

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

**UM MÉTODO DE LOCALIZAÇÃO PARA VEÍCULOS
SUBAQUÁTICOS BASEADO EM IMAGENS VISÍVEIS AÉREAS
E ACÚSTICAS SUBAQUÁTICAS.**

por

Matheus Machado dos Santos

Tese de Doutorado
em Modelagem Computacional

Rio Grande, 3 Agosto de 2022

Matheus Machado dos Santos

“Um Método para Localização de Veículos Subaquáticos Baseado em Imagens Visíveis Aéreas e Acústicas Subaquáticas”

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande - FURG, como requisito parcial para obtenção do Grau de Doutor. Área de concentração: Modelagem Computacional.

Aprovado em: 03/08/2022

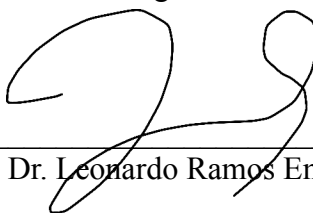
BANCA EXAMINADORA



Profa. Dra. Silvia Silva da Costa Botelho (Orientadora – FURG)



Prof. Dr. Paulo Lilles Jorge Drews Junior (Coorientador – FURG)



Prof. Dr. Leonardo Ramos Emmendorfer (UFSM)



Prof. Dr. Renan de Queiroz Maffei (UFRGS)



Prof. Dr. Rodrigo da Silva Guerra (FURG)

Ficha Catalográfica

S237m Santos, Matheus Machado dos.

Um método para localização de veículos subaquáticos baseado em imagens visíveis aéreas e acústicas subaquáticas / Matheus Machado dos Santos. – 2022.

100 f.

Tese (doutorado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Modelagem Computacional, Rio Grande/RS, 2022.

Orientadora: Dra. Silvia Silva da Costa Botelho.

Coorientador: Dr. Paulo Lilles Jorge Drews Junior.

1. Fusão Sensorial Multimodal 2. Localização Subaquática
3. Redes Neurais Convolucionais I. Botelho, Silvia Silva da Costa
II. Drews Junior, Paulo Lilles Jorge III. Título.

CDU 681.5.01

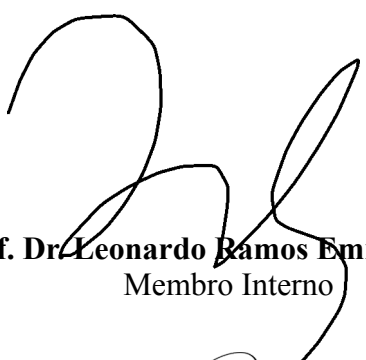
Catálogo na Fonte: Bibliotecário José Paulo dos Santos CRB 10/2344

ATA DE DEFESA DE TESE**ATA Nº 7/2022**

Ao terceiro dia do mês de agosto de dois mil e vinte e dois, através de videoconferência, realizou-se a 17ª sessão de defesa de tese do Programa de Pós-Graduação em Modelagem Computacional. O doutorando **Matheus Machado dos Santos** apresentou e defendeu a tese “**Um Método para Localização de Veículos Subaquáticos Baseado em Imagens Visíveis Aéreas e Acústicas Subaquáticas**” como requisito parcial para a obtenção do grau de Doutor em Modelagem Computacional. A sessão foi aberta às nove horas pela Profa. Dra. Silvia Silva da Costa Botelho, orientadora da tese e presidente da Comissão de Avaliação. A referida Comissão esteve integrada, além da presidente e orientadora, pelo Prof. Dr. Paulo Lilles Jorge Drews Junior da FURG (coorientador), Prof. Dr. Leonardo Ramos Emmendorfer da UFSM (docente do PPGMC/FURG), Prof. Dr. Renan de Queiroz Maffei da UFRGS e Prof. Dr. Rodrigo da Silva Guerra da FURG. Após cada integrante da Banca ter arguido o doutorando, e este ter tido a oportunidade de responder a cada um, os examinadores atribuíram os seguintes conceitos ao trabalho:

Prof. Dr. Leonardo Ramos Emmendorfer: conceito AProf. Dr. Renan de Queiroz Maffei: conceito AProf. Dr. Rodrigo da Silva Guerra: conceito A

Tendo sido o candidato APROVADO, com conceito A, neste requisito parcial para a obtenção do grau de Doutor em Modelagem Computacional.


Prof. Dr. Leonardo Ramos Emmendorfer
Membro Interno
Prof. Dr. Renan de Queiroz Maffei
Membro Externo
Prof. Dr. Rodrigo da Silva Guerra
Membro Interno
Profa. Dra. Silvia Silva da Costa Botelho
Orientadora
Prof. Dr. Paulo Lilles Jorge Drews Junior
Coorientador

AGRADECIMENTOS

Gostaria de agradecer à minha orientadora Prof. Dra. Silvia Botelho e o meu coorientador, Prof. Dr. Paulo Drews-Jr por aceitarem serem meu guia nesse trabalho, sempre ajudando, mesmo nos momentos mais críticos e mais difíceis. Sempre me convencendo a superar os meus limites de forma a alcançar os maiores objetivos. Muito obrigado por todas as oportunidades que vocês proporcionaram, me permitindo adquirir experiência internacional e conhecer importantes centros de pesquisa e grandes pesquisadores ao redor do mundo.

Agradeço ao Eng. de Comp. e Pesquisador Giovanni de Giacomo, por aceitar realizar seu trabalho de iniciação científica junto comigo, apoiando esse projeto, sempre dando bons conselhos, incentivando e contribuindo para que esse trabalho atingisse os seus objetivos. Muito obrigado Giovanni.

Agradeço a minha família pelo apoio constante na realização deste trabalho, sempre proporcionando as melhores condições e orientando pelos melhores caminhos em especial, meu pai Itamar e mãe Marta. Um agradecimento especial a minha irmã Mônica, na qual convivi dividindo a mesma morada durante toda realização deste trabalho, muito obrigado pela paciência e por todo apoio durante essa trajetória. Também agradeço a todos os amigos, pela paciência e por todos incentivos que deram forças para concluir este trabalho.

Agradeço aos membros da banca, Prof. Dr. Leonardo Ramos Emmendorfer, Prof. Dr. Rodrigo da Silva Guerra e Prof. Dr. Renan de Queiroz Maffei por aceitarem compartilharem o seu tempo para revisar e avaliar este trabalho, e por todas críticas e sugestões que contribuíram de forma positiva para a melhoria deste trabalho. Também agradeço ao Pesquisador Dr. Nuno Gracias pelo seu interesse e por todas as suas contribuições e diretivas de pesquisa durante a sua participação como membro da banca de qualificação deste trabalho.

Finalmente, gostaria de agradecer a Universidade Federal do Rio Grande (FURG), ao Centro de Ciências Computacionais (C3), ao laboratório NAUTEC, ao Programa de Pós Graduação em Modelagem Computacional (PPGMC) por fornecerem toda infraestrutura e as condições necessárias para realização deste trabalho. Em especial, gostaria de agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento desta pesquisa por tempo integral através de bolsas de pesquisa.

RESUMO

Localização subaquática é um problema desafiador devido às limitações sensoriais impostas pelo ambiente aquoso. Entre elas, o uso de Sistema de Posicionamento Global (GPS) e de sensores ópticos são severamente limitados. Apesar de sonares multi-feixe serem eficientes, suas imagens são de difícil interpretação. Novas tecnologias de sensoriamento remoto por drones ou satélite facilitaram o acesso a imagens aéreas no espectro visível e de alta resolução, permitindo um melhor entendimento da superfície terrestre e de regiões costeiras e portuárias. Estes fatores motivaram a proposta do primeiro método multi-domínio e multi-perspectiva de estimativa de localização subaquática para regiões parcialmente estruturadas, com estruturas parcialmente submersas, como portos e marinas. A localização é estimada utilizando a odometria do veículo, imagens acústicas subaquáticas, e imagens ópticas aéreas, através da percepção das estruturas parcialmente submersas no ambiente. As imagens aéreas são utilizadas como referência para localização, e são carregadas previamente no veículo subaquático. São apresentados os principais desafios para o tratamento das imagens de cada domínio, e os desafios de encontrar as suas relações. O método proposto combina Modelos Probabilísticos e Redes Neurais Convolucionais para estimar a similaridade entre as imagens dos dois domínios, explorando estruturas estacionárias como píeres, paredes e linhas litorâneas. A similaridade é utilizada na etapa de atualização do peso das partículas do algoritmo de localização de Monte Carlo. O sistema é avaliado utilizando dados simulados e reais, capturados em uma marina. Os resultados obtidos mostram a redução do erro de localização acumulado durante a navegação subaquática, representando uma alternativa viável para navegação subaquática autônoma em portos e marinas.

Palavras-chaves: Fusão Sensorial Multimodal, Localização Subaquática, Redes Neurais Convolucionais

ABSTRACT

Underwater localization is a challenging problem due to the sensory limitations imposed by the environment. Among them, the use of a Global Positioning System (GPS) and optical sensors are severely limited. Although multibeam sonars are effective, their images are difficult to interpret. New technologies for remote sensing by drones or satellites have facilitated access to aerial images in the visible and high-resolution spectrum, allowing a better understanding of the earth's surface and coastal and port regions. These facts motivated the proposal of the first multi-domain and multi-perspective method for underwater localization in partially structured regions, with partially submerged structures, as found in ports and marinas. The localization is estimated using vehicle odometry, underwater acoustic images, and aerial optical images, through the perception of partially submerged structures in the environment. The aerial images are previously loaded into the underwater vehicle and adopted as a reference for the localization. The main challenges for the treatment of images in each domain are presented, as well as the challenges of finding their relationships. The proposed method combines Probabilistic Models and Convolutional Neural Networks to estimate the similarity between the images of the two domains, exploring stationary structures such as piers, walls and shorelines. The similarity updates the particles' weight of the Monte Carlo localization algorithm. The system is evaluated using simulated and real data of a marina. The results show a reduction in the localization error accumulated during underwater navigation, representing a viable alternative for autonomous underwater navigation in ports and marinas.

Palavras-chaves: Multimodal Sensor Fusion, Underwater Localization, Convolutional Neural Networks

ÍNDICE

1	INTRODUÇÃO	18
1.1	Descrição do Problema	22
1.2	Hipótese Científica	26
1.3	Objetivos Gerais	26
1.4	Objetivos Específicos	27
1.5	Aplicações	27
1.6	Contribuições	27
1.7	Organização do Texto	29
2	TRABALHOS RELACIONADOS	30
2.1	Localização Terrestre com Base em Imagens Aéreas	31
2.1.1	Métodos de Localização de Imagens Terrestres Baseados em Características Similares em Diferentes Domínios	31
2.1.2	Métodos de Localização de Imagens Terrestres Baseado em Imagens Semânticas	32
2.1.3	Métodos de Localização de Imagens Terrestres Baseados em Aprendizagem Profundo	34
2.1.4	Localização de Veículos Terrestres com Base em Imagens Aéreas	38
2.1.5	Considerações aos Trabalhos Relacionados a Localização e Correspondência de Imagens Terrestres e Aéreas	41
2.2	Segmentação Semântica de Imagens Aéreas	41
2.2.1	Trabalhos relacionados a Segmentação de Imagens Aéreas no Espectro Visível RGB em Áreas Costeiras e Portuárias	43
2.2.2	Considerações Relacionadas aos Trabalhos de Segmentação de Imagens Aéreas	44
2.3	Melhoramento de Imagens Acústicas	44
2.3.1	Considerações aos Trabalhos Relacionados ao Melhoramento de Imagens Acústicas	49
2.4	Considerações Finais	52
3	METODOLOGIA	53
3.1	Etapa A: Processamento de Imagens Aéreas	55
3.2	Etapa B: Processamento de Imagens Acústicas	58
3.3	Etapa C: Comparação de Imagens Multi-Domínio e Multi-Perspectiva	60
3.3.1	Modelo de Observação Baseado em Campos de Probabilidade	61
3.3.2	Modelo de Observação Baseado em Aprendizagem Profunda	64
3.4	Etapa D: Estimativa de Localização	67
3.4.1	Atualização do Estado das Partículas - Modelo de Transição	67
3.4.2	Atualização dos pesos das Partículas - Modelo de Observação	68
3.4.3	Re-amostragem das Partículas	69
4	RESULTADOS	70

4.1	Conjunto de Dados	70
4.1.1	Imagens Acústicas	70
4.1.2	Imagens Aéreas	70
4.2	Segmentação Semântica de Imagens Aéreas	71
4.3	Localização Subaquática com Dados Acústicos Simulados e Imagens Aéreas Reais	78
4.4	Localização Subaquática com Imagens Aéreas e Acústicas Reais	82
5	CONCLUSÕES	89

LISTA DE FIGURAS

Figura 1.1: Imagens ópticas subaquáticas	19
Figura 1.2: Imagem acústica de um sonar multibeam de imageamento frontal.	20
Figura 1.3: Imagens Aéreas	22
Figura 1.4: Representação do problema.	23
Figura 1.5: Desafios da segmentação semântica de imagens aéreas RGB em marinas	25
Figura 1.6: Problema de Localização Subaquática	26
Figura 1.7: Diagrama proposto para lidar com o problema.	28
Figura 2.1: Diagrama de trabalhos correlacionados.	30
Figura 2.2: Etapas do método proposto por Wolff; Collins; Liu	32
Figura 2.3: Correspondência de imagens terrestres e aéreas proposto por CASTALDO et al.	33
Figura 2.4: Métodos propostos por LIN; BELONGIE; HAYS	34
Figura 2.5: Método proposto por LIN et al.	36
Figura 2.6: Etapas do método proposto por TIAN; CHEN; SHAH	37
Figura 2.7: Método proposto por LEUNG; CLARK; HUISSOON	38
Figura 2.8: Método de localização <i>indoor</i> proposto por KARKUS; HSU; LEE	40
Figura 3.1: Pipeline do método proposto.	53
Figura 3.2: Desafios de localização em ambientes dinâmicos.	56
Figura 3.3: Arquitetura da RNC de segmentação semântica.	57
Figura 3.4: Composição de segmentações semântica.	58
Figura 3.5: Transformação Imagens Acústicas em Distâncias.	59
Figura 3.6: Parâmetros do sistema de transformação de imagens acústicas em distância.	60
Figura 3.7: Metodologia proposta.	61
Figura 3.8: Contornos de estruturas estacionárias.	62
Figura 3.9: Transformação da segmentação semântica em campos de distância.	63
Figura 3.10: Pipeline do Modelo de Observação baseado em Aprendizagem Profunda.	65
Figura 3.11: Exemplo de <i>Triplets</i>	66
Figura 3.12: Exemplo de <i>Quadruplet</i>	66
Figura 3.13: Diagrama de Execução do Filtro de Partículas.	67
Figura 4.1: Conjunto de dados ARACATI 2017	71
Figura 4.2: Imagens aéreas Yacht Clube de Rio Grande.	71
Figura 4.3: Imagens aéreas do Japão e da Austrália.	72
Figura 4.4: Resultados qualitativos da segmentação semântica de imagens de satélite com <i>ground truth</i>	74
Figura 4.5: Resultados qualitativos da segmentação semântica de imagens de satélite sem <i>ground truth</i>	75
Figura 4.6: Cenário de simulação do Yacht Clube de Rio Grande.	79
Figura 4.7: Conversão de uma imagem de satélite em mapa.	80

Figura 4.8: Desempenho do método de localização proposto.	80
Figura 4.9: Highlights do experimento simulado de localização.	81
Figura 4.10: Trajetória estimada pelo método de localização.	84
Figura 4.11: Erro de localização em metros no experimento real	85
Figura 4.12: Mapa de Mínimas Distâncias vs Imagem Aérea Segmentada	87

LISTA DE TABELAS

Tabela 1:	Trabalhos Relacionados a Segmentação Semântica de Imagens Aéreas . . .	44
Tabela 2:	Relação de trabalhos com aplicações utilizando FLS	51
Tabela 3:	Modelo de Transição e o algoritmo de atualização das partículas.	68
Tabela 4:	Matriz de Confusão Aumentada	76
Tabela 5:	Resultados Estatísticos do Método de Segmentação Semântico	77
Tabela 6:	Valor dos parâmetros adotados no experimento	83

LISTA DE SÍMBOLOS E VARIÁVEIS

- α_1 - Especifica o ruído esperado na estimativa de rotação relacionado a componente rotacional do movimento do robô, este é um parâmetro utilizado pelo modelo de transição do algoritmo de filtro de partículas. 68
- α_2 - Especifica o ruído esperado na estimativa de rotação relacionado a componente translacional do movimento do robô, este é um parâmetro utilizado pelo modelo de transição do algoritmo de filtro de partículas. 68
- α_3 - Especifica o ruído esperado na estimativa de translação relacionado a componente translacional do movimento do robô, este é um parâmetro utilizado pelo modelo de transição do algoritmo de filtro de partículas. 68
- α_4 - Especifica o ruído esperado na estimativa de translação da odometria do componente rotacional do movimento do robô, este é um parâmetro utilizado pelo modelo de transição do algoritmo de filtro de partículas. 68
- α_5 - Especifica o ruído esperado relacionado à translação transversal do veículo (deslocamento deslocamento omnidirecional), este é um parâmetro utilizado pelo modelo de transição do algoritmo de filtro de partículas. 68
- B - Representa um *beam* acústico do sonar. Cada sonar possui um número fixo de *beams* que depende da sua fabricação, e da abertura do campo de visão. B_t^k indica a leitura do *beam* k no instante t . Um *beam* é composto por diversas leituras discretizadas em *bins* percebidas ao longo da mesma direção.. 58
- b - Representa um *bin* acústico do sonar. Uma intensidade de sinal acústico refletida pelo ambiente e detectada pelo sonar em uma determinada direção e distância. $b_{k,l}$ representa a intensidade acústica capturada na direção do *beam* k , pelo *bin* l . 59
- d_{max} - Parâmetro que indica a distância máxima até um objeto para ser considerada no mapa de mínimas distâncias.. 62
- δ_θ - Rotação estimada do veículo a partir do sinal de controle ou odometria. 68
- δ_{trans} - Deslocamento estimado do veículo a partir do sinal de controle ou odometria. 68
- f - Grau de similaridade entre uma observação realizada pelo veículo e o ambiente de navegação. A similaridade é calculada com relação ao estado de uma partícula k , e uma observação o , realizada no instante t . Por esse motivo a variável é exibida da forma f_t^k indicando a similaridade da observação no instante t para a partícula k com relação ao mapa de navegação $\mathbb{M}$. 54, 55

- I^a - Imagem óptica aérea de drone ou satélite. 53, 54
- $\hat{I}^a$ - Representa uma imagem aérea segmentada semanticamente nas três categorias: objetos estáticos, objetos móveis e água. 61
- I^s - Imagem acústica subaquática. A variável é utilizada em conjunto com t da forma I_t^s , onde t indica o instante de captura da imagem pelo sonar do veículo. 53, 55, 58, 79
- K - Número total de partículas do algoritmo de filtro de partículas. 53
- $\mathbb{M}$ - Representação do mapa do ambiente de navegação originado a partir das estruturas estacionárias das imagens aéreas. 53–55, 61, 64, 67, 68, 83
- u_t - Estimativa de deslocamento no instante t a partir do sinal de controle dos propulsores, ou de um sistema de odometria visual. 53, 54, 67
- o - Observação realizada pelo veículo, no contexto deste trabalho a observação é uma imagem acústica posteriormente processada e convertida em distâncias relativa ao obstáculo mais próximo do sonar. 54, 55, 58, 67, 68, 79
- s - Estado de uma partícula representado em três graus de liberdade sendo posição no plano XoY e orientação θ . 53, 54
- σ_θ - Desvio padrão esperado na estimativa de rotação do modelo de transição do filtro de partículas. 68
- σ_{trans} - Desvio padrão esperado na estimativa de translação do modelo de transição do filtro de partículas. 68
- t - Representa um instante de tempo, geralmente utilizado com outra variável para associação temporal de dados. 53–55
- w - Peso ou fator de importância associado a uma partícula. Geralmente associado a variável k de forma que w_k indica o peso da partícula k . 53, 54
- X_t - Representação do estado do veículo no instante t utilizando 3 graus sendo posição no plano XoY e orientação θ . 53

LISTA DE ABREVIATURAS

- AMCL** algoritmo de localização adaptativo de Monte Carlo, do inglês *Adaptive Monte Carlo Localization* (AMCL). Nesta versão do algoritmo de Monte Carlo ocorre uma adaptação dinâmica do número de partículas empregado de acordo com a confiança na estimativa do estado do veículo.. 67, 78, 80, 81
- AUV** robô subaquático autônomo, do inglês *Autonomous Underwater Robot*. 18, 22, 46, 91
- CPU** Unidade de Processamento Computacional, do inglês *Computer Processing Unit*, processador. 86
- DGPS** Do inglês *Differential Global Positioning System*, é um sistema de posicionamento global (GPS) capaz de obter precisões superiores ao do GPS tradicional, por utilizar múltiplos GPS para a correção do erro de posição. 70, 71, 82
- DIDSON** do inglês *Dual-frequency, IDentification SONar*. 45–47, 50
- DoF** Graus de liberdade de um sistema, do inglês *Degrees of Freedom*. 54
- DVL** do inglês *Doppler Velocity Log*. 21, 46, 48, 49, 53
- EKF** Filtro de Kalman Estendido do inglês *Extended Kalman Filter*. 46
- FLS** Sonar de imageamento frontal do inglês *Forward Looking Sonar*. 19, 20, 23–26, 45–48, 61, 70, 79, 86, 87, 89, 91
- GPS** Sistema de Localização Global, do inglês *Global Positioning System*. 20, 89
- GPU** Unidade de Processamento Gráfica, do inglês *Graphics Processing Unit*. 85, 86
- ILSVRC** Desafio de reconhecimento de imagens em larga escala, do inglês *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC). 42
- IMU** Unidade de Medição Inercial, do Inglês *Inertial Measurement Unit* (IMU), sensor de aceleração, velocidade angular e muitas vezes campo magnético e pressão atmosférica. 48, 49
- ISPRS** Sociedade Internacional de Fotometria e Sensoriamento Remoto, do inglês *International Society for Photogrammetry and Remote Sensing* (ISPRS). 42
- K-NN** K-vizinhos mais próximo, do inglês *K-nearest neighbor* (K-NN). 33, 35

- LBL** do inglês *Long Baseline*. 20
- LBV** Pequeno veículo subaquático, do inglês *Little Benthic Vehicle*. 70
- LSH** Hash Sensível a Localidade, do inglês *locality-sensitive hashing* (LSH), método de clusterização de hashes. 49
- MCL** algoritmo de localização de Monte Carlo, do inglês *Monte Carlo Localization* (MCL). 21, 27, 55, 91
- MSIS** do inglês *Mechanical Scanning Image sonar*. 19, 46
- NDT** do inglês *Normal Distribution Transform*. 47
- NDWI** índice de água com diferença normalizada, do inglês *normalized difference water index* (NDWI). 42
- NIR** Banda próxima a radiação infravermelho, do inglês *near-infrared radiation* (NIR). 42
- RANSAC** do inglês *Random Sample Consensus*. 45
- RGB** Padrão de representação de cores em imagens digitais constituído por vermelho, verde e azul, do inglês *Red Green Blue*. 10, 24–26, 41–44, 57
- RNC** Rede Neural Convolucional. Também conhecido do inglês *Convolutional Neural Network* (CNN). 10, 42, 43, 55, 56, 71, 78, 79, 90
- ROS** *Middleware* Sistema Operacional de Robos, do inglês *Robot Operating System* (ROS). 67, 78
- ROV** robô operado remotamente, do inglês *Remotely Operated Vehicle*. 22, 54, 90
- RPN** Rede de Propostas de Região, do inglês *Region Proposal Network*. 44
- SAR** Radar de abertura sintética, do inglês *synthetic aperture radar*. 42
- SIFT** do inglês *Scale-invariant feature transform*. 47
- SIG** Sistema de Informação Geográfica. Também conhecido como *Geographic Information System* (GIS). 21, 32
- SLAM** do inglês *Simultaneous Localization and Mapping*. 46
- SURF** do inglês *Speeded up robust features*, (BAY; TUYTELAARS; GOOL, 2006; BAY et al., 2008). 39
- SVD** Decomposição em valor singular, do inglês *Singular Value Decomposition*. 45
- SVM** Uma máquina de vetores de suporte (SVM, do inglês: *support-vector machine*). 33, 89

UAUV Veículo Aéreo Subaquático não tripulado, do inglês *Unmanned Aerial-Underwater Vehicle*. 22

USBL Sistema de localização subaquático baseado em comunicação acústica mais compacto que o SBL, do inglês *Ultra Short Baseline*. 20, 70

1 INTRODUÇÃO

Os ambientes subaquáticos compõem um ecossistema complexo e importante para a vida no planeta, tornando o seu estudo relevante tanto para ciência quanto para economia. No entanto, apenas uma pequena parte desse domínio foi explorado (GOONESEKERE; GUO, 2018). Tendo em vista o seu caráter inóspito e sua extensão, exige-se a construção de engenhos e tecnologias capazes de tratar seus desafios e restrições. Entre as possibilidades de exploração subaquática, os Veículos Subaquáticos Autônomos (AUVs) se destacam como uma forma eficaz e segura de intervenção. Esses veículos requerem algoritmos sofisticados capazes de oferecer condições de percepção, tomada de decisões e atuações autônomas reduzindo os riscos ao ambiente e ao ser humano.

A navegação subaquática autônoma, dependendo da sua aplicação e das suas restrições, continua sendo um problema em aberto e desafiador (WU et al., 2019b; CHAME; SANTOS; BOTELHO, 2018; SANTOS et al., 2019b; JALAL; NASIR, 2021). Esse fato se deve à natureza dos ambientes subaquáticos, que atuam como limitadores da comunicação e da percepção sensorial que são essenciais para a navegação autônoma.

No contexto da percepção subaquática para navegação autônoma surgem questões importantes relacionadas à localização do veículo na cena, bem como dos seus elementos constituintes. Para estas tarefas o uso de sensores ganham destaque.

ROMERO et al.(2014) e MURPHY(2019) classificam os sensores em proprioceptivo, quando coletam informações internas do veículo; exteroceptivos, quando percebem informações sobre o ambiente, e exproprioceptivos, quando providenciam informações do veículo relativas ao ambiente. Esta classificação será utilizada para apresentar os principais sensores subaquáticos, as suas vantagens e desvantagens.

Sensores Proprioceptivos

Os sensores proprioceptivos como giroscópios, acelerômetros, e de velocidade dos propulsores do veículo são fundamentais para a localização subaquática. A sua utilização é vantajosa pela sua operação não depender das condições externas do meio aquoso, e também pelo fato da navegação subaquática ser lenta e suave, com movimentos que raramente excedem os limites de medição dos sensores, reduzindo as chances de medições erradas. Apesar de suas vantagens, os erros de medições continuam acontecendo devido a variações de temperatura, problemas com calibração e devido à qualidade dos sensores utilizados (MILLER; MILLER; MILLER, 2021).

Os dados proprioceptivos são utilizados em modelos cinemáticos e dinâmicos que a partir de uma posição inicial e uma sequência de medições de velocidade, aceleração e orientação, conseguem manter uma estimativa da localização do veículo. Esses métodos, também conhecidos como *dead reckoning*, são essenciais para a utilização dos dados proprioceptivos. No entanto, eles não são suficientes para localizar um veículo por um longo período de tempo devido ao acúmulo

de erro, sendo necessária a sua correção através da percepção do ambiente, utilizando sensores exteroceptivos (CORKE et al., 2007).

Sensores Exteroceptivos

A percepção de ambientes subaquáticos é desafiadora. A Figura 1.1 retrata uma situação tipicamente encontrada em portos e marinas. É possível observar que muitas vezes as condições da água limitam a visibilidade, ou até impossibilitam a utilização de sensores exteroceptivos ópticos como lasers e câmeras (CODEVILLA et al., 2015).

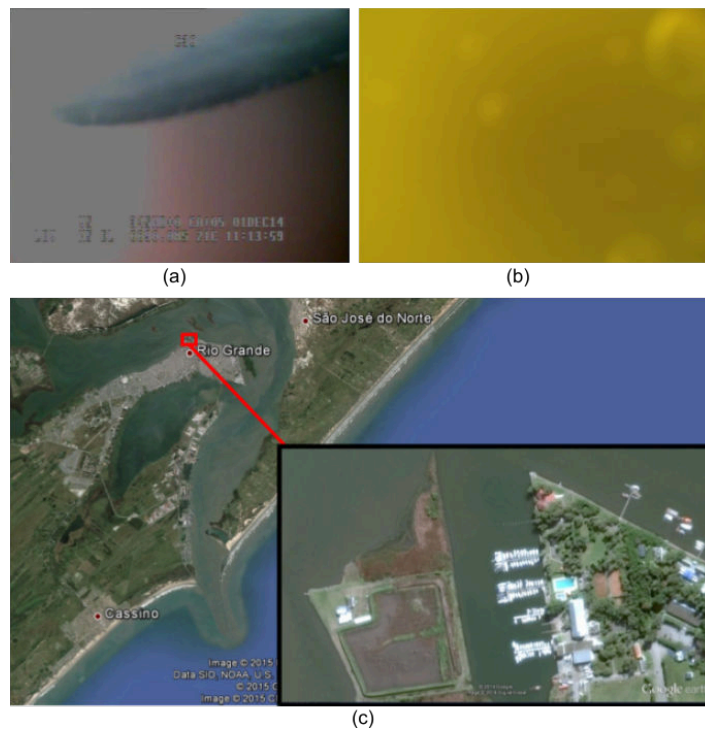


Figura 1.1: Imagens ópticas subaquáticas em regiões com baixa condição de visibilidade. Essas características geralmente são encontradas em regiões portuárias e marinas. (a) Imagem de um remo capturada por uma câmera subaquática analógica. (b) Imagem em resolução 4k capturada por uma câmera GoPro. (c) Localização da marina Yacht Clube de Rio Grande, onde foram capturadas as imagens (a) e (b), e representa uma região de estudo deste trabalho.

Por outro lado, os sensores exteroceptivos que utilizam o sinal acústico são eficientes e capazes de cobrir uma ampla área do ambiente, conforme demonstrado na Figura 1.2. Este fato se deve pela eficiência da propagação das ondas sonoras na água. Dentre os principais sensores acústicos pode-se citar os sonares de escaneamento mecânicos, do inglês *Mechanically Scanned Imaging Sonar (MSIS)*, os sonares de varredura lateral, do inglês *Side Scan Sonar*, e os sonares multi feixe de imageamento frontal, do inglês *Forward Looking Sonar (FLS)* que são capazes de medir distância e fornecer imagens do ambiente.

Apesar de suas vantagens, que os levam a serem amplamente utilizados na navegação autônoma, existem algumas limitações que precisam ser consideradas. Como a grande maioria dos sensores exteroceptivos são ativos, quando combinados, podem gerar interferência mútua. As imagens

acústicas capturadas pelo sonar são ruidosas, de baixa resolução e de difícil interpretação quando comparadas às imagens ópticas (HURTOS et al., 2013; SANTOS et al., 2017). Por exemplo, a Figura 1.2 mostra uma imagem capturada por um sonar de imageamento frontal (FLS). A imagem possui apenas um canal que representa a intensidade do retorno acústico. Essas informações se resumem a formas geométricas e distâncias dos objetos imageados na marina. Por esse motivo, muitas das técnicas convencionais para o tratamento de imagens coloridas adquiridas por sensores ópticos não se aplicam a imagens acústicas (HURTOS et al., 2013), sendo necessário o desenvolvimento de métodos específicos para imagens acústicas.

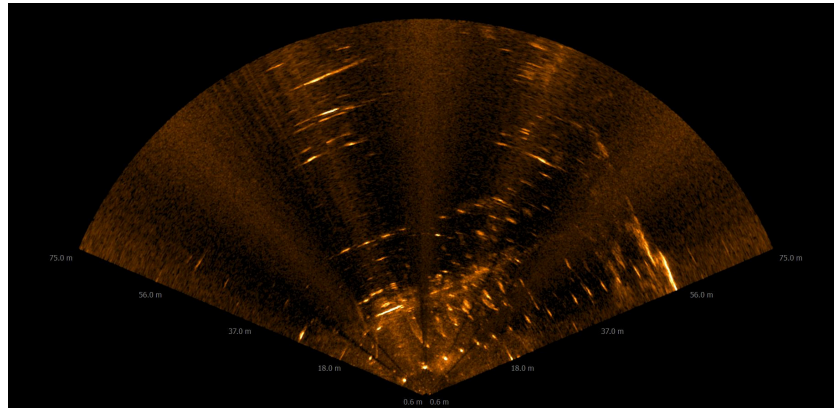


Figura 1.2: Imagem acústica capturada na mesma marina da Figura 1.1. A imagem cobre uma região de 75 metros à frente do sonar. É possível identificar as estruturas que sustentam os píeres da marina. Entretanto as informações são limitadas a formas geométricas representadas por um único canal.

Sensores Exproprioceptivos

A bússola ou magnetômetro é um sensor exproprioceptivo indispensável na navegação subaquática. Sua principal função é a correção da orientação do veículo. Apesar dos magnetômetros serem confiáveis, eles podem sofrer distúrbios magnéticos gerados por estruturas metálicas próximas ao veículo. Os sistemas de localização GPS também são sensores exproprioceptivos utilizados para estimar a localização, porém devido à rápida atenuação das ondas eletromagnéticas pela água, o seu uso é restrito apenas quando o veículo encontra-se na superfície da água. Para compensar essa limitação, foram desenvolvidos sensores exproprioceptivos de localização subaquática como o *Long Baseline* (LBL), e o *Ultra-Short Baseline* (USBL) que utilizam comunicação acústica entre transdutores para estimar a localização (MILLER; MILLER; MILLER, 2021). Entretanto, possuem limitações quanto à sua área de cobertura, e requerem procedimentos de instalação dos transdutores, exigindo muitas vezes uma embarcação de suporte para a sua operação. Além disso, possuem dificuldades para operar em águas muito rasas, como as encontradas em portos e marinas.

Outro sensor exproprioceptivo fundamental para navegação subaquática é o sensor de pressão, utilizado para medir a profundidade do veículo. Na maioria das aplicações este sensor é suficiente para a obtenção desta estimativa. Não possui limitações consideráveis como os outros

sensores, e geralmente a principal fonte de erro de medição é provocada pela calibração errada do sensor.

Entretanto, apenas a estimativa da profundidade não é suficiente para algumas aplicações que coletam dados do fundo marinho e precisam manter o veículo a uma distância fixa com relação a ele, independente do seu relevo. Para isso são utilizados os ecobatímetros, outro tipo de sensor exproprioceptivo ativo que mede a altitude do veículo através da emissão de ondas acústicas. O sensor de velocidade conhecido como *Doppler Velocity Log* (DVL) também é muito utilizado tanto para medir a altitude quanto a velocidade do veículo relativa ao fundo marinho, ou a velocidade relativa à água. A principal limitação dos ecobatímetros e DVLs é o alcance, limitando a utilização para apenas quando o veículo está próximo ao fundo. Existe uma relação de ganho e perda envolvendo a precisão e qualidade dos dados, com o alcance máximo do sensor.

Localização Subaquática

Devido às dificuldades e limitações de percepção dos ambientes subaquáticos, torna-se necessária a utilização de técnicas de fusão sensoriais que combinam informações proprioceptivas e exteroceptivas, explorando as vantagens e superando as limitações individuais de cada sensor (MILLER; MILLER; MILLER, 2021). Grande parte das técnicas de fusão sensorial realizam a inferência do próximo estado do veículo com base na confiança das medições sensoriais atuais, geralmente representada através de uma matriz de covariância. Os métodos mais conhecidos são baseados no método de localização de Markov (THRUN; BURGARD; FOX, 2006), no filtro de Kalman Estendido (EKF) (MURPHY, 2019), ou no método de localização de Monte Carlo, do inglês - *Monte Carlo Localization* MCL – também conhecido como filtro de partículas (THRUN; BURGARD; FOX, 2006).

