 75, mostra que a técnica TITAN apresentou menos detecções falso negativas em comparação com a técnica STANet e menos detecções falso positivas, em comparação com a técnica FLS.

Figura 75 – Resultados qualitativos sobre amostra de base **LEVIR-CD**.



Fonte: Elaborado pelo autor.

Em (a) tem-se as imagens alvo de detecção (a) I^t , (b) I^{t+1} , a respectiva (c) máscara rotulada, além dos resultados produzidos por (d) FLS, (e) STANet e (f) TITAN.

5.2.3 Comparação em Base PetrobrasROUTES

Na Tabela 12 são apresentados os resultados quantitativos médios em termos de IOU, Medida-F e PWC, obtidos de acordo com o treinamento das técnicas TITAN, DMINet, MU-Net1 e FgSegNet_v2 sobre a base de dados PetrobrasROUTES, sendo essas técnicas as que exibiram os melhores resultados quantitativos nos Estudos de Comparação das Subseções 5.2.1 e 5.2.2.

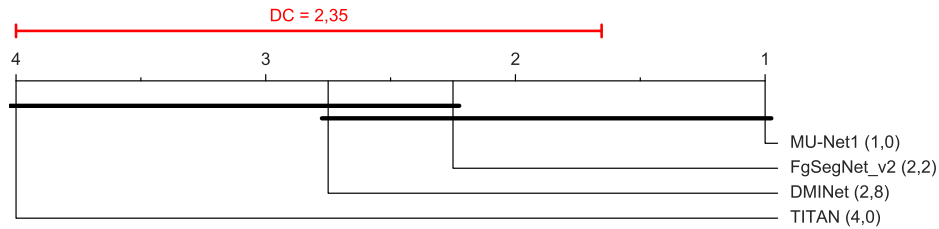
Com base nos resultados da Tabela 12, para um valor de significância $\rho = 0,05$, é possível dizer, num primeiro momento, que existem diferenças significativas entre as técnicas comparadas, tendo em vista que aplicando-se o teste estatístico de Friedman (1937) sobre as medidas IOU, Medida-F e PWC são obtidos os valores de significância

Tabela 12 – Resultados quantitativos médios de comparação de técnicas sobre imagens de obstrução de rotas de escape.

Métodos	Medida-F	IOU	PWC
TITAN	0,5579	0,3877	0,1633
DMINet	0,6164	0,4465	0,1467
MU-Net1	0,6871	0,5237	0,1340
FgSegNet_v2	0,6280	0,4584	0,1662

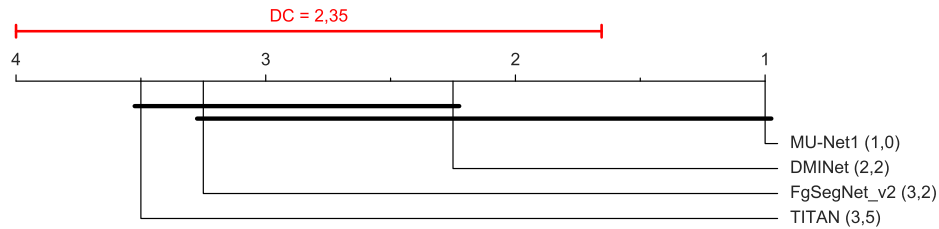
Fonte: Elaborada pelo autor.

$\rho_F = 9,51e^{-5}$, $\rho_F = 9,51e^{-5}$ e $\rho_F = 5,65e^{-3}$, respectivamente.

Figura 76 – Ranqueamento de técnicas treinadas sobre base de sobre imagens de obstrução de rotas de escape.

Fonte: Elaborada pelo autor.

