Aprendizado por Reforço Profundo para Navegação Sem Mapa de um Veículo Híbrido Aéreo-Aquático usando Imagens

Junior D. Jesus[1], **Paulo L. J. Drews-Jr**[1], **Rodrigo S. Guerra**[1]

[1]Programa de Pós-Graduação em Computação/PPGCOMP
Universidade Federal do Rio Grande/FURG

dranaju@gmail.com, paulodrews@furg.br, tioguerra@gmail.com

Abstract. *Reinforcement Learning (RL) has shown impressive performance in video games and continuous control tasks. However, RL has poor performance with high-dimensional observations such as raw pixel images. It is generally accepted that RL policies based on physical state, such as laser sensor measurements, provide more efficient sampling results than pixel-based learning. This work presents a new approach that extracts information from a depth map estimate and raw pixel images to train an RL agent to perform map-less navigation of a Hybrid Unmanned Aerial-Underwater Vehicle (HUAUV). The proposed approach is called Unsupervised Prioritized Contrastive Representations of Depth and Pixel Images in Reinforcement Learning (CUPRL and Depth-CUPRL) which estimates depth from images and uses raw pixel images with a prioritized replay memory. A combination of RL and Contrastive Learning is used to address the problem of RL based on image observations. Contrastive Learning allows to create a latent space that is capable of mapping pixel and depth images in a way that, even using only pixel images, it is possible to create efficient representations to solve navigation problems in complex environments. From the results obtained with the HUAUV, it can be concluded that the proposed CUPRL and Depth-CUPRL approach is effective for decision making and outperforms state-of-the-art pixel-based approaches in map-less navigation.*

Resumo. *O Aprendizado por Reforço (RL) tem se mostrado altamente eficaz em jogos eletrônicos e tarefas de controle contínuo. Entretanto, RL apresenta dificuldades em lidar com observações de alta dimensionalidade, como imagens brutas de pixels. É amplamente aceito que políticas de RL baseadas em estado físico, como medições de sensores a laser, geralmente produzem amostragens mais eficientes do que o aprendizado com base em pixels. Neste trabalho, é proposta uma nova abordagem que combina informações de uma estimativa de mapa de profundidade e imagens brutas de pixels para ensinar um agente de RL a realizar a navegação sem mapa em um Veículo Híbrido Aéreo-Aquático (HUAUV). Esta abordagem, denominada Representações Priorizadas Contrastivas Não Supervisionadas de Imagens de Profundidade e Pixel em Aprendizado por Reforço (CUPRL e Depth-CUPRL), estima a profundidade de imagens e utiliza imagens brutas de pixels com uma memória de repetição priorizada. É utilizada uma combinação de RL e Aprendizagem Contrastiva para lidar com o desafio de aprender com base em observações de imagens. A Aprendizagem Contrastiva permite criar um espaço latente que é capaz de mapear imagens de pixel e profundidade de tal forma que, mesmo ao utilizar apenas imagens*

de pixels, é possível criar representações eficientes para solucionar problemas de navegação em ambientes complexos. Os resultados obtidos com o HUAUV indicam que a abordagem proposta é eficaz na tomada de decisão e supera as abordagens baseadas em pixels existentes na capacidade de navegação sem mapa.

Dissertação de mestrado [1] apresentada ao Programa de Pós Graduação em Engenharia de Computação (PPGComp) da Universidade Federal de Rio Grande (FURG) em fevereiro de 2023 sobre a orientação do Prof. Dr. Paulo L. J. Drews-Jr e co-orientação do Prof. Dr. Rodrigo S. Guerra.

1. Introdução

Os veículos não tripulados, conhecidos como UVs (*Unmanned Vehicles*, em inglês), desempenham papéis importantes em diversas missões aéreas e aquáticas, abrangendo áreas como ciências marinhas, extração de petróleo e gás, inspeção, exploração e resgate [Cho et al. 2015]. Existem três categorias de UVs: terrestres (UGVs), aéreos (UAVs) e aquáticos (UUVs). Esses veículos têm a capacidade de transportar instrumentos, coletar amostras e conduzir pesquisas, permitindo que cientistas acompanhem o progresso de suas missões de forma segura à distância.

A navegação de veículos híbridos aéreo-aquáticos, conhecidos como HUAUVs (*Hybrid Unmanned Aerial Underwater Vehicles*, em inglês), em ambientes contrastantes é um desafio significativo. Embora UAVs e UUVs tenham suas vantagens em seus respectivos ambientes, eles não conseguem realizar um monitoramento completo em situações que exigem a atuação em ambos os meios, como inspeção de plataformas de petróleo. Assim, há uma crescente necessidade de desenvolver HUAUVs capazes de operar tanto em ambientes aéreos quanto subaquáticos, a fim de unir as vantagens desses veículos e explorar seus recursos em conjunto.

O problema reside na ineficiência das redes de Aprendizado por Reforço Profundo ou Deep Reinforcement Learning (Deep-RL) ao lidar com imagens [Kaiser et al. 2019], [Laskin et al. 2020] e na falta de estudos sobre a navegação de HUAUVs em ambientes aéreos e subaquáticos utilizando imagens. Portanto, o objetivo deste trabalho é propor abordagens baseadas em Deep-RL combinadas com imagens e informações de profundidade para a navegação de HUAUVs em ambientes híbridos. O objetivo é melhorar a estimativa do ambiente e obter respostas de ação mais precisas, permitindo a realização eficiente e segura de tarefas de navegação.