Este trabalho busca explorar um método de localização que combina informações que vão além do domínio subaquático. A facilidade de acessar imagens aéreas no espectro visível e de alta resolução aumentou ao longo dos anos graças à popularização dos drones, das imagens de satélites, e do desenvolvimento de Sistemas de Informações Geográficas (SIG). Conforme apresentado na Figura 1.3, essas imagens são ricas em informações da superfície terrestre e vêm sendo utilizadas com sucesso para auxiliar na localização de veículos terrestres (GAO et al., 2018). Os bons resultados obtidos pelo uso de imagens aéreas em sistemas integrados para localização de veículos terrestres em missões *outdoor* (GAO et al., 2018), inspira a possibilidade de seu uso em outras classes de problemas.

Motivado por esses resultados e pelos desafios da localização subaquática, surge de forma original neste trabalho o interesse em estudar a combinação de imagens aéreas e informações acústicas subaquáticas para auxiliar na localização de veículos submersos.

Admitindo-se que em determinados tipos de tarefas, próximas a portos e marinas, dispõe-se de imagens acústicas frontais do cenário, e vistas superiores de imagens aéreas da região de

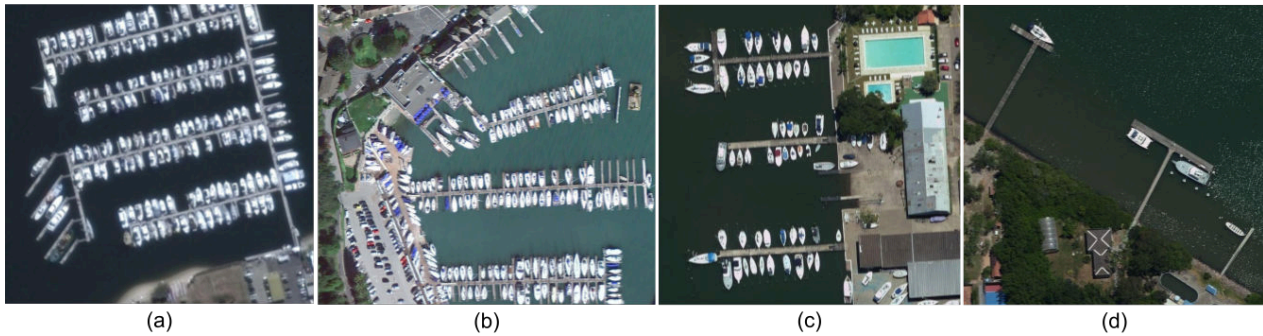


Figura 1.3: Exemplo de imagens ópticas aéreas. (a) e (b)- Imagens capturadas por satélite com resolução espacial de 50 cm. (c) e (d) - Imagens aéreas capturadas por drone com resolução espacial de 10 cm (GONÇALVES, 2020). Ambas as imagens de drone e de satélite são similares, porém geralmente as imagens de drones possuem maior resolução.

operação, este trabalho propõe um método de localização baseado em imagens multi-domínio (imagens acústicas e imagens ópticas) e multi-perspectiva (vista sensorial frontal e superior) para ambientes subaquáticos parcialmente estruturados e com estruturas parcialmente submersas. A localização subaquática é estimada combinando imagens acústicas com imagens aéreas de drones ou de satélite da mesma região que o veículo está navegando. O método pode ser aplicado em veículos subaquáticos como os operados remotamente –*Remotely Operated Vehicle*– ROVs, veículos subaquáticos autônomos –*Autonomous Underwater Vehicle*– AUVs, e veículos híbridos –*Unmanned Aerial Underwater Vehicle*– UAUVs, atuando próximos a estruturas parcialmente submersas, para auxiliar na localização e navegação autônoma em regiões como portos e marinas.

1.1 Descrição do Problema

O problema estudado neste trabalho é representado pela Figura 1.4. Percebe-se que em operações próximas à superfície, as imagens do domínio aéreo e do domínio subaquático compartilham o imageamento de alguns objetos da cena que estão presentes no nível da superfície da água. Com isto, é possível encontrar a correspondência entre essas estruturas e estimar se as imagens de ambos os domínios (aéreo e subaquático) tenham sido capturadas no mesmo lugar. Postes, pilares, píeres e o formato da orla são exemplos de estruturas parcialmente submersas estacionárias relevantes que podem ser utilizadas como referência para a identificação do local imageado. As imagens no domínio aéreo geralmente são geo-referenciadas, ou seja, a sua localização é conhecida. Portanto a correspondência entre as estruturas de ambos domínios pode ser utilizada para estimar a localização de um veículo subaquático. Porém a estimativa da localização e o tratamento dos dados representam um grande desafio o qual esse trabalho se propõe a resolver.

Neste contexto, estabelece-se como *domínio* o cenário a ser imageado associado a um determinado tipo de sensor. Mais precisamente, tem-se neste trabalho dois domínios: i) o domínio subaquático, percebido por sonares multi-feixe de imageamento frontal; ii) e o domínio aéreo,

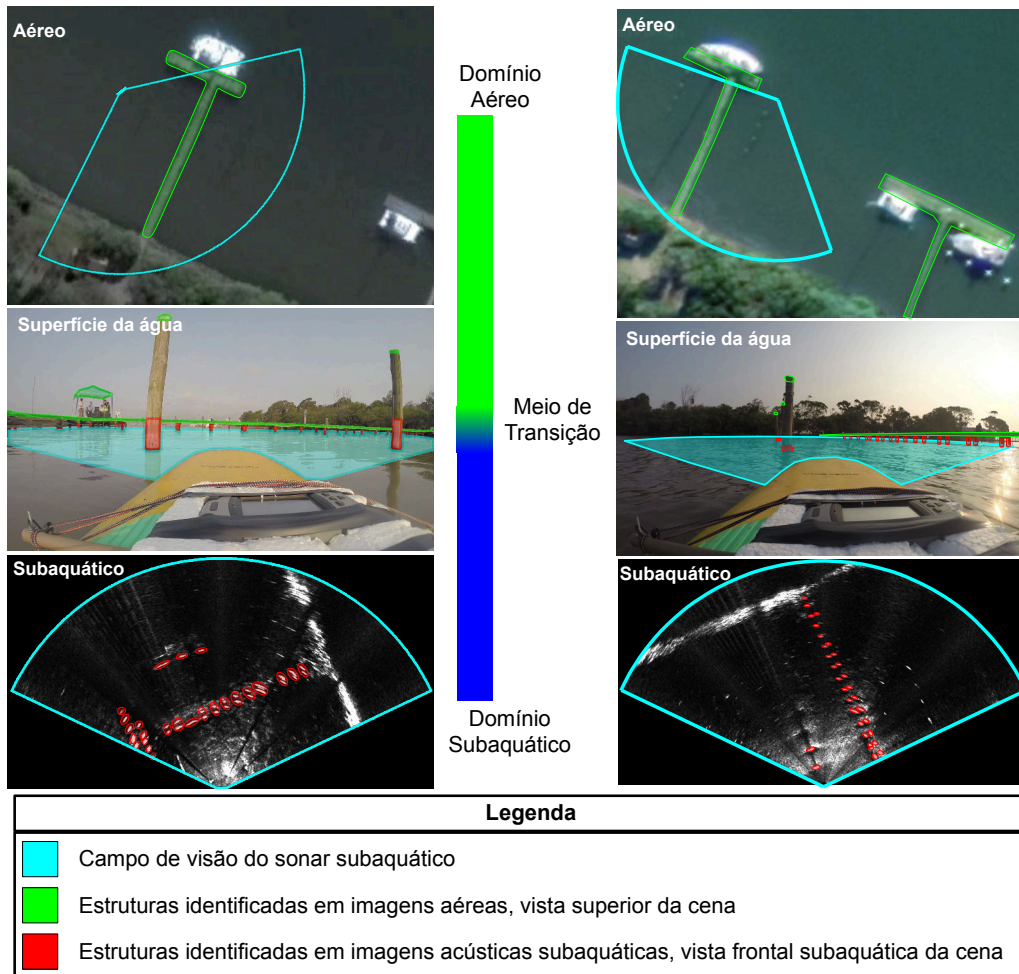


Figura 1.4: Exemplos de imagens acústicas subaquáticas e aéreas no espectro visível em um ambiente parcialmente estruturado. O campo de visão da imagem acústica é representado em ciano. Um píer é dividido em duas partes, a parte superior destacada em verde percebida pela imagem aérea, e a parte inferior destacada em vermelho e percebida pela imagem acústica subaquática. Estruturas parcialmente submersas como píeres podem ser percebidas em ambos os domínios aéreos e subaquáticos e utilizadas para inferir a localização de um veículo subaquático.

percebido por câmeras de drones ou satélites.

Encontrar a correspondência entre as vistas de diferentes domínios não é trivial (GAO et al., 2018). Apesar do problema ser atenuado pelo fato das imagens de ambos os domínios dispensarem transformações perspectivas, a Figura 1.4 mostra que a correspondência entre essas estruturas não é diretamente encontrada. Essa diferença decorre do fato de que as imagens aéreas capturam a vista superior e as imagens acústicas subaquáticas capturam as vistas inferior ou frontal da cena, entre diversos outros fatores.

Os sonares multi-feixe de imageamento frontal –*Forward Looking Sonar*– FLS são muito utilizados em veículos subaquáticos principalmente pela sua elevada taxa de captura de imagens, e por ser mais robusto a distorções acústicas provocadas pelo deslocamento do veículo. Por esses motivos, e pela sua disponibilidade no veículo de pesquisa do laboratório, o FLS foi adotado neste estudo.

As imagens acústicas de um sonar FLS apresentam características diferentes das imagens ópticas, que são inerentes às propriedades físicas da água e do som e que dificultam a sua percepção e utilização. Dentre elas, destacam-se:

- Baixa resolução - As imagens acústicas se resumem à informações de distância e forma dos objetos imageados. Informações de cor são inexistentes e de textura são muito pobres, fornecendo portanto poucos detalhes da cena imageada.
- Ruído - Por serem dispositivos ativos que operam com uma baixa razão sinal-ruído, ocorrem diversos tipos de ruídos oriundos de interferência entre as ondas acústicas transmitidas (reverberação acústica, também conhecido como *–multipath–*, causado pela reflexão de ondas na superfície, no fundo, ou por obstáculos), variações na densidade da água (bolhas de ar, variação de temperatura, pressão), e por sons gerados por outros sensores acústicos, animais ou veículos (motores).
- Sombras acústicas - Gera oclusão de parte da cena. São regiões não detectadas pelo sonar devido a obstrução do sinal provocado por um objeto posicionado na frente da região imageada. As sombras se deslocam conforme o movimento do sensor, alterando a região de oclusão da imagem.
- Ambiguidade - Devido à forma que os FLS são construídos, sua percepção é uma projeção 2D de uma cena tridimensional. Por esse motivo a informação de elevação dos objetos é perdida, gerando um problema de ambiguidade na imagem que impede a distinção da altura dos objetos detectados.
- Distribuição sonora não uniforme - Causado pela atenuação do sinal na água. A assinatura acústica de um objeto varia conforme a sua posição e orientação com relação ao sonar.

Por outro lado, a identificação de estruturas estacionárias em imagens aéreas no espectro visível, representado pelo padrão de cores RGB, também é desafiadora. Diversos fenômenos da natureza como as estações do ano, o clima, a posição do sol e das nuvens provocam alterações em diferentes aspectos das imagens como: mudança de tonalidade e de textura na água, no solo e nas árvores, assim como a ocorrência de oclusão, provocada pelo movimento das sombras (SANTOS et al., 2019c).

A Figura 1.5 ilustra alguns dos desafios comparando duas imagens de satélite da mesma região, adquiridas em datas distintas. É possível identificar diferenças significativas na textura da água e da terra devido à variações climáticas que provocam ventos fortes e inundações. Esses fenômenos dificultam a identificação das estruturas estacionárias.

A classificação de cada pixel da imagem é uma das formas de identificar as estruturas estacionárias nas imagens aéreas. Esse processo é conhecido como segmentação semântica (CHEN et al., 2018b). Neste caso, busca-se a identificação de três classes: estruturas estacionárias, objetos móveis e a água, sendo que as estruturas estacionárias possuem maior relevância para



Figura 1.5: Alguns desafios da segmentação semântica de imagens aéreas RGB em marinas: enchentes, ondas na superfície da água provocadas pelo vento, cor da água variável e objetos de tamanhos variados.

localização por não se movimentarem ao longo do tempo, portanto são mais estáveis para o reconhecimento da região.

Os desafios no tratamento de imagens acústicas e de imagens aéreas se somam quando busca-se encontrar a correlação entre as imagens dos dois domínios e estimar a localização subaquática. Além disso, as imagens aéreas possuem um campo de visão maior do que o das imagens acústicas, exigindo uma forma inteligente de buscar nas imagens aéreas, as estruturas identificadas nas imagens acústicas. Outros fatores que dificultam a comparação são: imagens aéreas providenciam informações ricas com cores e texturas, e as imagens acústicas consistem em um conjunto de distâncias e formas dos objetos; imagens aéreas providenciam uma vista superior da cena, as imagens acústicas providenciam uma vista frontal da cena.

A Figura 1.6 exemplifica o problema de localização subaquática estudado neste trabalho. Um veículo subaquático realiza inspeção e vigília autônoma em uma marina seguindo uma sequência de rotas de navegação pré programadas. Informações sensoriais de um magnetômetro, sinais de controle dos motores, imagens acústicas de um FLS e imagens aéreas coletadas por drones ou satélite são os dados disponíveis para navegação. As imagens aéreas são carregadas na memória do veículo antes do início da missão. A posição inicial do veículo é aproximadamente conhecida. As estruturas parcialmente submersas presentes na marina são detectadas nas imagens de ambos domínios e utilizadas para corrigir o erro de localização acumulado pelo veículo durante a sua navegação.

Portanto, neste trabalho busca-se tratar do problema de encontrar a correlação entre imagens de múltiplos domínios, óptico aéreo e acústico subaquático, e múltiplas perspectivas, superior e frontal, para estimar a localização subaquática de um veículo que navega em um ambiente parcialmente estruturado com estruturas parcialmente submersas, como os encontrados em marinas ou portos. As estruturas semi-submersas são percebidas em ambos os domínios e utilizadas para a estimativa da correlação das imagens. Sendo assim, pode-se dividir o problema tratado neste trabalho em três partes:

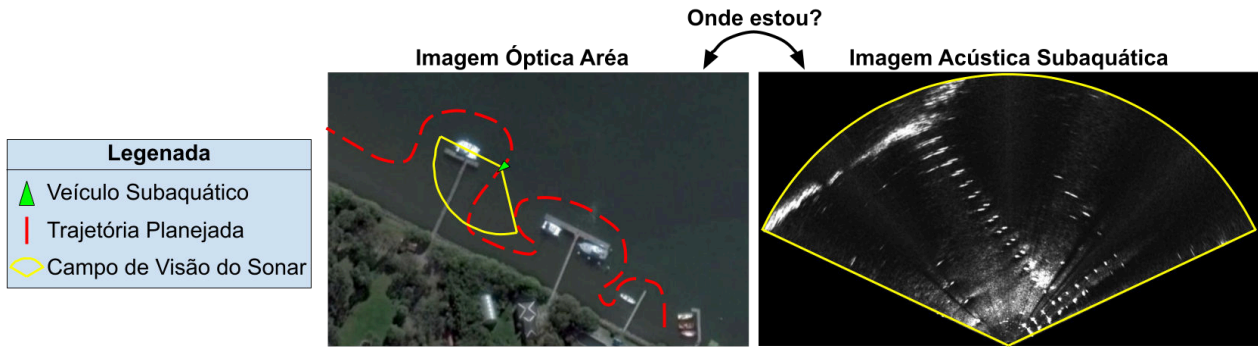


Figura 1.6: Um veículo subaquático realiza atividade autônoma em uma marina. Sua localização é estimada com base nas estruturas presentes nas imagens ópticas aéreas e nas imagens acústicas subaquáticas.

- Identificação de estruturas em imagens ópticas aéreas coloridas.
- Identificação de estruturas em imagens acústicas subaquáticas.
- Localização subaquática a partir da correlação das estruturas identificadas nas imagens óptica aéreas coloridas e nas imagens acústicas subaquáticas.

Com isto, espera-se que o método proposto para estimar localização utilize os momentos de navegação em regiões com estruturas para corrigir o erro de localização acumulado durante a navegação em regiões sem estruturas. O método seria capaz de operar em ambientes parcialmente estruturados, mas não funcionaria em ambientes completamente monótonos, como em mar aberto, em águas profundas, ou em regiões sem cobertura de imagens aéreas.

1.2 Hipótese Científica

Este trabalho parte da hipótese de que é possível estimar a localização de um veículo subaquático combinando dados sensoriais provenientes de dois domínios distintos, subaquático e aéreo, através da percepção de estruturas parcialmente submersas, utilizando técnicas de aprendizagem de máquina profunda associadas a abordagens tradicionais aplicadas à visão computacional e à localização.

1.3 Objetivos Gerais

Desenvolver um método para auxiliar na localização de veículos subaquáticos a partir de imagens aéreas visíveis, representadas pelo padrão de cores vermelho, verde e azul (RGB), e imagens acústicas subaquáticas de um FLS em ambientes parcialmente estruturados com elementos parcialmente submersos como portos e marinas.

1.4 Objetivos Específicos

- Investigar e desenvolver um método de segmentação semântica de imagens aéreas no espectro visível para regiões costeiras como portos e marinas.
- Investigar e desenvolver um método para identificação de estruturas parcialmente submersas relevantes para localização em imagens acústicas subaquáticas.
- Investigar e desenvolver um método para encontrar correlação entre imagens aéreas e imagens acústicas subaquáticas.
- Adaptar para o problema proposto um algoritmo de localização como o de Monte Carlo (MCL).
- Desenvolver experimentos reais e simulados no ambiente do problema proposto.
- Validar o método em experimentos reais e simulados.

1.5 Aplicações

A localização subaquática, tratada pelo método proposto, é um elemento essencial para a execução de tarefas autônomas. O método poderia ser empregado em veículos subaquáticos que navegam em ambientes parcialmente estruturados povoado por estruturas parcialmente submersas para desempenhar diversas atividades como:

- Vigilância autônoma de portos e marinas;
- Controle e rastreamento subaquático de barcos e navios em portos ou marinas;
- Inspeção periódica autônoma de estruturas submersas;
- Busca e recuperação de objetos em regiões como portos e marinas; e
- Localização de objetos e mapeamento de regiões inundadas.

O método também possibilita o planejamento de rotas subaquáticas utilizando imagens aéreas, onde a configuração prévia das rotas de um sistema autônomo seria definida a partir das imagens aéreas da região.

1.6 Contribuições

Este trabalho propõe o estudo e o desenvolvimento de um novo método de localização para veículos em missões que envolvam ambientes subaquáticos. Além da contribuição da proposta do método, foram realizadas contribuições em cada uma de suas etapas, que incluem:

- novas formas de identificação de estruturas em imagens aéreas e acústicas,

- formas de encontrar a correspondência entre as estruturas detectadas nos dois domínios, aéreo e subaquático. Onde foi proposta uma abordagem utilizando Redes Neurais Profundas, e outra utilizando um Modelo Probabilístico.
- adaptação do algoritmos de localização de Monte Carlo, ([THRUN; BURGARD; FOX, 2006](#)), também conhecido como filtro de partículas, para o problema proposto,
- realização de experimentos reais e o desenvolvimento de ambientes de simulação para a validação do método proposto.

O diagrama da Figura 1.7 mostra um panorama geral do método. É realizado o processamento *offline* das imagens aéreas de satélite ou de drones da região onde o veículo subaquático atua. Estruturas estacionárias e a água são identificadas e transformadas em um mapa de referência para a localização. O mapa é carregado na memória do veículo subaquático antes do mergulho. Durante a navegação, imagens acústicas são adquiridas e comparadas com o mapa de referência, alimentando o modelo de observação de um filtro de partículas que estima a localização subaquática do veículo. Maiores detalhes sobre a metodologia serão apresentados no Capítulo 3.

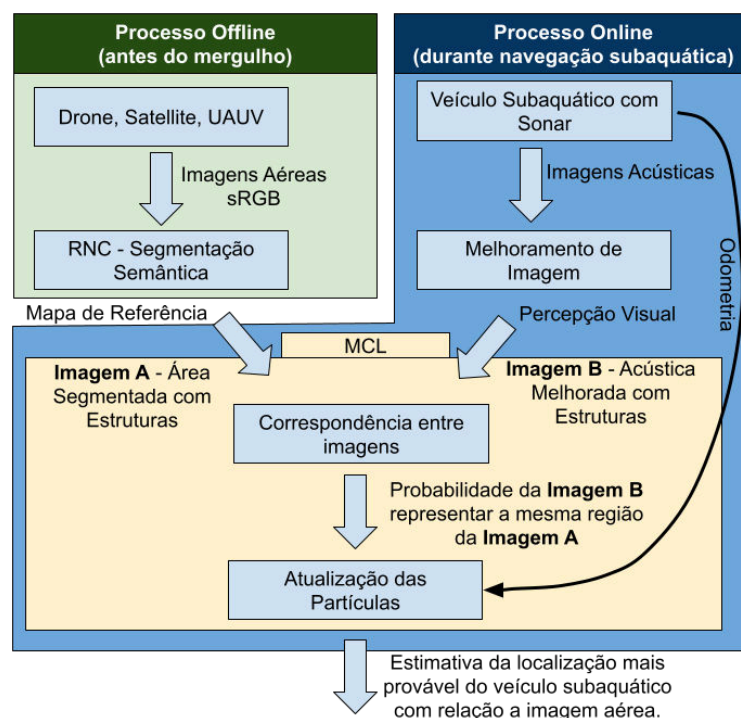


Figura 1.7: Diagrama proposto para lidar com o problema.

Algumas contribuições relacionadas a este trabalho foram parcialmente publicadas em conferências e revistas:

- Segmentação semântica de imagens aéreas ([SANTOS et al., 2019c](#)) (qualis A4, H5 14),
- Correlação de imagens ópticas aéreas e acústicas subaquáticas ([SANTOS et al., 2019a](#)) (qualis B1, H5 7), ([SANTOS et al., 2020](#)) (qualis A1, scopus percentil 94%, fator de

impacto (JIF) percentil 65%), (GIACOMO et al., 2020) (qualis B1, H5 7), (GIACOMO et al., 2022) (qualis A2, scopus percentil 81%),

- Localização subaquática utilizando dados simulados (SANTOS et al., 2019b) (qualis A3, H5 13),
- Localização subaquática utilizando dados reais (SANTOS et al., 2022) (qualis A1, scopus percentil 94%, fator de impacto (JIF) percentil 65%).

Maiores detalhes sobre essas publicações e os resultados obtidos serão apresentados no Capítulo 4.

1.7 Organização do Texto

O trabalho está organizado da seguinte forma: O próximo capítulo apresenta os trabalhos relacionados ao problema de correspondência e localização utilizando imagens de diferentes vistas, ao problema de segmentação semântica de imagens aéreas, e aos métodos de tratamento de imagens acústicas. Logo após, no Capítulo 3 é apresentada em detalhes a proposta do trabalho. Os resultados alcançados são apresentados no Capítulo 4 e finalmente o Capítulo 5 conclui o este trabalho apresentando as suas perspectivas futuras.

2 TRABALHOS RELACIONADOS

A localização de imagens subaquáticas em imagens aéreas é um novo domínio de aplicação e por este motivo não são encontrados trabalhos diretamente relacionados ao assunto. No entanto, foram propostas abordagens que estimam a localização de imagens terrestres a partir de imagens aéreas. Apesar de não abordarem imagens subaquáticas, o problema possui similaridades por tratar das correspondências entre imagens capturadas por vistas diferentes, ou seja, a vista aérea e a vista terrestre.

A Figura 2.1 mostra as principais linhas de pesquisa relacionadas a proposta deste trabalho. O domínio de aplicação é representado pelas cores do fundo em azul para subaquático, verde para aéreo e amarelo para terrestre. Os círculos em roxo representam as pesquisas diretamente relacionadas a este trabalho, referentes à localização de veículos autônomos. Estas pesquisas apresentam métodos que podem ser utilizados no pré-processamento das imagens acústicas subaquáticas ou das imagens ópticas aéreas. Os círculos em verde representam as contribuições deste trabalho, estando associados a métodos que não possuem similaridade direta com outros propostos na literatura. Os círculos em azul representam os trabalhos que não se aplicam diretamente ao problema deste trabalho, mas possuem semelhanças que podem contribuir com o desenvolvimento do método proposto.

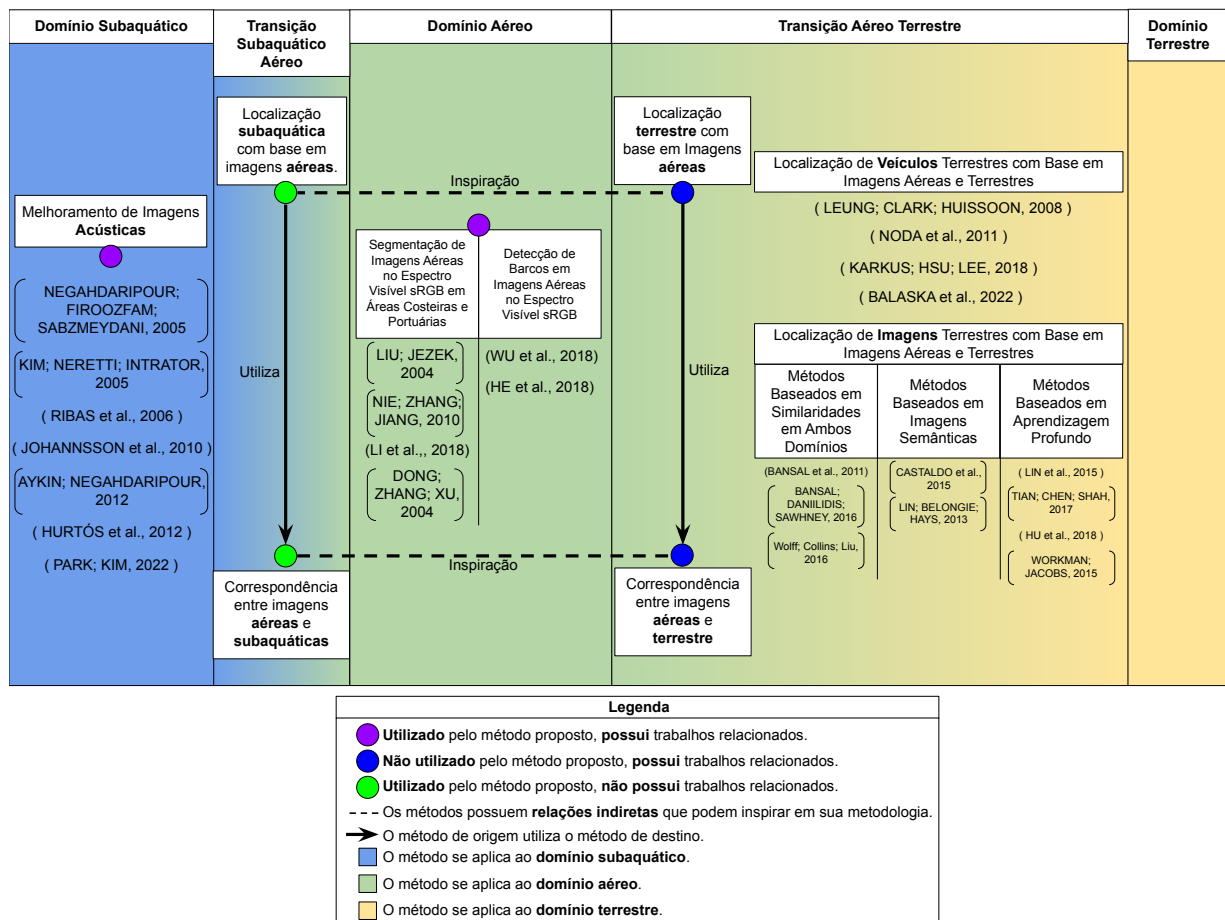


Figura 2.1: Diagrama de trabalhos correlacionados.

As informações de correspondência entre as imagens aéreas e terrestres são utilizadas pelos algoritmos de localização terrestre para estimar a localização com base em imagens aéreas. Esta relação é representada pelas setas na Figura 2.1.

A mesma relação se aplica para a localização subaquática utilizando imagens aéreas como referência. A distinção entre métodos de localização e métodos de correlação entre imagens é apresentada porque um método de localização pode fazer o uso de diferentes métodos de correlação entre imagens.

As próximas seções apresentam os trabalhos relacionados conforme o panorama apresentado na Figura 2.1.

2.1 Localização Terrestre com Base em Imagens Aéreas

Segundo GAO et al. (2018), os métodos que estimam a localização de imagens terrestres a partir de imagens aéreas, também conhecido como geo-localização de imagens terrestres, podem ser classificados como métodos baseados em imagens ou métodos baseados em estruturas.

Os métodos de correspondência baseados em imagens exploram as características das imagens em ambos os domínios aéreos e terrestres para encontrar suas correlações. As características extraídas dos domínios podem ser *similares entre si*, ou seja, as mesmas características são detectadas em ambas vistas (BANSAL; DANIILIDIS; SAWHNEY, 2016; BANSAL et al., 2011; Wolff; Collins; Liu, 2016). Elas podem ser características semânticas (CASTALDO et al., 2015; LIN; BELONGIE; HAYS, 2013) ou ainda podem usar técnicas de aprendizado profundo (LIN et al., 2015; TIAN; CHEN; SHAH, 2017; HU et al., 2018; WORKMAN; JACOBS, 2015), ou seja uma rede neural aprende a utilizar as características mais relevantes para estimar a sua correspondência, podendo incluir características como cor, forma, textura, informações semânticas entre outras.

Já os métodos baseados em estruturas buscam identificar a presença de estruturas no ambiente que podem ser observadas em ambas as vistas aéreas e terrestre e realizam a comparação das estruturas para encontrar correspondências entre as imagens (LEUNG; CLARK; HUISSOON, 2008; LI; MORARIU; DAVIS, 2014; Majdik; Albers-Schoenberg; Scaramuzza, 2013; NODA et al., 2011; Viswanathan; Pires; Huber, 2014).

2.1.1 Métodos de Localização de Imagens Terrestres Baseados em Características Similares em Diferentes Domínios

Os métodos que buscam encontrar as mesmas características em imagens de diferentes domínios (BANSAL et al., 2011; BANSAL; DANIILIDIS; SAWHNEY, 2016; Wolff; Collins; Liu, 2016) tiram proveito de prédios urbanos como fachadas de arranha-céus, que podem ser observados em ambas as vistas, terrestre e aérea. As imagens aéreas são combinadas com suas correspondentes terrestres através de recursos compartilhados, como textura, cor e formas.

A Figura 2.2, mostra o método proposto por Wolff; Collins; Liu (2016) para encontrar correspondências entre imagens aéreas e terrestres. O método trata de áreas urbanas e utiliza as fachadas dos prédios como referência. Inicialmente as fachadas são identificadas. Características similares entre si são extraídas de blocos de imagens (*patches*) das fachadas em ambos os domínios, utilizando o método proposto em (PARK et al., 2011). A similaridade entre os *patches* são estimadas com base em quatro características: (1) o contexto das formas identificadas (BELONGIE; MALIK; PUZICHA, 2002), (2) a cor, (3) a textura, e (4) a proximidade dos *patches*. O problema de encontrar a correspondência final entre os *patches* é transformado no problema de atribuição com custo mínimo, *assignment problem* (BURKARD; DELL'AMICO; MARTELLO, 2012). As relações entre todos os *patches* são representadas em um grafo bipartido onde os vértices representam as imagens, e as arestas representam o erro de similaridade (custo). O algoritmo conhecido como *Hungarian Algorithm* (KUHN, 2005) é utilizado para encontrar a solução ótima, isto é, os pares de imagens que possuem o menor custo, ou seja, a maior similaridade.

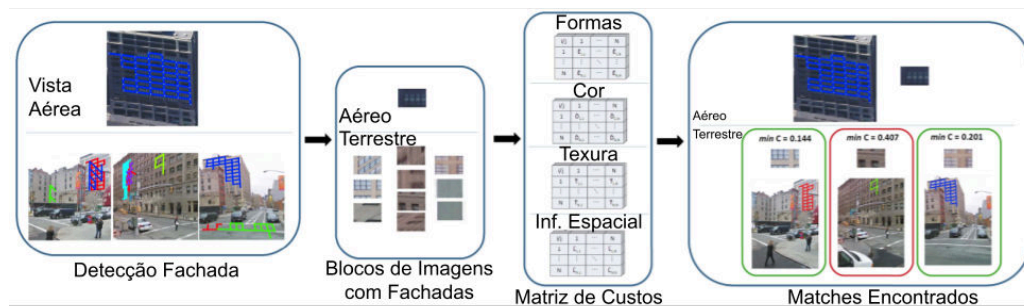


Figura 2.2: Etapas do método proposto por Wolff; Collins; Liu em (Wolff; Collins; Liu, 2016). Inicialmente são identificadas as treliças das fachadas dos prédios, posteriormente as fachadas identificadas são recortadas em blocos de imagem, *patches*. Os blocos são descritos de acordo com suas características. As correspondências finais são apresentadas na última coluna. Imagens com contorno verde são correspondentes, imagens com contorno vermelho não possuem correspondência. Figura adaptada de (Wolff; Collins; Liu, 2016).

2.1.2 Métodos de Localização de Imagens Terrestres Baseado em Imagens Semânticas

Os métodos baseados em semântica consideram informações sobre a cena observada para realizar a correspondência das imagens terrestres com a imagem aérea. Por exemplo, (CASTALDO et al., 2015) geo-localiza imagens do solo explorando mapas de Sistemas de Informações Geográficas (SIG). As imagens do solo são segmentadas e retificadas em uma vista superior. Os autores propuseram o descritor *Semantic Segment Layout* (SSL), que extrai as informações utilizadas para encontrar a correspondência entre as imagens retificadas e o mapa SIG. A Figura 2.3 mostra uma visão geral do método proposto por CASTALDO et al. (2015).