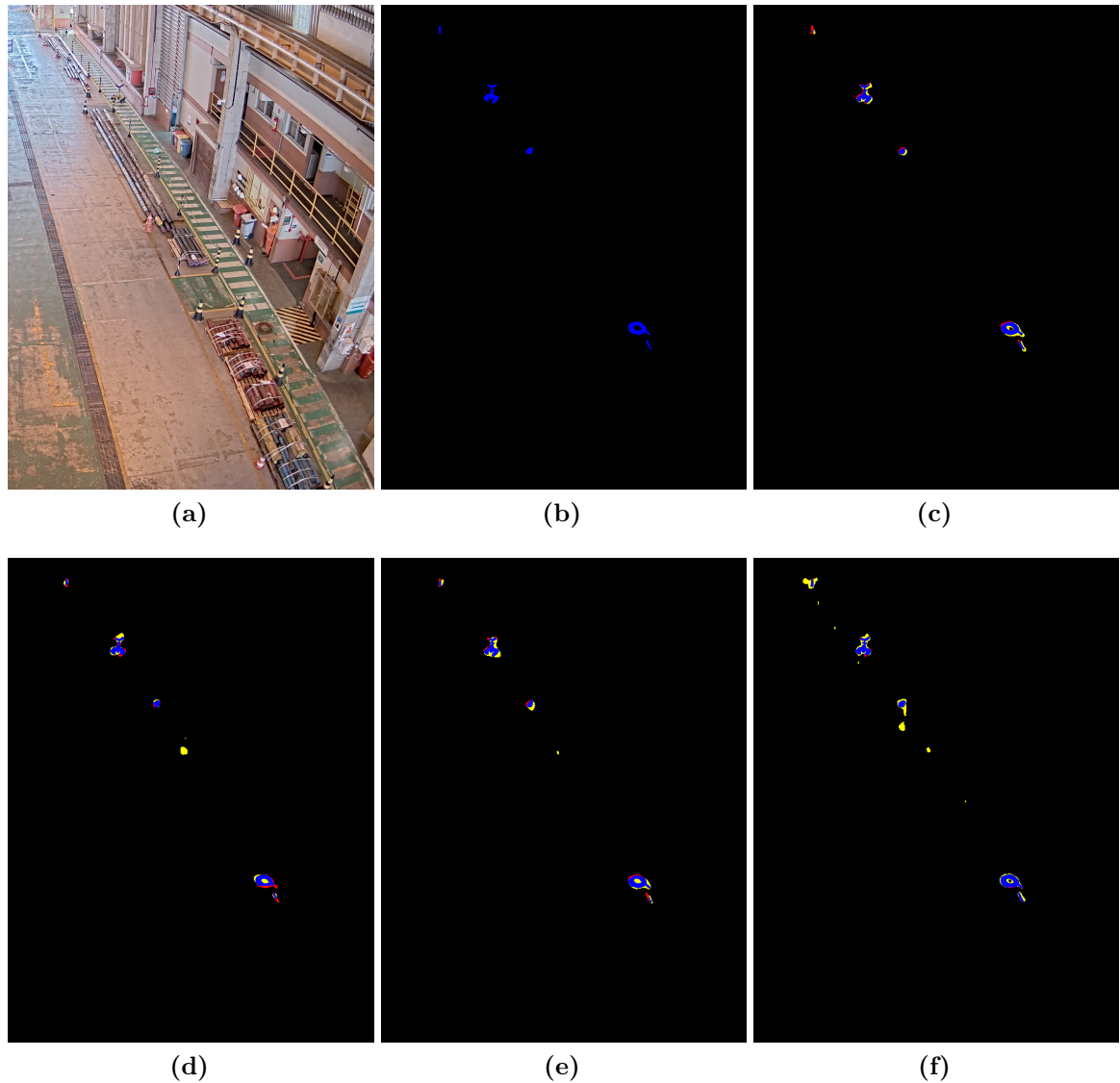
Consideram-se as medidas Medida-F e IOU representadas pelo mesmo gráfico, tendo elas exibido resultados estatísticos equivalentes.

Figura 77 – Ranqueamento de técnicas treinadas sobre imagens de obstrução de rotas de escape. Considera-se a medida PWC.

Fonte: Elaborada pelo autor.

As Figuras 76 e 77, geradas mediante aplicação do teste estatístico post-hoc de Nemenyi (1963), para um limiar de significância $\rho = 0,05$, nos permite afirmar conclusivamente apenas que a técnica proposta TITAN teve um pior desempenho quantitativo em comparação com a técnica MU-Net1.

Figura 78 – Exemplo de detecções de obstruções de rota de escape para base PetrobrasROUTES.



Fonte: Elaborado pelo autor.

Em (a) tem-se um exemplo pertencente ao grupo Macae e (b) Máscara rotulada. Além disso, têm-se os resultados produzidos pelas técnicas (c) TITAN, (d) DMINet, (e) FgSegNet_v2 e (f) MU-Net.

Analisando o desempenho qualitativo das técnicas TITAN, DMINet, FgSegNet_v2 e MU-Net1 sobre o grupo de imagens Macae, como mostra a Figura 78, observa-se uma tendência ao cometimento de erros de detecção falso positivos. Isso fica bastante evidente nas Figuras 78(d) à 78(f), que mostram resultados de detecção obtidos pelas técnicas DMINet, FgSegNet_v2 e MU-Net1, respectivamente. Como mostra a Figura 78(a), as variações de luminosidade ao longo de toda a região ao redor e pertencente à rota de escape

são muito intensas, fazendo com que as técnicas acabem atribuindo indevidamente rótulos de mudanças para regiões ocupadas por sombras e reflexos luminosos, em detrimento de objetos depositados sobre a rota de escape.