Este trabalho busca contribuir para a área explorando técnicas de aprendizado por reforço profundo em HUAUVs que operam tanto em ambientes aéreos quanto subaquáticos. Pretende-se analisar a capacidade de evitar colisões em cenários específicos, comparar diferentes abordagens de Deep-RL para a navegação de HUAUVs. Além disso, serão investigadas estratégias de navegação em ambientes 3D, utilizando apenas informações de localização e visão obtidas por câmera, sem a necessidade de um mapa aéreo ou subaquático. Essas contribuições visam aprimorar a eficiência e a autonomia dos HUAUVs em ambientes desafiadores.

[1] https://argo.furg.br/?BDTD13708

Até o presente momento, este trabalho já publicou os resultados da navegação em ambientes simples, que introduzem o conceito da Depth-CUPRL na Conferência Internacional de Robótica e Sistemas Inteligentes (IEEE IROS 2022: **Qualis-A1**) [de Jesus et al. 2022] e contribuiu para os trabalhos de [Jesus et al. 2021], [Grando et al. 2021a] e [Grando et al. 2022].

2. Trabalhos Relacionados

O estudo da navegação sem mapa tem sido amplamente pesquisado para robôs móveis terrestres [Tai and Liu 2016]. No entanto, a navegação autônoma de robôs móveis aéreos usando abordagens de Deep-RL é menos comum e geralmente se concentra em evitar o uso de informações visuais [Grando et al. 2022, Grando et al. 2020] ou em utilizar informações simplificadas sem o recurso de aprendizado contrastivo [Rodriguez-Ramos et al. 2018, Sampedro et al. 2019, He et al. 2020, Li et al. 2020].

[Rodriguez-Ramos et al. 2018] propõe uma abordagem baseada em Deep-RL para resolver o problema de pouso de um UAV em uma base que pode estar em movimento ou estática. [Sampedro et al. 2019] sugere que uma abordagem baseada em uma técnica Deep-RL poderia realizar uma tarefa de Busca e Resgate em cenários fechados, utilizando o algoritmo Deep Deterministic Policy Gradient (DDPG) [Lillicrap et al. 2015]. No trabalho de [Jesus et al. 2021], é demonstrado que abordagens baseadas no algoritmo Soft Actor-Critic (SAC) [Haarnoja et al. 2018b] são mais eficazes para a navegação de robôs do que abordagens baseadas no algoritmo DDPG.

O uso de técnicas de aprendizado profundo, como o Deep-RL, para a navegação de sistemas robóticos, como HUAUVs, enfrenta desafios devido à necessidade de grandes quantidades de dados para treinamento e à ineficiência ao lidar com observações de alta dimensão, como imagens de pixel bruto [Kaiser et al. 2019]. O estudo realizado por [Grando et al. 2021a] explorou técnicas de Deep-RL para HUAUVs, porém, com foco em observações de baixa dimensão, como sensores *lasers*. Para superar a ineficiência do Deep-RL com observações de alta dimensão, [Laskin et al. 2020] propôs a técnica CURL, capaz de extrair características úteis das imagens brutas e melhorar o desempenho do controle de redes Deep-RL. Embora a CURL tenha mostrado resultados promissores em ambientes simulados semelhantes a jogos, este projeto busca adaptá-la para ambientes mais próximos das condições reais, com visão de primeira pessoa e usando imagens monoculares.

Para navegação de veículos aéreos móveis, é possível encontrar trabalhos como o de [Thomas et al. 2021] que apresenta um algoritmo baseado em RL utilizando modelos de auto-atenção para controlar um UAV autônomo. O trabalho de [Thomas et al. 2021] mostrou que ele pode efetivamente completar a navegação do veículo mesmo quando submetido a diferentes entradas no algoritmo, destacando como o algoritmo é capaz de lidar com dados de estado com ruídos ou modificações. Outro trabalho, [He et al. 2020], usou um Lobula Giant Moment Detector (LGMD) para simplificar as informações de visão para Deep-RL na navegação e prevenção de obstáculos de um UAV. Ele realizou missões em um ambiente complexo com 80% de taxa de sucesso.

Este trabalho se destaca de outros estudos relacionados ao utilizar informações de mapas de profundidade em uma abordagem baseada em aprendizado contrastivo para lidar com o desafio das observações de alta dimensão. Ele também propõe um sistema de

memória de repetição de experiências priorizado que aumenta a eficiência da abordagem proposta, e é o primeiro a propor um sistema de aprendizado contrastivo com repetição de experiências priorizadas. A utilização do aprendizado contrastivo neste trabalho permite desenvolver um espaço latente que pode relacionar imagens de pixels e profundidade de forma que, mesmo usando somente imagens de pixels, é possível criar representações eficientes para resolver problemas de navegação em ambientes complexos.