Nesta mesma linha, (LIN; BELONGIE; HAYS, 2013) propõem uma abordagem de conversão cruzada de características de imagens para localizar imagens no solo. O método explora imagens aéreas, imagens terrestres e mapas de atributos da região para geo-localizar imagens terres-

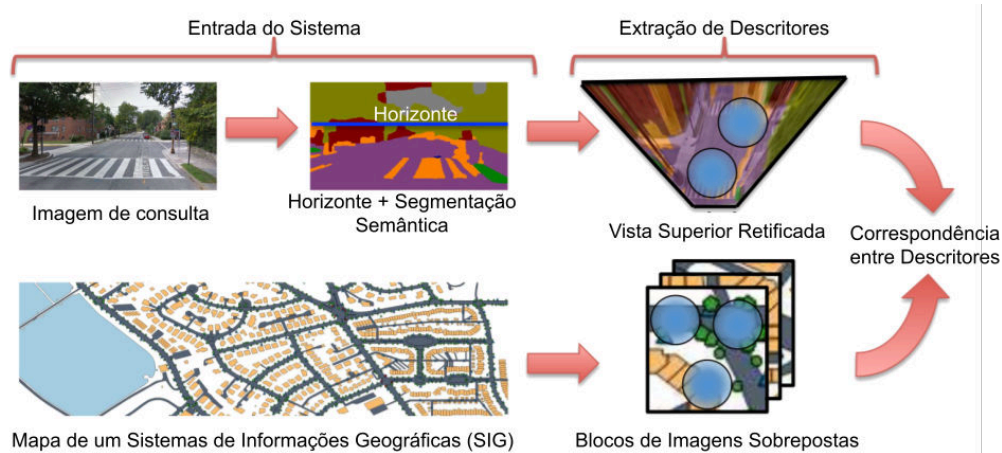


Figura 2.3: Método de correspondência de imagens terrestres e aéreas proposto por [CASTALDO et al.](#). A imagem terrestre de consulta é utilizada com sua respectiva segmentação semântica, a identificação do seu plano terrestre (atitude), ou linha do horizonte, representado pela linha azul. O plano terrestre é utilizado para retificar a imagem e minimizar as distorções entre o mapa SIG e a imagem terrestre de consulta. O mapa SIG é dividido em blocos e descrito utilizando o método de extração de layout proposto pelos autores. A vista superior da imagem terrestre, retificada, também é descrita. Os círculos em azul representam as regiões descritas em ambas as imagens. A distância entre pares de descritores é utilizada para determinar os blocos de imagens que possuem maior similaridade com a imagem de consulta. Imagem adaptada de ([CASTALDO et al., 2015](#)).

tres. As imagens são descritas utilizando Histogramas de Orientação de Gradientes, *Histogram of Oriented Gradients* (HoG) ([DALAL; TRIGGS, 2005](#)), Auto-similaridade ([SHECHTMAN; IRANI, 2007](#)), forma ([OLIVA; TORRALBA, 2001](#)) e histograma de cores. Esses descritores são representados em forma de vetores K-dimensionais. Um banco de dados de *triplets* que associa uma imagem terrestre, a sua imagem aérea e a sua imagem de mapa de atributos correspondente é mantido pelo método. Diferentes abordagens orientadas a dados, *data driven*, são utilizadas e avaliadas pelos autores para geo-localizar imagens terrestres.

Dentre elas, os autores propõem duas abordagens: a média de vetores de característica (AVG), e a tradução discriminativa (DT). Na média de vetores de características (AVG), a imagem terrestre de consulta é comparada com todas imagens terrestres do banco de dados. As imagens aéreas e imagens de atributos associadas às K imagens terrestres mais similares são selecionadas utilizando K-NN nas *triplets* do banco de dados. É realizada a média do vetor de características de todas as imagens aéreas e de atributos que foram selecionadas. Então o vetor médio resultante é utilizado para localizar a imagem terrestre de consulta na imagem aérea da região de interesse.

A abordagem de tradução discriminativa (DT) realiza um processo similar ao da média de vetores (AVG), porém além de selecionar as K-*triplets* mais similares, também são selecionadas as K-*triplets* menos similares a imagem terrestre de consulta. Uma máquina de vetores de suporte (SVM, do inglês: *support-vector machine*) é treinada com as *triplets* positivas e negativas. Então o SVM é utilizado para estimar o local da imagem terrestre na imagem aérea da região

de interesse.

O sistema foi avaliado utilizando diversas imagens de consultas em uma região de interesse de 40 km por 40 km em Charleston, na Califórnia do Sul. A Figura 2.4 mostra alguns detalhes do método proposto por LIN; BELONGIE; HAYS (2013) e alguns resultados obtidos. O método tem como saída um mapa de calor que indica a região mais provável de captura da imagem terrestre de consulta (entrada).

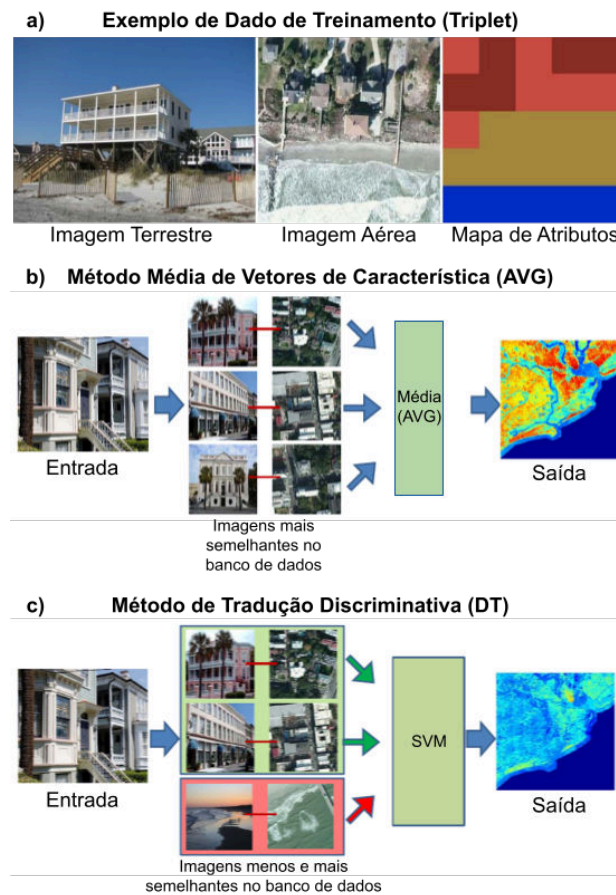


Figura 2.4: Métodos propostos por LIN; BELONGIE; HAYS (LIN; BELONGIE; HAYS, 2013). Em (a) um exemplo de uma *triplet* que associa as imagens terrestre, aérea (recorte) e mapa de atributos (recorte) correspondentes, armazenado no banco de dados do método. Em (b) o pipeline do método de Média de Vetores de Características (AVG). Em (c) o pipeline do método de Tradução Discriminativa (DT). As imagens com contorno verde são positivas, correspondem a imagem de consulta (entrada), e as imagens com contorno vermelho são negativas, não possui correspondência. Como resultado, é estimada a região em que a imagem foi capturada através de um mapa de calor (direita), calculado a partir do banco de imagens do método. A localização mais provável é representada pelos tons de vermelho mais intenso. Imagem adaptada de (LIN; BELONGIE; HAYS, 2013).

2.1.3 Métodos de Localização de Imagens Terrestres Baseados em Aprendizagem Profundo

Métodos para encontrar correspondência entre imagens terrestres e aéreas e localizar imagens terrestres utilizando aprendizado profundo também foram propostos (LIN et al., 2015; TIAN;

CHEN; SHAH, 2017; HU et al., 2018; WORKMAN; JACOBS, 2015).

O trabalho de LIN et al. (LIN et al.) propôs a rede neural *Where-CNN*, uma das primeiras abordagens de aprendizado profundo para este problema. Uma rede *Siamese*, inicialmente proposta por (CHOPRA; HADSELL; LECUN, 2005), é usada para aprender a incorporar as características relevantes das imagens e encontrar as correspondências entre as vistas terrestres e aéreas.

A Figura 2.5 mostra a abordagem proposta em (LIN et al., 2015). Um conjunto de imagens terrestres do *Google Street View* e imagens aéreas com 45 graus de inclinação com relação ao solo são coletadas em sete cidades de quatro países distintos. As imagens aéreas são recortadas em blocos de 256x256 pixels, cobrindo aproximadamente 15 x 15 metros da região coberta pela imagem terrestre. O processo resulta nos dados de treinamento com pares de imagens terrestre e aéreo conforme mostrado na Figura 2.5(a).

Duas redes *Siamese* idênticas, A e B, são treinadas sem compartilhar pesos. Desta forma, elas podem incorporar características específicas de cada domínio. A rede A para imagens terrestres e a rede B para imagens aéreas. Os parâmetros das redes foram pré-treinados em um conjunto de dados para reconhecimento de cenas e objetos disponibilizado em (ZHOU et al., 2014). A última camada *fully-connected* das redes *Siamese* é eliminada, utilizando assim, a penúltima camada de 4096 dimensões como representação da imagem, conhecido do inglês como *feature-vector*.

Após o treinamento das redes, as imagens aéreas são descritas pelo seu conjunto de *feature-vector* obtido de forma offline pela rede B, e a rede A é utilizada sob demanda para criar a representação (*feature-vector*) de cada nova imagem terrestre de consulta. A localização final da imagem de consulta é estimada encontrando as imagens aéreas mais semelhantes do banco de dados, utilizando o algoritmo K-NN, e a distância Euclidiana entre os *feature-vectors* das imagens como métrica. A Figura 2.4(b) mostra o método proposto por (LIN et al., 2015) e a Figura 2.5(c) mostra alguns resultados obtidos pelo método.

No trabalho de WORKMAN; JACOBS (2015) foi realizado um estudo avaliando a capacidade das redes neurais profundas de representarem imagens aéreas e imagens terrestres aplicado ao problema de geo-localização de imagens terrestres (WORKMAN; JACOBS, 2015). Os autores avaliaram duas redes neurais distintas, a ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) para classificar objetos, e a rede proposta em (ZHOU et al., 2014) para classificar cenas. As duas redes foram avaliadas sem ajustes de pesos, ou seja, utilizando os pesos disponibilizados pelos autores originais. Também não houve distinção com relação ao domínio aéreo ou domínio terrestre, a mesma rede é utilizada para descrever imagens de ambos os domínios. WORKMAN; JACOBS (2015) avaliaram individualmente a capacidade de descrição de imagens de cada uma das três últimas camadas das duas redes neurais estudadas. A técnica utilizada para localizar as imagens é semelhante a AVG proposta por LIN; BELONGIE; HAYS (2013) e representada na Figura 2.4. Os vetores de características da imagem terrestre de consulta são comparados

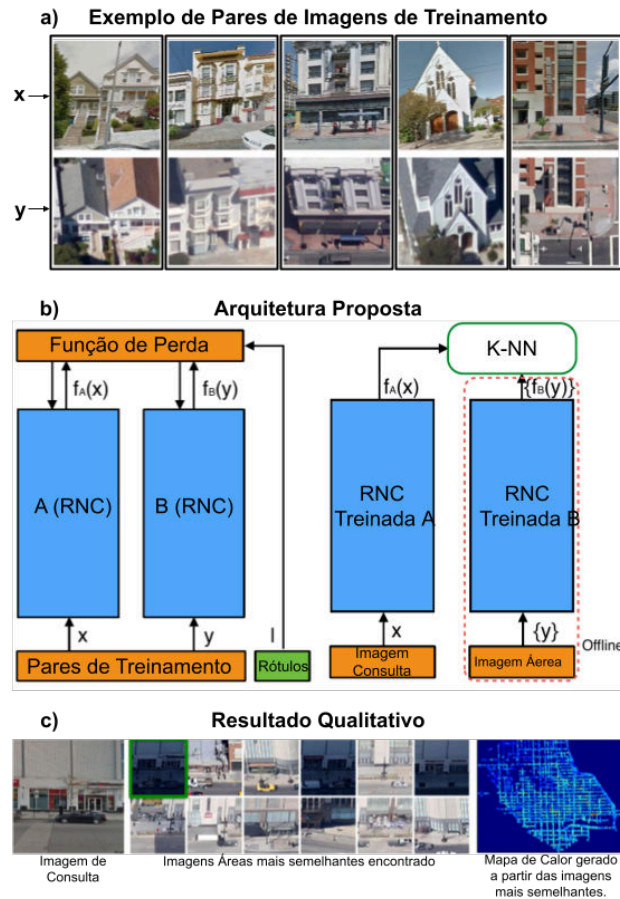


Figura 2.5: Método de aprendizagem profunda de representação de imagens proposto por LIN et al.. Em (a) exemplo de imagens de treinamento onde $x \in X$ são imagens terrestres e $y \in Y$ imagens aéreas com inclinação de 45 graus. Em (b) o método proposto por LIN et al.. Duas redes *Siamese* idênticas sem pesos compartilhados aprendem a representação das imagens terrestres e aéreas. Em (c) resultados obtidos com uma imagem terrestre coletada na cidade de Charleston - Carolina do Sul - Estados Unidos. Imagem adaptada de (LIN et al., 2015).

com os vetores de características de um banco de dados que mantém pares de imagem terrestre a sua imagem aérea correspondente. São selecionadas as 30 imagens terrestres mais similares à imagem de consulta, logo após, utiliza-se a média dos vetores de características das imagens aéreas associadas às 30 imagens terrestres selecionadas. Finalmente o vetor de característica médio é utilizado para comparar com a imagem aérea que cobre a região de interesse, resultando na estimativa da localização da imagem de consulta. A métrica utilizada para comparar os vetores de características é a distância Euclidiana.

Os resultados revelaram que a última camada *fully-connected* 'fc8' da rede proposta em (ZHOU et al., 2014) para classificação de cenas obteve o melhor desempenho para o problema de geo-localização de imagens terrestres, superando o método estado da arte de LIN; BELONGIE; HAYS (2013) no conjunto de dados de Charleston - Carolina do Sul, e descrito na Figura 2.4.

TIAN; CHEN; SHAH (2017) propõem um método para localizar imagens terrestres em áreas urbanas utilizando imagens aéreas como referência. O método tem como base a identificação dos edifícios nas vistas aérea e terrestre utilizando a rede *Faster R-CNN* (REN et al., 2015). A partir

da identificação dos edifícios, ocorre um processo de comparação e detecção de correspondência utilizando uma rede *Siamese* (CHOPRA; HADSELL; LECUN, 2005). A rede *Siamese* utiliza duas redes AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) sem compartilhamento de pesos para aprender a representação das características dos edifícios nas vistas aéreas e terrestre. A Figura 2.6 apresenta o método proposto. Os quatro prédios da imagem de consulta são identificados utilizando Faster R-CNN. Em seguida é estabelecido um *cluster* para cada um dos quatro prédios de consulta, reunindo os K prédios mais semelhantes encontrados nas imagens aéreas de referência, conforme representado na Figura 2.6(a). A comparação entre os prédios da imagem terrestre de consulta com os prédios das imagens aéreas de referência é realizada utilizando a rede *Siamese*. A Figura 2.6(b) mostra a próxima etapa do método, que consiste na criação de um grafo acíclico não direcionado a partir dos *cluster*. Cada vértice representa um prédio, e as arestas conectam apenas os prédios que não estão no mesmo *cluster*. O peso das arestas considera a distância geográfica entre os prédios ligados, e a soma da similaridade desses prédios com os da imagem de consulta. Com isso os autores priorizam aqueles prédios que além de serem mais similares através da imagem, também estão mais próximos geograficamente, portanto possuem uma maior expectativa de serem os mesmos prédios da imagem terrestre de consulta. A última etapa do método, representado na Figura 2.6(c), encontra o grupo de prédios dominante do grafo. A posição geográfica da imagem de consulta é estimada através da média das posições geográficas de cada prédio do grupo dominante.

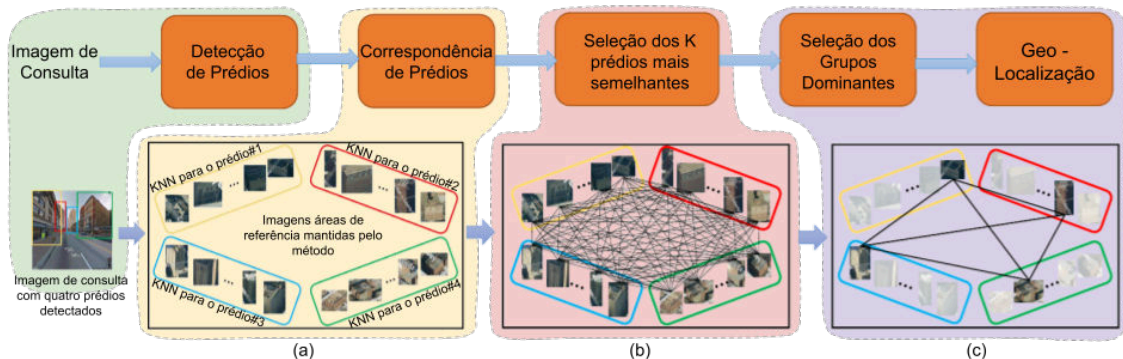


Figura 2.6: Etapas do método de aprendizagem profunda para geo-localização de imagens terrestres em áreas urbanas proposto por TIAN; CHEN; SHAH. Inicialmente são detectados os prédios da imagem terrestre de consulta e das imagens aéreas de referência. Em (a) um *cluster* com os K prédios mais semelhantes das imagens de referência com relação aos prédios de consulta é criado para cada prédio identificado na imagem de consulta. Em (b) é criado um grafo que associa a similaridade entre os prédios e as suas distâncias geográficas. Em (c) os prédios com maior similaridade e menor distância geográfica são selecionados como sendo os mais prováveis de serem os prédios da imagem de consulta. A localização final da imagem de consulta é dada pela média dos prédios encontrados. Imagem adaptada de (TIAN; CHEN; SHAH, 2017)

2.1.4 Localização de Veículos Terrestres com Base em Imagens Aéreas

O trabalho de [LEUNG; CLARK; HUISSOON \(2008\)](#) propõe uma abordagem de localização de imagens no solo usando imagens aéreas como referência. O método é baseado em estruturas ortogonais em áreas urbanas. Um mapa de características – *feature map* – é gerado a partir de orto-imagens aéreas. A *Progressive Probabilistic Hough Transform* (PPHT) ([MATAS; GALAMBOS; KITTLER, 2000](#)) é adotada para identificar os limites de construção. As linhas e os pontos de fuga detectados são analisados para determinar as orientações das paredes. As paredes são utilizadas como percepções visuais para alimentar um filtro de partículas ([THRUN; BURGARD; FOX, 2006](#)) que localiza as imagens do solo. A Figura 2.7 mostra o ambiente de estudo e como as linhas são extraídas das imagens aéreas e terrestres.

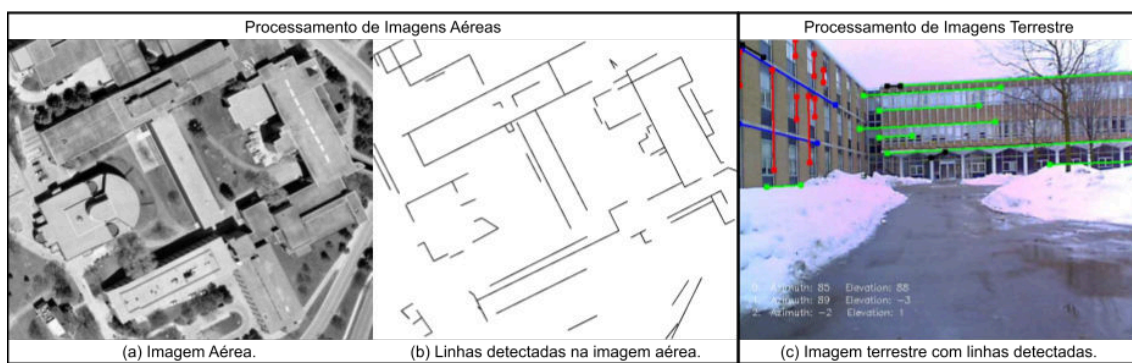


Figura 2.7: Imagens capturadas na Universidade de Waterloo, Ontário, Canadá, e estudadas por [LEUNG; CLARK; HUISSOON](#). Em (a) a imagem aérea de estudo, em (b) as linhas extraídas da imagem (a) utilizando o método proposto por [LEUNG; CLARK; HUISSOON](#), em (c) uma imagem terrestre do mesmo local com as linhas identificadas em destaque. Imagem adaptada de ([LEUNG; CLARK; HUISSOON, 2008](#))

O método PPHT pode identificar linhas que não pertencem às paredes de construções. Por exemplo, as linhas podem pertencer a vegetação do ambiente e, assim, prejudicar o sistema de localização, que é baseado nas estruturas dos prédios. Para tratar este problema, alguns algoritmos são propostos como por exemplo o uso do método de [HARRIS; STEPHENS \(HARRIS; STEPHENS, 1988\)](#) de identificação de cantos nas imagens. As regiões da imagem com vegetações possuem uma densidade de cantos diferente das regiões com construções. Outra técnica utilizada é selecionar os segmentos de linhas horizontais que são cortados ou estão próximos de segmentos de linhas verticais. Também é aplicada a técnica de unir linhas paralelas que estão muito próximas.

No trabalho de [NODA et al. \(2011\)](#) um veículo terrestre é geo-localizado através das imagens da sua câmera embarcada, utilizando imagens aéreas como referência de localização. O método estima a localização com base na correspondência entre os sinais de trânsito, como faixa de pedestre, pintados nas rodovias. Esses sinais podem ser identificados em ambas imagens aéreas e terrestres.

Inicialmente, um processo offline constrói um mapa de características utilizando apenas as ima-

gens aéreas. As rodovias são segmentadas manualmente, e os sinais de trânsito são identificados através da operação de limiarização das imagens, seguida da detecção dos cantos, utilizando o método de [HARRIS; STEPHENS \(1988\)](#). Todos os cantos são descritos utilizando o descritor SURF ([BAY et al., 2008](#)). O processo resulta em um mapa de características composto pela posição $\mathbf{x}_i$ e a descrição $\mathbf{f}_i$ de cada sinal de trânsito identificado, da forma $(\mathbf{x}_1, \mathbf{f}_1), \dots, (\mathbf{x}_N, \mathbf{f}_N)$.

As imagens capturadas pelo veículo terrestre são transformadas na perspectiva de visão aérea. Posteriormente, os sinais de trânsito pintados nas rodovias são descritos utilizando o mesmo descritor SURF. Para reduzir problemas com a oclusão dos sinais das rodovias que podem ocorrer devido aos outros veículos, é realizado o rastreamento dos sinais utilizando múltiplas imagens terrestres. O método de [LUCAS; KANADE et al. \(1981\)](#) é utilizado para o cálculo do fluxo óptico esparso. As descrições dos sinais de trânsito pintados nas rodovias, rastreados através de múltiplas imagens terrestres, são comparadas com o mapa de características de imagens aéreas e assim obtendo-se uma estimativa da localização do veículo na rodovia.

O trabalho de [BALASKA et al. \(2022\)](#) propõem um descritor de imagens para localização de veículos através da descrição de imagens terrestres e aéreas em regiões urbanas. O método combina informações semânticas e métricas de uma câmera estéreo para construir um grafo que descreve o ambiente. O descritor baseado em grafo foi chamado de SemMetric. Cada vértice representa o centro de uma entidade semântica, e as arestas são ponderadas com a distância euclidiana entre elas.

As entidades semânticas detectadas nas imagens aéreas são comparadas com as detectadas nas imagens terrestres utilizando a projeção da visão estéreo em um mapa aéreo. As informações de ambos domínios são fundidas no mesmo grafo gerando um mapa de memórias com informações mais completas da região do que as detectadas apenas com imagens de satélite. O método foi avaliado no dataset *Cityscapes* ([CORDTS et al., 2016](#)) com imagens da cidade de Frankfurt. Para identificar as informações semânticas nas imagens aéreas de satélite foi utilizada a rede neural SegNet ([BADRINARAYANAN; KENDALL; SEGNET, 2015](#)), e para identificar as informações semânticas das imagens terrestres foi adotada a rede ICNet ([ZHAO et al., 2018](#)). Ainda foi utilizado o algoritmo *Moore-Neighbor* proposto em ([LIU et al., 2004](#)) para instância de entidades semânticas.

Inicialmente é construído o mapa de memórias através da fusão das imagens aéreas e terrestres. Neste momento o grafo é construído considerando a posição do veículo estimado através de odometria e sinal de GPS. As informações semânticas utilizadas para a fusão são estrada, sinal de trânsito, calçadas, vegetação, edifícios e cerca. Em um segundo momento, o veículo é localizado utilizando apenas as leituras da sua câmera estéreo, através de consultas ao mapa de memórias da região previamente construído e de odometria visual, utilizando ORB-SLAM2 ([MUR-ARTAL; TARDÓS, 2017](#)). Os autores validam o método mostrando a acurácia de localização das consultas realizadas ao mapa de memória em duas ruas na cidade de Frankfurt.

O trabalho de [KARKUS; HSU; LEE \(2018\)](#) propõem um método de localização para áreas

internas *–indoor–* que utiliza o mapa 2D da construção como referência e imagens ópticas frontais capturadas pelo veículo que navega dentro da construção, conforme apresentado na Figura 2.8. Os autores propõem uma adaptação do algoritmo de Filtro Partículas que o torna diferenciável por completo. Desta forma é possível treinar uma rede neural de ponta a ponta *–end-to-end–* capaz de aprender todas as etapas do algoritmo de Filtro de Partículas, incluindo o modelo de transição de estados, e o modelo de observação. A vantagem desta abordagem está na forma de treinamento da rede, pois não são necessários pares de imagens anotadas para aprender o modelo de observação que compara as imagens ópticas frontais do veículo com o mapa 2D da construção. Utiliza-se diretamente as imagens coletadas pelo veículo e a sua localização verdadeira. O método é treinado e avaliado nos dados simulados *House3D* disponibilizados em (WU et al., 2018). Os autores avaliam a utilização de mapas 2D das construções de três formas: utilizando apenas informações de portas e paredes; utilizando apenas informações semânticas dos ambientes, como cozinha, quarto, banheiro; e finalmente utilizando ambas informações combinadas.

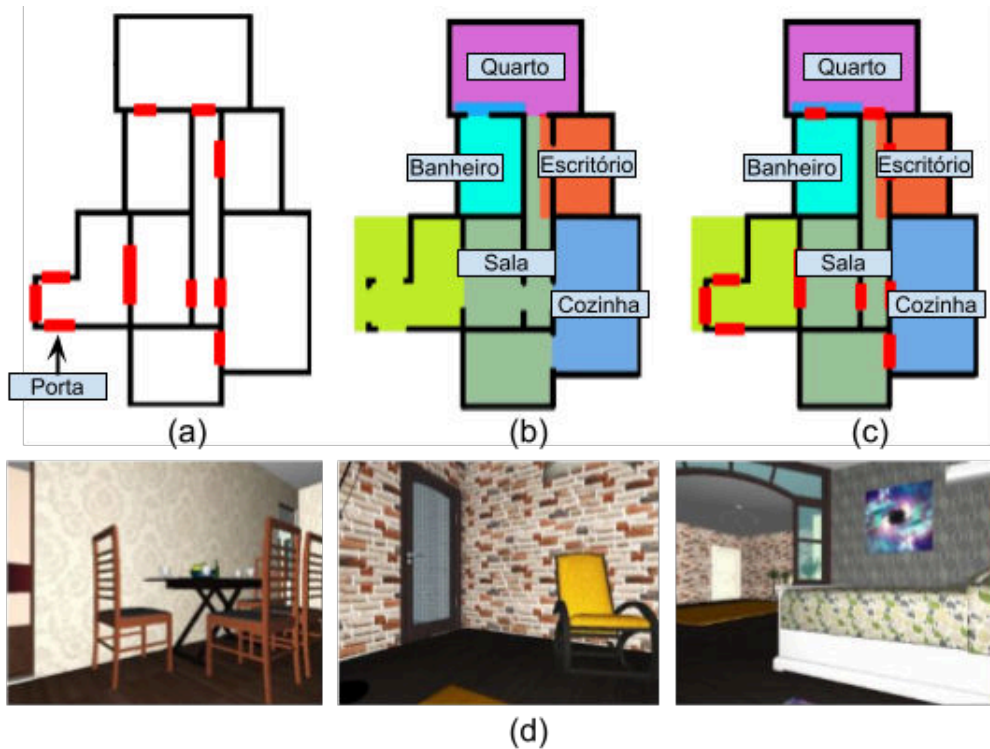


Figura 2.8: Exemplo de imagens utilizadas por KARKUS; HSU; LEE para treinamento e avaliação da rede neural para localização *indoor* de veículos. Essas imagens fazem parte do dataset *House3D* (WU et al., 2018). Em (a) informações de paredes e portas da construção que o veículo navega. Em (b) um exemplo de informações semânticas da construção utilizada. Em (c) as informações combinadas. Em (d) exemplo de imagens capturadas pelo veículo no ambiente de simulação. Com o método de KARKUS; HSU; LEE, o treinamento da rede é realizado utilizando imagens do tipo (a), (b), ou (c), combinado com imagens do tipo apresentando em (d) e a localização verdadeira do veículo. Para localização do veículo *indoor*, a rede neural necessita de uma imagem do tipo (a), (b), ou (c) e as imagens capturadas pelo veículo. Imagem adaptada de (KARKUS; HSU; LEE, 2018)

2.1.5 Considerações aos Trabalhos Relacionados a Localização e Correspondência de Imagens Terrestres e Aéreas

A localização geográfica de imagens terrestres em imagens aéreas têm aspectos semelhantes aos do problema de localização subaquática proposto nesta pesquisa. No entanto, conforme descrito na Seção 1.1, as imagens acústicas subaquáticas possuem características especiais, como a ausência de texturas e cor, maior presença de ruído, e outros efeitos inerentes à natureza acústica que não ocorrem em imagens ópticas, (HURTOS et al., 2013). Por esse motivo, o problema requer uma solução específica.

Os métodos baseados na extração direta de características das imagens como os apresentados em (BANSAL; DANIILIDIS; SAWHNEY, 2016; BANSAL et al., 2011; Wolff; Collins; Liu, 2016), não são adequados para este problema devido às grandes diferenças entre as imagens aéreas RGB e as imagens acústicas subaquáticas.

As imagens acústicas consistem em um conjunto de distâncias e formas dos objetos projetadas em um plano. Isto ocorre porque as informações de altura dos objetos são perdidas (HURTOS et al., 2013), resultando em uma imagem da vista frontal da cena pelo veículo, que se assemelha à uma vista aérea incompleta. Portanto, problemas relacionados à distorção de perspectiva relatados em (CASTALDO et al., 2015; NODA et al., 2011) devido à transformação entre vistas são menos críticos no caso das imagens acústicas de sonar. Um exemplo do campo de visão de uma imagem aérea RGB e o campo de visão de uma imagem acústica da mesma região pode ser visualizada na Figura 1.4.

Este trabalho busca encontrar a correspondência entre as imagens do domínio aéreo e o domínio subaquático utilizando as estruturas presentes na cena e posteriormente estimar a localização subaquática. Diferente do problema de localização terrestre utilizando imagens aéreas, a localização subaquática utilizando imagens aéreas envolve imagens com muita informação, RGB aéreas, e imagens acústicas subaquáticas com pouca informação. Apesar das diferenças que impossibilitam a utilização direta dos métodos apresentados, métodos baseado em aprendizagem profunda como o proposto em (TIAN; CHEN; SHAH, 2017), que utiliza a rede *Siamese*, e métodos de descrição de imagens combinado com o algoritmo de localização de Monte Carlo proposto em (LEUNG; CLARK; HUISSOON, 2008), podem ser avaliados e utilizados como inspiração nesta pesquisa.

2.2 Segmentação Semântica de Imagens Aéreas

A segmentação semântica de imagens aéreas requer a atribuição de uma categoria para cada pixel da imagem. É uma tarefa fundamental no sensoriamento remoto, pois permite a extração de informações importantes de uma vasta área do ambiente. A segmentação semântica vem contribuindo para o entendimento de fenômenos complexos em todo o mundo, como no estudo do impacto das mudanças climáticas e gestão ambiental (ZHONG et al., 2018), do crescimento

urbano, da biodiversidade e das tendências socioeconômicas da humanidade. A segmentação de imagens aéreas tem aplicações em sistemas para resposta a desastres (EZEQUIEL et al., 2014), no gerenciamento de recursos, na agricultura de precisão (CANDIAGO et al., 2015; GANCHENKO; DOUDKIN, 2019; ANAND et al., 2021), entre outros (WU et al., 2019a).

Desde que o trabalho de KRIZHEVSKY; SUTSKEVER; HINTON (2012) venceu o desafio *ImageNet Large Scale Visual Recognition Challenge*, ILSVRC-2012 (RUSSAKOVSKY et al., 2015), houve um crescente uso de Redes Neurais Convolucionais (RNCs) para solucionar problemas de visão computacional. Essa tendência ocorre devido às limitações dos métodos convencionais em representar características em imagens complexas. As RNCs superam essas limitações porque são capazes de aprender a melhor forma de representar as imagens através de um conjunto com milhares de imagens de treinamento.

A maioria das abordagens de segmentação semântica foi projetada para operar em áreas terrestres e urbanas (MURUGANANDHAM, 2016; PAISITKRIANGKRAI et al., 2016; KAMPFFMEYER; SALBERG; JENSSEN, 2016; MARMANIS et al., 2016; CHEN et al., 2018a; KEMKER; SALVAGGIO; KANAN, 2018) as quais envolvem a detecção e segmentação de veículos, vegetação, árvores, prédios, estradas, entre outros objetos. Para promover o desenvolvimento da área, foram criadas algumas competições como a da Sociedade Internacional de Fotogrametria e Sensoriamento Remoto, do inglês *International Society for Photogrammetry and Remote Sensing* ISPRS 2-D¹ (ROTTENSTEINER et al., 2014) e o desafio DeepGlobe (DEMIR et al., 2018)².

Entretanto, poucos trabalhos foram propostos para a segmentação semântica das regiões litorâneas e portuárias. O entendimento das imagens de sensoriamento remoto em áreas litorâneas é fundamental para o gerenciamento marinho, a detecção de navios, a vigilância dos portos (HE et al., 2018) e a atualização das informações geográficas (DONG; ZHANG; XU, 2014). Entre os trabalhos relacionados encontrados, a maioria realiza segmentação de áreas litorâneas (LI et al., 2018), detectam a orla (DONG; ZHANG; XU, 2014) ou detectam navios (BEŞBINAR; ALATAN, 2015; HE et al., 2018; WU et al., 2018).

Os desafios da segmentação semântica em regiões litorâneas dependem do sensor adotado, nos quais incluem câmeras hiper-espectrais, radar de abertura sintética (SAR) e câmeras ópticas.

A segmentação de regiões costeiras é facilmente obtida com imagens hiper-espectrais, porque a banda próxima à radiação infravermelha (NIR) é rapidamente absorvida pela água e eficientemente refletida pela vegetação terrestre e pelo solo árido (MCFEETERS, 1996). Portanto, a maioria das abordagens de segmentação costeira calcula o mapa do índice de água com diferença normalizada (NDWI) na banda NIR (KULELI et al., 2011).

Diferentemente das imagens multi e hiperespectrais, a segmentação semântica de imagens RGB

¹<http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>

²<http://deepglobe.org/challenge.html>

na banda visível em regiões litorâneas, foco deste trabalho, são mais desafiadoras. Enchentes, florestas e sombras apresentam situações ambíguas que podem confundir os algoritmos de segmentação (LI et al., 2018). A água sofre interferência de fatores atmosféricos, como nuvens, sombras e ondas geradas pelo vento. Além disso, a cor do mar muda dependendo do clima e da estação do ano. Alguns desses desafios foram previamente apresentados na Figura 1.5.

Neste trabalho, optou-se por utilizar imagens aéreas no espectro da luz visível representadas pelo sistema de cor RGB devido ao seu baixo custo e disponibilidade de imagens com mais alta resolução. Estas características tornam o sistema de visão aérea mais comumente utilizado em drones.

2.2.1 Trabalhos relacionados a Segmentação de Imagens Aéreas no Espectro Visível RGB em Áreas Costeiras e Portuárias

A segmentação semântica de imagens aéreas em regiões urbanas é bastante explorada (DEMIR et al., 2018). Por outro lado, as pesquisas relacionadas a segmentação em áreas costeiras e portuárias são bastante limitadas.