5.2.4 Comparação de Custo Computacional

Na Tabela 13 são apresentados os custos computacionais de execução dos métodos TITAN, DMINet, MU-Net1 e FgSegNet_v2, sendo esses métodos os que exibiram os melhores resultados nos Estudos Comparativos das Subseções 5.2.1 e 5.2.2. Os custos computacionais foram calculados por meio das medidas de Quantidade de Operações de Ponto Flutuante por Segundo (multiplicações e adições), do inglês *Floating-point operations* (FLOPs), quantidade de parâmetros treináveis e velocidade de processamento das imagens, do inglês *Frames Per Second* (FPS).

Tabela 13 – Custo computacional de detecção considerando dimensão espacial de 690×520 .

Métodos	FLOPs ($\times 10^9$)	Parâmetros ($\times 10^6$)	FPS
TITAN	52,62	3,72	59
DMINet	78,42	6,76	9
MU-Net1	130,98	17,80	39
FgSegNet_v2	249,94	9,23	23

Fonte: Elaborada pelo autor.

Como mostra a Tabela 13, a técnica proposta TITAN apresenta o menor custo computacional em comparação com todas as outras. Analisando os piores cenários a técnica TITAN gasta $25,8 \times 10^9$ FLOPs a menos do que a técnica DMINet, consumindo 55% menos memória. Além disso a técnica TITAN também consegue processar 20 quadros a mais do que a técnica MU-Net1 no mesmo intervalo de tempo de um segundo.

6 Conclusão

Nesse trabalho desenvolveram-se duas novas estratégias para modelagem computacional dos mecanismos de pré-atenção visual espacial sem e com filtragem seletiva de Posner (1980) e LaBerge e Brown (1989) denominadas por atenção residual implícita e atenção residual explícita, respectivamente. Como subproduto do mesmo trabalho foram criadas as técnicas CRCNN, MCRCNN e TITAN, sendo os métodos CRCNN e MCRCNN utilizados como instrumentos de análise do mecanismo de atenção residual implícita e o método TITAN do mecanismo de atenção residual explícita. Resultados de estudos de configuração das redes CRCNN e MCRCNN, considerando a base de dados CD2014, corroboraram a hipótese de que o mecanismo de atenção residual implícita seria benéfico ao processo de separação entre *pixels* de *foreground* e *background* em cenários envolvendo detecção de mudanças em vídeos, levando-se em conta a pouca disponibilidade de dados de treinamento coletados sob as mais variadas condições adversas. Por meio do mesmo processo de análise, observou-se também que o método MCRCNN foi capaz de superar o método CRCNN quantitativamente e qualitativamente. Resultados de estudos de configuração da rede TITAN, considerando as bases de dados **SZATAKI**, **Onera**, **LEVIR-CD** e **SYSU-CD**, corroboraram a hipótese de que o mecanismo de atenção residual explícita traria benefícios ao processo de detecção de mudanças em imagens aéreas de sensoriamento remoto. Resultados quantitativos de estudos comparativos utilizando base de dados CD2014, sem agrupamento de categorias de vídeos, indicaram que a técnica TITAN conseguiu superar as técnicas MCRCNN, FgSegNet_M, Cascade e RTSegNet em todas as cinco medidas analisadas, além de superar também a técnica FgSegNet_S, considerando medida PR. Qualitativamente o método TITAN proposto foi ainda capaz de superar até mesmo a técnica FgSegNet_v2. Resultados comparativos obtidos sobre a mesma base, porém considerando-se agrupamento de categorias de vídeos, foram também bastante favorável à técnica TITAN. Ela foi capaz de superar as técnicas MUNet2 e MCRCNN, considerando as cinco medidas analisadas, e também foi melhor que as técnicas MUNet1, DeepBS e SEUNet3 em ao menos quatro medidas analisadas. Ainda considerando o mesmo estudo, porém em termos qualitativos, verificou-se que a rede TITAN superou as técnicas MUNet1, SEUNet1 e SEUNet2 em vídeos coletados sob condições