3. Metodologia

Este trabalho propõe um sistema de controle para HUAUV utilizando informações visuais e de profundidade como entrada para redes de aprendizado por reforço profundo (Deep-RL) para construir uma política de movimento que evita obstáculos no ambiente. São geradas imagens de profundidade e imagens brutas de pixels que são enviadas como entrada para a rede neural similar a proposta por [Laskin et al. 2020], chamada CURL, que aprende a controlar a navegação nestes ambientes através de recompensas priorizadas. Os métodos propostos neste trabalho, denominados CUPRL e Depth-CUPRL, são baseados em uma rede CURL combinada com mapas de profundidade, imagens brutas de pixels e memórias priorizadas para a navegação de um veículo híbrido e são originados pelo próprio autor desse trabalho proposto. A equação de movimento para CUPRL e Depth-CUPRL é definida como:

$$v_t = f^X(I_t)\ , X \in \{CUPRL, Depth\text{-}CUPRL\} \tag{1}$$

onde I_t é a imagem de pixel bruta da câmera e v_t é a velocidade aplicada ao HUAUV. Para o treinamento do CUPRL, a Equação 1 recebe uma estimativa de profundidade de I_t e a informação da observação em pixel. A rede neural CUPRL extrai informações do mapa de profundidade e imagens de pixel para depois passar por uma rede SAC que retorna a velocidade ao robô. Na Figura 1 é mostrado o sistema de estados da CUPRL e Depth-CUPRL. Para CUPRL, I_t, I_{t-1}, I_{t-2} ou o_q e $D(I_t, I_{t-1}, I_{t-2})$ ou o_k que serão utilizados nas redes para o treinamento. $D(I_t, I_{t-1}, I_{t-2})$ é observação dos mapas de profundidades que foram gerados das imagens de pixel de I_t, I_{t-1}, I_{t-2}. Essas observações são passadas para seus codificadores f_{θ_q} e f_{θ_k} que retornam as representações de espaço latente q e k. O espaço latente q, a altura do veículo no eixo z e a posição do alvo no ambiente são as informações passada para o algoritmo Deep-RL. Já para a aprendizagem contrastiva, ambos espaços latentes q e k são utilizados no aprendizado das características de o_q e o_k.

A rede neural Depth-CUPRL, diferentemente da CUPRL, se concentra em extrair informações apenas a partir dos mapas de profundidade, como pode ser observado na Figura 1. As diferenças entre as duas redes propostas estão no tipo de informação que elas priorizam. No caso da Depth-CUPRL, as redes neurais recebem informações apenas de mapas de profundidades $D(I_t, I_{t-1}, I_{t-2})$ com modificações nessas observações para o treinamento. Por usar um contexto de modificações de recorte nas observações o_q e o_k, a Depth-CUPRL pode convergir mais rápido no contexto de aprender as características importantes do ambiente. Já o caso da CUPRL, uma obervação representa informações visuais e outra observações informações de profundidade. O caso da CUPRL toma mais tempo para aprender as características do ambiente.

4. Estrutura de rede CUPRL e Depth-CUPRL

A arquitetura CUPRL e Depth-CUPRL, utilizada neste trabalho, segue uma estrutura semelhante à rede SAC (Soft Actor-Critic) apresentada nos estudos de [Jesus et al. 2019]

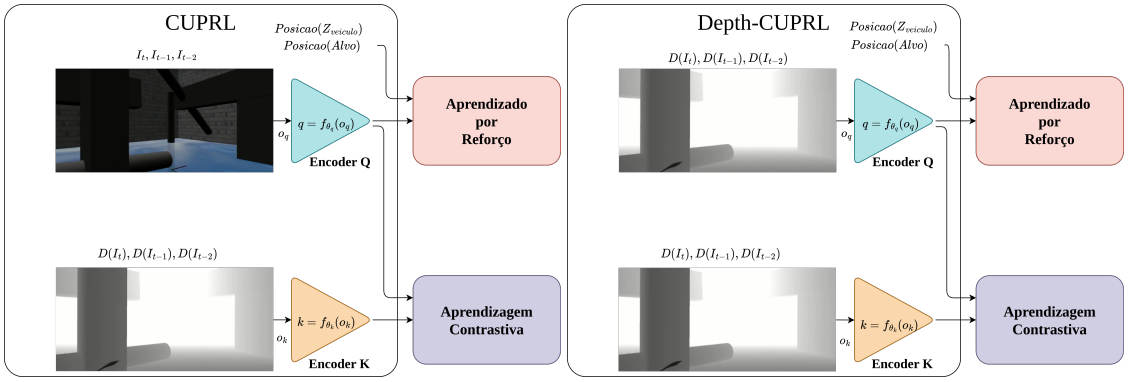


Figura 1. Sistema CUPRL e Depth-CUPRL.

e [Grando et al. 2021b]. Os encoders *query* e *key* da arquitetura proposta são implementados como camadas convolucionais. Essas redes, juntamente com as redes de comparação utilizadas neste trabalho, seguem ou se assemelham à estrutura desenvolvida em [Laskin et al. 2020].