O trabalho de LIU; JEZEK (2004) propôs um método de segmentação baseado em abordagens tradicionais, pois na época não existiam tecnologias para aplicação eficiente das RNCs. O pipeline de segmentação proposto é dividido em três estágios: pré segmentação, segmentação, e pós segmentação. A pré segmentação busca a redução de ruídos e o melhoramento das bordas dos elementos da imagem. São aplicados o filtro Gaussiano e o filtro de Lee (LEE, 1986). No estágio de segmentação, a imagem é particionada em regiões com água e regiões com terra. Nesse processo é adotado um método de binarização de imagens com limiarização localmente adaptativa. O estágio de pós segmentação busca a melhoria dos segmentos com base no uso de operadores morfológicos de dilatação e corrosão (PARKER, 2010). Os autores avaliam o método proposto em duas aplicações: a identificação da costa da Antártica utilizando imagens de radar do satélite canadense Radarsat-1 (JEZEK, 1999), e a identificação da costa do Golfo do México, utilizando imagens ópticas do satélite Landsat 7. Resultados qualitativos foram apresentados demonstrando o sucesso do método.

No trabalho de NIE; ZHANG; JIANG (2016) foi proposto um método de segmentação para regiões litorâneas aplicado em imagens de portos com base em *template matching*. O método identifica as linhas costeiras em portos usando o método limiarização adaptativa de imagens de OTSU (1979) para aprimorar a segmentação das imagens.

O trabalho de LI et al. (2018) propôs a DeepUNet, uma RNC baseada na arquitetura da U-Net (RONNEBERGER; FISCHER; BROX, 2015), para realizar a segmentação semântica de imagens ópticas de satélite. A *DeepUNet* segmenta as imagens em duas classes: terra e água. As principais contribuições do trabalho foram a proposta da rede *DeepUNet* que introduz *DownBlocks* e *UpBlocks* nas contração e expansão do fluxo de dados da arquitetura U-Net, e a criação de um novo conjunto de dados para o treinamento de redes neurais para o problema

de segmentação semântica de áreas marítimas.

O trabalho de HE et al. (2018) propõem um método de detecção de barcos que utiliza a rede neural Faster R-CNN (REN et al., 2015). Inicialmente uma rede realiza a extração de características da imagem. As características são compartilhadas com uma rede de propostas de região, do inglês *Region Proposal Network* (RPN), e com uma rede Faster R-CNN. As regiões de interesse identificadas pela RPN são utilizadas pela Faster R-CNN para a identificação final da localização dos barcos.

2.2.2 Considerações Relacionadas aos Trabalhos de Segmentação de Imagens Aéreas

Conforme representado na Tabela 1, propõe-se, de forma original, neste trabalho um método para segmentação que separa terra, água e barcos. Cada um desses elementos está dentro de uma das três classes propostas: Estruturas estacionárias (inclui a terra e todas estruturas que não mudam de posição), estruturas móveis (veículos de superfície como barcos e navios), e a água. A segmentação semântica pode ser usada como um mapa onde estruturas estacionárias se tornam obstáculos, estruturas móveis são eliminadas e a água se torna regiões livremente navegáveis. Até onde sabemos, essa aplicação de segmentação semântica, utilizando imagens RGB, nunca foi proposta antes.

Tabela 1: Trabalhos Relacionados a Segmentação Semântica de Imagens Aéreas

	A	B	C	D	E
(LIU; JEZEK, 2004)	X				
(DONG; ZHANG; XU, 2014)		X		X	
(BEŞBINAR; ALATAN, 2015)	X		X	X	
(NIE; ZHANG; JIANG, 2016)	X				
(LI et al., 2018)	X				X
(HE et al., 2018)			X		X
(WU et al., 2018)			X		X
Proposta deste trabalho	X		X		X
Legenda					
A	Segmentação de terra e água				
B	Detecção de orlas				
C	Detecção de barcos				
D	Utiliza imagem NIR				
E	Utiliza Aprendizagem profunda				

2.3 Melhoramento de Imagens Acústicas

Os métodos de pré-processamento e melhoramento de imagens acústicas subaquáticas tratam da redução de ruído das imagens, e da redução da distorção acústica. Os métodos propostos geralmente são apresentados como uma etapa de uma aplicação específica, como a criação de mosaicos, a identificação de similaridade entre imagens acústicas, reconstrução 3D, e localização

subaquática. Sendo que os trabalhos de construção de mosaico são os que apresentam os métodos mais sofisticados de pré-processamento e melhoramento de imagens acústicas.

No trabalho de [NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI \(2005\)](#), é realizado um estudo onde são apresentadas as principais características e os principais desafios dos sonares multifeixe de imageamento frontal, FLS. Experimentos foram realizados utilizando o sonar de alta resolução DIDSON, do inglês *Dual-frequency Identification Sonar*, onde os autores avaliam técnicas de pré-processamento das imagens visando a correção da distorção acústica, correção radiométrica das imagens e a redução de ruído.

O método de calibração apresentado para correção das distorções presentes nas imagens acústicas utiliza um objeto de calibração plano, em forma de *grid*. Um modelo de correção cúbico é utilizado para correção das imagens, cujo os coeficientes são estimados através da redução do erro de reprojeção dos cantos detectados no *grid*. Os autores apresentam o resultado de uma calibração utilizando 22 imagens acústicas.

A correção radiométrica é a compensação do decaimento da sensibilidade da lente acústica do sonar. Segundo os autores, o decaimento da sensibilidade acústica aumenta conforme os *beams* se afastam do centro da imagem. A correção da sensibilidade pode ser realizada através da média de uma grande sequência de imagens acústicas.

O tratamento de ruído proposto pelos autores também consiste em realizar a média de uma sequência de imagens acústicas, porém utilizando apenas os pixels das imagens que possuem sobreposição de visão. Desta forma, é necessário o alinhamento da sequência de imagens, para o posterior cálculo da média dos pixels sobrepostos. Os autores apresentaram um experimento onde o DIDSON foi instalado em um poste, em um tanque de testes. O poste é rotacionado de forma controlada, enquanto as imagens acústicas são capturadas. Conhecendo o movimento de rotação, e realizando *matching* de *features* das imagens, foi possível realizar o alinhamento, a média dos pixels sobrepostos, e a correção do ruído das imagens.

Também é apresentado um método de registro de imagens acústicas baseado na extração de cantos, utilizando o detector de Harris ([HARRIS; STEPHENS, 1988](#)). O *match* das detecções entre as imagens é filtrado utilizando RANSAC-SVD. Para o alinhamento das imagens é utilizado o modelo de transformação affine de corpos rígidos (4 parâmetros). Este modelo é uma aproximação simplificada da geometria das imagens acústicas, que foi adotado principalmente porque são necessárias poucas dezenas de pares de *features* correlacionadas para estimar uma transformação confiável entre as imagens. O método foi validado no mesmo tanque de teste, os resultados obtidos foram considerados confiáveis pelos autores.

No trabalho de [KIM; NERETTI; INTRATOR \(2006\)](#) é proposto um método de melhoramento de imagens acústicas de um FLS, estendido do estudo proposto pelos mesmos autores em ([KIM; NERETTI; INTRATOR, 2005](#)). O método utiliza uma sequência de imagens acústicas que são processadas em quatro etapas, são elas: (1) estimativa do perfil de iluminância e do perfil

de reflectância do sonar, (2) registro entre imagens consecutivas e entre imagens a parte, (3) linearização das intensidades acústicas, (4) fusão das imagens para o melhoramento da imagem mais recente.

Os perfis de iluminância e reflectância do FLS são obtidos utilizando a parte de baixa e de alta frequência das imagens respectivamente. Na etapa de registro de imagens, é proposto um algoritmo que determina quantas imagens serão utilizadas para o melhoramento da imagem atual. A escolha é baseada na correlação entre as imagens e considera a extração de suas *features*, seguido de remoção de *outliers* e matching das *features*. O algoritmo avalia não apenas a correlação entre as imagens consecutivas, mas também a correlação de todas as imagens com relação à primeira de cada sequência. O alinhamento das imagens utiliza transformação affine, encontrado através da minimização do erro de sobreposição entre as imagens consecutivas e entre a primeira imagem com relação às imagens da sequência.

Após o registro das imagens, ocorre a linearização das intensidades acústicas. Este processo foi proposto com base no estudo de (ROBERT, 1983), e utiliza ruído de Rician ao invés do ruído Gaussiano no processo de linearização das imagens. Finalmente, a última etapa do método realiza a fusão das imagens através de uma soma ponderada de todas as imagens da sequência avaliada, e transformadas na perspectiva da imagem atual. O peso de cada imagem é definido em função da intensidade luminosa, e do espalhamento das *features* detectadas. Como resultado, é apresentado o melhoramento de uma imagem do sonar de alta resolução DIDSON utilizando 7 imagens consecutivas.

O trabalho de RIBAS et al. (2006) apresenta um método de localização e mapeamento simultâneo (do inglês *Simultaneous localization and mapping* SLAM), para AUVs em ambientes parcialmente estruturados equipados com um sensor de velocidade DVL e um sonar mecânico de varredura de 360° (do inglês *Mechanical Scanning Image sonar*, MSIS). As distorções das imagens acústicas geradas pelo movimento do robô são corrigidas pela estimativa de deslocamento de um filtro de Kalman Estendido (EKF) utilizando um sensor de velocidade DVL.

As imagens acústicas são descritas através da detecção de linhas usando a transformada de Hough (BALLARD, 1981). Antes de calcular a transformada de Hough, é realizada a remoção de ruído das imagens através da eliminação dos pixels de baixa intensidade e da seleção dos máximos locais. As linhas são extraídas da imagem utilizando um esquema de votação. Devido ao atraso do sonar MSIS para realizar a varredura completa da cena, o SLAM precisa ser atrasado para que as correções visuais possam ser realizadas. Para determinar quando duas linhas de duas imagens acústicas consecutivas representam o mesmo ambiente na cena, é utilizado um teste de compatibilidade de linhas, onde uma equação é utilizada como métrica de comparação. Um mapa é construído com base nas linhas detectadas. Um experimento real é realizado utilizando o AUV GARBI (EL-FAKDI; CARRERAS; RIDAO, 2006) da Universidade de Girona.

No trabalho de JOHANSSON et al. (2010) é apresentado um método de SLAM para a vi-

gilância de portos e inspeção de casco de navios autônoma. O método utiliza informações de odometria combinado com um sonar DIDSON de 1.8 MHz. As correções de posição pela visão são estimadas através do alinhamento das imagens acústicas, onde é realizada uma densa extração de *features* das imagens.

Inicialmente é utilizado um filtro mediano para eliminar o ruído das imagens afetando o mínimo possível as suas bordas. Logo após é calculado o gradiente da imagem. Os pixels com gradiente superior a um determinado limiar são considerados *features* da imagem. O limiar é adaptado automaticamente em função do número de *features* encontrado, de forma que sempre seja extraída a mesma quantidade de *features* de cada imagem. Depois que todas as *features* foram extraídas, é realizada uma etapa de eliminação de *features* ruins *clusterizando* todas as *features* e eliminando os *clusters* pequenos. O alinhamento das imagens é encontrado utilizando as *features* extraídas no algoritmo *Normal Distribution Transform*(NDT)(BIBER; STRASSER, 2003) através da minimização do erro de sobreposição das *features*. O alinhamento das leituras é utilizado para estimar o deslocamento do veículo (x, y, θ) .

O trabalho de NEGAHDARIPOUR; AYKIN; SINNARAJAH (2011) trata dos desafios associados à busca por imagens acústicas similares. Foi demonstrada a dificuldade de uso do detector SIFT (LOWE, 2004) na detecção de similaridade entre imagens acústicas. O detector SIFT era considerado estado da arte para detecção de semelhanças entre imagens ópticas na época.

No trabalho de AYKIN; NEGAHDARIPOUR (2012) é apresentado um método de extração de *features* de imagens acústicas de um sonar de imageamento frontal (FLS) considerando o modelo de projeção 3D do sonar. No trabalho, é estimada a elevação ϕ da imagem acústica de acordo com a inclinação do sonar e o tamanho das sombras acústicas dos objetos detectados na cena. Este método é aplicado em regiões planas.

Inicialmente é realizada uma compensação de sensibilidade dos hidrofones do sonar, esta compensação pode variar de acordo com o sonar utilizado, neste caso foi utilizado o sonar DIDSON 1.8 MHz. Logo após são selecionados os pixels de alta intensidade da imagem. Esses pixels são *clusterizados* utilizando o algoritmo K-Means adaptativo proposto pelo autor, onde *clusters* pequenos são eliminados.

Basicamente o *K-Means* adaptativo utiliza dois parâmetros, a distância mínima entre dois agrupamentos T_d , e o tamanho máximo de um agrupamento T_σ . Esses parâmetros são utilizados durante as iterações do algoritmo, quando a distância entre o centro de dois agrupamentos for menor que T_d os agrupamentos são agrupados, ou quando o tamanho de um agrupamento for maior que T_σ ele é dividido.

A detecção da sombra acústica de cada objeto é realizada primeiro determinando dois limiares de intensidade, um para determinar se um pixel pertence a sombra acústica ou se pertence ao fundo do cenário, e outro para determinar se um pixel pertence ao fundo do cenário, ou se representa um objeto da cena. Esses limiares são encontrados baseados na distribuição de

intensidade dos pixels ao redor do objeto detectado. A análise dos pixels que representam a sombra acústica é realizada nos perfis de intensidade, suavizados, de cada *beam* na imagem crua $I(\theta, \rho)$.

No trabalho de [HURTÓS et al. \(2012\)](#) é proposto um método de alinhamento de imagens acústicas de um (FLS) utilizando as imagens no domínio da frequência. É utilizado o algoritmo de correlação de fase ([REDDY; CHATTERJI, 1996](#); [CASTRO; MORANDI, 1987](#)). Este método de correlação de fase demonstra-se ser robusto ao ruído, diferença de iluminação e oclusão presentes nas imagens acústicas.

O trabalho de [RIBEIRO et al. \(2018\)](#) propõem um método baseado em aprendizagem profunda para consulta de imagens acústicas similares –*image retrieval*–, utilizando FLS. Uma versão modificada da rede neural PoseNet ([KENDALL; GRIMES; CIPOLLA, 2015](#)) é adotada para descrever as imagens acústicas. A última camada da PoseNet, uma camada *full connected*, é removida, e são utilizadas as duas últimas camadas restantes na função de perda –*loss function*–, minimizada durante a etapa de treinamento.

A rede é treinada utilizando o método de *Triples*, requerendo um conjunto de treinamento composto por grupos de três imagens acústicas. Sendo uma imagem de referência, âncora, uma imagem negativa, capturada em uma região diferente da imagem âncora, e uma imagem positiva, capturada na mesma região da imagem âncora, mas em um instante e perspectiva diferente. O treinamento do modelo utiliza a função de perda proposta em ([VO; HAYS, 2016](#)).

O método proposto é comparado com o de [LI et al. \(2016\)](#), que também propõem uma abordagem de aprendizagem profunda para o reconhecimento de regiões similares. Três conjuntos de dados são utilizados, sendo um simulado pelo simulador de [CERQUEIRA et al. \(2016\)](#), e dois conjuntos de dados com imagens reais, ARACATI 2014 e ARACATI 2017 ([SANTOS et al., 2019a](#)). Duas versões das duas redes comparadas foram treinadas, sendo uma utilizando apenas os dados simulados, e uma utilizando apenas os dados do conjunto de dados ARACATI 2017. Os desempenhos das quatro redes foram avaliados utilizando o dataset ARACATI 2014. Nenhuma etapa de pré-processamento foi mencionada pelos autores. Os resultados revelaram que as redes treinadas com dados reais obtiveram maior acurácia de reconhecimento de imagens similares, e a rede proposta por [RIBEIRO et al. \(2018\)](#) obteve maior acurácia que a proposta por [LI et al. \(2016\)](#).

O trabalho de [PARK; KIM \(2022\)](#) propõem um método de localização subaquática que funde a estimativa de localização a partir de imagens acústicas de um sonar FLS, com os dados de velocidade de uma DVL, e as informações de aceleração de uma IMU. O método é designado para regiões subaquáticas de intervenção, onde as estruturas da cena foram introduzidas por humanos, e a sua forma e posição são precisamente conhecidas. O método consiste em reproduzir o mesmo cenário em um ambiente virtual de forma que seja possível gerar imagens acústicas simuladas semelhantes às reais de quando o veículo está navegando na cena.

As Imagens simuladas são geradas a partir de diferentes locais da cena formando um *grid*. Para cada nova imagem acústica coletada, é realizada uma busca pela imagem simulada com maior semelhança. O processo de busca é otimizado utilizando os métodos *minhashing* e hash sensível a localidade, do inglês *locality-sensitive hashing* (LSH).

O Minhashing realiza uma redução dimensional das imagens acústicas preservando a similaridade de Jaccard, já LSH é uma técnica de clusterização para criar grupos com imagens semelhantes.

Para estimar a localização da imagem acústica real no espaço contínuo, é realizado o alinhamento entre a imagem acústica real e a imagem acústica simulada com maior similaridade. A busca pelo melhor alinhamento utiliza duas métricas: a distância de intensidade dos pixels, e a distância de Chamfer, que considera, de forma indireta, as diferenças geométricas espaciais dos dados das imagens.

Para calcular a distância de Chamfer são utilizadas as bordas detectadas em ambas imagens utilizando o método de [TRUJILLO-PINO et al. \(2013\)](#). Finalmente, se o processo de alinhamento obter sucesso, a pose estimada é fundida com os sensores de velocidade (DVL) e aceleração (IMU), utilizando o filtro de Kalman Estendido. O método é validado em um experimento real realizado em um tanque de testes adaptado com estruturas conhecidas.

2.3.1 Considerações aos Trabalhos Relacionados ao Melhoramento de Imagens Acústicas

Os trabalhos apresentados foram classificados na Tabela 2 com relação a sua aplicação, e às técnicas de pré-processamento empregadas. É possível observar que os métodos de criação de mosaico apresentam técnicas de melhoramento de imagem mais complexas com base no alinhamento de múltiplas imagens para o tratamento do ruído ([NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI, 2005](#); [KIM; NERETTI; INTRATOR, 2006](#); [HURTÓs et al., 2012](#)). Já os trabalhos que realizam apenas a localização subaquática utilizam técnicas mais simples, que consideram apenas a imagem que está sendo tratada ([RIBAS et al., 2006](#); [JOHANSSON et al., 2010](#)).

Como os métodos de construção de mosaicos utilizam todos os pixels da imagem em sua solução, a recuperação de toda imagem, inclusive de regiões de baixa intensidade, recebe maior importância. Já métodos que realizam apenas localização, geralmente exploram apenas as características mais relevantes das imagens, como regiões com retornos acústicos de maior intensidade ([JOHANSSON et al., 2010](#)), ou que representam estruturas do ambiente imageado ([RIBAS et al., 2006](#)). Essas regiões, por possuírem um sinal acústico mais intenso, são menos afetadas pelo ruído, e muitas vezes são suficientes para a estimativa do deslocamento do veículo. Por esse motivo, técnicas que utilizam filtro médio, ou mediano, são suficientes para suavizar a imagem e eliminar *outliers* para o problema de localização, mas não são suficientes para o problema de construção de mosaicos.

Também vale ressaltar que as técnicas de melhoramento que utilizam sequências de imagens necessitam que todas as imagens estejam alinhadas sob a perspectiva da imagem que está sendo tratada. Muitas vezes o alinhamento acaba recaindo no problema de encontrar a correlação entre as imagens acústicas sem o tratamento do ruído. Por exemplo, em [NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI \(2005\)](#) o alinhamento das imagens é conhecido porque o experimento é realizado em um tanque de testes com um sonar instalado em um poste que se movimenta de forma controlada através dos seus motores. Em [HURTÓŠ et al. \(2012\)](#) o alinhamento é estimado por um método robusto ao ruído acústico baseado na correlação da fase das imagens no domínio da frequência. O melhoramento das imagens acaba sendo realizado apenas para a construção dos mosaicos.

Uma outra diferença, é que métodos de construção de mosaico não possuem restrições temporais rígidas porque o processamento das imagens pode ser realizado de forma *offline*. Já os métodos de localização possuem restrições temporais mais rígidas, porque precisam fornecer informações em tempo real para a navegação do veículo.

Observa-se ainda que os trabalhos baseados em aprendizagem profunda, como os de reconhecimento de regiões através de imagens acústicas propostos por [LI et al. \(2016\)](#) e por [RIBEIRO et al. \(2018\)](#), não adotam uma etapa de pré-processamento específico para imagens acústicas, porque assumem-se que este tratamento pode ser aprendido pela rede neural. Ainda, muitas vezes ocorre a redução da resolução das imagens para que possam ser utilizadas como entrada da rede de forma eficiente .

Com relação aos métodos de correção da distorção das imagens acústicas propostos, observa-se que são atrelados principalmente ao tipo, e a tecnologia do sonar utilizado. Por exemplo, no trabalho de [NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI \(2005\)](#) é realizado correção das distorções das imagens do sonar *Dual-frequency identification sonar* (DIDSON), que é do tipo multi feixe e de alta resolução. O modelo de correção é calibrado através do imageamento de um objeto plano em forma de *grid*. Em ([RIBEIRO et al., 2018](#)) é utilizado um sonar do tipo varredura mecânico de um único feixe, neste sonar a formação da imagem depende da varredura mecânica do ambiente. A distorção das imagens ocorre devido a movimentação do veículo durante a varredura acústica. Os autores realizam a correção das distorções das imagens através da compensação do movimento do veículo. Entretanto esses métodos não se aplicam ao sonar, do tipo multi feixe, BlueView P900, adotado nesta pesquisa pelos seguintes motivos:

- A resolução das imagens do sonar DIDSON utilizado em ([NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI, 2005](#)) é muito maior (2.5mm por *bin*) do que as do sonar BlueView P900 (25.4mm por *bin*), de forma que seria necessário um objeto de calibração de proporções muito maior para reproduzir o experimento. Além disso, o campo de visão dos sonares também são significativamente diferentes, enquanto uma imagem do sonar DIDSON cobre uma distância de até 10 metros, a imagem do sonar BlueView P900 pode cobrir até 100 metros.

Tabela 2: Relação de trabalhos com aplicações utilizando FLS

	Aplicação			Pré-Processamento			
	A	B	C	D	E	F	G
1	X			X		X	X
2				X		X	
3		X			X		X
4		X			X		
5	X						
6	X					X	
7	X			X		X	
8			X				
9			X				
10		X					

Legenda	
Aplicações	
A	Mosaicos, Registro de Imagens
B	Localização
C	Similaridade, Identificação de Regiões Similares
Pré-Processamento	
D	Tratamento de ruído utilizando múltiplas imagens
E	Tratamento de ruído na mesma imagem
F	Correção de iluminação (isonificação)
G	Correção de distorções
Trabalhos	
1	(NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI, 2005)
2	(KIM; NERETTI; INTRATOR, 2006)
3	(RIBAS et al., 2006)
4	(JOHANNSSON et al., 2010)
5	(NEGAHDARIPOUR; AYKIN; SINNARAJAH, 2011)
6	(AYKIN; NEGAHDARIPOUR, 2012)
7	(HURTÓZ et al., 2012)
8	(RIBEIRO et al., 2018)
9	(SANTOS et al., 2019a)
10	(PARK; KIM, 2022)

- O tipo de sonar adotado em (RIBEIRO et al., 2018) é diferente do adotado neste trabalho, as distorções das imagens de um sonar mecânico de um único feixe são causadas por motivos diferentes das distorções das imagens dos sonares multi feixe.

Finalmente, com relação ao método proposto em (NEGAHDARIPOUR; FIROOZFAM; SABZMEYDANI, 2005) para correção da iluminação não uniforme das imagens acústicas, onde um perfil de iluminação do sonar é estimado através da média de um grande conjunto de imagens, observa-se uma melhoria na qualidade das imagens. Entretanto, a aplicação é focada na criação de mosaicos, onde o sonar está sempre imageando o chão. A aplicação proposta neste trabalho utiliza o sonar para identificar estruturas na coluna d'água, resultando em situações em que o fundo está menos presente nas imagens acústicas. Além disso, a correção da ilu-

minação acústica afeta principalmente as regiões mais escuras das imagens, não tendo efeito significativo em regiões de alta intensidade³, que geralmente são as mais relevantes para estimar a localização.

2.4 Considerações Finais

Este capítulo apresentou um panorama dos principais trabalhos relacionados ao tema de pesquisa desta tese. Apesar de não se ter conhecimento de métodos na literatura que utilizam imagens multi-domínio e multi-perspectiva para localização subaquática, foram apresentados trabalhos em aplicações diferentes que utilizam técnicas que inspiraram a proposta metodologia deste trabalho.

Foi apresentado que a localização de imagens terrestres utilizando imagens aéreas possui desafios similares aos do problema proposto nesta pesquisa, principalmente porque ambos os problemas tratam de imagens capturadas em perspectivas diferentes (visão superior, e frontal). A visão superior das imagens aéreas possui um campo de visão maior do que o da visão frontal das imagens terrestres, fazendo-se necessário o emprego de técnicas inteligentes para buscar informações das imagens terrestres de consulta, nas imagens aéreas de referência, e estimar a correlação entre elas.

Foram apresentados métodos que encontram esta correlação utilizando aprendizagem profunda, informações semânticas das regiões, e que se baseiam na descrição de estruturas presentes nas imagens. Para a localização terrestre, alguns trabalhos utilizaram o filtro de partículas, que é uma estratégia conveniente para este problema, porque além de permitir a fusão dos sensores do veículo, o estado das partículas podem guiar as buscas por informações das imagens terrestres nas imagens aéreas.

Em busca de inspiração para o tratamento individual das imagens, foram apresentados trabalhos relacionados a segmentação semântica de imagens aéreas em regiões costeiras e portuárias, e uma avaliação das técnicas de pré-processamento e melhoramento de imagens acústicas.

Baseado no estudo e na avaliação desses trabalhos, o Capítulo 3 apresenta uma metodologia para o tratamento específico do problema de localização subaquática utilizando imagens óptica aéreas e imagens acústicas subaquáticas.

³A diferença de sensibilidade entre as lentes acústicas não são de grandes magnitudes, resultando em uma correção de iluminação também de baixa magnitude, afetando principalmente a aparência do fundo da imagem, sem alterações significativas nas regiões de alta intensidade.

3 METODOLOGIA

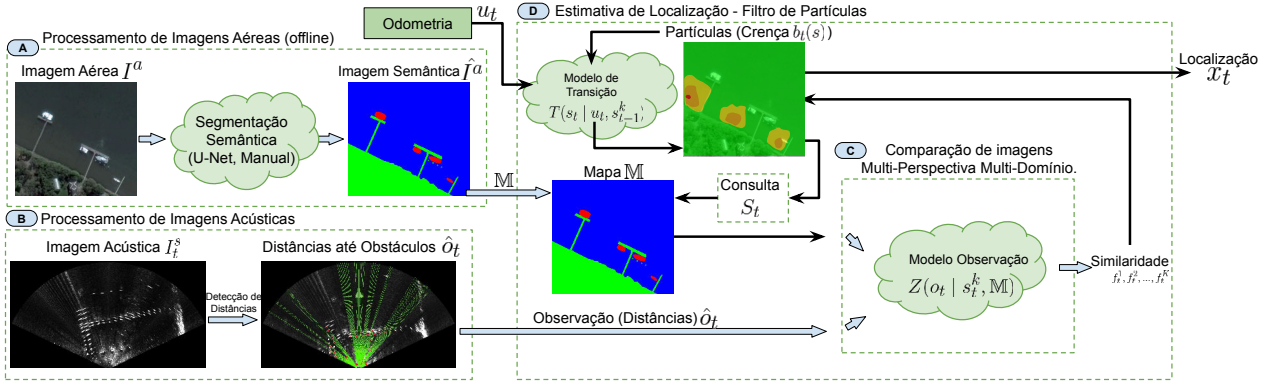


Figura 3.1: Pipeline do método proposto.

O método de estimativa de localização subaquático multi-perspectiva e multi-domínio proposto nesta tese combina imagens ópticas aéreas com imagens acústicas subaquáticas. As imagens ópticas aéreas, fornecidas por drones ou satélites, são utilizadas como mapa da região, enquanto que as imagens acústicas subaquáticas são observações captadas pelo veículo que está sendo localizado. A localização se dá pela estimativa da correlação entre as imagens acústicas e aéreas, utilizando o método probabilístico do filtro de partículas (THRUN; BURGARD; FOX, 2006).

Por esse motivo, o método é designado para operar em regiões subaquáticas parcialmente estruturadas, como portos e marinas. As estruturas parcialmente submersas encontradas nessas regiões, como píeres, cais, e a linha da costa, são detectadas nas imagens de ambos domínios e são utilizadas na estimativa da localização do veículo.

A Figura 3.1 apresenta o pipeline do método proposto. Os dados de entrada consistem na imagem aérea I^a de cobertura do ambiente navegável, imagens acústicas I_t^s observadas em cada instante t pelo veículo, e o sinal de controle u_t acionado no instante t pelo veículo. Alternativamente, o sinal u_t pode ser uma estimativa de deslocamento de um sistema de odometria, como odometria acústica, ou de um sensor DVL. Como resultado, o método estima a localização X_t mais provável do veículo no instante t .

De forma similar ao filtro de partículas aplicado a localização de robôs terrestres proposto por THRUN; BURGARD; FOX (2006), o método mantém um mapa estático do ambiente $\mathbb{M}$, um conjunto de K partículas $s_t^1, s_t^2, \dots, s_t^K \in S_t$ para cada instante t , assim como um modelo de transição $T()$ e um modelo de observação $Z()$.

A crença $b_t(s)$ de uma partícula s no instante t é dada pela distribuição de probabilidade,

$$b_t(s) \approx \langle s_t^k, w_t^k \rangle_{k=1:K}, \quad (3.1)$$

onde $\sum_k w_k = 1$ e K é o número total de partículas, s_k é o estado da partícula, w_k é o peso, ou

importância, da partícula, e t denota o tempo. O estado da partícula adotado neste trabalho consiste na posição x , y , e orientação θ , $s = [x \ y \ \theta]^T$ do veículo.

A representação simplificada do estado, utilizando três graus de liberdade, 3DoF, é adotada devido às limitações das imagens aéreas e subaquáticas acústicas que dificulta a estimativa dos ângulos de arfagem (*pitch*) e rolamento (*roll*), assim como a profundidade do veículo na água. O método proposto não é capaz de estimar estes graus de liberdade, porém, os ângulos *roll* e *pitch* podem ser facilmente estimados utilizando giroscópio e magnetrômetro, e a profundidade pode ser facilmente estimada por um sensor de pressão. Essas informações podem ser fundidas com a estimativa de localização do método proposto⁴.

As partículas são constantemente atualizadas de forma Bayesiana (THRUN; BURGARD; FOX, 2006), sendo amostradas através do modelo de transição da Equação 3.2,

$$s_t^k \sim T(s_t \mid u_t, s_{t-1}^k), \quad (3.2)$$

o qual considera o estado anterior da partícula s_{t-1}^k , e o último sinal de controle u_t . Após a amostragem, os pesos das partículas são atualizados através do modelo observação

$$f_t^k = Z(o_t \mid s_t^k, \mathbb{M}), \quad (3.3)$$

onde f_t^k é a estimativa da observação o_t estar no estado da partícula s_t^k , considerando o mapa de referência do ambiente $\mathbb{M}$. Este valor decorre da estimativa de similaridade proposta nesta tese que trata as observações das leituras do sonar realizadas pelo veículo durante a navegação, e do mapa construído a partir das imagens aéreas que cobrem o ambiente.

O peso, ou importância, das partículas são atualizados considerando a probabilidade f_t^k de forma que

$$w_t^k = \eta f_t^k, \quad (3.4)$$

Onde η é o fator de normalização dos pesos calculado de forma que $\eta^{-1} = \sum_{j=1:K} f_t^j$.

Desta forma, o pipeline proposto possui quatro etapas que possibilitam a execução do filtro de partículas no contexto do problema deste trabalho.

A Etapa A - Processamento de Imagens Aéreas— é realizada de forma offline, uma única vez, e consiste na segmentação semântica das imagens aéreas I^a . Por não depender dos dados de navegação, a segmentação pode ser realizada de forma manual, ou de forma automática

⁴Apesar do controle dos ângulos *pitch* e *roll* serem fundamentais para navegação de veículos autônomos do tipo torpedeiro, o veículo adotado neste estudo é um ROV *open-frame* que possui estabilização passiva de *pitch* e *roll* devido a suas características de flutuabilidade e centro de massa.

utilizando uma rede neural convolucional (RNC). As imagens são segmentadas em três classes: estruturas estacionárias, estruturas móveis e a água. A segmentação permite remover objetos móveis, como barcos e navios, e usar as estruturas estacionárias como um mapa de referência para a localização do veículo embaixo d'água.

A **Etapa B –Processamento de Imagens Acústicas**– é executada constantemente para cada imagem acústica I_t^s capturada no instante t ao longo da missão. Consiste no pré-processamento das imagens para o tratamento de ruídos de forma que somente as formas dos objetos com uma clara resposta acústica prossigam para o processo de correspondência da etapa C, na forma de uma observação o_t .

A **Etapa C –Comparação de Imagens Multi-Domínio e Multi-Perspectiva**– representa o modelo de observação do filtro de partículas da Equação 3.3. É comparada a observação atual o_t , originada de uma imagem acústica, com o mapa estático da cena $\mathbb{M}$, originado de imagens aéreas. A observação atual o_t é projetada no mapa estático $\mathbb{M}$ considerando o estado s_t^k da partícula k . Através da sobreposição das informações é estimada a similaridade f_t^k para a partícula k . Esse processo é realizado para todas as partículas K no instante t . Nesta fase propõem-se duas alternativas de implementação desta etapa: i) abordagem probabilística de baixa complexidade, e ii) abordagem conexionista baseada em dados. Nos dois casos, o resultado final desta etapa é a atualização dos pesos, ou importância, das partículas segundo a Equação 3.4.

Etapa D –Estimativa de localização– A estimativa de localização subaquática é realizada pelo algoritmo de localização de Monte Carlo (MCL), também conhecido como filtro de partículas, descrito em (THRUN; BURGARD; FOX, 2006). Diversas hipóteses da localização do veículo são espalhadas na forma de partículas no mapa $\mathbb{M}$. Conforme novas imagens acústicas são adquiridas, a confiança das partículas é atualizada através das correspondências calculadas na etapa C.

As etapas propostas serão discutidas em maiores detalhes nas próximas seções.

3.1 Etapa A: Processamento de Imagens Aéreas

A navegação em cenários dinâmicos é desafiadora devido às constantes mudanças nas posições dos objetos e das estruturas de uma mesma região. Uma vez que as estruturas estacionárias agem como um ponto de referência para a localização, elas precisam ser identificadas. São exemplos de estruturas estacionárias o cais e os seus postes de sustentação, as pedras e o formato da orla.