adversas de movimentação intermitente de conteúdo. Resultados obtidos por meio de estudos comparativos realizados bases de dados de imagens aéreas **SZATAKI**, **Onera**, **LEVIR-CD** e **SYSU-CD**, demonstraram que a técnica TITAN foi quantitativamente superior às técnicas STANet e FLS, podendo ser também equiparada às demais técnicas comparadas, DMINet e RSCDNet. Resultados comparativos obtidos para base de dados PetrobrasROUTES mostraram que a técnica TITAN não conseguiu superar o método de detecção MU-Net1, mas equiparou-se às técnicas FgSegNet_v2 e DMINet. Ainda sim, em termos qualitativos, foi possível observar casos em que a técnica TITAN foi capaz de superar as técnicas FgSegNet_v2 e DMINet em situações envolvendo condições adversas de luminosidade ambiente, algo muito comum nos vídeos pertencentes à base Petrobras-ROUTES. Por fim, em termos de custo computacional, os resultados obtidos mostram que a técnica TITAN superou as demais técnicas comparadas FgSegNet_v2, DMINet e MU-Net1, exigindo uma menor quantidade de execuções de operações por ponto flutuante (FLOPs), menor quantidade de parâmetros e maior velocidade de detecção de mudanças (FPS). Mesmo considerando os piores casos, a técnica TITAN realizou $25,8 \times 10^9$ menos FLOPs e precisou armazenar $3,04 \times 10^6$ menos parâmetros que a rede DMINet, enquanto conseguiu processar 20 quadros a mais que a técnica MU-Net1 no mesmo intervalo de um segundo.

6.1 Propostas para Trabalhos Futuros

Futuramente, pretende-se adaptar as estratégias computacionais dos mecanismos de atenção residual implícita e explícita propostos. Para o primeiro caso, relativo às redes CRCNN e MCRCNN, serão exploradas algumas alternativas de redução da quantidade de camadas, utilização de mais camadas de subamostragem espacial, assim como adaptações do módulo de processamento residual (RPM), visando aproximá-lo ainda mais de mecanismos de filtragem morfológica tais como dilatação, erosão, abertura e fechamento. No segundo caso, relativo à rede TITAN, pretende-se incorporar outras estratégias, além da espaço-temporal já utilizada, de aprimoramento do mecanismo proposto de atenção residual explícita. Dentre possíveis estratégias, como apontado por Kravitz e Behrmann (2011), tem-se mecanismos capazes de extrair representações perceptuais a partir de análises feitas sobre interações entre objetos e suas partes. Os trabalhos de LaLonde et al. (2021) e

Sabour et al. (2021) apontam para alternativas muito promissoras nessa mesma direção de pesquisa, onde utilizam-se estratégias de aprendizado baseadas em estruturas denominadas cápsulas. Essas estruturas, por sua vez, aumentam a capacidade de redes codificadoras-decodificadoras de aprenderem como distinguir diferentes grupos de objetos em imagens, sendo que para isso são combinadas metodologias de aprendizagem capazes de extrair não só características invariantes, mas também equivariantes, das partes que compõem tais objetos. Também serão exploradas alternativas de como tornar os métodos CRCNN, MCRCNN e TITAN menos dependentes de técnicas externas de modelagem de imagem de *background*, sendo que nosso ponto de partida seria utilizar de estratégias de modelagem à priori, à exemplo de Santos et al. (2021), para tentar gerar o *background* faltante utilizando uma Rede Profunda de Imagem à Priori, do inglês *Deep Image Prior*. Além disso, pretende-se adaptar as técnicas desenvolvidas na tentativa de conseguir solucionar também problemas de outros domínios. No caso da detecção de objetos seriam utilizados os mapas residuais das redes CRCNN e MCRCNN, ou mesmo os limitadores de atenção da rede TITAN, como indicadores de regiões contendo objetos detectáveis. Isso possivelmente contribuiria com a simplificação do processo de detecção em imagens, facilitando a etapa de localização dos objetos. Além da detecção de objetos as técnicas propostas também seriam possivelmente muito úteis em tarefas relacionadas à restauração de imagens. Nesse sentido, pretende-se utilizar o mapa residual gerado pelo módulo BCNN, pertencente às redes CRCNN e MCRCNN, para destacar regiões de alta frequência corrompidas por ruídos aditivos. Como ponto de partida temos o trabalho desenvolvido por Pires et al. (2020), que investiga justamente a capacidade de melhorar o processo de restauração de imagens, corrompidas por ruído aditivo, utilizando para tal um mecanismo de atenção residual.

6.2 Trabalhos Desenvolvidos

Trabalhos publicados:

- SANTOS, D. F.; PIRES, R. G.; COLOMBO, D.; PAPA, J. P. Video segmentation learning using cascade residual convolutional neural network. In: IEEE. *2019 32nd SIBGRPRI Conference on Graphics, Patterns and Images (SIBGRPRI)*. [S.l.], 2019. p. 1-7.