A CUPRL e a Depth-CUPRL têm como entrada a imagem bruta de pixels ou o mapa de profundidade atual, respectivamente, juntamente com duas imagens e estimativas anteriores do agente no ambiente. Essa entrada passa por quatro camadas convolucionais, representando o encoder, e três camadas totalmente conectadas, culminando na camada de saída. O número de camadas e nós foi escolhido com base na estrutura de rede proposta por [Laskin et al. 2020]. A camada de saída gera as velocidades angulares e lineares no eixo x, e a velocidade linear no eixo z, que são enviadas ao agente. É importante ressaltar que as ações são normalizadas para o intervalo $[-1, 1]$ devido à função tangente hiperbólica $(tanh)$ utilizada como função de ativação. Antes de serem enviadas ao veículo, as velocidades angulares e lineares no eixo x são restringidas aos intervalos de $-0,3$ a $0,3$ rad/s e de 0 a $0,3$ m/s, respectivamente. No caso da navegação no eixo z, a velocidade linear é limitada entre os valores de $-0,3$ a $0,3$ m/s [Grando et al. 2021b].

4.1. Função de recompensa

É necessário estabelecer um sistema de recompensa e penalidade para as redes CUPRL e Depth-CUPRL. Essas recompensas e penalidades são dadas ao agente com base em uma função de recompensa, que é criada com base em conhecimento empírico durante o processo de solução do problema. Dessa forma, a rede realiza uma etapa de *feedforward* e *backpropagation* para aprender os hiperparâmetros.

Para as tarefas propostas, o objetivo é fazer a navegação até um ponto alvo de um veículo híbrido que é capaz de navegar na água e no ar. A função de recompensa formulada para essa tarefa foi definida como:

$$r(s_t, a_t) = \begin{cases} r_{chegar} \text{ se } d_t < c_{d_t} \\ r_{colidir} \text{ se } min_x < c_o \\ r_{navegando} = c_{navegando}(d_{t-1} - d_t) \text{ se } min_x \geq c_o \end{cases} \quad (2)$$

onde uma recompensa negativa $(r_{colidir} = -1)$ é dada se a leitura miníma de distância do robô à um objeto min_x é menor que c_o. Como recompensa positiva e política que

se quer otimizar através do RL, foi usado $r_{chegar} = 1$, onde essa recompensa é dada se a distância atual do robô ao alvo d_t for menor que a distância mínima do robô até o alvo $c_{d_t} = 40\ cm$, considerado como "chegar"ao ponto alvo. Porém, tratando-se de um ambiente complexo e difícil de obter a recompensa r_{chegar}, foi definida uma recompensa que tem como objetivo incentivar a aproximação entre o agente e o alvo $r_{navegando}$, onde a distância anterior e atual do robô ao alvo são usadas para gerar esse incentivo. Esse valor de incentivo é multiplicado por um valor pequeno $c_{navegando} = 0{,}1$, visando diminuir o impacto dessa recompensa na política principal, que é chegar ao alvo no ambiente. Não foram empregadas técnicas de *grid-search* para a seleção de parâmetros de recompensa. O objetivo da função de recompensa era manter-se o mais simplificada e direta possível.

5. Configuração do Experimento

Neste trabalho, utilizou-se o ROS como base e o Gazebo como simulador para executar as simulações. A implementação foi principalmente em Python, com uso ocasional de C++ para otimização. Redes neurais foram criadas utilizando o PyTorch, enquanto a manipulação de imagens foi realizada com a biblioteca OpenCV. O robô Hydrone (HUAUV) também foi empregado na simulação [Grando et al. 2020].

Ambientes de treinamento foram desenvolvidos no Gazebo para demonstrar a capacidade das técnicas propostas em navegar sem colisões, incluindo transições entre meios aéreo e aquático, e atingir um objetivo específico. O primeiro ambiente contava com 4 obstáculos para aumentar o desafio do robô, utilizando a técnica de Deep-RL e uma função de recompensa que penaliza colisões. No segundo ambiente, a tarefa de navegação se tornava mais complexa, exigindo que o veículo evitasse obstáculos e alcançasse um ponto de destino.

Para avaliar a eficácia das técnicas propostas, foram criados dois ambientes de teste adicionais, que simulavam cenários semelhantes aos ambientes de treinamento. Esses ambientes foram utilizados para testar o desempenho das redes após o treinamento, verificando se elas realmente aprenderam uma política de navegação eficaz.