Já as estruturas móveis se resumem a barcos e navios. Embora as estruturas móveis possam ser detectadas nos dois domínios, aéreo e subaquático, não se tem garantia de sincronismo entre elas. Ou seja, um barco detectado no domínio das imagens aéreas pode não estar no mesmo local quando as imagens acústicas subaquáticas forem capturadas pelo veículo. A Figura 3.2 ilustra

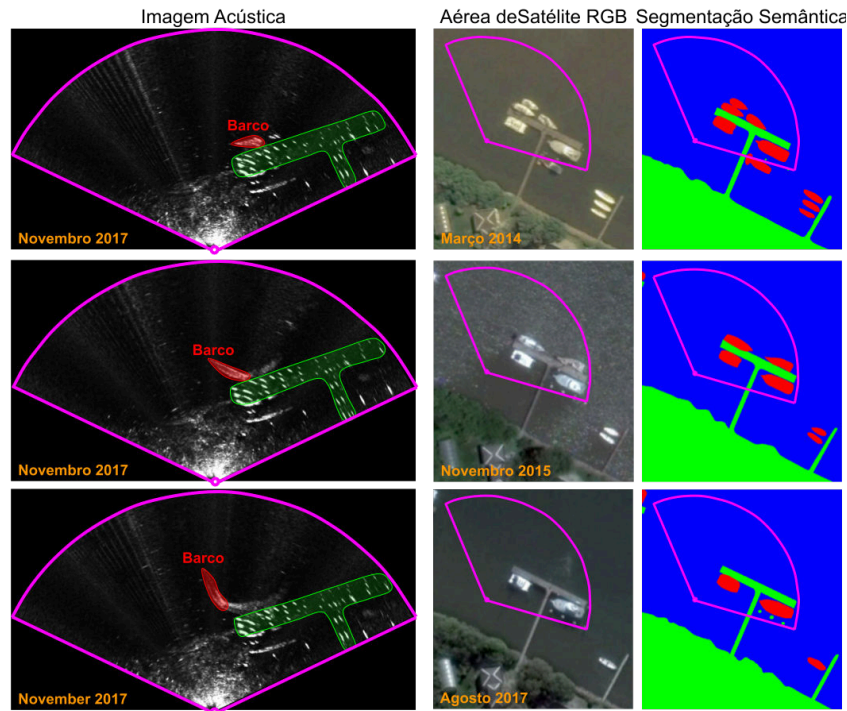


Figura 3.2: Comparativo das imagens acústicas subaquáticas e imagens aéreas. As estruturas móveis estão destacadas em vermelho, as estruturas estacionárias estão destacadas em verde. A coluna à esquerda mostra imagens acústicas capturadas em um curto intervalo de tempo por um veículo subaquático durante sua navegação. A coluna central mostra imagens aéreas de satélite capturadas na mesma região em datas diferentes, disponibilizadas por Google e Digital Globe. A coluna à direita mostra as respectivas segmentações semânticas das imagens aéreas. É possível observar que, diferentemente das estruturas estacionárias, a posição dos barcos mudam com o passar do tempo e portanto não são confiáveis para localização.

este problema comparando imagens aéreas de satélite com imagens acústicas subaquáticas.

A segmentação semântica das imagens aéreas permite filtrar as estruturas móveis, evitando correspondências equivocadas nas imagens subaquáticas. Assim, permitindo a extração das estruturas estacionárias que são importantes para o processo de correspondência e localização.

Graças ao sucesso das técnicas de aprendizagem profunda (DONG; WANG; ABBAS, 2021), uma RNC é proposta para realizar a segmentação semântica das imagens aéreas de maneira automatizada. A rede proposta é uma variação da arquitetura U-Net (REN et al., 2015), que usa *skip connections* para transmitir informações residuais. A U-Net (RONNEBERGER; FISCHER; BROX, 2015) foi proposta inicialmente para a segmentação de imagens biomédicas. Algumas arquiteturas baseadas nela foram utilizadas para aplicações, como classificação, reconstrução e tradução de imagens (GIACOMO et al., 2018).

Imagens aéreas com o tamanho $256 \times 256 \times 3$ são usadas como entrada da rede, enquanto a saída é uma imagem segmentada $256 \times 256 \times 3$ na codificação *one-hot*. Conforme exibido na Figura 3.3, a arquitetura consiste em cinco camadas de codificação e quatro camadas de decodificação, utilizando *padding* para preservar o tamanho original das imagens. A função

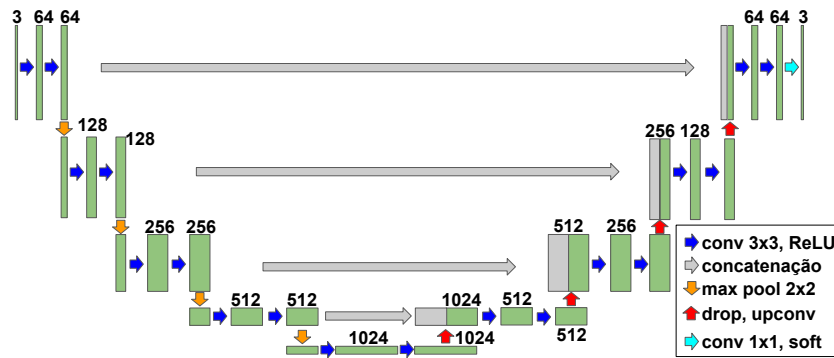


Figura 3.3: Arquitetura de rede neural baseada na U-Net utilizada para a segmentação de imagens aéreas RGB.

softmax é utilizada como função de ativação. Como função objetivo *loss function* para segmentação de múltiplas categorias (3 classes) é utilizado a entropia cruzada *softmax cross entropy*. O modelo é treinado utilizando o otimizador Adam (KINGMA; BA, 2014) e um tamanho de lote de 16, com a arquitetura sendo implementada no TensorFlow.

Dataset e Treinamento

Devido à inexistência de datasets associados ao problema se faz necessário o desenvolvimento de um conjunto de imagens anotadas conforme as estruturas e suas classes presentes na cena.

O conjunto de dados de treinamento foi criado anotando manualmente imagens de satélite de diferentes marinas capturadas em diferentes anos. Como a resolução das imagens varia conforme o sensor, uma operação de escala com interpolação foi aplicada para manter todas as imagens em uma resolução de 0,2 metros por pixel. Imagens de baixa resolução comumente encontradas na internet dificultam a segmentação devido à ausência de detalhes, que neste caso é preenchido pelo algoritmo de interpolação.

As imagens foram segmentadas manualmente e uma estratégia de aumento de dados *Data Augmentation* foi utilizada. Vários *patches* de 256×256 são recortados automaticamente das imagens de satélite originais, seguindo uma sequência de movimentos aleatórios com rotação e translação. O processo de extração dos recortes registra a ocorrência de cada classe em cada recorte, e realiza um controle estatístico de forma que o conjunto de dados de treinamento final seja equilibrado⁵.

Inferência de Imagens Aéreas RGB

Tipicamente uma imagem aérea cobre uma grande região da superfície terrestre e por esse motivo elas tendem a ser grandes. Assim, a avaliação da rede, ou seja, a segmentação semântica das imagens aéreas, é realizada em três etapas: *divisão*, *avaliação* e *combinação*.

Inicialmente uma imagem aérea de qualquer tamanho passa por uma operação de *divisão*, onde

⁵Uma demonstração em vídeo do método de aumento de dados está disponível em <https://bit.ly/2kyC3UZ>.

uma janela deslizante 256×256 percorre as linhas e colunas da imagem com um passo de 128 pixels realizando vários recortes. O tamanho da imagem original é ajustado através do preenchimento de borda espelhada *-padding-* caso o seu tamanho não seja múltiplo dos passos da janela deslizante. Desta forma, como passo é de metade do tamanho da janela, todos os recorte se interseccionam quatro vezes.

Todos os recortes são *avaliados* individualmente pela rede, resultando em diversas segmentações.

Na combinação, as segmentações são projetadas em um campo de votação com o mesmo tamanho da imagem original. A classe final de cada pixel da imagem original é determinada pelo vencedor de uma eleição que considera os quatro votos originados das intersecções entre os recortes⁶, conforme representado na Figura 3.4.

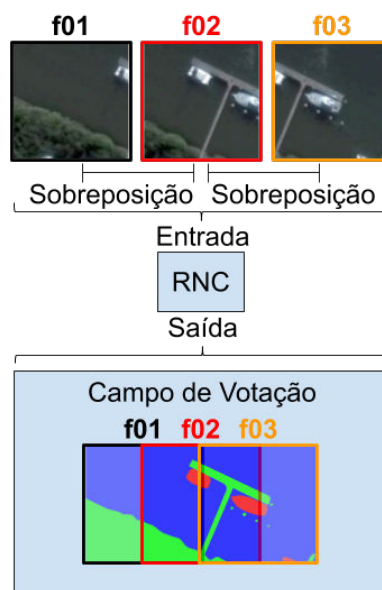


Figura 3.4: A avaliação de imagens aéreas de qualquer tamanho é realizada através da operação de divisão, avaliação e combinação. A combinação utiliza a sobreposição dos recortes projetados em um campo de votação para determinar a classe final de cada pixel da imagem.

Essa abordagem permite avaliar qualquer tamanho de imagem e reduzir classificações imprecisas da rede.

3.2 Etapa B: Processamento de Imagens Acústicas

Esta etapa transforma uma imagem acústica I_t^s em uma observação o_t para o filtro de partículas. O método proposto é inspirado na ideia de transformar os *beams* de uma imagem acústica em distâncias, obtendo-se uma leitura similar às de um sensor *laser scanner* 2D.

Cada *beam* B_t^k , onde $k = 1:K$, e K é o número total de *beams* da imagem acústica I_t^s , cobre uma direção θ_k do campo de visão do sonar, e é composto por um conjunto de *bins*.

⁶Em caso de empate, o algoritmo escolhe a classe de menor número, sendo 1 para objetos móveis, 2 para água e 3 para estruturas estacionárias.

O método analisa a intensidade de cada *bin* $b_{k,l}$ de um *beam* k , mantendo a intensidade médias dos último w_{sz} *bins*, sendo que $l = l_{ini}:L$, l_{ini} é o *bin* inicial da busca, L é o número total de *bins* ao longo da direção θ_k do *beam* k .

O objetivo da análise é encontrar o $\hat{l}$ -*bin* com intensidade acústica compatível a de um objeto. O $\hat{l}$ -*bin* é encontrado quando a intensidade média supera a intensidade limiar i_{min} . Obtém-se a distância ρ_k do objeto detectado no *beam* k com relação ao sonar através Equação 3.5.

$$\rho_k = \hat{l}/L * \rho_{max}. \quad (3.5)$$

Onde $\hat{l}$ é a posição do primeiro *bin* com intensidade média superior ao limiar i_{min} ao longo do *beam* k , L é o número total de *bins* de um *beam*, e ρ_{max} é a cobertura máxima do sonar em metros. Quando nenhum *bin* supera o limiar i_{min} , assume-se $\rho_k = \infty$.

A intensidade média dos w_{sz} últimos *bins* é utilizada para reduzir o problema com ruído acústico, e evitar que um *bin outlier* seja identificado como um objeto.

A Figura 3.5 mostra como as imagens acústicas são convertidas em distâncias. Os *beams* são representados pelas linhas verdes, o $\hat{l}$ -*bin* de cada *beam* são representados pelos círculos vermelhos.

Como resultado, a observação o_t deixa de ser uma imagem, e passa a ser composta por um conjunto de distâncias $z_k = (\theta_k, \rho_k)$ dos objetos detectados na imagem acústica, relativo ao referencial local do sonar, conforme ilustrado na Fig. 3.5⁷.

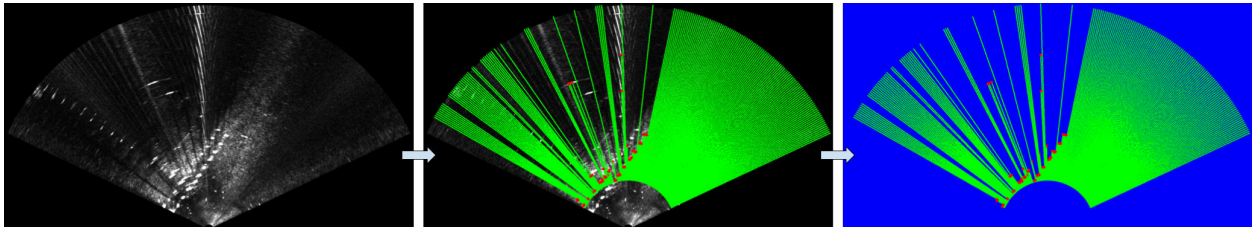


Figura 3.5: Processo de transformação de imagens acústicas subaquáticas em distâncias. Cada *beam*, representado pela linha verde, é analisado individualmente. A intensidade do primeiro *bin* superior ao limiar i_{min} é considerado como sendo o retorno acústico de um objeto é transformado em distância. Se nenhum *bin* obtiver intensidade suficientemente grande, é considerado que a distância é infinita.

A Figura 3.6 mostra o efeito dos parâmetros utilizados no método. Como existem diversos tipos de sonar, com diferentes campos de visão (HURTÓS, 2014), é necessário especificar essas propriedades para que o método analise corretamente a imagem acústica.

⁷Um vídeo demonstrado o processo de transformação está disponível em <https://youtu.be/XEUujtveE4w>.

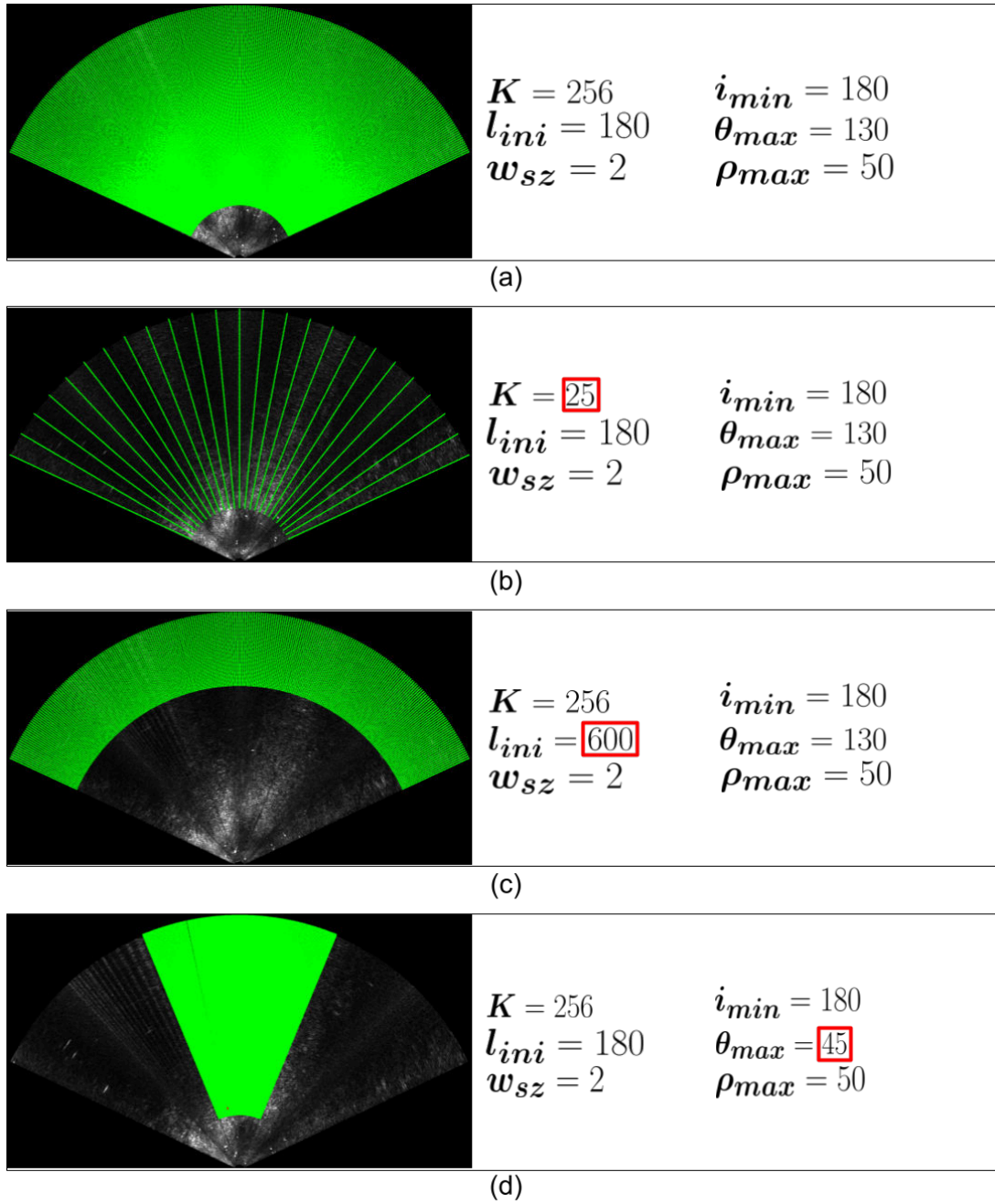


Figura 3.6: Exemplo do efeito provocado pelos parâmetros durante a transformação de imagens acústicas para distâncias. As linhas em verde representam os *beams* analisados na imagem. Em (a) é mostrado os parâmetros ideais para o sonar Teledyne Blue View P900-130, adotado nesta pesquisa. Em (b) é reduzido o número total de *beams* K de 256 para 25. Em (c) é alterado o *bin* inicial de análise, i_{min} , do pixel 180 para o pixel 600, finalmente em (d) é reduzida a abertura do sonar, θ_{max} , de 130 graus para 45 graus.

3.3 Etapa C: Comparação de Imagens Multi-Domínio e Multi-Perspectiva

Nesta etapa é realizada a fusão dos dados multi-perspectiva e multi-domínio. As informações geradas pelas imagens ópticas aéreas são comparadas com as informações das imagens acústicas subaquáticas, resultando no ajuste do peso, ou fator de importância, das partículas.

Nesta fase foram propostas e avaliadas duas abordagens distintas com o objetivo de se obter um pipeline genérico e um mais efetivo. A primeira abordagem é baseada em um modelo probabilístico específico para o processamento de imagens aéreas e acústicas de um FLS. O modelo possui baixa complexidade computacional podendo ser executado em dispositivos com baixo poder de processamento. A segunda abordagem é um modelo baseado em dados, ancorando-se em técnicas de *deep learning* para fornecer uma função de comparação que, em função dos dados disponíveis, pode ser utilizado em outros contextos envolvendo diferentes sensores, vistas e representações do ambiente.

As próximas seções descrevem cada uma das abordagens propostas.

3.3.1 Modelo de Observação Baseado em Campos de Probabilidade

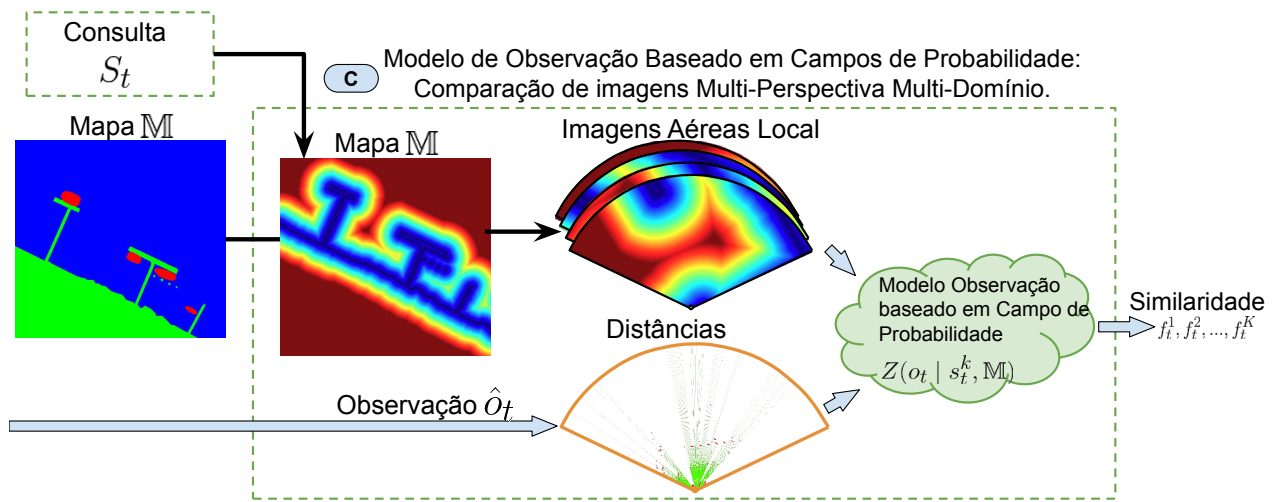


Figura 3.7: Modelo de Observação Baseado em Campos de Probabilidade.

Nesta primeira abordagem se propõe um método probabilístico de comparação entre as imagens ópticas aéreas e as imagens acústicas de um FLS capturadas pelo veículo subaquático.

A comparação utiliza o modelo de campos de probabilidade, conhecido como *Likelihood Field*, muito utilizados em sensores *laser scanner*, descrito em (THRUN; BURGARD; FOX, 2006). O método pode ser dividido em três etapas: preparação de dados aéreos, preparação de dados subaquáticos, e correspondência, conforme apresentado na Figura 3.7.

Preparação de dados aéreos A preparação dos dados aéreos ocorre uma única vez na inicialização do sistema. As imagens aéreas segmentadas $\hat{I}^a$, Figura 3.9-a, são transformadas no Mapa de Mínimas Distâncias M , Figura 3.9-c.

Inicialmente são extraídos os contornos dos segmentos das estruturas estacionárias, representado em verde na Figura 3.9-a. A identificação dos contornos pode ser obtida aplicando o operador de Sobel (IRWIN et al., 1968), resultando no Mapa Estacionário da Figura 3.9-b. A imagem do Mapa Estacionário utiliza o conceito de ocupação de espaço –Occupancy Grid– (MORAVEC; ELFES, 1985), onde cada pixel indica a presença ou não de uma estrutura na região.

A utilização dos contornos é inspirada pelo fato de que o sonar detecta as superfícies frontais das estruturas imaginadas, enquanto que as imagens aéreas capturam apenas as superfícies no topo das estruturas. Ao extrair o contorno dos segmentos aéreos, obtém-se uma aproximação do formato da superfície frontal das estruturas. Por exemplo, a Figura 3.8 compara a leitura acústica (b), com um recorte do mapa que utiliza contornos das estruturas da imagem aérea em (a), e um recorte do mapa que utiliza a estrutura completa em (c). Percebe-se que a imagem acústica detecta apenas o contorno da orla e não existem retornos acústicos posteriores na região que corresponde a terra.

Ao projetar a imagem acústica nos dois mapas, de forma a maximizar a sobreposição das detecções, percebe-se que o mapa com contornos da Figura 3.8-a é uma representação mais próxima e menos ambígua da imagem acústica do que o mapa que utiliza o segmento completo apresentado na Figura 3.8-c.

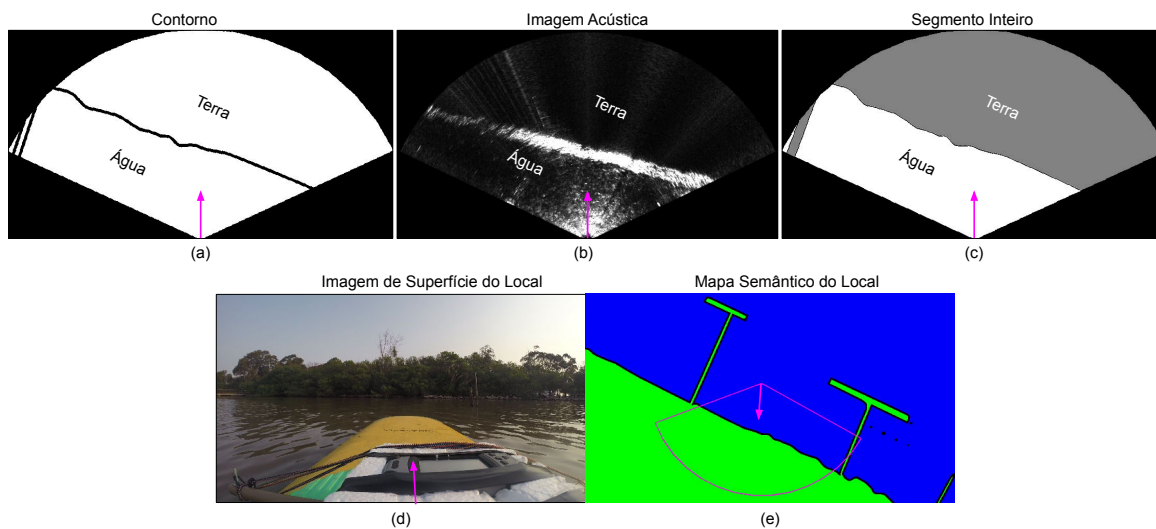


Figura 3.8: Comparação das percepções aéreas e subaquáticas. Em (a) um exemplo de mapa de ocupação utilizando o contorno das estruturas estacionárias das imagens aéreas. Em (b) uma leitura acústica na mesma região. Em (c) um exemplo de mapa de ocupação utilizando o segmento completo das estruturas estacionárias detectadas em imagens aéreas. Em (d) uma imagem capturada na superfície da água, no mesmo local. A imagem não é utilizada pelo método e foi incluída para facilitar a compreensão do leitor. Finalmente em (e) uma imagem de satélite segmentada do mesmo local. A seta em rosa representa a orientação de captura da imagem acústica no ambiente. O sonar captura apenas a superfície frontal das estruturas imageadas, já as imagens aéreas capturam a superfície do topo das estruturas, sendo necessária a utilização dos seu contorno para comparar com os dados acústicos.

A partir do Mapa Estacionário da Figura 3.9-(b), é calculado o Mapa de Mínimas Distâncias da Figura 3.9-(c). Neste mapa, o valor de cada pixel representa a distância até a região ocupada mais próxima, ou infinita, quando não existirem obstáculos dentro do alcance de busca determinado pelo parâmetro d_{max} . Para o caso da Figura 3.9-(c) foi utilizado $d_{max} = 20$ metros⁸.

⁸Um vídeo mostrando a imagem acústica, um recorte da imagem aérea e do Mapa de Mínimas Distâncias está disponível em: <https://youtu.be/fAjTXJn4gDU>

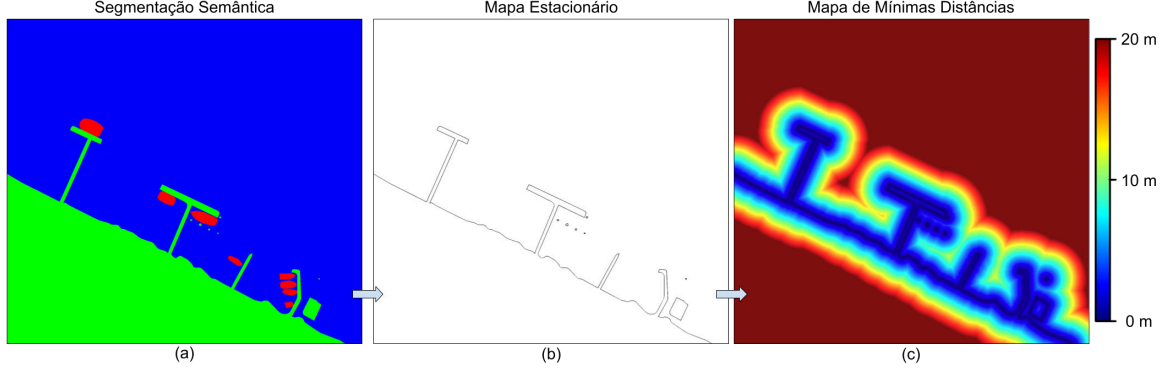


Figura 3.9: Processo de transformação da segmentação semântica de imagens aéreas em Mapa de Mínimas Distâncias $\mathbb{M}$. O canal verde, que representa as estruturas estacionárias, é utilizado para a identificação dos contornos. Posteriormente os contornos são utilizados para calcular as distâncias até os obstáculos mais próximos.

Preparação de dados subaquáticos Como as leituras de distâncias o_t , relativas aos objetos detectados na imagem acústica, estão no referencial de coordenadas do veículo, elas precisam ser transformadas para o sistema de coordenadas globais do mapa $\mathbb{M}$ antes de realizar a comparação.

O estado da partícula $s_t^k = (x \ y \ \theta)^T$ denota uma hipótese da posição do veículo no mapa $\mathbb{M}$, no instante t , e portanto pode ser utilizada na conversão de sistema de coordenadas. A localização relativa do sonar no sistema de coordenadas local do veículo é $(x_{son} \ y_{son})^T$, e a medição do sonar é indicada como $z_t^k = (\theta_t^k, \rho_t^k)$. A transformação das medições acústicas para o sistema de coordenadas global é alcançada por

$$\begin{pmatrix} x_{z_t^k} \\ y_{z_t^k} \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_{son} \\ y_{son} \end{pmatrix} + \rho_t^k \begin{pmatrix} \cos(\theta + \theta_t^k) \\ \sin(\theta + \theta_t^k) \end{pmatrix}, \quad (3.6)$$

onde $[x_{z_t^k} \ y_{z_t^k}]^T$ denota a posição da estrutura observada na imagem acústica, projetada no sistema de coordenadas global do mapa $\mathbb{M}$. Desta forma é possível comparar a projeção da estrutura detectada pelo sonar, com as detecções da imagem aérea, e estimar a confiança da hipótese representada pela partícula s_t^k .

Correspondência de dados A última etapa do modelo de observação é a correspondência de dados. Esta etapa atualiza o peso w_k^t de cada partícula $k = 1 : K$ com base na correspondência entre os dados aéreos, representados pelo mapa $\mathbb{M}$, e os dados subaquáticos, representados pela observação o_t .

Seja $p_t^j = [x \ y]^T$ a posição do objeto j no sistema de coordenadas do mapa, resultante da projeção de o_t em $\mathbb{M}$, utilizando a hipótese da partícula s_t^k , obtida pela Equação 3.6.

Seja z_t^j , a distância até o obstáculo mais próximo da posição do objeto p_t^j no mapa $\mathbb{M}$ definido como:

$$z_t^j = \mathbb{M}(p_t^j). \quad (3.7)$$

A probabilidade f_k^t da partícula k estar no local da observação do veículo, $Z(o_t|s_t^k, \mathbb{M})$, é dada pelo modelo de observação:

$$f_k^t = \prod_j z_{hit} e^{-\frac{z_t^{j2}}{2\sigma_{hit}^2}} + z_{rand} \frac{1}{d_{max}}, \quad (3.8)$$

onde J é o número total de objetos detectados na observação acústica o_t , σ_{hit} é o desvio padrão do ruído de medição do sonar, d_{max} é a distância máxima aceita no mapa $\mathbb{M}$, e z_{hit} e z_{rand} são coeficientes de balanceamento do modelo.

O modelo apresentado na Equação 3.8 utiliza uma distribuição gaussiana centrada em zero para modelar o ruído do sonar, e uma distribuição uniforme para modelar medições acústicas inexplicáveis, como leituras fantasmas ou leituras provocadas por interferência *cross-talk*.

Finalmente, a Equação 3.4 atualiza o peso das partículas normalizando a probabilidade de todas as partículas.

3.3.2 Modelo de Observação Baseado em Aprendizagem Profunda

Alternativamente, um Modelo de Observação baseado em Aprendizagem Profunda é proposto. Uma rede neural é utilizada para aprender a Equação 3.3 do Modelo de Observação, onde as imagens acústicas subaquáticas são comparadas com as imagens aéreas pela rede neural, e como resultado são estimados os *scores* de similaridade f_t^k entre as imagens.

A Figura 3.10 mostra o pipeline do Modelo de Observação utilizando aprendizagem profunda. Nesta abordagem, as observações $\hat{o}_t$ deixam de ser vetores de distância, e passam a ser a própria imagem acústica. As imagens aéreas segmentadas semanticamente também passam a ser utilizadas diretamente. Para cada imagem acústica $\hat{o}_t$ percebida no instante t , é realizada uma consulta ao mapa, considerando o estado de todas as partículas S_t no mesmo instante t . Como resultado da consulta, é gerado um recorte da imagem aérea segmentada para cada partícula k , considerando o seu estado s_t^k e o campo de visão do sonar em formato de cone. Os recortes são comparados com a imagem acústica $\hat{o}_t$, resultando nos *scores* f_t^k , utilizado posteriormente na atualização do peso das partículas.

Rede Neural QuadrupletNet A arquitetura da rede neural foi proposta em (GIACOMO et al., 2020) pensando na problemática deste trabalho. Foram considerados principalmente os desafios de comparar imagens multi-domínio e multi-perspectivas, e também a dificuldade de se obter dados com imagens acústicas reais e rotuladas para o treinamento de redes neurais.

Devido a grande diferenças entre as imagens multi-domínio e multi-perspectiva, a arquitetura

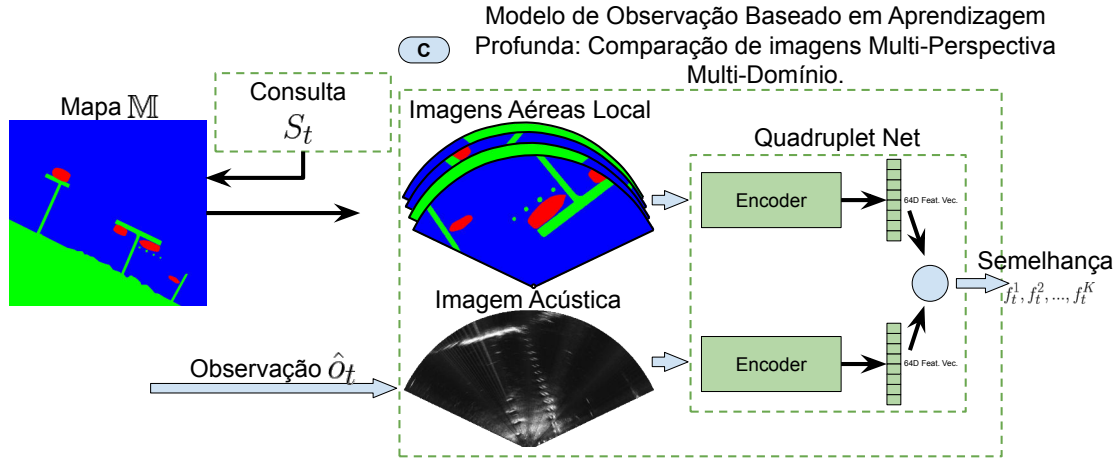


Figura 3.10: Pipeline do Modelo de Observação baseado em Aprendizagem Profunda.

QuadrupletNet utiliza uma rede neural para cada domínio de imagem. Desta forma, cada rede se especializa em aprender as características específicas de cada domínio. No caso do problema deste trabalho, é utilizada uma rede para as imagens acústicas, e outra para as imagens aéreas segmentadas. Essas redes são denominadas *Encoders*.

A arquitetura da rede *Encoder* foi baseada na rede *U-Net Encoder* (RONNEBERGER; FISCHER; BROX, 2015) e é composta por camadas de convolução dilatada, *max pooling* e ReLu. A convolução dilatada é uma operação definida em (STEFFENS et al., 2019; STEFFENS et al., 2020) que consiste em um bloco com quatro convoluções de *kernel* 3×3 , e taxas de dilatação de 1, 2, 4 e 8. Após a aplicação das convoluções, os mapas de características resultantes são concatenados antes de uma convolução final ser aplicada. O objetivo da convolução final é fundir as informações contidas nos mapas de características individuais. Expandir o campo receptivo da rede neural é o principal objetivo desta convolução (YU; KOLTUN, 2015).

Como entrada para a rede *Encoder*, é utilizada uma imagem com dimensões 256×128 e um número arbitrário de canais. Após a propagação da imagem, é produzido um vetor de 64 dimensões na saída armazenando a representação da imagem de entrada.

Treinamento Cooperativo Visando as dificuldades para preparar dados de treinamento com imagens acústicas subaquáticas, e a melhor performance do treinamento das redes *Encoders*, é proposta uma abordagem de aprendizagem cooperativa que explora a técnica das redes neurais *triplets* (HOFFER; AILON, 2015).

A ideia das *triplets* é que as redes *Encoders* são treinadas a partir de três exemplos anotados: um exemplo âncora, um exemplo positivo e um exemplo negativo. A ideia é que, à medida que a rede atualiza seus pesos, a saída produzida para a âncora se aproxime da saída do exemplo positivo e se distancie do exemplo negativo. O conceito de aproximação é baseado em uma métrica de similaridade. Neste trabalho é utilizado a distância L2.