- SANTOS, D. F.; PIRES, R. G.; COLOMBO, D.; PAPA, J. P. Scene change detection using multiscale cascade residual convolutional neural networks. In: IEEE. *2020 33rd SIBGRPRI Conference on Graphics, Patterns and Images (SIBGRPRI)*. [S.l.], 2020. p. 108-115.
- SANTOS, D. F.; PIRES, R. G.; PASSOS, L. A.; PAPA, J. P. Ddipnet and ddipnet+: Discriminant deep image prior networks for remote sensing image classification. In: IEEE. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. [S.l.], 2021. p. 2843-2846.
- SANTOS, D. F. S.; PAPA, J. P. Titan: A lightweight temporal attention network for remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 20, p. 1-5, 2023.
- SANTOS, D. F. S.; PIRES, R. G.; PAPA, J. P. Deblur Capsule Networks. In: *2023 26th Iberoamerican Congress on Pattern Recognition, Patterns and Images (SIBGRPRI)*. [S.l.], 2020. p. 1-15.
- PIRES, R. G.; SANTOS, D. F.; PAPA, J. P. Image denoising using attention-residual convolutional neural networks. In: IEEE. *2020 33rd SIBGRPRI Conference on Graphics, Patterns and Images (SIBGRPRI)*. [S.l.], 2020. p. 101-107.
- PIRES, R. G.; SANTOS, D. F.; PASSOS, L. A.; PAPA, J. P. Deep Regressor Networks for Blind Image Deblurring. In: IEEE. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. [S.l.], 2021. p. 5390-5393.
- ROSA, G. H.; RODER, M.; SANTOS, D. F. S.; COSTA, K. A. P. Enhancing Anomaly Detection Through Restricted Boltzmann Machine Features Projection. *International Journal of Information Technology*, Springer, p. 49-57, 2021.

REFERÊNCIAS

- BABAEI, M.; DINH, D. T.; RIGOLL, G. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, Elsevier, v. 76, p. 635–649, 2018.
- BARKUR, R.; SURESH, D.; LAL, S.; REDDY, C. S.; DIWAKAR, P. et al. Rscdnet: A robust deep learning architecture for change detection from bi-temporal high resolution remote sensing images. *IEEE Transactions on Emerging Topics in Computational Intelligence*, IEEE, v. 7, n. 2, p. 537–551, 2022.
- BENEDEK, C.; SZIRÁNYI, T. A mixed markov model for change detection in aerial photos with large time differences. In: IEEE. *2008 19th International Conference on Pattern Recognition*. [S.l.], 2008. p. 1–4.
- BENEDEK, C.; SZIRÁNYI, T. Change detection in optical aerial images by a multilayer conditional mixed markov model. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 47, n. 10, p. 3416–3430, 2009.
- BENEDEK, C.; SZIRÁNYI, T. *SZTAKI AirChange Benchmark set*. 2013. Disponível em: <http://mplab.sztaki.hu/remotesensing/airchange_benchmark.html>. Acesso em: 6 ago. 2023.
- BEVILACQUA, A.; STEFANO, L. D.; LANZA, A. A simple self-calibration method to infer a non-parametric model of the imaging system noise. In: IEEE. *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*. [S.l.], 2005. v. 2, p. 229–234.
- BILODEAU, G.-A.; JODOIN, J.-P.; SAUNIER, N. Change detection in feature space using local binary similarity patterns. In: IEEE. *2013 International conference on computer and robot vision*. [S.l.], 2013. p. 106–112.
- BOUWMANS, T.; JAVED, S.; SULTANA, M.; JUNG, S. K. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks*, Elsevier, v. 117, p. 8–66, 2019.
- BOVIK, A. C. *Handbook of Image and Video Processing*. [S.l.]: Academic Press, 2000.
- BRAHAM, M.; DROOGENBROECK, M. V. Deep background subtraction with scene-specific convolutional neural networks. In: IEEE. *2016 international conference on systems, signals and image processing (IWSSIP)*. [S.l.], 2016. p. 1–4.
- CHEN, H.; SHI, Z. *LEVIR-CD*. 2020. Disponível em: <<https://chenhao.in/LEVIR/>>. Acesso em: 14 ago. 2023.
- CHEN, H.; SHI, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, MDPI, v. 12, n. 10, 2020. Disponível em: <<https://doi.org/10.3390/rs12101662>>. Acesso em: 12 ago. 2023.

- CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 40, n. 4, p. 834–848, 2017.
- DAUDT, R. C.; SAUX, B. L.; BOULCH, A.; GOUSSEAU, Y. *Onera Satellite Change Detection Dataset*. 2018. Disponível em: <<https://rcdaudt.github.io/oscd/>>. Acesso em: 14 ago. 2023.
- DAUDT, R. C.; SAUX, B. L.; BOULCH, A.; GOUSSEAU, Y. Urban change detection for multispectral earth observation using convolutional neural networks. In: IEEE. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.], 2018. p. 2115–2118.
- DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd international conference on Machine learning*. [S.l.: s.n.], 2006. p. 233–240.
- DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, JMLR. org, v. 7, p. 1–30, 2006.
- DONG, Y.; DESOUZA, G. N. Adaptive learning of multi-subspace for foreground detection under illumination changes. *Computer Vision and Image Understanding*, Elsevier, v. 115, n. 1, p. 31–49, 2011.
- DUNNETT, C. W. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, Taylor & Francis, v. 50, n. 272, p. 1096–1121, 1955.
- FANG, K.; XIANG, Y.; LI, X.; SAVARESE, S. Recurrent autoregressive networks for online multi-object tracking. In: IEEE. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. [S.l.], 2018. p. 466–475.
- FENG, Y.; JIANG, J.; XU, H.; ZHENG, J. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 61, 2023. Disponível em: <<https://doi.org/10.1109/TGRS.2023.3241257>>. Acesso em: 12 ago. 2023.
- FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937.
- FUKUSHIMA, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, Springer, v. 36, n. 4, p. 193–202, 1980.
- GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 3. ed. Upper Saddle River, NJ, EUA: Prentice-Hall, Inc., 2006.
- GROSS, C. G.; ROCHA-MIRANDA, C. d.; BENDER, D. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of neurophysiology*, v. 35, n. 1, p. 96–111, 1972.

- HAN, J.; MORAGA, C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: SPRINGER. *International Workshop on Artificial Neural Networks*. [S.l.], 1995. p. 195–201.
- HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778.
- HOLSCHNEIDER, M.; KRONLAND-MARTINET, R.; MORLET, J.; TCHAMITCHIAN, P. A real-time algorithm for signal analysis with the help of the wavelet transform. In: SPRINGER. *Wavelets: Time-Frequency Methods and Phase Space Proceedings of the International Conference, Marseille, France, December 14–18, 1987*. [S.l.], 1990. p. 286–297.
- HUBEL, D. H.; WIESEL, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, Wiley Online Library, v. 160, n. 1, p. 106–154, 1962.
- HUBEL, D. H.; WIESEL, T. N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, Wiley Online Library, v. 195, n. 1, p. 215–243, 1968.
- IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- JAIN, A. K. *Fundamentals of Digital Image Processing*. USA: Prentice-Hall, Inc., 1989.
- JAVED, S.; JUNG, S. K.; MAHMOOD, A.; BOUWMANS, T. Motion-aware graph regularized rpca for background modeling of complex scenes. In: IEEE. *2016 23rd International Conference on Pattern Recognition (ICPR)*. [S.l.], 2016. p. 120–125.
- KAEWTRAKULPONG, P.; BOWDEN, R. An improved adaptive background mixture model for real-time tracking with shadow detection. In: *Video-based surveillance systems*. [S.l.]: Springer, 2002. p. 135–144.
- KIM, K.; CHALIDABHONGSE, T. H.; HARWOOD, D.; DAVIS, L. Real-time foreground–background segmentation using codebook model. *Real-time Imaging*, Elsevier, v. 11, n. 3, p. 172–185, 2005.
- KRAVITZ, D. J.; BEHRMANN, M. Space-, object-, and feature-based attention interact to organize visual scenes. *Attention, Perception, & Psychophysics*, Springer, v. 73, p. 2434–2447, 2011.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012.
- LABERGE, D.; BROWN, V. Theory of attentional operations in shape identification. *Psychological review*, American Psychological Association, v. 96, n. 1, p. 101, 1989.
- LALONDE, R.; XU, Z.; IRMAKCI, I.; JAIN, S.; BAGCI, U. Capsules for biomedical image segmentation. *Medical image analysis*, Elsevier, v. 68, p. 101889, 2021.