6. Resultados

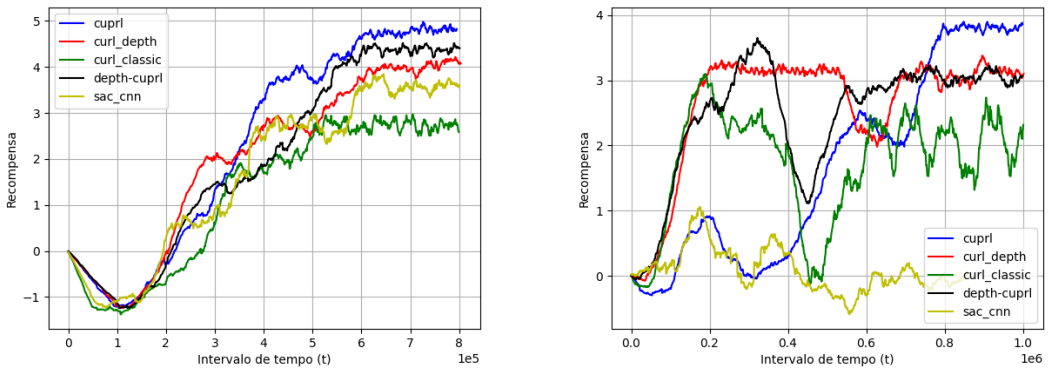
É importante observar que diferentes agentes foram treinados para cada ambiente proposto: um simples e outro mais complexo. Foram comparados os métodos propostos CUPRL e Depth-CUPRL com o CURL [Laskin et al. 2020] tendo como entrada a imagem bruta de pixel, e com uma versão modificada do CURL chamada de CURL (Depth) que utiliza mapas de profundidade como entrada, mas sem a utilização de memória priorizada. Também foram comparados com uma rede SAC (CNN prio.) como adotado em [Grando et al. 2021a], mas com camadas convolucionais. A SAC (CNN prio.) utiliza mapas de profundidade como entradas da rede e memória priorizada. Todas as redes testadas nesse trabalho seguiram a arquitetura proposta. A posição inicial do veículo é alterada para testes de transição ar-água e água-ar. Em testes de transição ar-água, a posição inicial do veículo é $[3{,}6; -3{,}2; 2{,}5]$ e para casos de transição água-ar a posição inicial do veículo é $[3{,}6; -3{,}2; -0{,}5]$.

A navegação em um ambiente 3D apresenta desafios distintos devido à locomoção do veículo híbrido em três dimensões, incluindo o eixo z. Além disso, um ponto alvo é estabelecido para ser alcançado pelo agente nos ambientes de treinamento e teste. No

primeiro ambiente de treinamento, o alvo é selecionado aleatoriamente e permite que o veículo navegue sem entrar em conflito com obstáculos. Esse alvo é alternado entre regiões no ar e na água de forma alternada. Desta forma, se o alvo começar no ar e o veículo conseguir chegar lá, uma recompensa é concedida à rede e o alvo é trocado para uma região na água. Se o veículo chegar ao alvo na água, o alvo é trocado para uma região no ar. Isso foi feito para forçar as transições de ambiente que o veículo precisa aprender de forma mais eficiente. O episódio de treinamento só termina em caso de colisão ou quando o número máximo de intervalos de tempo é alcançado.

Para treinar no primeiro ambiente, foi configurado um ambiente com uma memória de repetição de 100.000 amostras para todas as redes treinadas. Isso foi necessário, pois ao adicionar a dimensão de locomoção no eixo z, a tarefa de alcançar um ponto alvo se tornou mais complexa. Cada episódio é composto por 1000 intervalos de tempo t, com um intervalo de ação de $0{,}025ms$ ou $40Hz$. As redes neurais passam por 1000 intervalos de tempo, e então é realizada uma avaliação das redes propostas e de comparação com este trabalho. A avaliação das redes é feita utilizando apenas a resposta determinística da técnica, sem qualquer ruído, por 10 episódios.

Os resultados do treinamento da função de recompensa do primeiro ambiente, para as redes propostas e testadas, são mostrados na Figura 2a. É possível notar que, nos primeiros episódios de treinamento das redes neurais, há valores negativos para as recompensas das ações tomadas pelo agente. No entanto, para esse treinamento não há um valor máximo de recompensa estabelecido, pois isso depende da posição dos alvos escolhidos de forma aleatória. A recompensa adquirida nos episódios de avaliação mostra o grau de aprendizado da técnica Deep-RL e, em média, quantos alvos foram encontrados na avaliação. Isso é possível pois $r_{chegar} = 1$ é o maior valor da função de recompensa e $r_{navegando}$ é muito pequeno para impactar significativamente na média da recompensa em uma avaliação. Todas as redes foram treinadas por aproximadamente 800.000 intervalos de tempo. É possível concluir que os algoritmos propostos, CUPRL e Depth-CUPRL, obtiveram a maior média de recompensa para o primeiro ambiente de treinamento. Em seguida, as maiores recompensas foram obtidas pelas abordagens CURL (Depth), SAC (CNN prio) e CURL (Clássica).



(a) Primeiro ambiente **(b) Segundo ambiente**

Figura 2. Recompensas do primeiro e segundo ambiente simulado

Na Tabela 1 é mostrado os resultados obtidos para todas as técnicas avaliadas. É importante lembrar que as informações do *encoder key*, da Figura 1, são usadas apenas durante o treinamento da rede contrastiva e não são utilizadas durante a etapa de avaliação. Dessa forma, para o caso da CUPRL, a avaliação é realizada apenas com base em imagens brutas de pixels. O mesmo ocorre para a Depth-CUPRL, mas suas entradas incluem apenas informação de profundidade.