A Figura 3.11 mostra como é realizado o treinamento da rede de *Encoder* para imagens acústicas

e da rede de *Encoder* para imagens aéreas segmentadas. A função de perda *Triplet Loss* representada pelo bloco azul garante que os pesos da rede no bloco amarelo sejam ajustados de forma que sua saída se aproxime do exemplo positivo e se afaste do exemplo negativo. Esta estratégia melhora a performance de treinamento das redes, e facilita a preparação dos dados, já que não é preciso criar pares de imagens com o *score* de similaridade anotado, basta selecionar as imagens como exemplo positivo, negativo ou âncora.

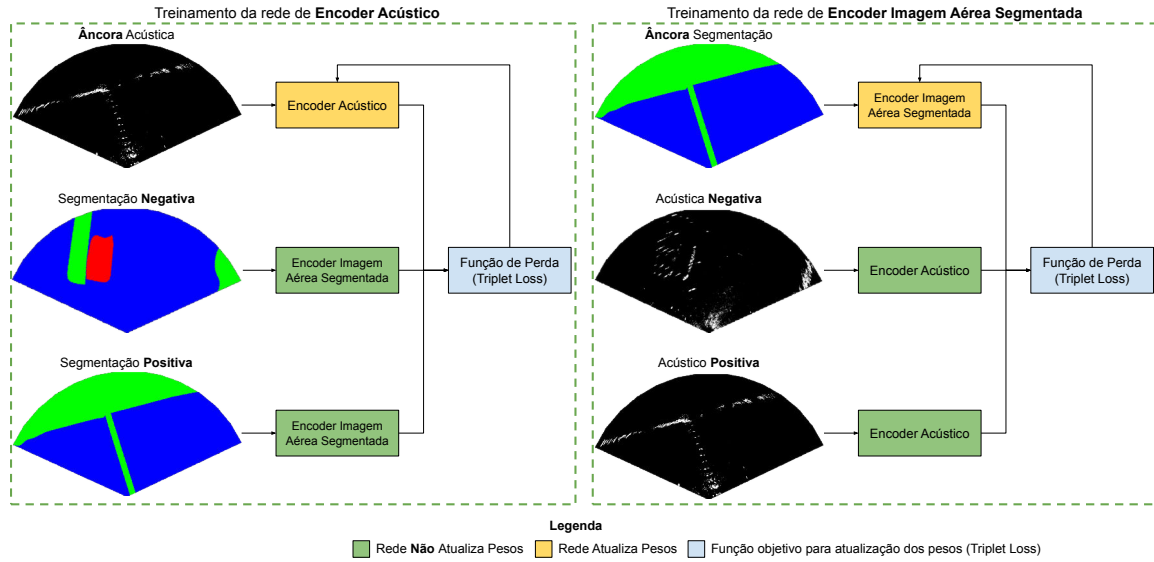


Figura 3.11: Exemplo de *Triplets* para o treinamento da rede de *Encoder* de imagens acústicas, e para a rede *Encoder* de imagens aéreas segmentadas.

A ideia das *Triplets* foi estendida para *Quadruplets*, onde os conjuntos com seis imagens utilizadas para treinar as duas redes de *Encoder* são reorganizadas em grupos de quatro imagens, formando dois pares. Onde cada par é composto por uma imagem acústica subaquática e uma imagem aérea segmentada ambas do mesmo local. O cruzamento dos pares forma as duas *Triplets* originais, conforme apresentado na Figura 3.12.

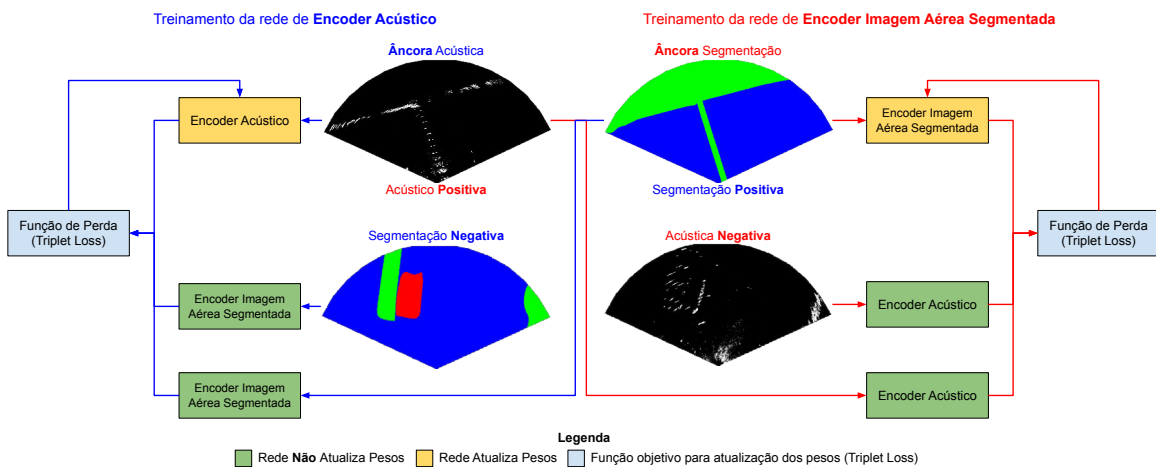


Figura 3.12: Exemplo de *Quadruplet* para o treinamento da rede de *Encoder* de imagens acústicas, e para a rede *Encoder* de imagens aéreas segmentadas.

3.4 Etapa D: Estimativa de Localização

O filtro de partículas estima a localização do veículo subaquático a cada instante. Este estudo utiliza o algoritmo Adaptativo de Monte Carlo (AMCL) descrito em (THRUN; BURGARD; FOX, 2006), adaptando a implementação disponível no pacote amcl do *framework Robot Operated System* (ROS), (QUIGLEY et al., 2009) para o problema proposto. O Diagrama da Figura 3.13 mostra as principais etapas do algoritmo.

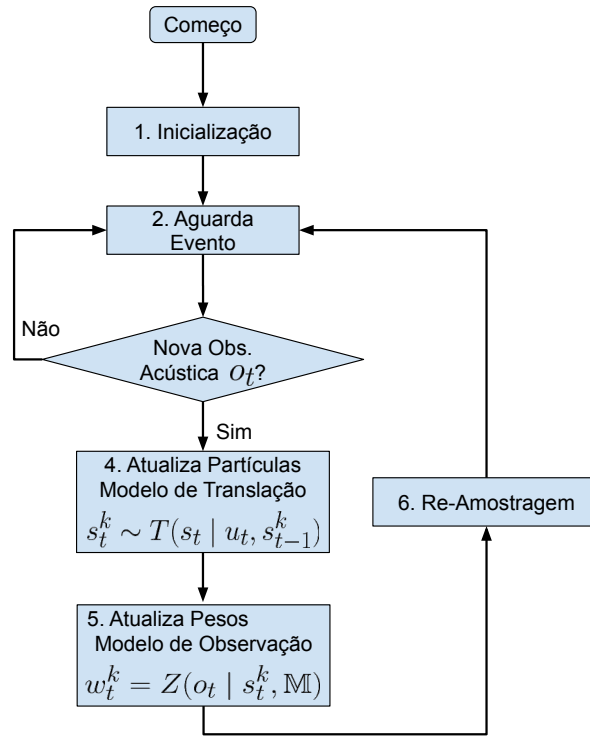


Figura 3.13: Diagrama de Execução do Filtro de Partículas.

Inicialmente as partículas são amostradas no mapa $\mathbb{M}$ utilizando uma distribuição normal, $G(\mu_x, \mu_y, \mu_\theta, \sigma_x, \sigma_y, \sigma_\theta)$, onde μ_x, μ_y, μ_θ são as médias da distribuição, centrada na última posição do veículo antes do mergulho, e $\sigma_x, \sigma_y, \sigma_\theta$ o desvio padrão. Nesta etapa também é calculado o mapa de mínimas distâncias $\mathbb{M}$, conforme descrito na Seção 3.3.1.

Segundo o diagrama, após a inicialização, o algoritmo entra no loop principal que aguarda uma nova leitura acústica o_t . Para cada nova leitura, ocorre a atualização das partículas utilizando o modelo de transição, seguido da atualização dos pesos das partículas, utilizando o modelo de observação. Finalmente, as partículas são re-amostradas, onde as partículas com menor peso são eliminadas e criadas próximas das partículas com maior peso.

3.4.1 Atualização do Estado das Partículas - Modelo de Transição

O modelo de transição utiliza a pose das partículas considerando seu estado anterior s_{t-1}^k , e o sinal de controle atual u_t . A nova pose é calculada utilizando a estimativa de deslocamento

acrescido de um ruído Gaussiano. Seja $u_t = (\sigma_x, \sigma_y, \sigma_\theta)^T$ o deslocamento do veículo entre os instantes $t-1$ e t , o modelo de transição atualiza as partículas segundo o algoritmo da Tabela 3.

1.	Modelo de Transição (u_t, x_{t-1})
2.	$\delta_{trans} = \sqrt{\delta_x^2 + \delta_y^2}$
3.	$\delta_{rot} = \delta_\theta$
4.	$\sigma_{trans} = \sqrt{\alpha_3 * \delta_{trans}^2 + \alpha_4 * \delta_{rot}^2}$
5.	$\sigma_{rot} = \sqrt{\alpha_1 * \delta_{rot}^2 + \alpha_2 * \delta_{trans}^2}$
6.	$\sigma_{lat} = \sqrt{\alpha_4 * \delta_{rot}^2 + \alpha_5 * \delta_{trans}^2}$
7.	Para cada partícula $s \in S$
8.	$\delta_{rumo} = \text{angdif}(\text{atan2}(\delta_y, \delta_x), \theta_{t-1}) + s_\theta$
9.	$\hat{\delta}_{trans} = \delta_{trans} + \text{NumGaus}(\sigma_{trans})$
10.	$\hat{\delta}_{rot} = \delta_{rot} + \text{NumGaus}(\sigma_{rot})$
11.	$\hat{\delta}_{lat} = 0 + \text{NumGaus}(\sigma_{lat})$
12.	$s_x = s_x + \hat{\delta}_{trans} \cdot \cos(\delta_{rumo}) + \hat{\delta}_{lat} \cdot \text{sen}(\delta_{rumo})$
13.	$s_y = s_y + \hat{\delta}_{trans} \cdot \text{sen}(\delta_{rumo}) + \hat{\delta}_{lat} \cdot \cos(\delta_{rumo})$
14.	$s_\theta = s_\theta + \hat{\delta}_{rot}$

Tabela 3: Modelo de Transição e o algoritmo de atualização das partículas.

Onde δ_{trans} é o deslocamento estimado para o veículo no plano desde a última atualização, δ_θ é o movimento angular estimado do veículo, σ_{trans} é o desvio padrão esperado no erro de deslocamento, σ_θ é o desvio padrão esperado do erro de estimativa de rotação, σ_{lat} é o desvio padrão esperado para estimativa de deslocamentos transversais do veículo. α_1, α_2 são parâmetros do modelo relacionado ao ruído de rotação, α_3 e α_4 são parâmetros do modelo relacionado ao ruído de translação, e α_5 especifica o ruído esperado em deslocamentos transversais.

As linhas 2 e 3 da Tabela 3 calculam os deslocamentos totais de translação e rotação. As linhas 4 a 6 calculam o desvio padrão esperado segundo os coeficientes de erro α_i . As linhas 7 a 14 são executadas para cada partícula. A linha 8 calcula o rumo, direção do deslocamento do veículo, as linhas 9-11 adicionam ruídos Gaussiano ao movimento estimado. A função $\text{NumGaus}(\sigma)$ retorna um valor aleatório que segue uma distribuição normal com média zero e desvio padrão σ . Finalmente as linhas 12-14 atualizam o estado da partícula considerando o deslocamento estimado acrescido de ruído Gaussiano. O modelo de transição adotado é tipicamente empregado em veículos omnidirecionais, maiores detalhes do modelo e suas derivações podem ser acessados em (THRUN; BURGARD; FOX, 2006).

3.4.2 Atualização dos pesos das Partículas - Modelo de Observação

A atualização dos pesos das partículas ocorre utilizando o Modelo de Observação descrito na Seção 3.3.1. O modelo considera a observação atual o_t , o mapa $\mathbb{M}$ e o estados das partículas atualizado na etapa anterior.

3.4.3 Re-amostragem das Partículas

A última etapa é a re-amostragem das partículas. Esta etapa utiliza o princípio de sobrevivência do mais apto –survival of the fittest– ([THRUN; BURGARD; FOX, 2006](#)), onde as partículas improváveis são substituídas pelas mais prováveis. Uma roleta é criada com o peso acumulado de cada partícula de forma que as partículas com maior peso ocupam uma área maior na roleta. As partículas selecionadas passam para o novo grupo de partículas, as não selecionadas são re-amostradas. A re-amostragem gera uma pose aleatória uniforme no mapa livre de obstáculos.

4 RESULTADOS

O método de localização subaquática proposto foi avaliado utilizando dados reais e simulados. A Seção 4.1 apresenta os conjuntos de dados utilizados para avaliação, a Seção 4.2 avalia o método de segmentação semântica de imagens aéreas proposto na Seção 3.1. A Seção 4.3 avalia a performance do algoritmo de Filtro de Partículas para localizar um veículo em uma marina utilizando dados acústicos simulados. Finalmente, a Seção 4.4 apresenta a evolução deste estudo comparando abordagens iniciais que foram desenvolvidas e testadas com o método de localização proposto. Os métodos são avaliados utilizando dados reais e simulados.

4.1 Conjunto de Dados

A apresentação dos conjuntos de dados está dividida em duas partes. Inicialmente são apresentados os dados de imagens acústicas disponíveis para pesquisa na Seção 4.1.1. Logo em seguida, na Seção 4.1.2 são apresentados os dados disponíveis para a segmentação semântica de imagens de satélite.

4.1.1 Imagens Acústicas

O conjunto de dados ARACATI 2017⁹, (SANTOS et al., 2019b) foi coletado por um veículo subaquático, *Seabotix Little Benthic Vehicle* (LBV) 300-5, equipado com um sonar (FLS) BlueView P900-130, configurado para cobrir 50 metros.

Uma prancha foi fixada ao veículo para permitir a coleta de dados de localização por um sistema diferencial de posicionamento global (DGPS), conforme apresentado na Figura 4.1. O veículo navega na marina do Yacht Clube de Rio Grande, Brasil, submerso próximo a superfície da água. Os dados de localização do DGPS são de alta precisão e podem ser utilizados como referência para localização.

O experimento durou 29 minutos e o veículo se desloca por 485 metros atingindo a velocidade máxima de 0.6 m/s. Também foram coletados dados de sua orientação por uma bússola interna, de sua localização por uma USBL e imagens da superfície por uma câmera GoPro instalada na prancha. O dataset foi publicado em (SANTOS et al., 2022) e está disponível para download¹⁰.

4.1.2 Imagens Aéreas

Atualmente imagens aéreas de praticamente toda superfície terrestre estão disponíveis na internet. As imagens utilizadas neste trabalho foram adquiridas através do aplicativo *Google Earth* e foram manualmente segmentadas utilizando o aplicativo GIMP nas três classes: estrutura estacionárias, estruturas móveis e água.

⁹O vídeo da coleta dos dados está disponível em <https://youtu.be/RaooXyLv3-s>

¹⁰O dataset ARACATI 2017 está disponível em <https://github.com/matheusbg8/aracati2017>

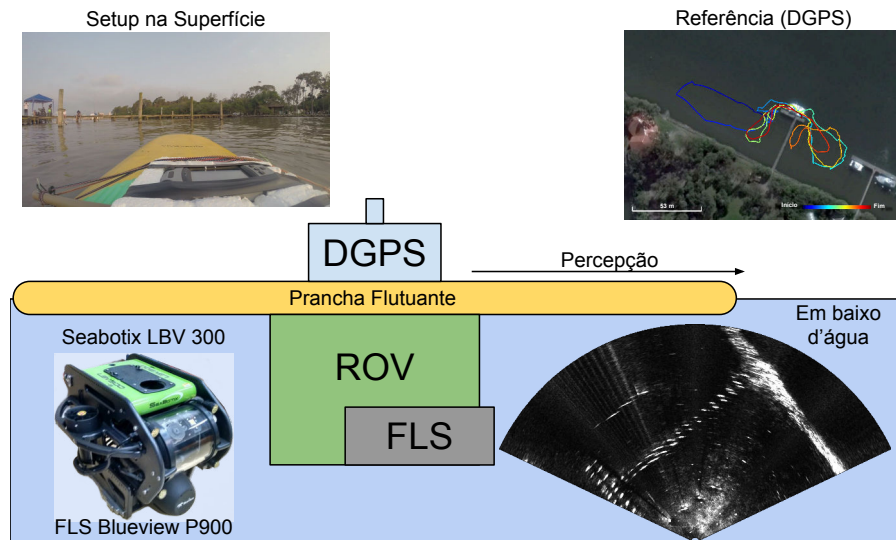


Figura 4.1: Dados do DGPS mostrando o caminho percorrido pelo veículo ao coletar as imagens acústicas do *dataset* ARACATI 2017.

Ao total foram segmentadas nove imagens, sendo cinco imagens do Yacht Clube de Rio Grande, Figura 4.2, duas imagens do *Kansai Yacht Clube* no Japão e duas no *Southport Yacht Clube* na Austrália, Figura 4.3.

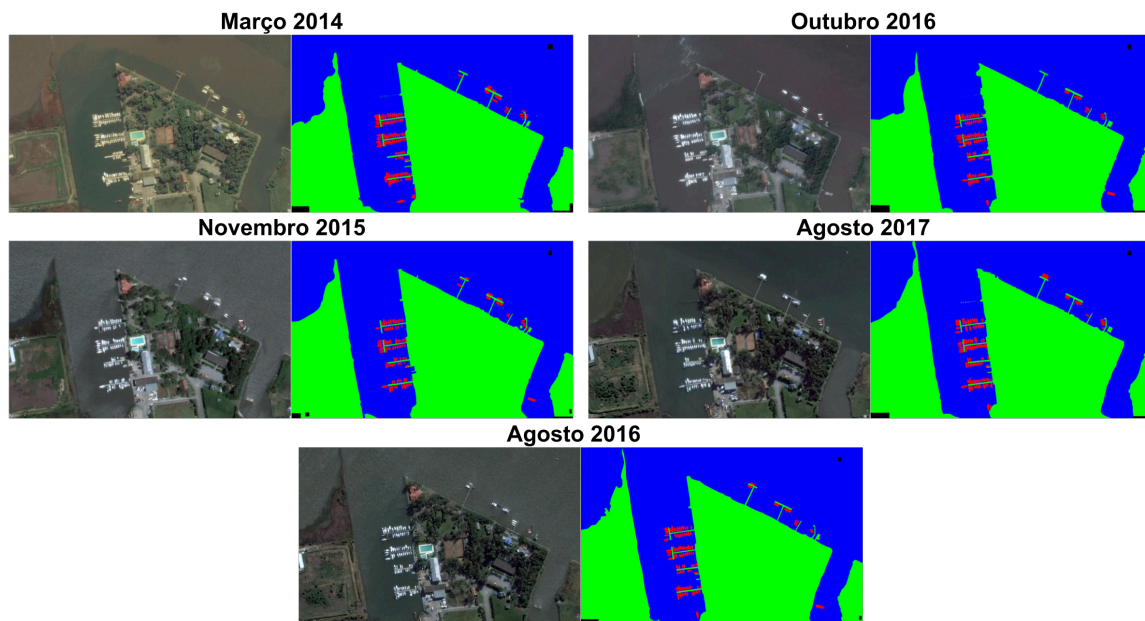


Figura 4.2: Imagens aéreas do Yacht Clube de Rio Grande com suas respectivas segmentações semânticas manualmente anotadas. A cor verde representa estruturas estacionárias, a cor vermelho representa estruturas móveis e a cor azul representa a água.

4.2 Segmentação Semântica de Imagens Aéreas

O primeiro estudo realizado neste trabalho teve o objetivo de avaliar a capacidade da RNC de segmentação, descrita na Seção 3.1, de segmentar imagens de satélite de uma mesma marina,

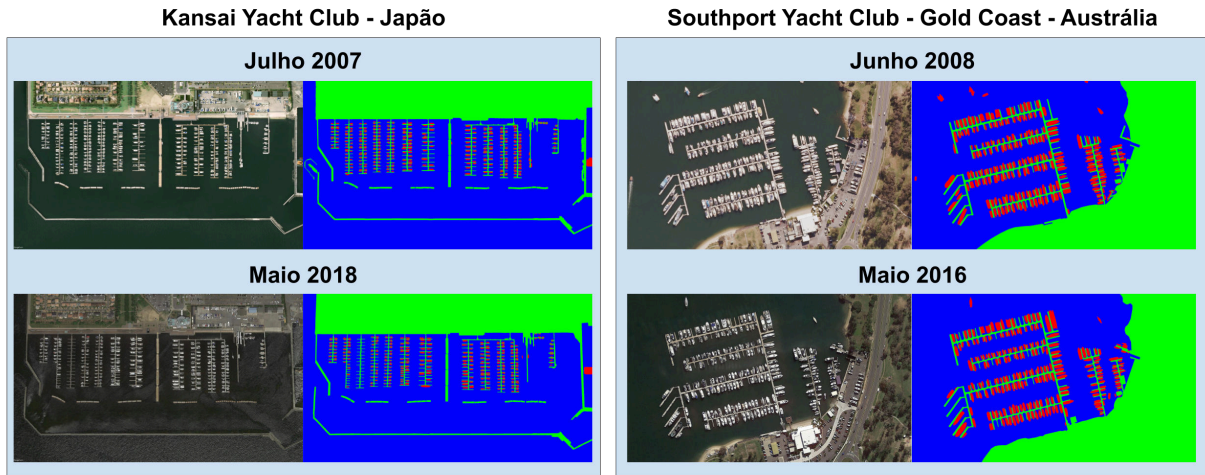


Figura 4.3: Imagens aéreas do Japão e da Austrália, com suas respectivas segmentações semânticas manualmente anotadas.

porém adquiridas em diferentes datas e condições climáticas. Seis imagens ópticas de satélite foram avaliadas, sendo cinco com *Ground Truth* de segmentação disponível.

Treinamento

Para o treinamento da rede foram utilizadas cinco imagens do Yacht Clube de Rio Grande coletadas nas seguintes datas: Março 2014, Novembro 2015, Agosto 2016, Outubro 2016 e Agosto 2018. Todas as imagens possuem tamanho 2944×1920 . A estratégia de *data augmentation* descrita na Seção 3.1 foi empregada para extrair 10.000 recortes, sendo 2.000 de cada uma das cinco imagens. Os recortes possuem rotação aleatória e foram extraídos de regiões aleatórias das imagens originais¹¹.

Validação

Na validação foram utilizados os cinco grupos de imagens de treinamento, porém recortadas de uma forma diferente, utilizando as operações de separação, validação e combinação descrita na Seção 3.1. Nesta operação os recortes são realizados linha a linha sem rotação, resultando em recortes com uma perspectiva diferente dos utilizados no treinamento. Também foi avaliada uma imagem da mesma marina coletada em Outubro de 2009, uma imagem do *Southport Yacht Club*, coletada na Austrália em Agosto de 2014, e uma imagem do *San Francisco Yacht Club*, coletada nos Estados Unidos em Outubro de 2009, ambas imagens nunca foram vista pela rede.

A partir da validação foram gerados resultados qualitativos e quantitativos.

¹¹Uma demonstração em vídeo do método de *data augmentation* está disponível em <https://bit.ly/2kyC3UZ>.

Resultados Qualitativos

Os resultados qualitativos com *ground truth* são exibidos na Figura 4.4 que inclui a imagem de entrada, a segmentação obtida pela rede e a segmentação verdadeira/esperada realizada manualmente. A Figura 4.5 mostra os resultados qualitativos de imagens sem *ground truth*.

Conforme demonstrado na Seção 2.2, as imagens apresentam desafios de segmentação. Mudanças de textura e cor tanto na terra quanto na água podem ser, em alguns casos, confusos para a segmentação. Além disso, objetos dinâmicos, como barcos, são semelhantes a estruturas estáticas na terra ou até mesmo com as ondas na água.

As figuras 4.4 e 4.5 mostram que a rede proposta consegue segmentar imagens aéreas com estas características. Se tratam de imagens da mesma região do treinamento, porém coletado em uma data diferente, ou imagens coletadas em regiões diferentes, como no caso das figuras 5(c) e 5(e). Além disso, é possível identificar que as imagens da Figura 4.4, apesar de serem da mesma marina, apresentam texturas do solo mais clara ou escura, e a tonalidade da água varia com relação às demais imagens.

Apesar dos resultados mostrarem que ainda existe espaço para melhorias, pois ocorrem problemas principalmente com as bordas das segmentações, o modelo proposto alcança bons resultados segmentando corretamente a maior parte da água e das estruturas estacionárias da imagem.

Resultados Quantitativos

Os resultados apresentados na Figura 4.4 possuem *ground truth* e foram avaliados quantitativamente. A precisão –*precision*–, a revocação –*recall*– e a pontuação F1 –*F1 score*– são as métricas normalmente adotadas para avaliação da segmentação neste contexto. Por exemplo, essas métricas foram utilizadas para avaliação das soluções submetidas ao desafio de segmentação ISPRS 2-D¹².

Uma matriz de confusão aumentada que considera todos os pixels foi criada para cada imagem, conforme mostra a Tabela 4. Nesta matriz, a direção das linhas representa a classe esperada de cada pixel (referência), a direção das colunas representa a classe atribuída pela rede de segmentação (previsão). Os valores são contabilizados em número de pixels e então normalizados linha a linha independentemente.

Os elementos da diagonal principal fornecem as previsões verdadeiras positivas (VP) de cada classe, ou seja, quando a classe esperada é atribuída corretamente pela rede de segmentação. A previsão de falso positivo (FP), quando a rede erra a segmentação do pixel, é obtida com a soma dos elementos da coluna sem considerar o elemento da diagonal principal. O falso negativo (FN) é derivado da soma dos elementos da linha, excluindo a diagonal principal. Finalmente, o verdadeiro negativo (VN) é a soma dos elementos restantes. O resultado geral é calculado adicionando todos os VP, FP, FN, e VN parciais.

¹²<http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>

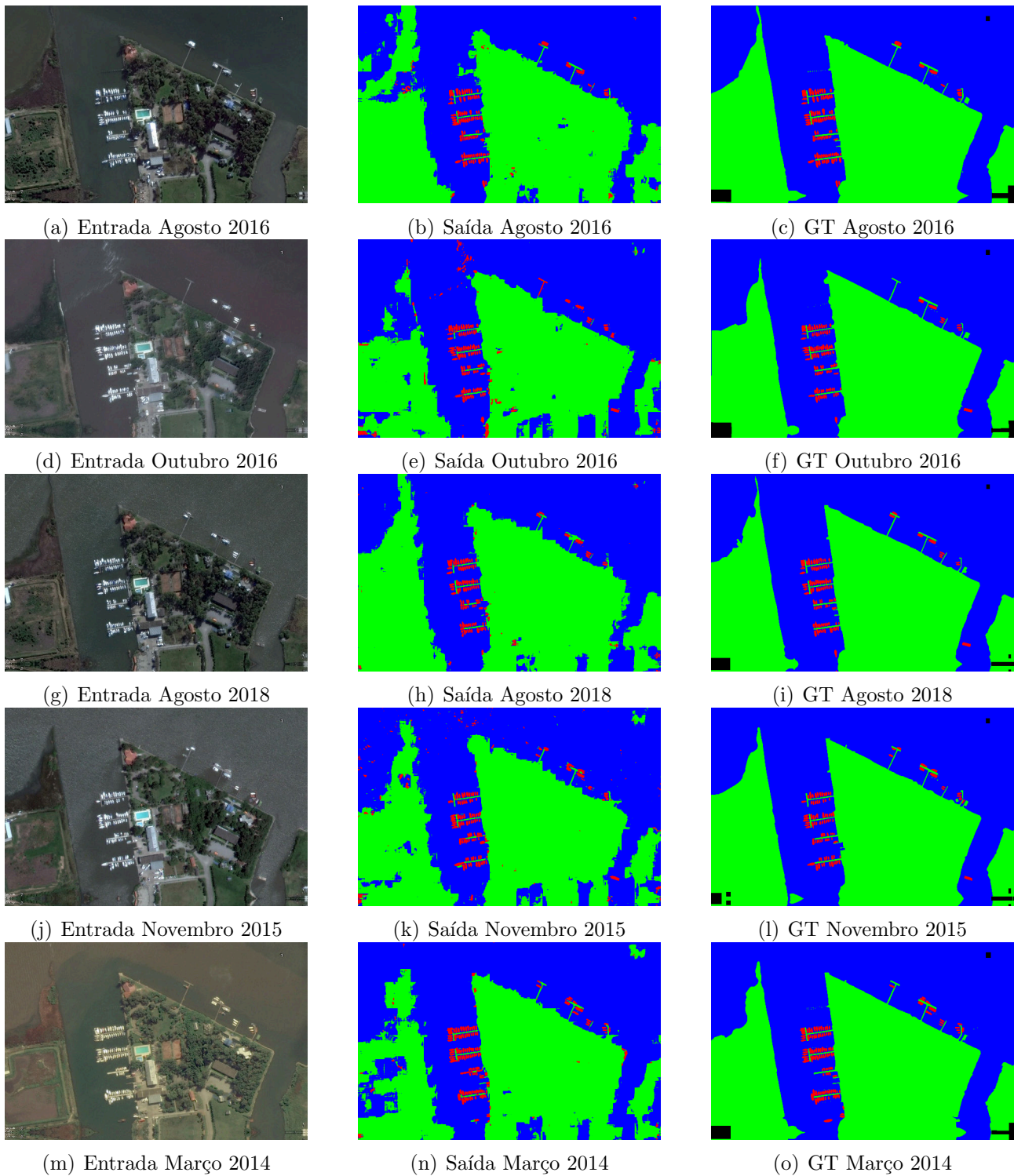


Figura 4.4: Resultados qualitativos da segmentação semântica de imagens de satélite do Yacht Clube de Rio Grande. As imagens da primeira coluna representam a imagem de satélite de entrada da rede de segmentação proposta, a segunda coluna apresenta a segmentação semântica realizada pela rede, e a terceira coluna apresenta a segmentação de referência, *ground truth*. A representação de cores das imagens segmentadas é verde para as estruturas estacionárias, vermelho para as estruturas móveis e azul para a água.

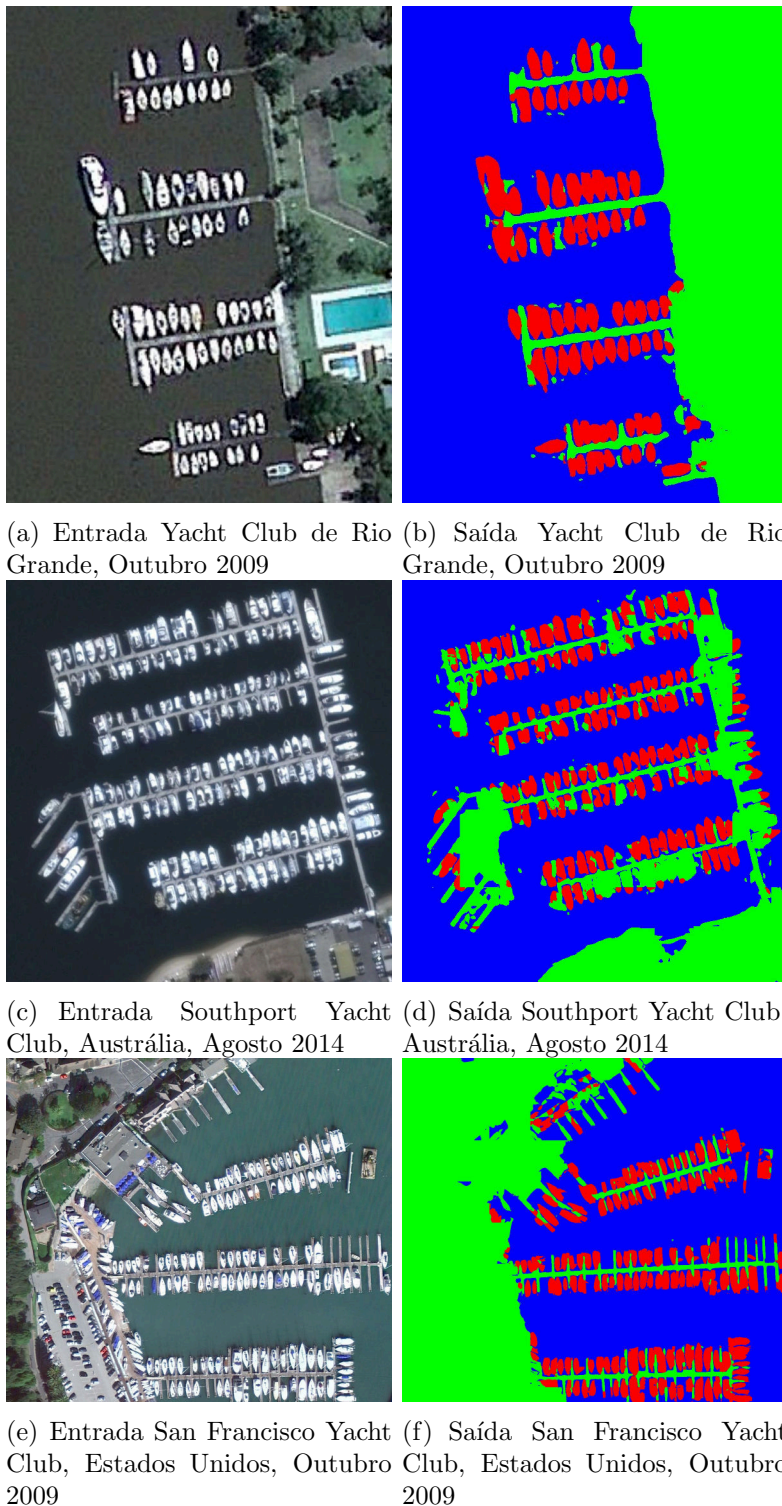


Figura 4.5: Resultados qualitativos da segmentação semântica de um imagens de satélite de imagens sem *ground truth*. A primeira coluna representa a imagem de entrada da rede de segmentação proposta, a segunda coluna representa a segmentação obtida. A primeira linha mostra o resultado obtido para uma imagem capturada no Yacht Club de Rio Grande, a segunda linha mostra o resultado obtido para uma imagem capturada na Austrália, e a última linha mostra o resultado obtido para uma imagem capturada nos Estados Unidos. As imagens segmentadas utilizam a mesma representação de cores: verde para as estruturas estacionárias, vermelho para as estruturas móveis e azul para a água.