- LANZA, A.; STEFANO, L. D. Statistical change detection by the pool adjacent violators algorithm. *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 33, n. 9, p. 1894–1910, 2011.
- LECUN, Y. Generalization and network design strategies. In: *Connectionism in perspective*. [S.l.]: Elsevier, 1989.
- LECUN, Y.; BOSER, B. E.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W. E.; JACKEL, L. D. Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*. [S.l.: s.n.], 1990. p. 396–404.
- LIM, L. A.; KELES, H. Y. Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognition Letters*, v. 112, p. 256–262, 2018.
- LIM, L. A.; KELES, H. Y. Learning multi-scale features for foreground segmentation. *Pattern Analysis and Applications*, Springer, p. 1–12, 2019.
- LIN, C.; ZHANG, S.; YOU, S.; LIU, X.; ZHU, Z. Real-time foreground object segmentation networks using long and short skip connections. *Information Sciences*, Elsevier, v. 571, p. 543–559, 2021.
- LIN, M.; CHEN, Q.; YAN, S. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco: Common objects in context. In: SPRINGER. *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. [S.l.], 2014. p. 740–755.
- LIU, S.; DENG, W. Very deep convolutional neural network based image classification using small training sample size. In: IEEE. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. [S.l.], 2015. p. 730–734.
- LIU, W.; ZHANG, M.; LUO, Z.; CAI, Y. An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors. *IEEE Access*, IEEE, v. 5, p. 24417–24425, 2017.
- LU, C.; FENG, J.; CHEN, Y.; LIU, W.; LIN, Z.; YAN, S. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 42, n. 4, p. 925–938, 2019.
- MANDAL, M.; VIPPARTHI, S. K. An empirical review of deep learning frameworks for change detection: Model design, experimental frameworks, challenges and research needs. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 23, n. 7, p. 6101–6122, 2021.
- NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *ICML*. [S.l.: s.n.], 2010.
- NEMENYI, P. B. *Distribution-free Multiple Comparisons*. Tese (Doutorado) — Princeton University, 1963.

- PASZKE, A.; GROSS, S.; CHINTALA, S.; CHANAN, G.; YANG, E.; DEVITO, Z.; LIN, Z.; DESMAISON, A.; ANTIGA, L.; LERER, A. Automatic differentiation in pytorch. 2017.
- PIRES, R. G.; SANTOS, D. F.; SANTOS, C. F.; SANTANA, M. C.; PAPA, J. P. Image denoising using attention-residual convolutional neural networks. In: IEEE. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.], 2020. p. 101–107.
- POSNER, M. I. Orienting of attention. *Quarterly journal of experimental psychology*, SAGE Publications Sage UK: London, England, v. 32, n. 1, p. 3–25, 1980.
- RAHMON, G.; BUNYAK, F.; SEETHARAMAN, G.; PALANIAPPAN, K. Motion u-net: Multi-cue encoder-decoder network for motion segmentation. In: IEEE. *2020 25th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2021. p. 8125–8132.
- RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. [S.l.], 2015. p. 234–241.
- SABOUR, S.; TAGLIASACCHI, A.; YAZDANI, S.; HINTON, G.; FLEET, D. J. Unsupervised part representation by flow capsules. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2021. p. 9213–9223.
- SANTANA, M. C.; JÚNIOR, L. A. P.; MOREIRA, T. P.; COLOMBO, D.; ALBUQUERQUE, V. H. C. de; PAPA, J. P. A novel siamese-based approach for scene change detection with applications to obstructed routes in hazardous environments. *IEEE Intelligent Systems*, IEEE, v. 35, n. 1, p. 44–53, 2019.
- SANTOS, D. F.; PIRES, R. G.; COLOMBO, D.; PAPA, J. P. Video segmentation learning using cascade residual convolutional neural network. In: IEEE. *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.], 2019. p. 1–7.
- SANTOS, D. F.; PIRES, R. G.; PASSOS, L. A.; PAPA, J. P. Ddipnet and ddipnet+: Discriminant deep image prior networks for remote sensing image classification. In: IEEE. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. [S.l.], 2021. p. 2843–2846.
- SANTOS, D. F. S.; PAPA, J. P. Titan: A lightweight temporal attention network for remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 20, p. 1–5, 2023.
- SANTOS, D. F. S.; VALEM, L. P.; SANTANA, M. C. S. S.; PIRES, R. G.; JÚNIOR, L. A. P.; OLIVEIRA, D. d. S.; PEDRONETTE, D. C. G.; MARANA, A. N.; GUILHERME, I. R.; PAPA, J. P.; COLOMBO, D. *SiSuSMS - Sistema Supervisorio de SMS para Sondas*. São Paulo, 2021.
- SHERVATOV, V.; ARGUNOV, B. I.; SKORNYAKOV, L.; BOLTYANSKII, V. *Hyperbolic Functions: with Configuration Theorems and Equivalent and Equidecomposable Figures*. [S.l.]: Courier Corporation, 2013.