Tabela 1. Testes no primeiro ambiente de avaliação

Métodos	**Imagem**	**Ar-água (%)**	**Água-ar (%)**
CUPRL [Autor]	pixel	100 %	24,5 %
Depth-CUPRL [Autor]	profundidade	97,7 %	30,1 %
CURL(Depth) [Laskin et al. 2020]	profundidade	0 %	0 %
CURL(Clássica) [Laskin et al. 2020]	pixel	0 %	15,3 %
SAC(CNN prio) [Haarnoja et al. 2018a]	profundidade	0 %	0 %

A avaliação da Tabela 1 mostra que os algoritmos CUPRL e Depth-CUPRL obtiveram os melhores resultados. Já as outras técnicas avaliadas, não conseguiram completar a navegação até um ponto alvo ou obtiveram resultados insuficientes, como a CURL (Clássica) na transição água-ar. A transição ar-água foi onde as redes propostas se saíram melhor de acordo com testes de navegação ao alvo. É evidente que as redes mostraram uma porcentagem baixa de sucesso para chegar ao alvo na transição água-ar. Um comportamento observado nas técnicas avaliadas, especialmente aquelas que utilizam apenas mapas de profundidade, foi a dificuldade de transicionar entre meios. Para contornar este problema, foi incluída uma recompensa adicional para ajudar na transição:

$$r_{transicao} = c_{transicao}(d_{t-1} - d_t) \text{ se } -0{,}05 > Z_{veiculo} > -0{,}23 \tag{3}$$

onde $r_{transicao}$ foi definido com o objetivo de incentivar a transição entre os meios ar-água. Esse valor de incentivo é multiplicado por um pequeno fator $c_{transicao} = 0{,}2$ que tem o dobro do valor de $c_{navegando}$.

Os caminhos percorridos durante a avaliação das redes foram registrados e estão ilustrados na Figura 3. Nesta representação, o ponto verde representa a posição inicial do veículo, o ponto vermelho representa o objetivo alvo e a linha azul representa a trajetória percorrida pelo veículo.

Os caminhos mostrados na Figura 3 ilustram como o veículo navegou pelo ambiente durante os 1000 episódios de avaliação. Os caminhos percorridos pela CUPRL apresentam maior estabilidade na sua avaliação quando comparado com a Depth-CUPRL. Os outros métodos colidiram várias vezes, como é possível observar. No entanto, há um caso que pode ser visto na Figura 3c, onde a CURL (Depth) consegue evitar obstáculos e chegar próximo ao alvo. Entretanto, ao chegar perto, a CURL (Depth) apresenta um comportamento de *hovering* (permanecer parado no ar). Mesmo com uma recompensa para ajudar na transição, este comportamento não foi completamente eliminado.

Os resultados do treinamento da função de recompensa do segundo ambiente são apresentados na Figura 2b. Todas as redes foram treinadas por aproximadamente 1 milhão de intervalos de tempo. Quando comparado com as funções de recompensa anteriores, na

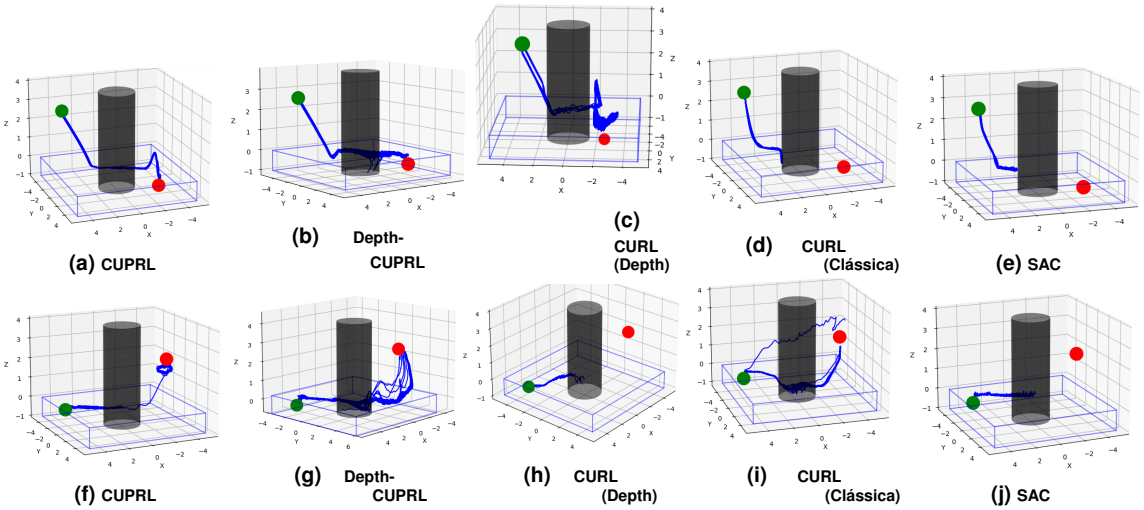


Figura 3. Caminho durante avaliação das técnicas no primeiro ambiente

Figura 2b, é possível observar uma recompensa mais instável para os algoritmos propostos e comparados. A CUPRL, apesar de ter um treinamento mais lento no final, superou a recompensa média adquirida em comparação com os outros algoritmos. A Depth-CUPRL e CURL (Depth) apresentaram no final do treinamento resultados semelhantes entre si. A CURL (Clássica) conseguiu obter recompensas médias boas, apesar de ter uma média instável. Já a SAC (CNN prio.), o único algoritmo que não utiliza aprendizagem contrastiva, não foi capaz de obter recompensas médias superiores a 1. Isso indica que a rede não conseguia alcançar o ponto de destino durante o treinamento da SAC (CNN prio.). Portanto, é possível concluir que a aprendizagem contrastiva apresenta uma vantagem na navegação através de imagens brutas de pixels e mapas de profundidade.