Março 2014, Fig. 4(n)			
GT\Entrada	Água	Estacionário	Móvel
Água	0.931	0.067	0.002
Estacionário	0.117	0.875	0.007
Móvel	0.059	0.021	0.92
Novembro 2015, Fig. 4(k)			
GT\Saída	Água	Estacionário	Móvel
Água	0.908	0.085	0.007
Estacionário	0.082	0.912	0.006
Móvel	0.037	0.012	0.951
Agosto 2016, Fig. 4(h)			
GT\Saída	Água	Estacionário	Móvel
Água	0.906	0.091	0.003
Estacionário	0.025	0.971	0.004
Móvel	0.068	0.033	0.899
Novembro 2016, Fig. 4(e)			
GT\Saída	Água	Estacionário	Móvel
Água	0.939	0.054	0.007
Estacionário	0.126	0.863	0.011
Móvel	0.052	0	0.948
Dezembro 2017, Fig. 4(b)			
GT\Saída	Água	Estacionário	Móvel
Água	0.898	0.1	0.002
Estacionário	0.043	0.951	0.006
Móvel	0.105	0.003	0.892
Resultado Final			
GT\Saída	Água	Estacionário	Móvel
Água	0.917	0.079	0.004
Estacionário	0.078	0.915	0.007
Móvel	0.065	0.014	0.921

Tabela 4: Matriz de Confusão Aumentada

A precisão, revocação e pontuação F1 são derivados de VP, FP e FN pelas seguintes equações:

$$Precisao = \frac{VP}{VP + FP}, \quad (4.1)$$

$$Revocacao = \frac{VP}{VP + FN}, \quad (4.2)$$

$$F1 = 2 * \frac{Precisao * Revocacao}{Precisao + Revocacao}. \quad (4.3)$$

March 2014, Fig. 4(n)							
	VP	VN	FP	FN	Revocação	Precisão	F1
Água	0.49131	0.41753	0.05448	0.03669	0.93052	0.90018	0.9151
Estacionário	0.40086	0.5064	0.03569	0.05704	0.87543	0.91824	0.89632
Móvel	0.01297	0.98121	0.00468	0.00113	0.91969	0.73473	0.81687
November 2015, Fig. 4(k)							
	VP	VN	FP	FN	Revocação	Precisão	F1
Água	0.4765	0.43664	0.03843	0.04844	0.90773	0.92537	0.91646
Estacionário	0.42289	0.49149	0.04467	0.04095	0.91171	0.90447	0.90807
Móvel	0.01067	0.98193	0.00685	0.00055	0.95079	0.60909	0.74252
August 2016, Fig. 4(h)							
	VP	VN	FP	FN	Revocação	Precisão	F1
Água	0.4775	0.46075	0.01224	0.0495	0.90607	0.975	0.93927
Estacionário	0.44794	0.49021	0.04857	0.01328	0.9712	0.90218	0.93542
Móvel	0.01059	0.98507	0.00316	0.00118	0.89945	0.7704	0.82994
November 2016, Fig. 4(e)							
	VP	VN	FP	FN	Revocação	Precisão	F1
Água	0.49955	0.41034	0.05794	0.03217	0.93949	0.89607	0.91727
Estacionário	0.39422	0.51481	0.02865	0.06232	0.86349	0.93226	0.89656
Móvel	0.01113	0.97975	0.00852	0.00061	0.94833	0.56655	0.70933
December 2017, Fig. 4(b)							
	VP	VN	FP	FN	Revocação	Precisão	F1
Água	0.46197	0.46403	0.0218	0.05219	0.89849	0.95493	0.92585
Estacionário	0.4496	0.47592	0.05132	0.02316	0.95101	0.89755	0.92351
Móvel	0.01166	0.98328	0.00364	0.00141	0.89188	0.76186	0.82176
Overall Results							
	VP	VN	FP	FN	Revocação	Precisão	F1
Água	0.4814	0.4379	0.0370	0.0438	0.9166	0.9287	0.9226
Estacionário	0.4231	0.4958	0.0418	0.0393	0.9149	0.9101	0.9125
Móvel	0.0114	0.9822	0.0054	0.0010	0.9210	0.6799	0.7823

Tabela 5: Resultados Estatísticos do Método de Segmentação Semântico

Os resultados são mostrados na Tabela 5. Observa-se uma alta taxa de revocação, indicando que a rede de segmentação encontra cerca de 91% da ocorrência de cada classe em todo o conjunto de dados avaliado. Além disso, a segmentação de água e de estruturas estacionárias

é confiável devido à alta taxa de precisão (maior que 90%). Portanto, pelo menos 90% da segmentação dessas estruturas estavam corretas.

Embora a maioria das estruturas móveis do conjunto de dados tenham sido segmentadas pela rede, o que justifica a taxa de revocação de 92%, ocorreram muitas segmentações erradas, reduzindo a precisão da segmentação de estruturas móveis para 68%.

Com estes resultados, conclui-se que a rede é confiável para segmentação de estruturas estacionárias e a água. Porém, a segmentação de estruturas móveis ainda precisa de melhorias. Acredita-se que a complexidade para segmentação de estruturas móveis é maior devido a similaridade com estruturas estacionárias no solo, e as diferenças de escala. Barcos, são menores que terra ou água, e a rede de segmentação não possui informações de textura suficientes para segmentá-las corretamente.

No entanto, os resultados são aceitáveis com o objetivo de auxiliar na localização marítima. Os elementos mais importantes, água e estruturas estacionárias, para criar um mapa são confiáveis. As estruturas móveis são removidas do mapa, *ou seja*, transformados em água. Pequenos erros de segmentação podem ser resolvidos pelo Modelo de Observação descrito na Seção 3.3.1 ou pelo algoritmo de Filtro de Partículas conforme será demonstrado na próxima seção. Os resultados de segmentação semântica obtidos foram publicados em (SANTOS et al., 2019c).

4.3 Localização Subaquática com Dados Acústicos Simulados e Imagens Aéreas Reais

Um estudo do comportamento do algoritmo de localização adaptativo de Monte Carlo (AMCL) foi realizado no contexto da aplicação deste trabalho. Um dos objetivos foi identificar o comportamento e a convergência do AMCL em situações ambíguas encontradas em marinas, como a presença de barcos que são detectados nas imagens acústicas mas não são detectadas nas imagens aéreas.

Para avaliar a performance do AMCL, criou-se um cenário adaptado para o simulador Stage (VAUGHAN, 2008) no *middleware* ROS (QUIGLEY et al., 2009). Para garantir uma topologia coerente com a realidade da marina, o cenário foi construído manualmente com base em uma imagem de satélite do Yacht Clube de Rio Grande, conforme apresentado na Figura 4.6.

Nas regiões do píer, foram incluídos os postes de sustentação, detectados apenas no meio subaquático pelo sonar. A linha que representa a costa foi criada à frente da percebida pela imagem de satélite. Por fim, foram adicionados alguns barcos no ambiente de simulação. Desta forma, os barcos são detectados pelo sonar, mas não estão presentes no mapa gerado pela RNC de segmentação semântica da Figura 4.7. Assim, é possível avaliar o desempenho do AMCL em situações onde o mapa não reflete exatamente o cenário percebido pelo veículo. Isto ocorre devido ao movimento dos barcos e as diferenças entre percepções aéreas e subaquáticas.

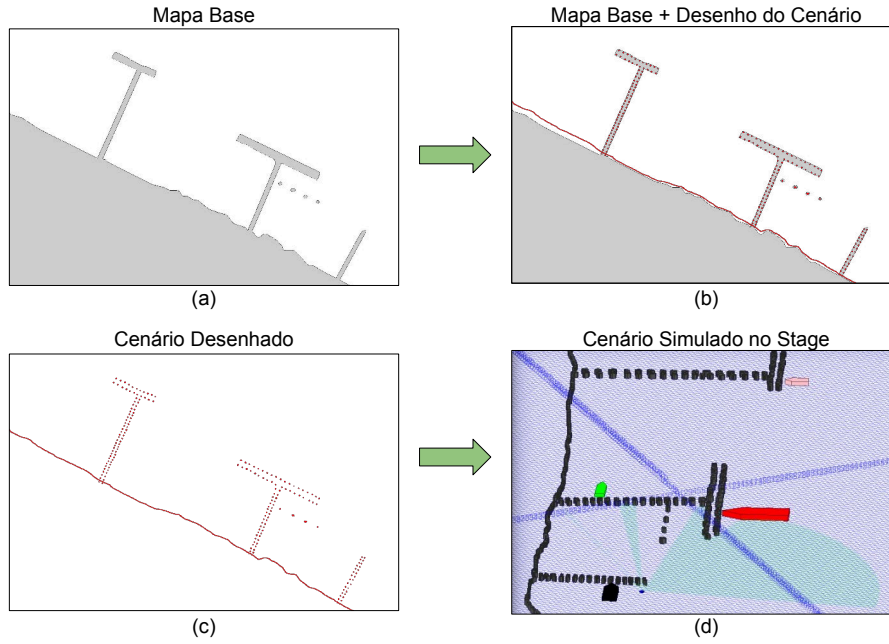


Figura 4.6: Cenário criado para simulação baseado em imagens de satélite do Yacht Clube de Rio Grande. Em (a) o mapa originado de uma imagem de satélite da marina utilizado como base para a criação do cenário. O desenho em vermelho apresentado nas figuras (b) e (c) foi criado manualmente e representa a posição dos obstáculos simulados pelo Stage. O cenário resultante é apresentado em (d). O veículo é representado pelo quadrado azul, o campo de visão do sonar FLS é representado pela cor verde. Alguns barcos foram adicionados no ambiente ao redor dos píeres.

O Stage simula a movimentação do veículo, e as leituras acústicas. O sonar foi configurado com uma abertura de 130 graus, 768 feixes e um alcance de 50 metros, as mesmas características do sonar real Teledyne BlueView P900-130 utilizado no dataset ARACATI 2017. A odometria é simulada com ruído gaussiano branco utilizando os desvios padrões $\sigma_x = 0.05$, $\sigma_y = 0.08$ e $\sigma_\theta = 0.03$.

Com relação à proposta original descrita no Capítulo 3, apenas a etapa B de conversão das imagens acústicas I_t^s em distâncias, representada pela observação o_t , não é executada. O Stage não é capaz de simular imagens acústicas e portanto as leituras acústicas são retornadas diretamente no formato de vetores de distâncias pelo simulador. Por esse motivo, a abordagem baseada em Aprendizagem Profunda, que depende das imagens acústicas, descrita na Seção 3.3.2 não foi avaliada, sendo portanto utilizado apenas o Modelo de Observação baseado em Campos de Probabilidades descrito na Seção 3.3.1 em conjunto com o Filtro de Partículas.

Resultados

Os resultados são obtidos utilizando uma imagem de satélite real do Yacht Clube de Rio Grande - Brasil, adquirida através da Digital Globe e Google Earth. A imagem coletada em 18 de Agosto de 2016 é convertida em um mapa através da segmentação semântica gerada pela RNC apresentada na Seção 3.1. Apenas as estruturas estacionárias são transformadas em obstáculos.

O resultado é mostrado na Figura 4.7.

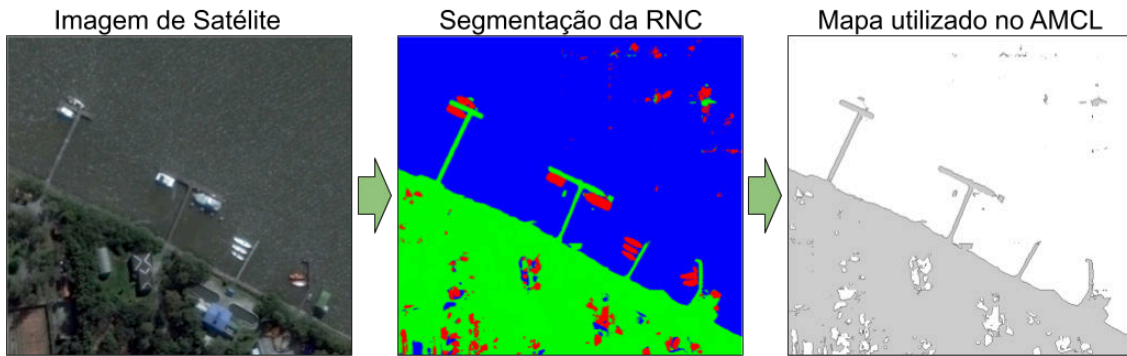


Figura 4.7: Conversão de uma imagem de satélite em mapa. Na imagem segmentada, a cor verde representa estruturas estacionárias, a cor vermelha representa estruturas móveis e a cor azul representa a água. No mapa, as áreas que podem ser navegadas estão na cor branca, os obstáculos estão na cor cinza.

A trajetória verdadeira percorrida pelo veículo, a trajetória estimada pelo AMCL, e a trajetória estimada pela odometria – *dead reckoning* –, assim como os seus respectivos erros relativos a trajetória verdadeira fornecida pelo simulador são mostrados na Figura 4.8.

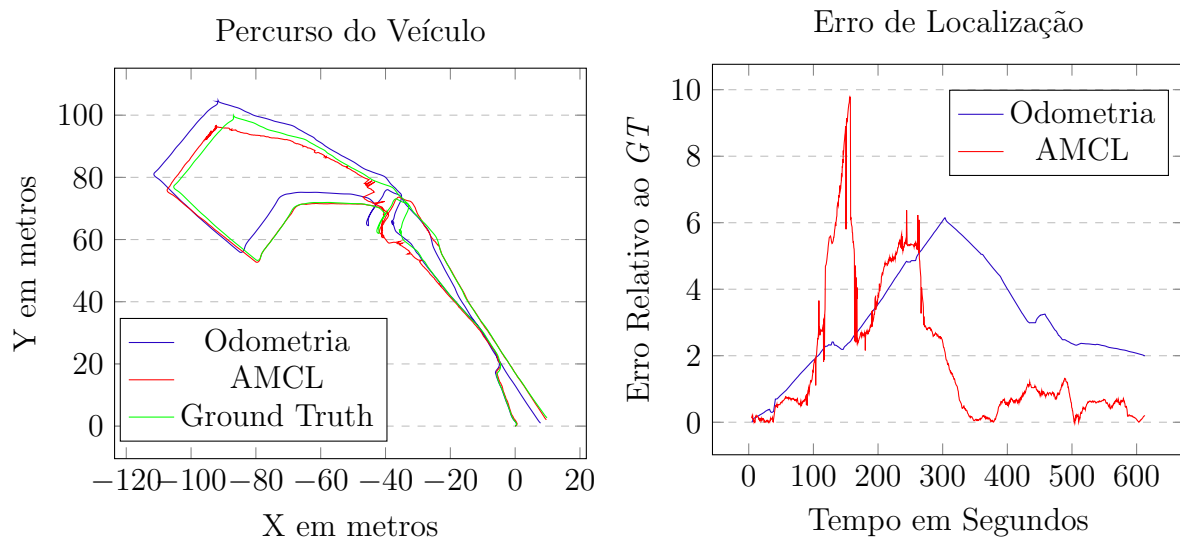


Figura 4.8: Os gráficos exibem as trajetórias estimadas e percorridas pelo veículo, e o erro de localização relativo à trajetória verdadeira.

Conforme apresentado na Figura 4.8, o método proposto utilizando AMCL obtém uma melhor estimativa de localização do que a odometria pura do veículo durante a maior parte da simulação¹³. O maior erro do AMCL ocorre entre 110 e 160 segundos devido o barco vermelho, que obriga o veículo a fazer um desvio, afastando-se das estruturas do cenário, conforme ilustrado nas figuras 4.9-(a) e 4.9-(b). Nesse intervalo nenhuma informação visual é obtida, fazendo com que as partículas se espalhem, aumentando a incerteza do método e o erro acumulado de odometria, conforme apresentado na Figura 4.9-(b). No entanto, quando o veículo

¹³Um vídeo da simulação está disponível em <https://youtu.be/CQrOskiSoc0>

retorna aos cais, o erro é significativamente reduzido, superando a odometria pura, conforme representado na Figura 4.9-(c). Esses resultados revelam que o AMCL é capaz de localizar um veículo subaquático mesmo com as diferenças de percepções do meio subaquático e aéreo. Apesar da existência das embarcações detectadas apenas nas leituras acústicas, o AMCL foi capaz de re-localizar o veículo após a aproximação do píer da marina. Os resultados obtidos nesse experimento foram publicados em (SANTOS et al., 2019b).

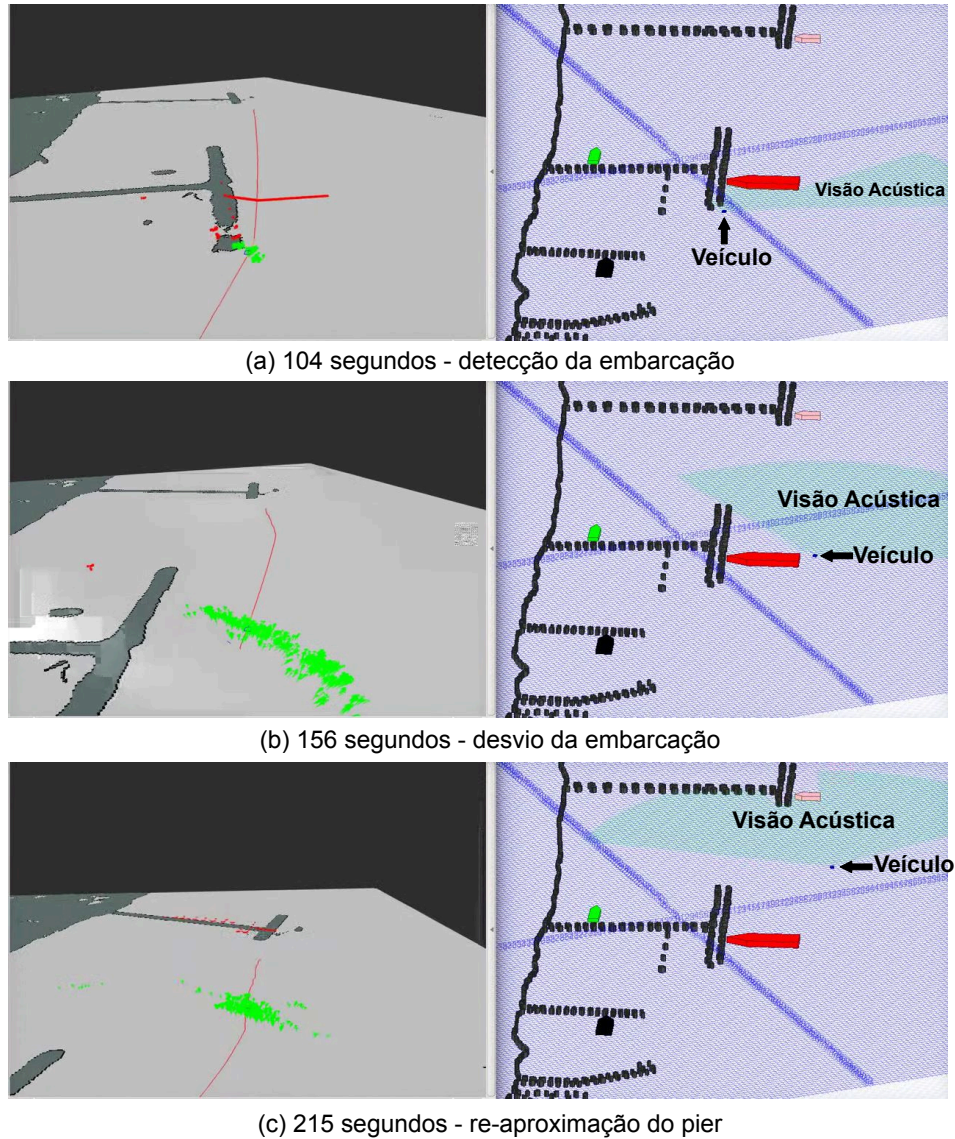


Figura 4.9: Destaques do experimento simulado. Em (a) aos 104 segundos, o veículo identifica uma embarcação que não está presente no mapa. Em (b), aos 156 segundos, o veículo desvia a embarcação em direção a uma área aberta. Nenhum objeto passa a ser identificado pelo sonar, as partículas se espalham e a incerteza de localização aumenta. Em (c), aos 215 segundos, o veículo passa a detectar o píer e uma embarcação, as partículas começam a convergir e o erro de localização é reduzido graças às estruturas identificadas nos dois domínios.

4.4 Localização Subaquática com Imagens Aéreas e Acústicas Reais

Durante o desenvolvimento desta pesquisa foram propostas duas abordagens para a estimar a similaridade entre as imagens multi-perspectiva e multi-domínio, resultando em diferentes Modelos de Observação para o algoritmo de localização de Monte Carlo, também conhecido como Filtro de Partículas (THRUN; BURGARD; FOX, 2006).

A abordagem baseada em aprendizagem profunda descrita na Seção 3.3.2 foi explorada em (SANTOS et al., 2021; SANTOS et al., 2022). Os resultados obtidos são comparados com o método que utiliza o Modelo de Observação baseado em Campos de Probabilidades descrito na Seção 3.3.1. Os algoritmos são avaliados utilizando os dados de imagens acústicas reais do conjunto de dados ARACATI 2017 descrito na Seção 4.1.1.

A Tabela 6 mostra os parâmetros adotados no experimento¹⁴. Os parâmetros da etapa de preparação de dados subaquáticos, relacionados ao sonar, como o número de *Beams* K , a posição inicial do *bin* i_{ini} , e o campo de visão θ_{max} e alcance ρ_{max} do sonar, foram definidos de acordo com as especificações do sonar Teledyne Blue View P900-130. Os demais parâmetros foram ajustados de acordo com o comportamento do veículo e as recomendações para o algoritmo do Filtro de Partículas, de forma a obter os melhores resultados.

Enquanto os métodos que utilizam Modelo de Observação baseado em aprendizagem profunda (SANTOS et al., 2021; SANTOS et al., 2022) espalha no máximo de 120 partículas, devido a limitações do hardware utilizado na avaliação, o método que utiliza Modelo de Observação baseado em campos de probabilidade descrito na Seção 3.3.1 utilizou 5.000 partículas, usando o mesmo hardware. Esta limitação ocorre principalmente devido à quantidade de memória da Unidade de Processamento Gráfica (GPU) utilizada pelos métodos baseados em redes neurais. A Figura 4.10 mostra a trajetória, em metros, obtida por cada método. A linha laranja representa a trajetória de referência –*ground truth*–, registrada por um DGPS, em todas as sub-figuras. As figuras 4.10-(a,b e c) mostram respectivamente, a trajetória estimada pela odometria a partir dos sinais de controle μ_t do veículo, a trajetória estimada pelo método proposto, utilizando o Modelo de Observação baseado em Aprendizagem Profunda descrito na Seção 3.3.2, e a trajetória estimada utilizando o Modelo de Observação baseado em Campos de Probabilidade descrita na Seção 3.3.1.

A Figura 4.11 mostra o erro de localização, em metros, da odometria e os três métodos avaliados em relação a localização do DGPS. A sombra representa a incerteza do método estimado de acordo com a convergência das partículas.

Os resultados mostraram que em geral o Filtro de Partículas melhorou a localização do veículo ao combinar imagens ópticas aéreas e acústicas submarinas. Conforme observado na Figura 4.11, na maioria do tempo do experimento o erro de localização é menor do que o da odometria pura do veículo. Note que a odometria pura plotada no gráfico é a mesma que alimenta o

¹⁴Um vídeo do experimento 8x mais rápido está disponível em <https://youtu.be/-nC0DkgszmE>

Parâmetro	Valor	Descrição
Preparação de Dados Aéreos		
d_{max}	20 metros	Distância Máxima ao Obstáculo
Preparação de Dados Subaquáticos (Sonar Teledyne BlueView P900-130)		
K	256	Num. de beams avaliados (quantidade)
l_{ini}	180	Posição inicial do bin (pixels)
w_{sz}	2	Tamanho da janela de suavização (pixels)
i_{min}	180	Threshold de intensidade para detectar objetos (valor do pixel)
θ_{max}	130	Campo de visão horizontal (graus)
p_{max}	50	Alcance máximo (metros)
Filtro de Partículas (AMCL) - Inicialização		
N	5000	Número de partículas (quantidade)
σ_{xx}	0,5	Covariância de posição no eixo x, inicialização das partículas (metros)
σ_{yy}	0,5	Covariância de posição no eixo y, inicialização das partículas (metros)
$\sigma_{\theta\theta}$	0,1	Covariância de orientação, inicialização das partículas (metros)
Re-amostragem - Modelo de Transição		
α_1	1,5	Erro de rotação esperado durante movimento de rotação do veículo
α_2	0,2	Erro de rotação esperado durante movimento de translação do veículo
α_3	2,6	Erro de translação esperado durante movimento de translação do veículo
α_4	0,2	Erro de translação esperado durante movimento de rotação do veículo
α_5	1,3	Erro de translação esperado durante movimento transversal do veículo
Correspondência - Modelo de Observação		
z_{hit}	0,5	Peso do erro de medição do sonar (valor normativo)
z_{rand}	0,5	Peso da medição aleatória do sonar (valor normativo)
σ_{hit}	0,2	Padrão Gaussiano. (metros)

Tabela 6: Valor dos parâmetros adotados no experimento

Modelo de Transição dos demais métodos apresentados.

Quando o veículo se move para uma área aberta, o algoritmo baseia-se apenas em odometria. Isso acontece porque as estruturas ficam fora do alcance do sonar e, portanto, não são detectadas pelo método. Assim, nenhuma correspondência de imagens aéreas e subaquáticas é encontrada, e o peso de todas as partículas é atualizado nas mesmas proporções. Neste momento, o Modelo de Transição passa a ter maior influência na estimativa da pose do veículo através da atualização do estado das partículas. As partículas começam a se espalhar ao longo do mapa $\mathbb{M}$ devido às incertezas modeladas pelo modelo de transição apresentado na Tabela 3.

No entanto, quando o veículo retorna à região estruturada, graças às correlações encontradas entre as imagens, o Modelo de Observação aumenta o peso das partículas próximas das estruturas e reduz o peso das partículas mais distantes. Na etapa de re-amostragem, apresentado na Figura 3.13, as partículas com menor peso tendem a ser re-amostradas próximo das partículas com maior peso, resultando na sua convergência e numa estimativa mais confiável da localização do veículo quando navegando próximo às estruturas parcialmente submersas. Este efeito é observado logo após os 500 segundos na Figura 4.11, principalmente pelo resultado da linha verde, que representa o método utilizando o Modelo de Observação probabilístico. Neste

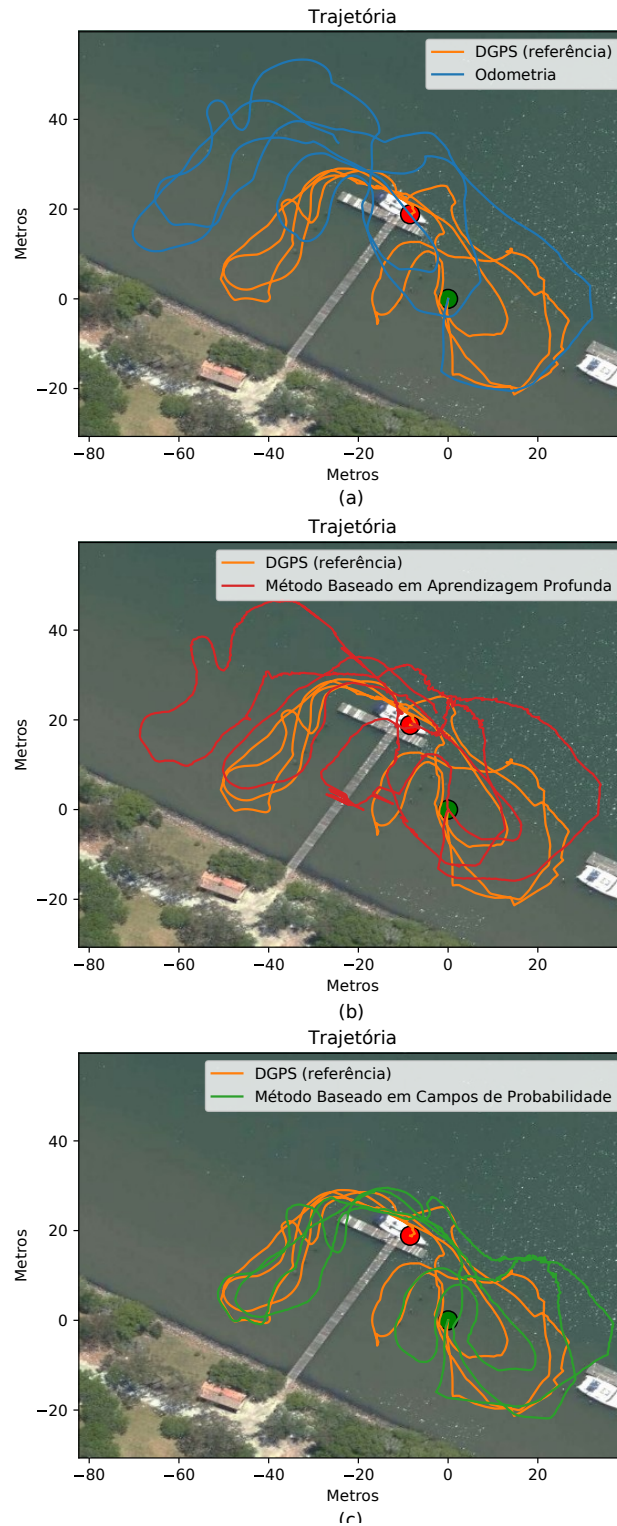


Figura 4.10: Trajeto do veículo, em metros, estimado por: (a) Odometria pura –*Dead Reckoning* –, (b) Filtro de Partículas utilizando o Modelo de Observação baseado em Aprendizado Profundo publicado em (SANTOS et al., 2022), (c) Filtro de Partículas utilizando o Modelo de Observação baseado em Campos de Probabilidade, descrito na Seção 3.3.2. A imagem de fundo mostra a imagem aérea do ambiente capturada por um drone, o norte aponta para cima e o leste aponta para a direita.

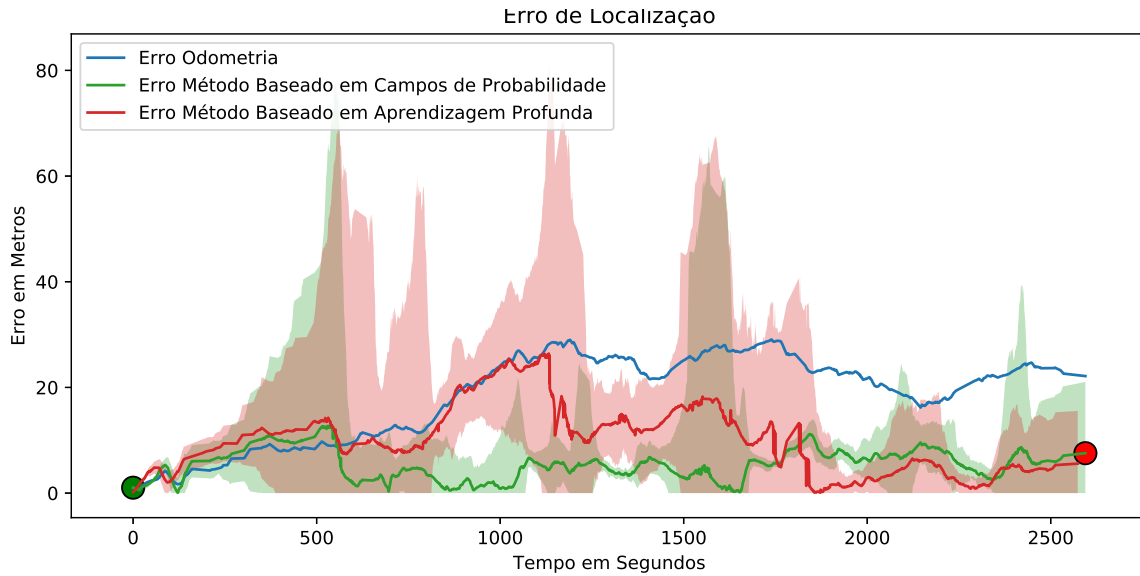


Figura 4.11: Erro de localização em metros, a sombra representa a incerteza de localização relacionado a convergência das partículas.

momento o veículo está chegando no píer oeste após ter saído do píer leste na Figura 4.10.

Os píeres são semelhantes e representam um desafio à nossa abordagem. Quando o veículo navega de um píer para o outro, grupos de partículas podem se espalhar entre o píer atual e o anterior. Devido ao problema de ambiguidade dos píeres, o Modelo de Observação pode acabar fortalecendo as partículas de ambos os grupos, gerando saltos na estimativa de localização entre os píeres e oscilações no gráfico de erro de localização.

Porém, assim que o veículo se movimenta e adquire novas imagens acústicas do píer, as partículas que estão no píer atual passam a ter um fator de importância maior, fazendo com que as partículas do píer anterior sejam re-amostradas para o píer correto. Os vales no gráfico de erro da Figura 4.11 mostram que o método proposto consegue melhorar a localização do veículo ao retornar a região estruturada, após períodos navegando em área aberta, sem detectar estruturas.

A Figura 4.11 também mostra que o método que adota o Modelo de Observação baseado em Campos de Probabilidade –*Likelihood Field Model*–, teve um desempenho melhor que o método que utilizam a abordagem com Aprendizagem Profundo no Modelo de Observação do Filtro de Partículas.

Performance Computacional

Os experimentos foram executados em um computador com Processador Ryzen 7 2700x e Unidade de Processamento Gráfico (GPU) NVIDIA RTX 3070. Os métodos baseados em rede neural (SANTOS et al., 2021; SANTOS et al., 2022) executam o Modelo de Observação em GPU. Utilizando a NVIDIA RTX 3070, sem otimizações, foi possível realizar até 120 inferências por segundo, tempo de inferência de 8,3ms, utilizando 6GB dos 8GB de memória da placa

gráfica, também conhecido como VRAM. Desta forma é possível atualizar os pesos de até 120 partículas por segundo. O uso da memória se deve pelo fato de que é necessário carregar o modelo e as imagens na memória da GPU.

O Modelo de Observação que utiliza Campos de Probabilidades descrito na Seção 3.3.1, é executado na Unidade de Processamento (CPU) em 0,003ms. O processo de transformação da imagem acústica em distâncias, descritos na Seção 3.2, sem otimizações, é executado em 2,85 ms. Assim, a abordagem probabilística proposta neste estudo é mais rápida do que a abordagem baseada em Aprendizado Profundo. Este fato permite elevar o número de partículas utilizadas pelo método, que consequentemente acaba avaliando uma amostragem maior do cenário, aumentando as chances de obter melhores resultados. Note que a transformação de imagem acústica em distâncias é executada apenas uma vez para cada imagem acústica, a inferência da rede neural para atualizar o peso das partículas precisa ser realizada uma vez para cada partícula, a cada nova imagem acústica capturada.

A quantidade de imagens acústicas capturadas por um FLS pode variar dependendo da distância de cobertura da imagem. No caso do sonar Blue View P900, a taxa de aquisição de imagens pode variar entre 4 e 15 imagens por segundo. No dataset ARACATI 2017 são adquiridas 4 imagens acústicas por segundo, porém devido a baixa velocidade de deslocamento do veículo, (máximo de 0.6 m/s), e o raio de cobertura de 50 metros das imagens acústica, pode-se considerar a avaliação de uma imagem acústica por segundo como sendo suficiente para navegação.

Considerações Finais

O método proposto utilizando o Modelo de Observação baseado em Campos de Probabilidade obteve melhores resultados de localização e desempenho computacional do que os que são baseados em Aprendizado Profundo. Algumas considerações com relação aos resultados obtidos pelos métodos baseado em Aprendizado Profundo são:

- modelos baseado em Aprendizado Profundo necessitam de uma grande quantidade de dados, geralmente rotulados, para o treinamento. No contexto deste trabalho, a aquisição de dados acústicos é desafiadora (SINGH; VALDENEGRO-TORO, 2021). Os experimentos exigem esforços logísticos consideráveis, e nem sempre é possível obter o rótulo para as imagens *–ground truth–*. Os locais para realização de experimentos controlados são limitados, e os simuladores capazes de simular imagens acústicas de um FLS também são limitados (RIBEIRO et al., 2018). Portanto a quantidade de dados utilizado no treinamento das abordagens baseadas em Aprendizado Profundo pode ter sido o maior fator limitante para obtenção de melhores resultados,
- a performance computacional limitou significativamente os modelos baseado em Aprendizado Profundo quando comparado ao número de partículas utilizado na abordagem probabilística,

- o campo de avaliação do cenário utilizando o Mapas de Mínimas Distâncias, descrito na Seção 3.3.1, é maior do que o das abordagens com aprendizagem profunda que utilizam diretamente recortes da imagem aérea de satélite. A Figura 4.12 mostra uma situação em que esta estratégia obtém vantagem quando comparado com a abordagem baseada em Aprendizagem Profunda. Por utilizar o mapa de mínimas distâncias, o espaço de avaliação do cenário é expandido além do campo de visão da imagem acústica.