- SHI, Q.; LIU, M.; LI, S.; LIU, X.; WANG, F.; ZHANG, L. *SYSU-CD*. 2020. Disponível em: <<https://github.com/liumency/SYSU-CD>>. Acesso em: 14 ago. 2023.
- SHI, Q.; LIU, M.; LI, S.; LIU, X.; WANG, F.; ZHANG, L. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE transactions on geoscience and remote sensing*, IEEE, v. 60, p. 1–16, 2021.
- ST-CHARLES, P.-L.; BILODEAU, G.-A. Improving background subtraction using local binary similarity patterns. In: IEEE. *IEEE winter conference on applications of computer vision*. [S.l.], 2014. p. 509–515.
- ST-CHARLES, P.-L.; BILODEAU, G.-A.; BERGEVIN, R. Subsense: A universal change detection method with local adaptive sensitivity. *Transactions on Image Processing*, IEEE, v. 24, n. 1, p. 359–373, 2014.
- ST-CHARLES, P.-L.; BILODEAU, G.-A.; BERGEVIN, R. A self-adjusting approach to change detection based on background word consensus. In: IEEE. *2015 IEEE winter conference on applications of computer vision*. [S.l.], 2015. p. 990–997.
- STAUFFER, C.; GRIMSON, W. E. L. Adaptive background mixture models for real-time tracking. In: IEEE. *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*. [S.l.], 1999. v. 2, p. 246–252.
- SUDHAKARAN, S.; ESCALERA, S.; LANZ, O. Gate-shift networks for video action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 1102–1111.
- SULTANI, W.; CHEN, C.; SHAH, M. Real-world anomaly detection in surveillance videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 6479–6488.
- TARMIZI, I. A.; AZIZ, A. A. Vehicle detection using convolutional neural network for autonomous vehicles. In: IEEE. *2018 International Conference on Intelligent and Advanced System (ICIAS)*. [S.l.], 2018. p. 1–5.
- TOMPSON, J.; GOROSHIN, R.; JAIN, A.; LECUN, Y.; BREGLER, C. Efficient object localization using convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 648–656.
- ULYANOV, D.; VEDALDI, A.; LEMPITSKY, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in Neural Information Processing Systems*, v. 30, 2017.
- WANG, R.; BUNYAK, F.; SEETHARAMAN, G.; PALANIAPPAN, K. Static and moving object detection using flux tensor with split gaussian models. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2014. p. 414–418.

- WANG, Y.; JODOIN, P.-M.; PORIKLI, F.; KONRAD, J.; BENEZETH, Y.; ISHWAR, P. *ChangeDetection.Net (CDNET)*. 2012. Disponível em: <<http://changedetection.net>>. Acesso em: 6 ago. 2023.
- WANG, Y.; JODOIN, P.-M.; PORIKLI, F.; KONRAD, J.; BENEZETH, Y.; ISHWAR, P. Cdnet 2014: An expanded change detection benchmark dataset. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. [S.l.: s.n.], 2014. p. 387–394.
- WANG, Y.; LUO, Z.; JODOIN, P.-M. Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters*, Elsevier, v. 96, p. 66–75, 2017.
- WANG, Z.; PENG, C.; ZHANG, Y.; WANG, N.; LUO, L. Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss. *Neurocomputing*, Elsevier, v. 457, p. 155–167, 2021.
- WILCOXON, F. Individual comparisons by ranking methods. In: *Breakthroughs in Statistics: Methodology and Distribution*. [S.l.]: Springer, 1992. p. 196–202.
- WU, M.; PENG, X. Spatio-temporal context for codebook-based dynamic background subtraction. *AEU-International Journal of Electronics and Communications*, Elsevier, v. 64, n. 8, p. 739–747, 2010.
- YANG, L. W.; SU, C. Y. Low-cost cnn design for intelligent surveillance system. In: IEEE. *2018 International Conference on System Science and Engineering (ICSSE)*. [S.l.], 2018. p. 1–4.
- ZHANG, K.; ZUO, W.; CHEN, Y.; MENG, D.; ZHANG, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, IEEE, v. 26, n. 7, p. 3142–3155, 2017.
- ZHANG, M.; XU, G.; CHEN, K.; YAN, M.; SUN, X. Triplet-based semantic relation learning for aerial remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 16, n. 2, p. 266–270, 2018.
- ZHAO, X.; WANG, G.; HE, Z.; JIANG, H. A survey of moving object detection methods: A practical perspective. *Neurocomputing*, Elsevier, v. 503, p. 28–48, 2022.
- ZIVKOVIC, Z. Improved adaptive gaussian mixture model for background subtraction. In: IEEE. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. [S.l.], 2004. v. 2, p. 28–31.
- ZIVKOVIC, Z. Improved adaptive gaussian mixture model for background subtraction. In: IEEE. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. [S.l.], 2004. v. 2, p. 28–31.