Após o treinamento das redes neurais no segundo ambiente, uma etapa de avaliação do modelo final de treinamento das redes neurais foi realizada por 1000 episódios. Os resultados da avaliação para todas as técnicas comparadas neste trabalho são apresentados na Tabela 2.

Tabela 2. Testes no segundo ambiente de avaliação

Métodos	**Imagem**	**Ar-Água (%)**	**Água-Ar (%)**
CUPRL [Autor]	pixel	100 %	14,6 %
Depth-CUPRL [Autor]	profundidade	100 %	0 %
CURL (Depth) [Laskin et al. 2020]	profundidade	0 %	0 %
CURL (Clássica) [Laskin et al. 2020]	pixel	0 %	0 %
SAC(CNN prio) [Haarnoja et al. 2018a]	profundidade	0 %	0 %

Na avaliação da Tabela 2, é possível observar que os algoritmos propostos, CUPRL e Depth-CUPRL, obtiveram novamente os melhores resultados na avaliação. Já as outras técnicas avaliadas zeraram em todos os casos. Com base nos resultados da avaliação apresentados no primeiro e segundo ambientes, é possível ver que os resultados das redes propostas se mostraram eficazes na transição ar-água. No entanto, nenhuma das redes testadas obteve bom desempenho na transição água-ar. Isso pode ser apon-

tado como um limite dessas redes no esquemático de simulação, função de recompensa e treinamento estabelecidos.

Os caminhos percorridos durante a avaliação das redes no segundo ambiente foram registrados e apresentados na Figura 4.

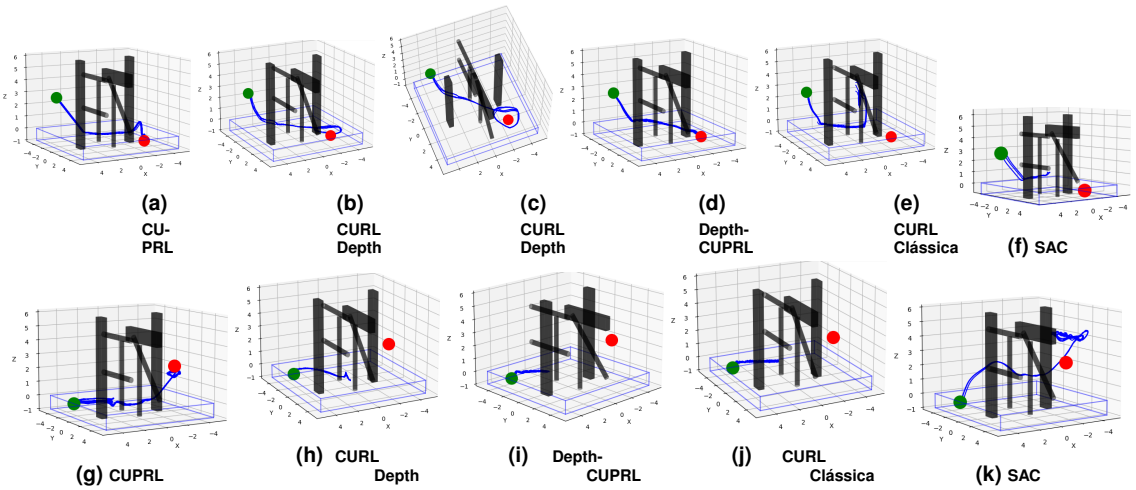


Figura 4. Caminho durante avaliação das técnicas no segundo ambiente

Os caminhos apresentados na Figura 4 mostram como o veículo navegou pelo segundo ambiente, que é mais complexo, durante os 1000 episódios de avaliação. A CUPRL conseguiu chegar ao destino 100% das vezes na transição ar-água e algumas vezes na transição água-ar. Já a Depth-CUPRL conseguiu navegar com sucesso somente nas transições ar-água. As outras técnicas avaliadas, CURL (Depth), CURL (Clássica) e SAC (CNN prio.), apresentaram comportamento de colisão ou ”hovering”sobre a água. Este comportamento de *hovering* pode ser visto novamente na CURL (Depth). Outro comportamento interessante é o observado na avaliação da SAC (CNN prio.) na transição água-ar. Durante o treinamento, a SAC (CNN prio.) não conseguia recompensas positivas e também não apresentava uma média negativa que indicaria colisões. Esse comportamento de *hovering* próximo ao alvo pode indicar que a SAC (CNN prio.) não conseguiu compreender a política ótima para a navegação até o ponto de destino.