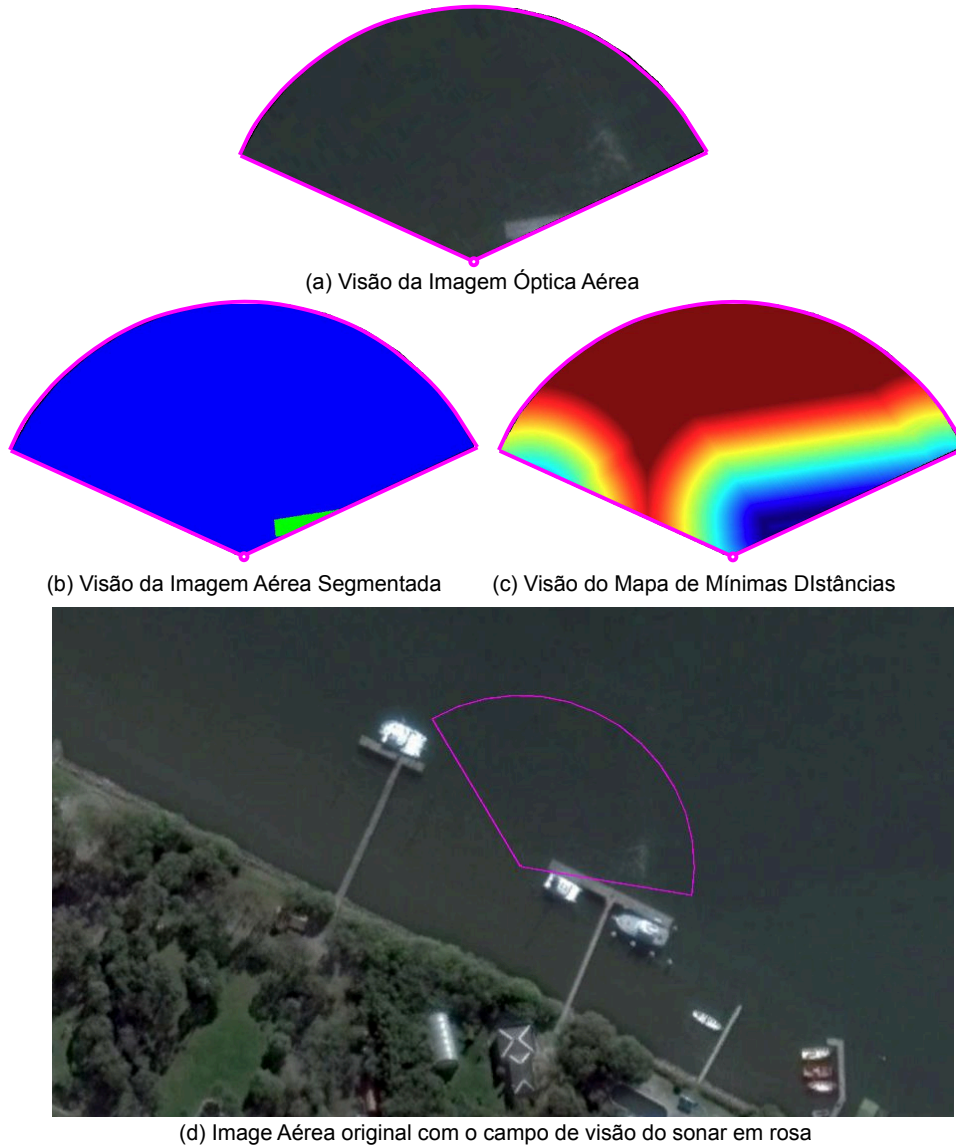


Figura 4.12: Comparação entre Mapa de Mínimas Distâncias e Imagem Aérea Segmentada recortadas no campo de visão de um sonar multi-feixe FLS. Diferente da imagem segmentada da Figura (b), a representação utilizando o Mapa de Mínimas Distâncias da Figura (c) consegue representar as estruturas que não aparecem no campo de visão do sonar, podendo trazer benefícios para o Modelo de Observação baseado em Aprendizagem Profunda.

Com relação ao Modelo de Observação baseado em Campos de Probabilidade, pode-se destacar as seguintes vantagens:

- possibilidade de avaliação de um espaço amostral maior, mais partículas no ambiente,

- menor exigência de hardware embarcado, o que pode significar uma redução nos custos do projeto de um veículo autônomo,
- obteve os melhores resultados nas avaliações deste trabalho.

Entretanto, existem algumas vantagens ao utilizar as abordagens baseada em aprendizagem profundo que justificam a continuação de sua pesquisa, são elas:

- flexibilidade de uso com diferentes sensores e imagens aéreas. A rede pode aprender recursos de diferentes sensores e imagens aéreas, com por exemplo: pode aprender a localizar utilizando cartas náuticas, mapas, imagens multi espectrais, imagens com mais classes semânticas, por exemplo, como a delimitação de regiões que geralmente são avistados barcos, ou até mesmo o treinamento de redes utilizando o Mapa de Mínimas Distância descrito na Seção 3.3.1 e abordado na Figura 4.12;
- as limitações devido à performance computacional podem ser compensada com as novas tecnologias de processamento embarcado, como a da NVIDIA Jetson Xavier, que possui 32 GB de VRAM, os modelos também podem ser otimizados, reduzindo o seu tamanho e tempo de inferência;
- a capacidade da rede neural identificar a correlações entre imagens multi-domínio pode ser generalizada para dados multi-domínio, expandido a sua utilização em outras aplicações, como por exemplo, identificar se duas sentenças escrita em línguas diferente possuem o mesmo significado, (GIACOMO et al., 2022).

5 CONCLUSÕES

Este trabalho apresentou um estudo de sistemas para o tratamento do problema de localização de veículos subaquáticos que navegam em regiões semi estruturadas, com estruturas parcialmente submersas, baseado nas imagens acústicas subaquáticas e ópticas aéreas do ambiente.

Foi apresentada uma revisão do estado da arte associada aos métodos de localização a partir de imagens multi-domínios e multi-perspectivas. O estudo conclui que as abordagens atuais não tratam do problema proposto neste trabalho, mas utilizam técnicas que podem ser adaptadas para as imagens nos domínios deste problema. Em busca desta adaptação, uma revisão dos métodos associados à segmentação semântica de imagens aéreas e de técnicas de melhoramento de imagens acústicas foi apresentada.

O estudo conclui que a maioria das abordagens atuais realizam localização terrestre através da correlação entre as imagens aéreas e terrestres. Observou-se um foco maior em regiões urbanas, principalmente por ser um ambiente estruturado, que exige métodos de localização que tratam o problema do bloqueio do sinal de GPS provocado pelas construções da cidade. As técnicas para estimar a correlação entre as imagens variam desde a utilização de descritores de imagens, técnicas de aprendizagem de máquinas tradicionais como Vetores de Suporte de Máquina SVM, até técnicas mais sofisticadas e recentes como aprendizagem profunda, geralmente empregando variações da rede neural *Siamese* (CHOPRA; HADSELL; LECUN, 2005). Já com relação aos métodos de segmentação semântica, o estudo conclui que a maioria classifica áreas urbanas ou agrícolas com poucos métodos aplicados às regiões costeiras e portuárias, foco deste trabalho. As técnicas mais recentes utilizam aprendizagem profunda, geralmente empregando variações da rede *U-Net* (RONNEBERGER; FISCHER; BROX, 2015), como em (LI et al., 2018; SANTOS et al., 2019c). Com relação às técnicas de melhoramento de imagens acústicas, o estudo revelou métodos para correção de distorções e de iluminação de imagens acústicas, assim como técnicas para redução do ruído baseado em sequência de imagens, essas técnicas geralmente são empregadas em métodos de construção de mosaico a partir das imagens acústicas.

A partir do levantamento bibliográfico realizado foi possível propor um *pipeline* inédito para o tratamento do problema de localização subaquático através de imagens ópticas aéreas e acústicas subaquáticas de um FLS. O *pipeline* envolveu quatro etapas principais.

A primeira etapa é responsável pelo entendimento das imagens ópticas aéreas através da segmentação semântica das imagens nas três classes: estruturas estacionárias, objetos móveis, e água. As estruturas estacionárias são utilizadas como referência para a correspondência com as imagens acústicas e posterior estimativa da localização do veículo subaquático. A segmentação semântica é realizada por uma rede *U-Net* (RONNEBERGER; FISCHER; BROX, 2015) combinado com uma técnica de recortes das imagens aéreas que permite a avaliação de imagens de qualquer tamanho.

Em seguida, o *pipeline* apresenta o tratamento das imagens acústicas. Para isso foi proposto um método que estima as distâncias das estruturas à frente do veículo a partir das imagens acústicas através da análise de intensidade dos seus *beams*.

A terceira etapa trata da estimativa da similaridade entre as imagens acústicas subaquáticas e as imagens ópticas aéreas, e foi construída a partir de duas alternativas metodológicas:

1. a estimativa de similaridade por um modelo probabilístico baseado na menor distância entre as estruturas identificadas nos dois domínios de imagens.
2. a estimativa de similaridade por uma Rede Neural Convolucional (RNC) chamada *QuadrupletNet*, que utiliza uma rede *Encoder* independente, baseada na arquitetura U-Net, para aprender as características específicas de cada domínio de imagem. As redes *Encoders* são treinadas utilizando uma abordagem de aprendizagem cooperativa baseado em *Triples*.

Finalmente a quarta etapa utiliza o algoritmo de Monte Carlo para estimar a localização subaquática do veículo. O algoritmo espalha hipóteses da localização do veículo –partículas–, no mapa do ambiente construído a partir das imagens aéreas. Uma das metodologias da terceira etapa é adotada como Modelo de Observação para a atualização do peso das partículas com base na similaridade entre as imagens.

O *pipeline* foi testado e analisado em experimentos simulados e reais propostos nesta tese. Para a validação em simulação foi desenvolvido um ambiente virtual baseado em uma marina real, com a presença de estruturas estacionárias parcialmente submersas como píeres, postes e a orla da marina. Também foram incluídos objetos móveis como barcos que ficaram ancorados nos píeres durante a simulação.

Por outro lado, a validação em ambiente real foi realizada utilizando um ROV com um sonar multifeixe BlueView Teledyne P900-130 e um mecanismo que acompanha o veículo na superfície, permitindo a leitura de dados de DGPS, utilizados como referência de localização. O veículo navega em uma marina, deslocando-se entre os píeres do local e passando por regiões com e sem estruturas.

Os resultados apresentados com relação a segmentação semântica de imagens aéreas mostram que foi possível obter uma revocação de **91%** para segmentação da água, estruturas estacionárias e móveis, uma precisão de **91%** para segmentação de água e estruturas estacionárias e uma precisão de apenas **68%** para objetos móveis. Apesar das dificuldades com a segmentação de objetos móveis, o método de segmentação apresentado se demonstrou aceitável principalmente pelo fato do filtro de partículas conseguir lidar com imperfeições no mapa.

Como trabalhos futuros para a segmentação semântica pode-se explorar as seguintes alternativas:

- Uso de técnicas recentes de *fine tuning*.

- Aumentar os dados de treinamento através da segmentação manual de mais imagens.
- Alterar o balanceamento do conjunto de treinamento, inserindo mais casos de estruturas móveis e menos casos de água e estruturas estacionárias.
- Remover do conjunto de dados de treinamento os barcos na terra, considerados estruturas estacionárias, e assim evitando conflito com os barcos na água, considerados estruturas móveis.
- Avaliar o comportamento da rede ao alterar a escala das imagens, atualmente de 0.2 metros por pixel, garantido por uma operação de *resize* com interpolação linear, conforme descrito na Seção 3.1.
- Realizar comparações com outras redes de segmentação como SegNet ([BADRINARAYAN; KENDALL; CIPOLLA, 2017](#)) e DeepLab ([CHEN et al., 2018b](#)).

Com relação a avaliação do algoritmo de Monte Carlo (MCL), obteve-se resultados satisfatórios com um erro de localização inferior ao da odometria pura –*dead reckoning*– na maior parte do tempo tanto em simulação quanto no experimento real. O resultado mostra que o MCL é capaz de operar em situações onde o mapa não reflete exatamente as percepções visuais do veículo. Uma vez que o mapa é gerado por imagens aéreas capturadas em épocas diferentes, as diferenças entre o mapa gerado e as percepções visuais do veículo serão corriqueiras. Principalmente por causa das estruturas móveis que, apesar de eliminadas no mapa, continuam existindo nas percepções acústicas. Também foi possível observar que o método consegue manter o rastreamento e localizar o veículo mesmo quando navegando por curtos intervalos em regiões não estruturadas, ou em locais com estruturas semelhantes como os píeres da marina.

Como perspectivas futuras relacionadas ao método de localização proposto, vislumbra-se:

- Embarcar o sistema desenvolvido em um veículo subaquático autônomo AUV e verificar o comportamento do sistema proposto junto com o sistema de navegação e planejamento de rotas do veículo.
- Verificar o comportamento do sistema em diferentes cenários utilizando diferentes modelos de FLS,
- Explorar novas possibilidades de utilização do método, como por exemplo, a construção de mapas híbridos que une informações aéreas e subaquáticas e um único mapa. Essas informações podem ser espaciais, na forma da reconstrução completa de estruturas parcialmente submersas, ou na forma de informações semânticas, onde são marcados no mesmo mapa objetos que podem estar submersos ou fora da água. Uma abordagem utilizando mapa híbrido aplicado a regiões urbanas utilizando imagens aéreas e terrestres foi explorada em ([BALASKA et al., 2022](#)).

Referências

- ANAND, T. et al. Agrisegnet: Deep aerial semantic segmentation framework for iot-assisted precision agriculture. *IEEE Sensors Journal*, IEEE, v. 21, n. 16, p. 17581–17590, 2021.
- AYKIN, M.; NEGAHDARIPOUR, S. On feature extraction and region matching for forward scan sonar imaging. In: IEEE. *Oceans, 2012*. [S.l.], 2012. p. 1–9.
- BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2017.
- BADRINARAYANAN, V.; KENDALL, A.; SEGNET, R. C. A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, v. 5, 2015.
- BALASKA, V. et al. Generating graph-inspired descriptors by merging ground-level and satellite data for robot localization. *Cybernetics and Systems*, Taylor & Francis, p. 1–19, 2022.
- BALLARD, D. H. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, Elsevier, v. 13, n. 2, p. 111–122, 1981.
- BANSAL, M.; DANIILIDIS, K.; SAWHNEY, H. Ultrawide baseline facade matching for geo-localization. In: _____. *Large-Scale Visual Geo-Localization*. Cham: Springer International Publishing, 2016. p. 77–98. ISBN 978-3-319-25781-5. Disponível em: https://doi.org/10.1007/978-3-319-25781-5_5.
- BANSAL, M. et al. Geo-localization of street views with aerial image databases. In: *Proceedings of the 19th ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2011. (MM '11), p. 1125–1128. ISBN 978-1-4503-0616-4. Disponível em: <http://doi.acm.org/10.1145/2072298.2071954>.
- BAY, H. et al. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, v. 110, n. 3, p. 346 – 359, 2008. ISSN 1077-3142. Similarity Matching in Computer Vision and Multimedia. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1077314207001555>.
- BAY, H.; TUYTELAARS, T.; GOOL, L. V. SURF: Speeded up robust features. In: *In ECCV*. [S.l.: s.n.], 2006. p. 404–417.
- BELONGIE, S.; MALIK, J.; PUZICHA, J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 4, p. 509–522, 2002.
- BESİBINAR, B.; ALATAN, A. A. Inshore ship detection in high-resolution satellite images: approximation of harbors using sea-land segmentation. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Image and Signal Processing for Remote Sensing XXI*. [S.l.], 2015. v. 9643, p. 96432D.
- BIBER, P.; STRASSER, W. The normal distributions transform: A new approach to laser scan matching. In: IEEE. *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*. [S.l.], 2003. v. 3, p. 2743–2748.
- BURKARD, R.; DELL'AMICO, M.; MARTELLO, S. *Assignment problems: revised reprint*. [S.l.]: SIAM, 2012.

- CANDIAGO, S. et al. Evaluating multispectral images and vegetation indices for precision farming applications from uav images. *Remote sensing*, MDPI, v. 7, n. 4, p. 4026–4047, 2015.
- CASTALDO, F. et al. Semantic cross-view matching. In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. [S.l.: s.n.], 2015.
- CASTRO, E. D.; MORANDI, C. Registration of translated and rotated images using finite fourier transforms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE, n. 5, p. 700–703, 1987.
- CERQUEIRA, R. et al. Custom shader and 3d rendering for computationally efficient sonar simulation. In: *29th Conference on Graphics, Patterns and Images-SIBGRAPI*. [S.l.: s.n.], 2016.
- CHAME, H. F.; SANTOS, M. M. dos; BOTELHO, S. S. da C. Neural network for black-box fusion of underwater robot localization under unmodeled noise. *Robotics and Autonomous Systems*, v. 110, p. 57 – 72, 2018. ISSN 0921-8890. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0921889018302926>.
- CHEN, K. et al. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 15, n. 2, p. 173–177, 2018.
- CHEN, L.-C. et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *ECCV*. [S.l.: s.n.], 2018.
- CHOPRA, S.; HADSELL, R.; LECUN, Y. Learning a similarity metric discriminatively, with application to face verification. In: IEEE. *null*. [S.l.], 2005. p. 539–546.
- CODEVILLA, F. et al. Achieving turbidity robustness on underwater images local feature detection. In: *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2015. p. 154.1–154.13. Disponível em: <https://dx.doi.org/10.5244/C.29.154>.
- CORDTS, M. et al. The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016.
- CORKE, P. et al. Experiments with underwater robot localization and tracking. In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2007. p. 4556–4561.
- DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. [S.l.: s.n.], 2005. v. 1, p. 886–893 vol. 1.
- DEMIR, I. et al. Deepglobe 2018: A challenge to parse the earth through satellite images. In: IEEE. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. [S.l.], 2018. p. 172–17209.
- DONG, S.; WANG, P.; ABBAS, K. A survey on deep learning and its applications. *Computer Science Review*, Elsevier, v. 40, p. 100379, 2021.
- DONG, Y.; ZHANG, J.; XU, F. Auto localization for coastal satellite imagery based on curve matching. In: IEEE. *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*. [S.l.], 2014. p. 1–6.

- EL-FAKDI, A.; CARRERAS, M.; RIDAO, P. Towards direct policy search reinforcement learning for robot control. In: IEEE. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. [S.l.], 2006. p. 3178–3183.
- EZEQUIEL, C. A. F. et al. Uav aerial imaging applications for post-disaster assessment, environmental management and infrastructure development. In: IEEE. *2014 International conference on unmanned aircraft systems (ICUAS)*. [S.l.], 2014. p. 274–283.
- GANCHENKO, V.; DOUDKIN, A. Image semantic segmentation based on convolutional neural networks for monitoring agricultural vegetation. In: ABLAMEYKO, S. V.; KRASNOPROSHIN, V. V.; LUKASHEVICH, M. M. (Ed.). *Pattern Recognition and Information Processing*. Cham: Springer International Publishing, 2019. p. 52–63. ISBN 978-3-030-35430-5.
- GAO, X. et al. Ground and aerial meta-data integration for localization and reconstruction: A review. *Pattern Recognition Letters*, 2018. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865518303544>.
- GIACOMO, G. et al. Sonar-to-satellite translation using deep learning. In: IEEE. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.], 2018. p. 454–459.
- GIACOMO, G. G. D. et al. Cooperative training of triplet networks for cross-domain matching. In: IEEE. *2020 Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics (SBR) and 2020 Workshop on Robotics in Education (WRE)*. [S.l.], 2020. p. 1–6.
- GIACOMO, G. G. D. et al. Matching cross-domain data with cooperative training of triplet networks: A case study on underwater robotics. *Journal of Intelligent & Robotic Systems*, Springer, v. 104, n. 3, p. 1–13, 2022.
- GONÇALVES, G. A. *Rio Grande Yacht Club: Memorial Descritivo da Área em Solicitação de Disponibilidade*. [S.l.], 2020. 1-2 p.
- GOONESEKERE, R.; GUO, Y. Unmanned underwater drone design for ocean exploration. In: AMERICAN SOCIETY OF MECHANICAL ENGINEERS. *ASME International Mechanical Engineering Congress and Exposition*. [S.l.], 2018. v. 52187, p. V013T05A034.
- HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In: CITESEER. *Alvey vision conference*. [S.l.], 1988. v. 15, p. 50.
- HE, Y. et al. Ship detection without sea-land segmentation for large-scale high-resolution optical satellite images. In: IEEE. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.], 2018. p. 717–720.
- HOFFER, E.; AILON, N. Deep metric learning using triplet network. In: SPRINGER. *International Workshop on Similarity-Based Pattern Recognition*. [S.l.], 2015. p. 84–92.
- HU, S. et al. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2018.
- HURTOS, N. et al. Evaluation of registration methods on two-dimensional forward-looking sonar imagery. In: *OCEANS - Bergen, 2013 MTS/IEEE*. [S.l.: s.n.], 2013. p. 1–8.
- HURTÓS, N. V. *Forward-Looking Sonar Mosaicing for Underwater Environments*. Tese (Doutorado) — Universitat de Girona, 2014.

- HURTÓ, N. et al. Fourier-based registrations for two-dimensional forward-looking sonar image mosaicing. In: *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. [S.l.: s.n.], 2012. p. 5298–5305. ISSN 2153-0858.
- IRWIN, F. et al. An isotropic 3x3 image gradient operator. *Presentation at Stanford AI Project*, v. 2014, n. 02, 1968.
- JALAL, F.; NASIR, F. Underwater navigation, localization and path planning for autonomous vehicles: A review. In: IEEE. *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. [S.l.], 2021. p. 817–828.
- JEZEK, K. Glaciologic properties of the antarctic ice sheet from spaceborne synthetic aperture radar observations. *Annals of Glaciology*, v. 29, p. 286–290, 1999.
- JOHANSSON, H. et al. Imaging sonar-aided navigation for autonomous underwater harbor surveillance. In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. [S.l.: s.n.], 2010. p. 4396–4403. ISSN 2153-0858.
- KAMPFMEYER, M.; SALBERG, A.-B.; JENSSEN, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. [S.l.: s.n.], 2016. p. 1–9.
- KARKUS, P.; HSU, D.; LEE, W. S. Particle filter networks with application to visual localization. In: BILLARD, A. et al. (Ed.). *Proceedings of The 2nd Conference on Robot Learning*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 87), p. 169–178. Disponível em: <https://proceedings.mlr.press/v87/karkus18a.html>.
- KEMKER, R.; SALVAGGIO, C.; KANAN, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 145, p. 60–77, 2018.
- KENDALL, A.; GRIMES, M.; CIPOLLA, R. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 2938–2946.
- KIM, K.; NERETTI, N.; INTRATOR, N. Mosaicing of acoustic camera images. *Radar, Sonar and Navigation, IEE Proceedings -*, v. 152, n. 4, p. 263–270, Aug 2005. ISSN 1350-2395.
- KIM, K.; NERETTI, N.; INTRATOR, N. Video enhancement for underwater exploration using forward looking sonar. In: SPRINGER. *International Conference on Advanced Concepts for Intelligent Vision Systems*. [S.l.], 2006. p. 554–563.
- KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105.
- KUHN, H. W. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, v. 52, n. 1, p. 7–21, 2005. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.20053>.
- KULELI, T. et al. Automatic detection of shoreline change on coastal ramsar wetlands of turkey. *Ocean Engineering*, Elsevier, v. 38, n. 10, p. 1141–1149, 2011.

- LEE, J.-S. Speckle Suppression And Analysis For Synthetic Aperture Radar Images. *Optical Engineering*, SPIE, v. 25, n. 5, p. 636 – 643, 1986. Disponível em: <https://doi.org/10.1117/12.7973877>.
- LEUNG, K. Y. K.; CLARK, C. M.; HUISSOON, J. P. Localization in urban environments by matching ground level video images with an aerial image. In: *2008 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2008. p. 551–556. ISSN 1050-4729.
- LI, A.; MORARIU, V. I.; DAVIS, L. S. Planar structure matching under projective uncertainty for geolocation. In: FLEET, D. et al. (Ed.). *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014. p. 265–280. ISBN 978-3-319-10584-0.
- LI, J. et al. Utilizing high-dimensional features for real-time robotic applications: Reducing the curse of dimensionality for recursive bayesian estimation. In: IEEE. *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. [S.l.], 2016. p. 1230–1237.
- LI, R. et al. Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 11, n. 11, p. 3954–3962, Nov 2018.
- LIN, T.-Y.; BELONGIE, S.; HAYS, J. Cross-view image geolocalization. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2013.
- LIN, T.-Y. et al. Learning deep representations for ground-to-aerial geolocalization. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015.
- LIU, H.; JEZEK, K. Automated extraction of coastline from satellite imagery by integrating canny edge detection and locally adaptive thresholding methods. *International journal of remote sensing*, Taylor & Francis, v. 25, n. 5, p. 937–958, 2004.
- LIU, T. et al. An investigation of practical approximate nearest neighbor algorithms. *Advances in neural information processing systems*, v. 17, 2004.
- LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Kluwer Academic Publishers, Hingham, MA, USA, v. 60, n. 2, p. 91–110, 2004. ISSN 0920-5691.
- LUCAS, B. D.; KANADE, T. et al. An iterative image registration technique with an application to stereo vision. In: VANCOUVER. [S.l.], 1981.
- Majdik, A. L.; Albers-Schoenberg, Y.; Scaramuzza, D. Mav urban localization from google street view data. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. [S.l.: s.n.], 2013. p. 3979–3986. ISSN 2153-0858.
- MARMANIS, D. et al. Semantic segmentation of aerial images with an ensemble of cnns. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, v. 3, p. 473, 2016.
- MATAS, J.; GALAMBOS, C.; KITTLER, J. Robust detection of lines using the progressive probabilistic hough transform. *Computer Vision and Image Understanding*, Elsevier, v. 78, n. 1, p. 119–137, 2000.
- MCFEETERS, S. K. The use of the normalized difference water index (ndwi) in the delineation of open water features. *International journal of remote sensing*, Taylor & Francis, v. 17, n. 7, p. 1425–1432, 1996.

- MILLER, A.; MILLER, B.; MILLER, G. Navigation of underwater drones and integration of acoustic sensing with onboard inertial navigation system. *Drones*, v. 5, n. 3, 2021. ISSN 2504-446X. Disponível em: <https://www.mdpi.com/2504-446X/5/3/83>.
- MORAVEC, H.; ELFES, A. High resolution maps from wide angle sonar. In: *Proceedings. 1985 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 1985. v. 2, p. 116–121.
- MUR-ARTAL, R.; TARDÓS, J. D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, IEEE, v. 33, n. 5, p. 1255–1262, 2017.
- MURPHY, R. R. *Introduction to AI robotics*. [S.l.]: MIT press, 2019.
- MURUGANANDHAM, S. *Semantic segmentation of satellite images using deep learning*. 2016.
- NEGAHDARIPOUR, S.; AYKIN, M.; SINNARAJAH, S. Dynamic scene analysis and mosaicing of benthic habitats by fs sonar imaging-issues and complexities. In: IEEE. *OCEANS 2011*. [S.l.], 2011. p. 1–7.
- NEGAHDARIPOUR, S.; FIROOZFAM, P.; SABZMEYDANI, P. On processing and registration of forward-scan acoustic video imagery. In: *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*. [S.l.: s.n.], 2005. p. 452–459.
- NIE, S.; ZHANG, H.; JIANG, Z. Sea-land segmentation based on template matching. In: ACM. *Proceedings of the Fifth International Conference on Network, Communication and Computing*. [S.l.], 2016. p. 32–36.
- NODA, M. et al. Vehicle ego-localization by matching in-vehicle camera images to an aerial image. In: KOCH, R.; HUANG, F. (Ed.). *Computer Vision – ACCV 2010 Workshops*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 163–173. ISBN 978-3-642-22819-3.
- OLIVA, A.; TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, Springer, v. 42, n. 3, p. 145–175, 2001.
- OTSU, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, IEEE, v. 9, n. 1, p. 62–66, 1979.
- PAISITKRIANGKRAI, S. et al. Semantic labeling of aerial and satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, IEEE, v. 9, n. 7, p. 2868–2881, 2016.
- PARK, J.; KIM, J. Robust underwater localization using acoustic image alignment for autonomous intervention systems. *IEEE Access*, v. 10, p. 58447–58457, 2022.
- PARK, M. et al. Translation-symmetry-based perceptual grouping with applications to urban scenes. In: KIMMEL, R.; KLETTE, R.; SUGIMOTO, A. (Ed.). *Computer Vision – ACCV 2010*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 329–342. ISBN 978-3-642-19318-7.
- PARKER, J. R. *Algorithms for image processing and computer vision*. [S.l.]: John Wiley & Sons, 2010.
- QUIGLEY, M. et al. Ros: an open-source robot operating system. In: KOBE, JAPAN. *ICRA workshop on open source software*. [S.l.], 2009. v. 3, n. 3.2, p. 5.

- REDDY, B. S.; CHATTERJI, B. N. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, v. 5, n. 8, p. 1266–1271, 1996.
- REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: CORTES, C. et al. (Ed.). *Advances in Neural Information Processing Systems 28*. [S.l.]: Curran Associates, Inc., 2015. p. 91–99.
- RIBAS, D. et al. SLAM using an imaging sonar for partially structured underwater environments. In: *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. [S.l.: s.n.], 2006. p. 5040–5045.
- RIBEIRO, P. O. C. S. et al. Underwater place recognition in unknown environments with triplet based acoustic image retrieval. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2018. p. 524–529.
- ROBERT, W. F. Statistics of speckle in ultrasound b-scans. *IEEE Trans. Sonics & Ultrason.*, v. 30, n. 3, p. 156–163, 1983. Disponível em: <https://cir.nii.ac.jp/crid/1573668923839083776>.
- ROMERO, R. F. et al. *Robótica Móvel*. [S.l.]: LTC - GRUPO GEN, 2014.
- RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N. et al. (Ed.). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015. p. 234–241. ISBN 978-3-319-24574-4.
- ROTTENSTEINER, F. et al. Isprs semantic labeling contest. *ISPRS: Leopoldshöhe, Germany*, v. 1, p. 4, 2014.
- RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, Springer, v. 115, n. 3, p. 211–252, 2015.
- SANTOS, M. D. et al. Object classification in semi structured enviroment using forward-looking sonar. *Sensors*, v. 17, n. 10, 2017. ISSN 1424-8220. Disponível em: <http://www.mdpi.com/1424-8220/17/10/2235>.
- SANTOS, M. M. et al. Underwater place recognition using forward-looking sonar images: A topological approach. *Journal of Field Robotics*, Wiley Online Library, v. 36, n. 2, p. 355–369, 2019.
- SANTOS, M. M. et al. Underwater place recognition using forward-looking sonar images: A topological approach. *Journal of Field Robotics*, v. 36, n. 2, p. 355–369, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21822>.
- SANTOS, M. M. D. et al. A framework for underwater vehicle localization based on cross-view and cross-domain acoustic and aerial images. In: IEEE. *2021 Latin American Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE)*. [S.l.], 2021. p. 204–209.
- SANTOS, M. M. D. et al. Cross-view and cross-domain underwater localization based on optical aerial and acoustic underwater images. *IEEE Robotics and Automation Letters*, v. 7, n. 2, p. 4969–4974, 2022.
- SANTOS, M. M. D. et al. Matching color aerial images and underwater sonar images using deep learning for underwater localization. *IEEE Robotics and Automation Letters*, v. 5, n. 4, p. 6365–6370, 2020.

- SANTOS, M. M. dos et al. Satellite and underwater sonar image matching using deep learning. In: *SBR-LARS 2019*. [S.l.: s.n.], 2019.
- SANTOS, M. M. dos et al. Underwater sonar and aerial images data fusion for robot localization. In: *ICAR 2019*. [S.l.: s.n.], 2019.
- SANTOS, M. M. dos et al. Semantic segmentation of static and dynamic structures of marina satellite images using deep learning. In: *BRACIS 2019*. [S.l.: s.n.], 2019.
- SHECHTMAN, E.; IRANI, M. Matching local self-similarities across images and videos. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2007. p. 1–8.
- SINGH, D.; VALDENEGRO-TORO, M. The marine debris dataset for forward-looking sonar semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2021. p. 3741–3749.
- STEFFENS, C. R. et al. Cnn based image restoration. *Journal of Intelligent & Robotic Systems*, Springer, v. 99, n. 3, p. 609–627, 2020.
- STEFFENS, C. R. et al. Contrast enhancement and image completion: A cnn based model to restore ill exposed images. In: IEEE. *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. [S.l.], 2019. v. 1, p. 226–232.
- THRUN, S.; BURGARD, W.; FOX, D. Probabilistic robotics (intelligent robotics and autonomous agents series). *Intelligent robotics and autonomous agents, The MIT Press (August 2005)*, 2006.
- TIAN, Y.; CHEN, C.; SHAH, M. Cross-view image matching for geo-localization in urban environments. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017.
- TRUJILLO-PINO, A. et al. Accurate subpixel edge location based on partial area effect. *Image and Vision Computing*, v. 31, n. 1, p. 72–90, 2013. ISSN 0262-8856. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0262885612001850>.
- VAUGHAN, R. Massively multi-robot simulation in stage. *Swarm Intelligence*, v. 2, n. 2, p. 189–208, Dec 2008. ISSN 1935-3820. Disponível em: <https://doi.org/10.1007/s11721-008-0014-4>.
- Viswanathan, A.; Pires, B. R.; Huber, D. Vision based robot localization by ground to satellite matching in gps-denied situations. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. [S.l.: s.n.], 2014. p. 192–198. ISSN 2153-0858.
- VO, N. N.; HAYS, J. Localizing and orienting street views using overhead imagery. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 494–509.
- Wolff, M.; Collins, R. T.; Liu, Y. Regularity-driven building facade matching between aerial and street views. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 1591–1600. ISSN 1063-6919.
- WORKMAN, S.; JACOBS, N. On the location dependence of convolutional neural network features. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2015.
- WU, F. et al. Inshore ship detection based on convolutional neural network in optical satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, IEEE, v. 11, n. 11, p. 4005–4015, 2018.

- WU, M. et al. Towards accurate high resolution satellite image semantic segmentation. *IEEE Access*, IEEE, v. 7, p. 55609–55619, 2019.
- WU, Y. et al. Survey of underwater robot positioning navigation. *Applied Ocean Research*, v. 90, p. 101845, 2019. ISSN 0141-1187. Disponible em: <http://www.sciencedirect.com/science/article/pii/S0141118718305546>.
- WU, Y. et al. Building generalizable agents with a realistic and rich 3d environment. *CoRR*, abs/1801.02209, 2018. Disponible em: <http://arxiv.org/abs/1801.02209>.
- YU, F.; KOLTUN, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
- ZHAO, H. et al. Icnet for real-time semantic segmentation on high-resolution images. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018.
- ZHONG, Y. et al. Mini-uav-borne hyperspectral remote sensing: From observation and processing to applications. *IEEE Geoscience and Remote Sensing Magazine*, IEEE, v. 6, n. 4, p. 46–62, 2018.
- ZHOU, B. et al. Learning deep features for scene recognition using places database. Neural Information Processing Systems Foundation, 2014.