7. Considerações Finais

Este trabalho teve como objetivo desenvolver um sistema capaz de navegar um HUAUV utilizando características extraídas de imagens brutas de pixels e mapas de profundidade através de aprendizado por reforço. As abordagens propostas e outras três métodos foram desenvolvidas e aplicadas em duas tarefas, mostrando melhores resultados em comparação com outras abordagens que visam aumentar a eficácia dos métodos de aprendizado por reforço em imagens brutas de pixels. A proposta foi investigar o uso de imagens para a navegação de veículos híbridos, considerando que redes de aprendizado por reforço são ineficientes em termos de amostra com imagens. A abordagem CUPRL utiliza imagens brutas de pixels e mapas de profundidade para aprender representações a partir de sequências de imagens, enquanto a abordagem Depth-CUPRL utiliza apenas as estimativas de profundidade. Os resultados mostraram que as redes propostas obtêm a maior recompensa média na navegação e foram capazes de completar a navegação em

ambientes de avaliação novos. A principal contribuição desse trabalho proposto foi desenvolver algoritmos capazes de aprender com informações visuais de pixel e profundidade, e sugere-se investigar a problemática da transição de meios água-ar e explorar outras técnicas para melhorar os resultados, além de avaliar os algoritmos em ambientes reais e superar os desafios da transferência de conhecimento entre a simulação e a realidade.

Referências

Cho, J., Lim, G., Biobaku, T., Kim, S., and Parsaei, H. (2015). Safety and security management with unmanned aerial vehicle (uav) in oil and gas industry. *Procedia manufacturing*, 3:1343–1349.

de Jesus, J. C., Kich, V. A., Kolling, A. H., Grando, R. B., Guerra, R. S., and Drews, P. L. (2022). Depth-cuprl: Depth-imaged contrastive unsupervised prioritized representations in reinforcement learning for mapless navigation of unmanned aerial vehicles. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10579–10586. IEEE.

Grando, R. B., de Jesus, J. C., and Drews-Jr, P. L. (2020). Deep reinforcement learning for mapless navigation of unmanned aerial vehicles. In *2020 Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics (SBR) and 2020 Workshop on Robotics in Education (WRE)*, pages 1–6. IEEE.

Grando, R. B., de Jesus, J. C., Kich, V. A., Kolling, A. H., Bortoluzzi, N. P., Pinheiro, P. M., Alves Neto, A., and Drews-Jr, P. L. J. (2021a). Deep reinforcement learning for mapless navigation of a hybrid aerial underwater vehicle with medium transition. In *IEEE ICRA*, pages 1088–1094.

Grando, R. B., de Jesus, J. C., Kich, V. A., Kolling, A. H., Bortoluzzi, N. P., Pinheiro, P. M., Neto, A. A., and Drews, P. L. J. (2021b). Deep reinforcement learning for mapless navigation of a hybrid aerial underwater vehicle with medium transition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1088–1094.

Grando, R. B., de Jesus, J. C., Kich, V. A., Kolling, A. H., and Drews-Jr, P. L. J. (2022). Double critic deep reinforcement learning for mapless 3d navigation of unmanned aerial vehicles. *Journal of Intelligent & Robotic Systems*, 104(2):1–14.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018a). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S. (2018b). Soft actor-critic algorithms and applications. *CoRR*, abs/1812.05905.

He, L., Aouf, N., Whidborne, J. F., and Song, B. (2020). Integrated moment-based LGMD and deep reinforcement learning for UAV obstacle avoidance. In *IEEE ICRA*, pages 7491–7497.

Jesus, J. C., Bottega, J. A., Cuadros, M. A., and Gamarra, D. F. (2019). Deep deterministic policy gradient for navigation of mobile robots in simulated environments. In *2019 19th International Conference on Advanced Robotics (ICAR)*, pages 362–367. IEEE.

Jesus, J. C. d., Kich, V. A., Kolling, A. H., Grando, R. B., Cuadros, M. A. d. S. L., and Gamarra, D. F. T. (2021). Soft actor-critic for navigation of mobile robots. *Journal of Intelligent & Robotic Systems*, 102(2):1–11.

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al. (2019). Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*.

Laskin, M., Srinivas, A., and Abbeel, P. (2020). Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR.

Li, B., Gan, Z., Chen, D., and Sergey Aleksandrovich, D. (2020). Uav maneuvering target tracking in uncertain environments based on deep reinforcement learning and meta-learning. *Remote Sensing*, 12(22):3789.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Rodriguez-Ramos, A., Sampedro, C., Bavle, H., Moreno, I. G., and Campoy, P. (2018). A deep reinforcement learning technique for vision-based autonomous multirotor landing on a moving platform. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1010–1017. IEEE.

Sampedro, C., Rodriguez-Ramos, A., Bavle, H., Carrio, A., de la Puente, P., and Campoy, P. (2019). A fully-autonomous aerial robot for search and rescue applications in indoor environments using learning-based techniques. *Journal of Intelligent & Robotic Systems*, 95(2):601–627.

Tai, L. and Liu, M. (2016). Towards cognitive exploration through deep reinforcement learning for mobile robots. *arXiv preprint arXiv:1610.01733*.

Thomas, D.-G., Olshanskyi, D., Krueger, K., Wongpiromsarn, T., and Jannesari, A. (2021). Interpretable uav collision avoidance using deep reinforcement learning. *arXiv preprint arXiv:2105.12254*.